REACTION-FREE PERSUASION MEASUREMENT VIA CONTRASTIVE LIKELIHOOD PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalable evaluation of persuasive messages is increasingly important as large language models (LLMs) generate persuasive content. Yet human experiments are costly, while common LLM-based evaluators rely on simulated audience reactions or holistic judgments that can be sensitive to seemingly innocuous changes in evaluation instructions. This sensitivity makes it difficult to distinguish message-level persuasiveness from measurement reactivity. We study the estimation of message-level persuasiveness using aggregate human treatment effects as an external criterion and introduce the Contrastive Stance Probe Estimator (CSPE), a reaction-free readout that measures how a message shifts the conditional likelihood of balanced third-party pro and con stance probes relative to a claim-matched neutral reference. We evaluate CSPE against Likert and Rubric-score baselines on 720 messages spanning ten policy issues and three evaluator models, and test stability transfer on an additional dataset of 5,052 messages. Across evaluator models and persona conditions, CSPE is substantially more stable to meaning-preserving changes in scoring instructions and consistently achieves the highest observed correspondence with human persuasion effects. A separate matched-quartet validation that independently manipulates stance strength and persuasion level further shows that CSPE remains sensitive to persuasion at both stance levels, whereas an entailment-based reference does not. Together, these results establish CSPE as a reproducible message-level estimator that combines high configuration stability with consistently stronger observed correspondence to aggregate human persuasion effects than the evaluated elicited-response baselines.

1 INTRODUCTION

Large language models (LLMs) can generate persuasive messages and dialogue at scale, creating growing demand for reliable measures of how persuasive a given message is (Singh et al., 2025; Bai et al., 2025; Salvi et al., 2025; Karinshak et al., 2023; Spitale et al., 2023; Goldstein et al., 2024). Such measures are important both for assessing beneficial communication and for understanding the risks of persuasive model behavior. Randomized human experiments provide the most direct evidence of persuasion by measuring changes in audience attitudes or behavior, but they are costly to conduct repeatedly across models, prompts, claims, and messages (Tappin et al., 2023; Simchon et al., 2024; Hackenburg & Margetts, 2024; Hackenburg et al., 2025b;a). Accordingly, human persuasion outcomes provide an external criterion for evaluating automated measures.

Existing automated approaches typically rely on an LLM to produce an explicit evaluative response. In *simulated-audience* evaluation, the model adopts a persona and reports its agreement or attitude on a Likert scale (Argyle et al., 2023; Aher et al., 2023; Santurkar et al., 2023; Bisbee et al., 2024; Antelmi et al., 2026; Bozdag et al., 2026). In *LLM-as-a-judge* evaluation, the model directly rates argument quality or expected persuasiveness (Zheng et al., 2023; Liu et al., 2023; Chiang & Lee, 2023; Wang et al., 2024; Koo et al., 2024; Panickssery et al., 2024). Related approaches learn from human comparisons of persuasive language (Pauli et al., 2025). These approaches operationalize persuasion through elicited reactions or evaluative judgments.

Explicit evaluation introduces a response-mapping step between the target construct and the observed score. Human measurement research shows that rating responses can depend on scale design and how judgments are mapped to response categories (Schwarz et al., 1991; Tourangeau et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

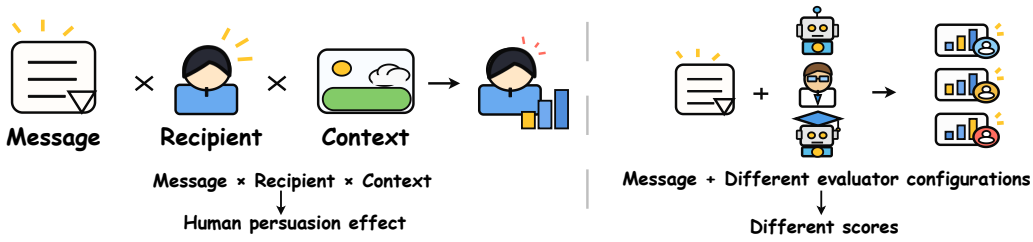

Figure 1: Human persuasion effects and automated message-level measurements. Human effects depend jointly on the message, recipient, and exposure context. Automated scores additionally depend on how the evaluator is configured.

2000). Evaluations of arguments can also be shaped by prior attitudes, with attitude-congruent arguments often judged more favorably than attitude-incongruent ones (Lord et al., 1979; Taber & Lodge, 2006). LLM-based evaluations exhibit analogous sensitivity to evaluator configuration, including prompting choices, response formats, and evaluator perspectives (Zhao et al., 2021; Lu et al., 2022; Sclar et al., 2024; Antelmi et al., 2026; Bozdag et al., 2026; Hwang et al., 2025; Zahraei et al., 2026). In this work, we focus on variation induced by meaning-preserving changes in evaluation instructions while holding the message and evaluator perspective fixed. We refer to this phenomenon as *measurement reactivity* (Figure 1). Such variation directly reduces the reproducibility of message-level rankings.

Our approach instead uses an LLM as a conditional probability model. We hypothesize that message persuasiveness is reflected in the message-induced shift in the conditional likelihood of balanced pro and con statements about the target claim. Based on this idea, we introduce the *Contrastive Stance Probe Estimator* (CSPE). Given a target claim and a persuasive message, CSPE measures this shift using matched third-party pro and con stance probes relative to a claim-matched neutral reference. The resulting likelihood contrast provides a direct message-level persuasiveness score without requiring the evaluator to generate an explicit rating or simulated attitude.

This design enables controlled evaluation of both reproducibility and external correspondence. Because the same stance probes and neutral reference are retained across evaluator configurations, we can directly evaluate reproducibility under meaning-preserving instruction changes and assess the consistency of the resulting rankings across personas, evaluator models, and reading frames. We then evaluate whether its message rankings correspond to aggregate persuasion effects from randomized human experiments and use a matched-quartet experiment to test sensitivity to persuasion independently of stance strength and semantic entailment.

Contributions.

- We frame automated persuasion evaluation around two complementary measurement criteria: reproducibility under controlled evaluator variations and correspondence with randomized human persuasion effects.
- We characterize *measurement reactivity* in elicited-response LLM evaluators by quantifying ranking sensitivity to semantically equivalent instruction wordings within fixed evaluator perspectives.
- We introduce *CSPE*, a likelihood-based estimator that measures message-induced shifts in balanced pro and con stance probes relative to a claim-matched neutral reference, without eliciting an explicit rating or simulated attitude.
- Across all evaluated model–persona combinations, CSPE substantially improves wording stability over Likert and Rubric scoring and achieves the highest human-outcome correspondence among the evaluated methods. Matched-quartet and component analyses further show that its signal tracks manipulated persuasion beyond stance strength and semantic entailment.

2 Related Work

Computational persuasion research has progressed from identifying linguistic factors associated with success in online arguments and donation dialogues (Persing & Ng, 2015; Habernal & Gurevych, 2016; Lukin et al., 2017; Carlile et al., 2018; Simpson & Gurevych, 2018; Li et al., 2020; Tan et al., 2016; Wang et al., 2019) to generating persuasive messages through learned strategies, planning, and LLMs (Chawla et al., 2023; Shi et al., 2021; Chen et al., 2022; Jin et al., 2024; Singh et al., 2025). As persuasive generation becomes increasingly scalable, reliable automated measurement of message-level persuasiveness becomes correspondingly important.

Human judgment is known to depend on both response-scale design and prior attitudes (Schwarz et al., 1991; Tourangeau et al., 2000; Lord et al., 1979; Taber & Lodge, 2006), providing a measurement perspective for studying configuration sensitivity in LLM-based evaluation. One line of work uses LLMs as simulated audiences. PMIYC tracks self-reported agreement during multi-turn persuasion (Bozdag et al., 2026), while persona-based evaluators measure endorsement from neutral or partisan perspectives (Antelmi et al., 2026). These approaches provide scalable audience-conditioned measurements by eliciting attitudes or agreement from the model. Their measurements therefore depend on the evaluator configuration used to elicit those responses.

A second line of work evaluates persuasive messages through direct judgments or learned scorers. PersuasionBench and PersuasionArena combine LLM judges with other evaluation signals (Singh et al., 2025), while Persuasive-Pairs trains a scorer from human comparisons (Pauli et al., 2025). Rubric-based assessment can improve scoring reliability (Jonsson & Svingby, 2007), but LLM judgments remain sensitive to rhetorical features, prior beliefs, prompt format, and example order (Hwang et al., 2025; Zahraei et al., 2026; Zhao et al., 2021; Lu et al., 2022; Sclar et al., 2024). These findings motivate evaluating persuasion measures not only by their scores, but also by the reproducibility of the rankings they induce.

Randomized human experiments provide a complementary standard by directly measuring attitude or behavior change following exposure to policy messages, personalized appeals, or interactive debates (Anthropic, 2024; Bai et al., 2025; Matz et al., 2024; Salvi et al., 2025). Recent large-scale studies extend such evidence across hundreds of messages and policy issues (Hackenburg et al., 2025b;a), providing a strong external criterion for evaluating automated message-level measures.

These lines of work leave open how to obtain a scalable message-level measure that combines reproducibility across evaluator configurations with correspondence to randomized human persuasion outcomes. CSPE addresses this problem through a likelihood-based contrast over balanced stance probes, allowing configuration stability and human-outcome correspondence to be evaluated within a single measurement framework.

3 Method

3.1 Motivating Observations: Prompt Sensitivity and Ranking Reversal in Elicited Persuasion Scores

Table 1: Two motivating observations for elicited-response persuasion measures. **Left:** Ranking sensitivity under semantically equivalent prompts. **Right:** A ranking reversal on Medicaid work requirements. Likert and Rubric scores are averaged across three no-persona prompts; n denotes the number of human participants per message.

Prompt sensitivity

Method	w_1–w_2	w_1–w_3	w_2–w_3	Mean
Likert	.411	.696	.365	.491
Rubric	.755	.472	.507	.578

Ranking reversal

	Human effect	n	Likert	Rubric
Message A	**14.80**	92	1.00	2.33
Message B	−4.47	71	**2.33**	**3.67**

Before introducing our estimator, we first examine two empirical properties of elicited-response persuasion measures that motivate the CSPE design (Table 1). Under a fixed no-persona GPT-OSS-120B evaluation, semantically equivalent instruction wordings produce substantial variation

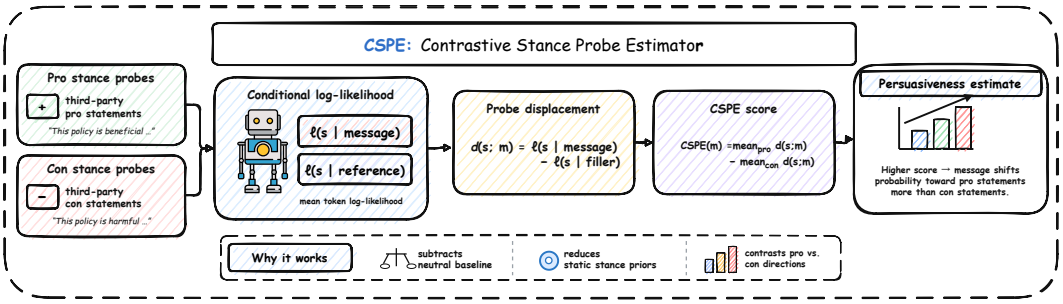

Figure 2: Overview of CSPE. Pro and con stance probes are scored under both a persuasive message and a claim-matched neutral reference. CSPE contrasts their mean likelihood displacements to quantify the message-induced change in directional support.

in Likert and Rubric message rankings. A same-claim comparison further shows disagreement between the elicited-response rankings and the ordering observed in randomized human effects. Together, these observations motivate a likelihood-based measurement strategy that reads directional persuasion signal directly from the evaluator's conditional distribution.

3.2 Problem Formulation

Let $t \in \mathcal{T}$ denote an issue, q an advocated target claim associated with t, and $\mathcal{M}_q$ the set of messages intended to persuade in favor of q. Given an evaluator model E and evaluator configuration a, our goal is to estimate a message-level persuasiveness score

$$S_{E,a}(m; q) \in \mathbb{R}, \qquad m \in \mathcal{M}_q,$$

that quantifies the persuasive effect of message m on target claim q.

The evaluator configuration a specifies the fixed evaluation conditions used for a scoring run, including the persona when applicable and the surrounding prompt context. Holding a fixed ensures that message-level variation is evaluated under a common recipient configuration. The same construction can be repeated across evaluator configurations to measure persuasion under different simulated audience conditions without changing the scoring rule.

3.3 CSPE Estimator

We operationalize this target by measuring how a persuasive message changes the evaluator's conditional likelihood of stance-specific statements relative to a claim-matched neutral reference. For each target claim, we use matched sets of pro and con stance probes and quantify the message-induced likelihood displacement for each probe. We then aggregate these displacements separately for the two stance directions and define the persuasion score as the difference between the average pro and con displacements. This construction captures whether the message shifts the evaluator's conditional distribution preferentially toward the advocated position rather than producing a general change in issue-related likelihood.

For a message m targeting claim q, let $\mathcal{S}_q^+$ denote a set of pro probes, $\mathcal{S}_q^-$ a set of con probes, and f_q a claim-matched neutral reference. Let $x_{m,a}$ and $x_{f,a}$ denote evaluation contexts constructed using the same prompt wrapper and evaluator configuration a. The two contexts differ only in whether the message position contains the persuasive message m or the neutral reference f_q.

For a probe $s = (s_1, \ldots, s_{N_E(s)})$, we compute its mean teacher-forced conditional log-likelihood under evaluator E as

$$\ell_{E,a}(s \mid x) = \frac{1}{N_E(s)} \sum_{i=1}^{N_E(s)} \log p_E(s_i \mid x, s_{<i}), \tag{1}$$

where $N_E(s)$ is the number of tokens in s under the tokenizer of E. Averaging over tokens normalizes the score for probe length.

We define the message-induced likelihood displacement of probe s as

$$d_{E,a}(s; m) = \ell_{E,a}(s \mid x_{m,a}) - \ell_{E,a}(s \mid x_{f,a}). \tag{2}$$

The displacement measures how the conditional likelihood of a fixed stance statement changes when the persuasive message replaces the claim-matched neutral reference. Positive values indicate that the message increases the evaluator's compatibility with the probe relative to the neutral context, whereas negative values indicate a decrease.

We aggregate these displacements separately over the pro and con probe sets:

$$D^{+}_{E,a}(m; q) = \frac{1}{|\mathcal{S}^{+}_{q}|} \sum_{s \in \mathcal{S}^{+}_{q}} d_{E,a}(s; m), \tag{3}$$

$$D^{-}_{E,a}(m; q) = \frac{1}{|\mathcal{S}^{-}_{q}|} \sum_{s \in \mathcal{S}^{-}_{q}} d_{E,a}(s; m). \tag{4}$$

Here, $D^{+}_{E,a}(m; q)$ measures the average message-induced displacement for statements supporting the target claim, while $D^{-}_{E,a}(m; q)$ measures the corresponding displacement for statements opposing it.

The CSPE score is defined as the difference between these average pro and con displacements:

$$S_{E,a}(m; q) = \mathrm{CSPE}_{E,a}(m; q) = D^{+}_{E,a}(m; q) - D^{-}_{E,a}(m; q). \tag{5}$$

Comparing the two displacement terms isolates the direction of the induced shift. A message can increase the likelihood of issue-related statements on both sides by making the topic more salient in the evaluator's conditional distribution. CSPE captures the differential change between the two stance directions: positive values indicate a greater shift toward the advocated claim, whereas negative values indicate a greater shift toward the opposing position.

For interpretation, define the mean conditional log-likelihood over a probe set as

$$\bar{\ell}_{E,a}(\mathcal{S} \mid x) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \ell_{E,a}(s \mid x). \tag{6}$$

Algebraically, the same score can be rearranged as

$$\begin{aligned} \mathrm{CSPE}_{E,a}(m; q) = \; & \left[\bar{\ell}_{E,a}(\mathcal{S}^{+}_{q} \mid x_{m,a}) - \bar{\ell}_{E,a}(\mathcal{S}^{-}_{q} \mid x_{m,a})\right] \\ & - \left[\bar{\ell}_{E,a}(\mathcal{S}^{+}_{q} \mid x_{f,a}) - \bar{\ell}_{E,a}(\mathcal{S}^{-}_{q} \mid x_{f,a})\right]. \end{aligned} \tag{7}$$

This equivalent form shows that CSPE can also be interpreted as the change in the pro–con likelihood contrast induced by the persuasive message relative to the claim-matched neutral reference.

CSPE scores fixed probe continuations using teacher-forced likelihoods rather than eliciting a generated rating or judgment. For a fixed evaluator, configuration, probe set, and rendered context, the resulting score is deterministic.

3.4 Measurement Instrument

External stance probes. For each target claim, we construct a balanced panel of eight third-party declarative stance probes, with four supporting the claim and four opposing it. Pro and con probes are matched in linguistic form, specificity, and semantic scope, so that their likelihood contrast captures directional support for the target claim. Averaging over multiple probes in each direction reduces dependence on any individual probe formulation. For a given claim q, the same probe sets $\mathcal{S}^{+}_{q}$ and $\mathcal{S}^{-}_{q}$ are used for all messages in $\mathcal{M}_q$ and across evaluator configurations.

Neutral reference. For each claim q, we additionally construct a claim-matched neutral reference f_q that provides relevant claim context without advocating either the pro or con position. The reference approximately matches the characteristic length of the persuasive messages associated with the claim and is shared across all messages targeting q. It therefore provides a common claim-level baseline for measuring the message-induced displacement of the stance probes. GPT-5.6 generates both the primary stance probes and neutral references.

4 EXPERIMENTS

We evaluate CSPE as a measure of message-level persuasiveness through three research questions:

RQ1 — Stability and reproducibility. Does CSPE provide a reproducible measure of message-level persuasiveness across variations in evaluation settings and procedures?

RQ2 — Human-outcome correspondence. Do CSPE scores correspond to persuasion effects observed in randomized human experiments?

RQ3 — Construct validation. Does CSPE respond to persuasion beyond stance strength or semantic entailment?

4.1 EXPERIMENTAL SETUP AND EVALUATION

We compare CSPE with simulated-audience Likert and explicit Rubric scoring on 720 messages across ten policy issues with randomized human treatment-effect estimates (Hackenburg et al., 2025b). Likert elicits a simulated audience's reported agreement with the target claim on a 1–5 Likert-type item, whereas Rubric elicits a 1–5 holistic judgment of the message's persuasive ability. Each evaluator (Llama-2-7B-Chat, Llama-3.1-8B, and GPT-OSS-120B) uses three persona conditions (no audience prefix, politically left-leaning, or right-leaning) and three meaning-preserving instruction wordings. We include token length as an invariance control; ConvinceQA provides a second dataset for stability transfer. Prompts are in Appendix K.

For RQ1, wording stability within fixed evaluator–persona conditions provides the primary measure of prompt-level reproducibility, while joint configuration agreement and cross-evaluator agreement evaluate consistency under broader changes in evaluation configuration. For RQ2, we use within-issue Spearman correlation and weighted pairwise concordance (WPC), with Top-K enrichment and AUROC as complementary measures. Spearman measures whether messages receiving higher evaluator scores also exhibit larger human persuasion effects within each issue. WPC complements this rank-based criterion by weighting message pairs according to the estimated reliability of their human-effect ordering, giving greater influence to comparisons supported more precisely by the randomized experiment. Human-correspondence scores are computed from standardized configuration scores, with 95% confidence intervals from 2,000 issue-clustered bootstrap replicates. Full definitions and aggregation procedures are in Appendix C.

4.2 RQ1: STABILITY AND REPRODUCIBILITY

Table 2 summarizes the primary reproducibility and human-correspondence results across evaluators and persona conditions. We begin with wording stability and return to human-outcome correspondence in RQ2. Across all nine evaluator–persona combinations, CSPE substantially exceeds both Likert and Rubric in wording stability and remains above .835 in every condition.

The advantage persists when the elicited-response baselines use continuous soft-label readouts. Although these readouts improve both baselines, CSPE retains higher joint configuration agreement for all three evaluators, with all paired 95% confidence intervals above zero (Appendix G). Thus, output discretization contributes to baseline instability but does not account for the stability gap.

Robustness analyses show the same pattern across reading frame, evaluator model, dataset, and probe construction (Table 3). CSPE remains highly consistent under direct and stepwise framing, shows stronger cross-evaluator agreement than both elicited-response baselines, and transfers to ConvinceQA. Paired cross-evaluator comparisons are positive in all cases, with all six 95% confidence intervals above zero (Appendix D.2). Configuration agreement also remains high when the probe panel is independently regenerated or individual stance probes are removed (Appendix D.4). Together, these results support CSPE's reproducibility across the evaluated settings and procedures.

4.3 RQ2: HUMAN-OUTCOME CORRESPONDENCE

We next evaluate correspondence with persuasion effects observed in randomized human experiments. In Table 2, CSPE attains the highest observed Spearman and WPC in all nine evaluator–persona combinations. This advantage is also consistent at the individual-configuration level: across

Table 2: Primary results by evaluator, method, and persona. Human Spearman ρ and WPC aggregate three wordings; wording stability compares them within each evaluator–persona condition. Full results and CIs are in Appendix Table 9.

Evaluator	Method	Persona	Human Spearman ρ	WPC	Wording stability
Llama-2-7B-Chat	Likert	None	.051	.502	.072
		Left	$-.014$	.489	.074
		Right	.006	.498	.073
	Rubric	None	$.080^{\dagger}$	.501	.200
		Left	$.049^{\dagger}$	.502	.300
		Right	$.049^{\dagger}$	.502	.300
	CSPE	None	**.319**	**.641**	**.984**
		Left	**.320**	**.642**	**.986**
		Right	**.315**	**.641**	**.985**
Llama-3.1-8B	Likert	None	.156	.557	.349
		Left	.064	.519	.397
		Right	.214	.592	.478
	Rubric	None	.270	.602	.606
		Left	.139	.545	.699
		Right	.273	.613	.599
	CSPE	None	**.340**	**.648**	**.978**
		Left	**.289**	**.626**	**.976**
		Right	**.292**	**.628**	**.979**
GPT-OSS-120B	Likert	None	.184	.575	.255
		Left	.206	.589	.300
		Right	.248	.607	.313
	Rubric	None	.166	.569	.264
		Left	.163	.575	.371
		Right	.148	.566	.380
	CSPE	None	**.332**	**.645**	**.879**
		Left	**.300**	**.632**	**.835**
		Right	**.290**	**.628**	**.844**

all three evaluators, CSPE configurations cluster at higher WPC values than the elicited-response baselines, and even the lowest CSPE WPC exceeds the highest Likert or Rubric WPC within each evaluator (Figure 3; Appendix J).

Paired comparisons with the strongest elicited-response baseline are positive in point estimate for all three evaluators. For Llama-2, the improvements in both Spearman and WPC have 95% confidence intervals entirely above zero; complete paired differences and confidence intervals for all evaluators are reported in Appendix Table 11.

Complementary analyses show the same overall pattern for AUROC, while Top-K enrichment favors CSPE for two of the three evaluators (Appendix F). CSPE's human correspondence also changes little under stepwise framing (Table 3; Appendix D).

4.4 RQ3: Construct Validation — Separating Persuasion from Stance

RQ3 isolates persuasion-related variation from stance strength and compares CSPE with a semantic-entailment reference (NLI). We use 50 matched quartets in a 2×2 design that independently crosses

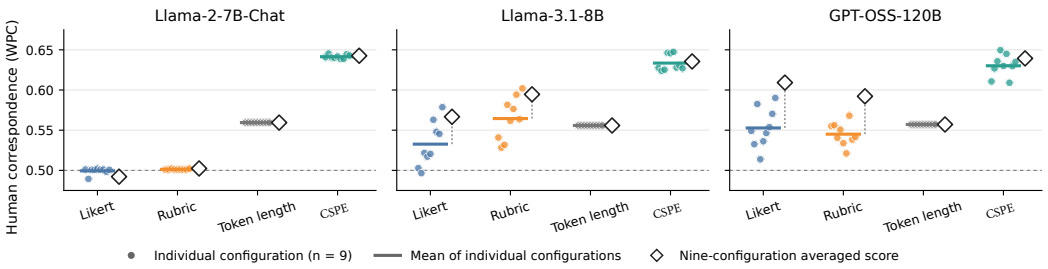

Figure 3: Human correspondence across individual evaluator configurations. Filled circles show configuration-level WPC, horizontal segments their mean, and open diamonds the nine-configuration composite. Dashed lines indicate chance-level WPC.

Table 3: Robustness and stability-transfer results. Panel A reports direct-to-stepwise changes in human correspondence together with within-frame configuration agreement for each reading frame; Panel B reports cross-evaluator agreement; Panel C reports ConvinceQA configuration agreement. Full confidence intervals are in Appendix Tables 6–8.

A. Reading-frame robustness (Llama-3.1)

Method	Δ Spearman ρ	Δ WPC	Joint configuration agreement Direct → stepwise
Likert	−.050	−.025	.316 → .341
Rubric	−.031	−.012	.532 → .541
CSPE	**−.005**	**−.002**	**.953 → .962**

B. Cross-evaluator agreement

Evaluator pair	Likert	Rubric	**CSPE**
L2–L3	.074	.158	**.647**
L2–GPT	.034	.145	**.531**
L3–GPT	.168	.246	**.436**

C. ConvinceQA stability transfer

Evaluator	Likert	Rubric	**CSPE**
Llama-3.1	.291	.155	**.913**
Llama-2	.172	.002	**.910**

stance strength with persuasion level, yielding 200 texts. This design allows the persuasion contrast to be evaluated while holding stance level fixed. Construction and inference details are provided in Appendix H.

CSPE scores increase with manipulated persuasion at both stance levels, with simple persuasion contrasts of 2.3886 at high stance and 1.5755 at low stance. The factorial analysis yields a persuasion main effect of 1.9820 (95% CI [1.7908, 2.1755]) and a positive interaction of 0.8132 (95% CI [0.6905, 0.9344]; Table 4); the positive stance main effect further shows sensitivity to stance strength. The NLI reference exhibits a different pattern, with a positive persuasion contrast at low stance but a slightly negative contrast at high stance. Together, these results show that CSPE captures persuasion-related variation within fixed stance levels that is not reproduced by the NLI entailment signal.

4.5 Estimator Component Validation

We finally examine the contribution of CSPE's main estimator components. We compare the full estimator with Pro only (D^+), Con only ($-D^-$), and a variant without neutral-reference subtraction.

Table 4: Matched-quartet factorial effects with 95% confidence intervals. CSPE and NLI use different numerical scales.

Method	Stance	Persuasion	Interaction
CSPE	0.8278 $[0.7676, 0.8863]$	1.9820 $[1.7908, 2.1755]$	0.8132 $[0.6905, 0.9344]$
NLI	0.8377 $[0.8106, 0.8665]$	0.3737 $[0.3441, 0.4022]$	-0.8096 $[-0.8666, -0.7494]$

Table 5 summarizes pooled cross-claim correspondence across the nine evaluator–persona combinations; Appendix I provides full results and inference details.

Table 5: Component ablations: pooled 720-message metrics, macro-averaged across evaluator–persona combinations. Deltas are relative to Full.

Configuration	Global human ρ	Δ vs. Full	Global WPC	Δ vs. Full
Full	.356	—	.660	—
Pro only	.145	$-.211$	.563	$-.097$
Con only	.043	$-.313$	.521	$-.139$
Without neutral reference	.204	$-.152$	.590	$-.070$

The full CSPE estimator achieves the strongest pooled human correspondence among all component configurations, with global Spearman $\rho = .356$ and WPC $= .660$ (Table 5); the same ordering holds for AUROC (Appendix I). Pro-only, con-only, and no-neutral-reference variants all yield lower correspondence, supporting complementary roles for the two main components: the pro–con contrast provides directional discrimination, while claim-matched neutral centering improves comparison across claims.

Across the three research questions, CSPE combines three complementary properties: highly stable message rankings under controlled evaluation variations, consistently stronger observed correspondence with randomized human persuasion effects than the evaluated elicited-response baselines, and sensitivity to persuasion when stance strength is experimentally controlled. Component analyses further show the strongest pooled human correspondence for the complete contrastive estimator. Together, these results establish a unified likelihood-based measurement framework that links reproducibility, external correspondence, and construct sensitivity in scalable message-level persuasion evaluation.

5 Conclusion and Future Work

We introduced CSPE, a message-level persuasiveness estimator that measures message-induced shifts in balanced pro and con probe likelihoods relative to a neutral reference.

Across 720 messages, ten issues, and three evaluator models, CSPE showed high wording stability within every persona condition (0.835–0.986) and the highest observed correspondence with human persuasion effects in all nine evaluator–persona combinations (Spearman 0.289–0.340; WPC 0.626–0.648). Component analyses show that the pro–con contrast is important for human correspondence, while neutral-reference subtraction reduces claim-specific baseline offsets when scores from multiple claims are considered jointly. Matched-quartet validation further shows that CSPE responds positively to experimentally manipulated persuasion at both levels of stance strength, supporting its sensitivity beyond stance variation alone.

Future work can extend CSPE to additional domains, languages, interaction settings, evaluator models, claims, and probe constructions, and use participant-level data to study heterogeneous treatment effects across recipient characteristics and contexts.

AI USE STATEMENT

Generative AI was used to generate the stance-probe texts and claim-matched neutral reference texts used as experimental inputs in the CSPE evaluation. Generative AI tools were also used to assist with language editing and improving manuscript readability. The research methodology, experimental design, scientific claims, and interpretation of results were developed by the authors. The authors reviewed the AI-generated experimental materials and take responsibility for the final content of the work.

ETHICS STATEMENT

This work studies automated measurement of persuasive messages and does not conduct new human-subject experiments; human outcomes are drawn from previously collected aggregate treatment-effect data. Because persuasion measurement can potentially be used to optimize manipulative or targeted persuasive content, including in politically sensitive settings, the proposed estimator may have dual-use risks. Our experiments focus on evaluating message-level measurement rather than generating persuasive content or profiling individual users.

REPRODUCIBILITY STATEMENT

We provide the estimator definition, experimental setup, evaluation metrics, and statistical procedures in the main text and appendices. Complete evaluation prompts are given in Appendix K, together with supplementary robustness analyses and full experimental results. The datasets and evaluator models used in all experiments are identified in the paper.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Alessia Antelmi, Alessia Galdeman, Lucio La Cava, Arianna Pera, and Giovanni Da San Martino. Political persuasion and endorsement in large language models, 2026. URL https://arxiv.org/abs/2606.05961.

Anthropic. Measuring the persuasiveness of language models. Anthropic Research, 2024. URL https://www.anthropic.com/news/measuring-model-persuasiveness. Accessed: 2026-09-01.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16:6037, 2025. doi: 10.1038/s41467-025-61345-5.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Nimet Beyza Bozdag, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tur. 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 Agentic Systems*, 2026. doi: 10.1145/3786335.3813181.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 621–631. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1058.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 750–766. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.53.

Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Seamlessly integrating factual information and social content with persuasive dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp. 399–413. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.aacl-main.31.

Cheng-Han Chiang and Hung-yi Lee. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8928–8942. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.599.

Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2):pgae034, 2024. doi: 10.1093/pnasnexus/pgae034.

Ivan Habernal and Iryna Gurevych. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1214–1223. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1129.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, 2024. doi: 10.1073/pnas.2403116121.

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, 2025a. doi: 10.1126/science.aea3884.

Kobi Hackenburg, Ben M. Tappin, Paul Röttger, Scott A. Hale, Jonathan Bright, and Helen Margetts. Scaling language model size yields diminishing returns for single-message political persuasion. *Proceedings of the National Academy of Sciences*, 122(10):e2413443122, 2025b. doi: 10.1073/pnas.2413443122.

Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. Can you trick the grader? adversarial persuasion of LLM judges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14632–14651, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.790.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. Persuading across diverse domains: A dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1678–1706. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.92.

Anders Jonsson and Gunilla Svingby. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2):130–144, 2007.

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):1–29, 2023. doi: 10.1145/3579592.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.29.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Building efficient universal classifiers with natural language inference, 2023. URL https://arxiv.org/abs/2312.17543.

Jialu Li, Esin Durmus, and Claire Cardie. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8905–8912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.716.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556.

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

Sandra C. Matz, Jacob D. Teeny, Shreya S. Vaid, Hannah Peters, Gabriella M. Harari, and Moran Cerf. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14: 4692, 2024. doi: 10.1038/s41598-024-53755-0.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. Measuring and benchmarking large language models' capabilities to generate persuasive language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10056–10075. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.506.

Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 543–552. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1053.

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.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Norbert Schwarz, Bärbel Knäuper, Hans-J Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. Rating scales numeric values may change the meaning of scale labels. *Public opinion quarterly*, 55(4):570–582, 1991.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3478–3492. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.295.

Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 3(2):pgae035, 2024. doi: 10.1093/pnasnexus/pgae035.

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

Somesh Kumar Singh, Yaman Kumar Singla, Harini S I, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NfCEVihkdC.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023. doi: 10.1126/sciadv.adh1850.

Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 613–624, 2016. doi: 10.1145/2872427.2883081.

Ben M. Tappin, Chloe Wittenberg, Luke B. Hewitt, Adam J. Berinsky, and David G. Rand. Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, 120(25):e2216261120, 2023. doi: 10.1073/pnas.2216261120.

Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. *The psychology of survey response*, volume 595. Cambridge New York, 2000.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1566.

Pardis Sadat Zahraei, Xiaoning Wang, Nimet Beyza Bozdag, Gokhan Tur, and Dilek Hakkani-Tür. Prior beliefs prejudice LLM-as-judge: Evidence from persuasion evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42049–42082, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.2087.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A LIMITATIONS

Human-outcome validation uses aggregate treatment effects from one English-language benchmark of 720 policy messages across ten issues. Accordingly, the present validation target is within-issue message ranking under aggregate human effects. Individual-level heterogeneity and absolute cross-claim calibration fall outside the present evaluation. CSPE also requires access to token-level likelihoods for evaluator scoring.

B ILLUSTRATIVE RANKING-REVERSAL CASE

Table 1 in Section 3.1 compares two messages advocating Medicaid work requirements. Message A (ID 692) has an estimated human persuasion effect of 14.80 (HC2 SE $= 3.16$, 95% CI $[8.60, 21.00]$; $n = 92$), whereas Message B (ID 573) has an estimated effect of -4.47 (HC2 SE $= 3.22$, 95% CI $[-10.78, 1.84]$; $n = 71$). The estimated difference is 19.27 ($z = 4.60$). Under GPT-OSS-120B with no persona, however, the mean scores across the three instruction wordings rank B above A for both Likert (2.33 vs. 1.00) and Rubric (3.67 vs. 2.33).

Message A (ID 692).

> Implementing a work requirement for Medicaid eligibility is a necessary step towards ensuring the sustainability and fairness of the program. By mandating that able-bodied individuals engage in work, job training, or community service, we can promote self-sufficiency and personal responsibility among beneficiaries. This approach not only helps recipients develop valuable skills and experience but also fosters a sense of pride and accomplishment. Moreover, a work requirement would safeguard taxpayer dollars by ensuring that Medicaid resources are directed towards those who genuinely need assistance while encouraging others to transition to financial independence. Critics argue that such a requirement would deprive needy individuals of healthcare, but provisions can be made for those with legitimate barriers to employment. Ultimately, a work requirement for Medicaid aligns with the program's intended purpose of providing a safety net for the most vulnerable while incentivizing self-reliance whenever possible. By striking this balance, we can create a more sustainable and equitable Medicaid system that benefits recipients, taxpayers, and society as a whole.

Message B (ID 573).

> Dear citizens,It is important for us to come together and consider the implementation of a requirement for people to work in order to receive Medicaid. While this proposal may be initially met with apprehension, I would like to take a moment to highlight the benefits that such a measure could bring to our society. Firstly, by encouraging work as a requirement for Medicaid, we can help break the cycle of poverty that exists in many low-income communities. People who are in need of Medicaid are often also struggling with underemployment or unemployment, and this requirement could provide them with the incentive they need to seek out job opportunities and become self-sufficient. Secondly, by incentivizing work, we could reduce the cost of Medicaid for taxpayers. Instead of the government shouldering the entire burden of providing health care to low-income individuals, we could instead spread the cost more evenly by connecting those in need with job opportunities that could help them become more financially independent. Finally, working can provide numerous benefits to individuals, including increased self-esteem, a sense of purpose, and improved physical and mental health. By encouraging work as a requirement for Medicaid, we could help those in need to lead happier, healthier lives, while also reducing the burden on taxpayers. In conclusion, the requirement for people to work in order to receive Medicaid is a proposal that is worth considering. It could provide numerous benefits to our society, including breaking the cycle of poverty, reducing the cost of Medicaid for taxpayers, and improving the health and well-being of those in need. I urge you to join me in support of this important measure.

C Evaluation Metrics and Statistical Inference

C.1 Uncertainty-Aware Human-Outcome Criterion

The human treatment-effect estimates for individual messages in the PNAS data carry sampling uncertainty. Messages with fewer observations or larger standard errors have less precisely estimated effects. The pairwise criterion therefore assigns greater influence to message pairs with more reliably estimated effect differences.

We construct an uncertainty-aware pairwise validation criterion within each issue. Let $\widehat{\theta}_A$ and $\widehat{\theta}_B$ be the estimated human treatment effects for messages A and B. Using their HC2 variance–covariance matrix, the estimated variance of their difference is

$$\widehat{V}_{AB} = \widehat{\mathrm{Var}}_{\mathrm{HC2}}(\widehat{\theta}_A) + \widehat{\mathrm{Var}}_{\mathrm{HC2}}(\widehat{\theta}_B) - 2\widehat{\mathrm{Cov}}_{\mathrm{HC2}}(\widehat{\theta}_A, \widehat{\theta}_B). \tag{8}$$

Under the normal approximation to the two effect estimates, the probability that A has a larger human treatment effect than B is

$$P_{AB} = P(\theta_A > \theta_B) = \Phi\left(\frac{\widehat{\theta}_A - \widehat{\theta}_B}{\sqrt{\widehat{V}_{AB}}}\right), \tag{9}$$

where Φ is the standard normal cumulative distribution function. We assign the pair the reliability weight

$$w_{AB} = |2P_{AB} - 1|. \tag{10}$$

When the human data cannot distinguish the pair, $P_{AB} \approx 0.5$ and its weight approaches zero. When one ordering is highly probable, P_{AB} approaches zero or one and the weight approaches one.

For an evaluator score s, let $c_{AB}(s)$ equal one when $s_A - s_B$ agrees with the more probable human ordering, zero when it reverses that ordering, and $1/2$ for an evaluator tie. The resulting WPC is

$$C_w(s) = \frac{\sum_{t\in\mathcal{T}} \sum_{(A,B)\in\mathcal{P}_t} w_{AB} c_{AB}(s)}{\sum_{t\in\mathcal{T}} \sum_{(A,B)\in\mathcal{P}_t} w_{AB}}, \tag{11}$$

where $\mathcal{P}_t$ contains only pairs of messages from issue t. Each pair's contribution to WPC is weighted by w_{AB}.

C.2 Primary Metrics

We separate variation across wording from variation across persona. For a method k, evaluator e, persona p, wording w, and message m, let $s^{(k)}_{mepw}$ denote the message's persuasiveness score. Let H_m be the corresponding human persuasion-effect point estimate and let $n_t = |\mathcal{M}_t|$.

Wording stability. Holding evaluator, method, persona, and issue fixed, the three wording variants produce three message-score vectors. We compute their three pairwise Spearman correlations and average first within issue and then across the ten issues:

$$\rho^{(k)}_{\mathrm{cfg}}(e,p) = \frac{1}{|\mathcal{T}|} \sum_{t\in\mathcal{T}} \left[\frac{1}{3} \sum_{1\le w<w'\le 3} \rho_S\Big(\{s^{(k)}_{mepw}\}_{m\in\mathcal{M}_t}, \{s^{(k)}_{mepw'}\}_{m\in\mathcal{M}_t} \Big) \right]. \tag{12}$$

Wording stability lies in $[-1, 1]$. Values near one indicate that instruction paraphrases preserve score-induced orderings, values near zero indicate little rank association, and negative values indicate reversed orderings. This statistic assesses wording robustness without human outcomes.

Persona-specific composite scores. For human correspondence, we standardize each configuration over all 720 messages and average the three standardized wording scores within each persona:

$$z^{(k)}_{mepw} = \frac{s^{(k)}_{mepw} - \mu^{(k)}_{epw}}{\sigma^{(k)}_{epw}}, \qquad \bar{s}^{(k)}_{mep} = \frac{1}{3}\sum_{w=1}^{3} z^{(k)}_{mepw}. \tag{13}$$

Standardization gives each wording a common scale. Every persona-specific composite score is evaluated within issue against the same aggregate human outcomes.

Spearman correlation. Within each issue, we correlate the evaluator composite with the human persuasion-effect point estimates:

$$\rho_t = \rho_S\Big(\{\bar{s}^{(k)}_{mep}\}_{m\in\mathcal{M}_t}, \{H_m\}_{m\in\mathcal{M}_t}\Big). \tag{14}$$

The ten issue correlations are pooled on the Fisher-z scale:

$$\rho = \tanh\left[\frac{\sum_{t\in\mathcal{T}}(n_t-3)\,\operatorname{arctanh}(\rho_t)}{\sum_{t\in\mathcal{T}}(n_t-3)}\right]. \tag{15}$$

All ten issues contain $n_t = 72$ messages, so the Fisher-transformed correlations receive equal weight. Spearman ρ lies in $[-1, 1]$ and measures whether higher persuasiveness scores correspond to larger human effects, independently of score scale.

WPC. We apply Equation 11 to the same issue-local composite scores. WPC gives greater weight to message pairs whose human-effect ordering is estimated more precisely. A value of 0.5 denotes chance agreement; values above 0.5 indicate that higher scores tend to accompany larger human effects among the better-supported comparisons.

Statistical inference. Where reported, 95% confidence intervals use 2,000 issue-clustered bootstrap replicates. Each replicate samples ten complete issues with replacement, recomputes the aggregate statistic, and contributes to the 2.5th and 97.5th percentiles. The intervals therefore preserve messages and dependent pairwise comparisons within an issue. The component ablations use a separate global analysis, defined in Appendix I, to evaluate components that operate on claim-level offsets.

C.3 Supplementary Metrics

Full-grid composite and joint configuration agreement. The robustness analyses in Appendix D and several supplementary experiments summarize the full nine-cell grid. Their human-correspondence and cross-model metrics use

$$\bar{s}^{(k)}_{me} = \frac{1}{9}\sum_{p}\sum_{w=1}^{3} z^{(k)}_{mepw}. \tag{16}$$

We call the mean within-issue Spearman correlation over all $\binom{9}{2} = 36$ configuration pairs *joint configuration agreement*. It describes agreement across the complete persona–wording grid, whereas Equation 12 isolates wording changes within a persona.

Top-K enrichment. For Top-K enrichment and AUROC, let E_m denote the relevant persona-specific or full-grid evaluator composite. We set the top fraction to $\alpha = 0.2$, giving $K = \alpha n_t = 14.4$ for each 72-message issue. Fractional membership handles boundary ties while ensuring that the total membership weight in each top set is exactly 14.4. Let M^H_{mt} and M^E_{mt} be message m's membership weights in the human and evaluator top sets. Their tie-aware overlap is

$$O_t = \sum_{m\in\mathcal{M}_t} \min\big(M^H_{mt}, M^E_{mt}\big). \tag{17}$$

We compute

$$\text{Precision}_{20\%} = \frac{\sum_t O_t}{\sum_t \alpha n_t}, \qquad \text{Top-K enrichment} = \frac{\text{Precision}_{20\%}}{0.2}. \tag{18}$$

An enrichment of one is the random-overlap expectation.

Top-versus-bottom AUROC. Within each issue, the top 25% of messages by human effect are positive cases, the bottom 25% are negative cases, and the middle 50% do not enter this metric. Boundary ties again receive fractional membership weights. For a positive message h and negative message l, let

$$a_{hl} = \begin{cases} 1, & E_h > E_l, \\ \frac{1}{2}, & E_h = E_l, \\ 0, & E_h < E_l. \end{cases} \tag{19}$$

If $M_{ht}^{H,+}$ and $M_{lt}^{H,-}$ denote their fractional human top- and bottom-quartile memberships, then

$$\text{AUROC} = \frac{\sum_t \sum_{h,l} M_{ht}^{H,+} M_{lt}^{H,-} a_{hl}}{\sum_t \sum_{h,l} M_{ht}^{H,+} M_{lt}^{H,-}}. \tag{20}$$

Thus AUROC measures whether messages in the human top quartile receive higher scores than those in the human bottom quartile; 0.5 denotes chance.

D Robustness and Transfer Details

These analyses test changes to reading frame, evaluator model, dataset, and probe panel using the full nine-cell persona–wording grid. Full-grid composites and joint configuration agreement are defined in Appendix C.

D.1 Direct and Stepwise Framing

We isolate the framing axis for Llama-3.1-8B. Within each framing condition, we standardize every configuration over all 720 messages and average the same nine matched persona–wording cells, using the composite in Equation 16. WPC uses the HC2 uncertainty weights from Section C.1; all 95% confidence intervals resample the ten issues 2,000 times. Joint configuration agreement is computed over all 36 configuration pairs within each framing condition.

Table 6: Llama-3.1-8B results under direct and stepwise framing. Entries are point estimates with 95% issue-bootstrap confidence intervals. Joint configuration agreement averages all 36 configuration pairs within each frame. Bold highlights CSPE.

Frame	Method	ρ	WPC	Joint configuration agreement	Top-K
Direct	Likert	0.160 [0.084, 0.240]	0.567 [0.529, 0.604]	0.316 [0.258, 0.369]	0.979 [0.813, 1.156]
Direct	Rubric score	0.231 [0.144, 0.322]	0.595 [0.555, 0.635]	0.532 [0.461, 0.600]	0.921 [0.708, 1.097]
Direct	**CSPE**	**0.310** [**0.230**, **0.389**]	**0.636** [**0.600**, **0.671**]	**0.953** [**0.935**, **0.967**]	**1.125** [**0.819**, **1.438**]
Stepwise	Likert	0.111 [0.044, 0.177]	0.542 [0.510, 0.576]	0.341 [0.286, 0.396]	0.778 [0.503, 1.045]
Stepwise	Rubric score	0.200 [0.117, 0.286]	0.583 [0.544, 0.625]	0.541 [0.480, 0.597]	0.898 [0.598, 1.201]
Stepwise	**CSPE**	**0.306** [**0.231**, **0.384**]	**0.634** [**0.600**, **0.668**]	**0.962** [**0.949**, **0.973**]	**1.139** [**0.847**, **1.451**]

Table 6 shows small changes in CSPE's composite human correspondence under the stepwise-reading prefix: direct-to-stepwise changes are -0.005 in Spearman, -0.002 in WPC, and $+0.014$ in top-K enrichment. Paired issue-bootstrap intervals include zero for these three differences. Joint configuration agreement increases by 0.010 [95% CI: 0.005, 0.016]. Under the tested framing perturbation, absolute changes across the four metrics are at most 0.014.

The elicited-response baselines show larger declines in human correspondence. Stepwise framing reduces Likert Spearman by 0.050 [95% CI: -0.110, -0.001] and WPC by 0.025 [-0.0523, -0.0001]; it reduces the Rubric score's Spearman by 0.031 [-0.052, -0.006] and WPC by 0.012 [-0.021, -0.002]. All four paired intervals exclude zero. Likert top-K enrichment also falls by 0.202 [-0.396, -0.0005], while top-K enrichment for the Rubric score changes by -0.023. Under this framing perturbation, the Likert and Rubric score composites have lower correspondence with aggregate human outcomes, while CSPE's correspondence remains similar.

D.2 Cross-Evaluator Agreement

We next assess the association between persuasiveness scores from different evaluator models. We use the nine-cell composite in Equation 16. For each evaluator pair and method, we compute Spearman correlation between the two composite score vectors over the 72 messages in each issue. The

ten issue correlations are pooled on the Fisher-z scale,

$$\rho_{e,e'}^{(k)} = \tanh\left[\frac{1}{10}\sum_{t=1}^{10}\operatorname{arctanh}\Big(\rho_{S,t}^{(k)}(e,e')\Big)\right]. \tag{21}$$

The 95% confidence intervals resample the ten issues 2,000 times. Comparisons between CSPE and a baseline use the same resampled issues, yielding paired intervals for the difference in pooled correlations.

Table 7: Cross-evaluator score agreement. Entries are Fisher-z pooled within-issue Spearman correlations with 95% issue-bootstrap confidence intervals. L2, L3, and GPT denote Llama-2-7B-Chat, Llama-3.1-8B, and GPT-OSS-120B. Bold highlights the proposed estimator; token length is the invariance control.

Evaluator pair	Likert	Rubric score	Token length	**CSPE**
L2–L3	0.074 [−0.008, 0.149]	0.158 [0.126, 0.193]	0.9943 [0.9923, 0.9958]	**0.647** [**0.536**, **0.730**]
L2–GPT	0.034 [−0.074, 0.138]	0.145 [0.117, 0.169]	0.9941 [0.9921, 0.9955]	**0.531** [**0.395**, **0.641**]
L3–GPT	0.168 [0.093, 0.235]	0.246 [0.148, 0.332]	0.9995 [0.9994, 0.9996]	**0.436** [**0.321**, **0.549**]

Table 7 shows that CSPE's cross-evaluator agreement is moderate to strong (0.436–0.647), well above Likert (0.034–0.168) and the Rubric score (0.145–0.246). CSPE's gains over Likert and the Rubric score are respectively 0.573 and 0.488 for L2–L3, 0.497 and 0.386 for L2–GPT, and 0.268 and 0.190 for L3–GPT. All six paired 95% confidence intervals lie above zero; the smallest contrast is CSPE minus the Rubric score for L3–GPT, 0.190 [0.012, 0.363].

Token length's cross-evaluator Spearman correlation is 0.9941–0.9995.

D.3 ConvinceQA Stability Transfer

We next examine agreement across the persona–wording grid on a second dataset. ConvinceQA was introduced to study prior-belief bias in persuasion judgments and contains controlled supporting and opposing arguments across subjective, misinformation, and harmful claims (Zahraei et al., 2026). We use its test split. After removing one malformed claim, the evaluation contains 1,263 claims and 5,052 messages: 473 subjective claims (1,892 messages), 442 misinformation claims (1,768 messages), and 348 harmful claims (1,392 messages). Each claim contributes supporting, strongly supporting, opposing, and strongly opposing messages. ConvinceQA does not provide human-outcome labels, so we use it to evaluate stability transfer through reproducibility across configurations.

Each dataset claim is paired with a balanced set of third-party pro and con probes and one neutral reference shared by its four messages. For a message advocating direction $y_m \in \{-1, +1\}$, we use $\mathrm{CSPE}^{\mathrm{aligned}}(m) = y_m \mathrm{CSPE}^{\mathrm{raw}}(m)$, so that larger scores consistently denote stronger support for the message's own stance. We evaluate Llama-3.1-8B and Llama-2-7B-Chat under the same nine persona–wording configurations as the primary experiment, with direct framing fixed throughout. Likert, Rubric score, and token length follow Section 3. For every claim and configuration pair, we compute Spearman correlation between the score vectors for its four messages, assigning zero when either vector is constant. We average first over the 36 configuration pairs and then over claims; 2,000 bootstrap replicates resample whole claims.

Table 8 reports Spearman correlations between scores for each claim's four messages, averaged over all persona–wording configuration pairs. CSPE's agreement is 0.913 for Llama-3.1-8B and 0.910 for Llama-2-7B-Chat, compared with 0.291 and 0.172 for Likert and 0.155 and 0.002 for the Rubric score.

D.4 Probe-Panel Sensitivity

Finally, we test whether the CSPE findings depend on one particular probe panel. These analyses use Llama-3.1-8B with direct framing, all 720 messages and ten issues, and the same nine persona–

Table 8: Within-claim joint configuration agreement on the ConvinceQA test split. Entries average claim-level Spearman agreements over the 36 configuration pairs and all claims, with 95% claim-bootstrap intervals. Bold highlights CSPE; token length is the invariance control.

Method	Llama-3.1-8B	Llama-2-7B-Chat
Likert	0.291 [0.278, 0.305]	0.172 [0.163, 0.182]
Rubric score	0.155 [0.145, 0.166]	0.002 [0.001, 0.004]
Token length	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
CSPE	**0.913 [0.906, 0.921]**	**0.910 [0.902, 0.918]**

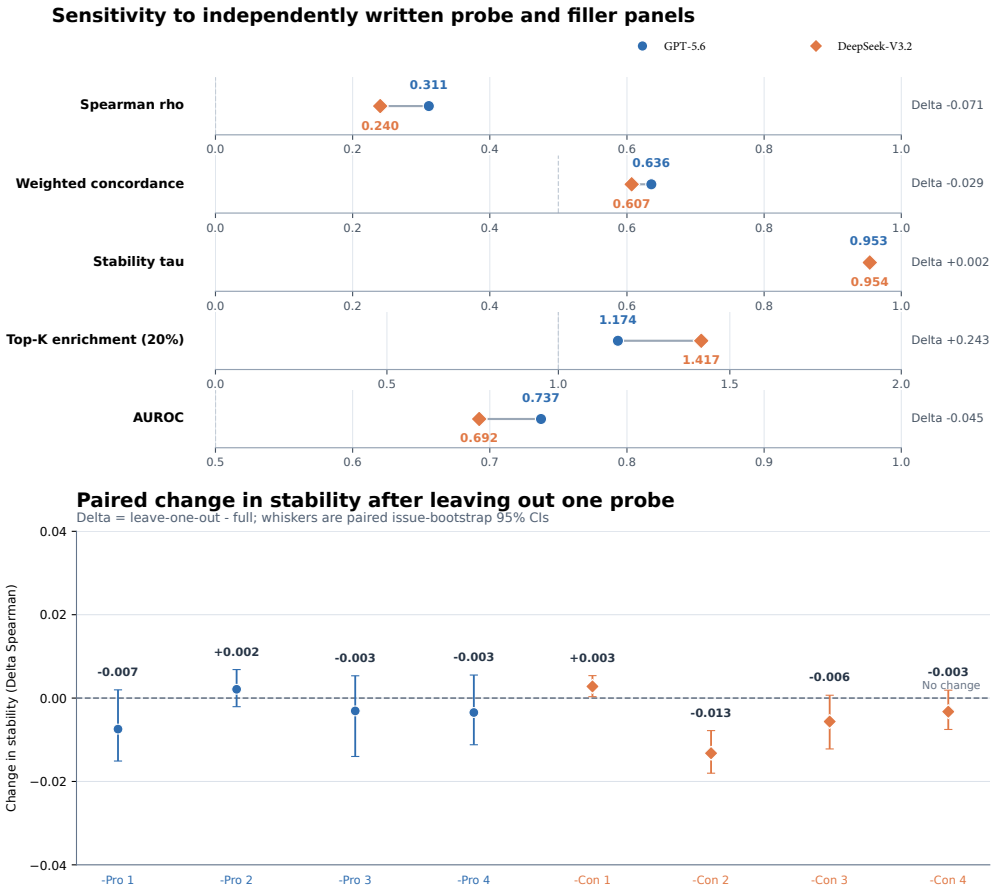

Figure 4: Sensitivity of CSPE to probe-panel composition. *Top*: changes from replacing the original GPT-5.6-generated probe panel with a DeepSeek-V3.2-generated panel (DeepSeek-V3.2 minus GPT-5.6). *Bottom*: changes in joint configuration agreement after leaving out each pro or con probe from the original panel. Whiskers show 95% paired issue-bootstrap intervals; zero denotes no change.

wording configurations. Metrics are computed within issue before aggregation. Spearman, top-K enrichment, and AUROC use the human persuasion-effect point estimates; WPC uses the pairwise ordering probabilities and HC2 uncertainty weights in Equation 11. Joint configuration agreement uses no human outcomes: within each issue it averages Spearman agreement over all 36 pairs of persona–wording configurations. All difference intervals use 2,000 paired issue-cluster bootstrap replicates.

The top panel of Figure 4 replaces the original GPT-5.6-generated probe panel with a panel generated directly by DeepSeek-V3.2; the corresponding neutral references are regenerated at the same time. Joint configuration agreement changes from 0.953 to 0.954. Human Spearman decreases

from 0.311 to 0.240, WPC from 0.636 to 0.607, and AUROC from 0.737 to 0.692, while top-20% enrichment increases from 1.174 to 1.417. All five paired 95% intervals for the difference (DeepSeek-V3.2 minus GPT-5.6) include zero. The regenerated neutral references contribute constant within-issue offsets, so these rank-based comparisons isolate changes in the probe panel. The sensitivity analysis therefore separates highly stable configuration agreement from greater variation in human-correspondence estimates across independently generated probe panels.

The bottom panel asks a narrower question by dropping one of the eight stance probes at a time while holding the rest of the measurement panel fixed. Relative to full-panel joint configuration agreement of 0.9525, six of eight paired intervals include zero. Dropping Con 1 produces a small increase of 0.0028 [95% CI: 0.0003, 0.0054], while dropping Con 2 produces the largest decline, −0.0132 [−0.0180, −0.0078]. All remaining point changes lie between −0.0074 and 0.0021.

E Complete Persona-Specific Results

Table 9: Complete persona-specific results on 720 messages across ten policy issues. Parentheses contain 95% issue-cluster bootstrap confidence intervals. Wording stability ρ_{cfg} is computed across wording variants with evaluator model and persona fixed. Human-correspondence metrics use the same aggregate human criterion for all persona conditions.

Evaluator	Method	Persona	Spearman ρ	WPC	Wording stability ρ_{cfg}	Top-K enrichment	AUROC
Llama-2	Likert	no persona	0.051 (−0.019, 0.118)	0.502 (0.492, 0.512)	0.072 (0.000, 0.217)	1.040 (1.014, 1.063)	0.508 (0.494, 0.523)
		left	−0.014 (−0.109, 0.075)	0.489 (0.459, 0.513)	0.074 (0.000, 0.202)	0.990 (0.882, 1.064)	0.492 (0.445, 0.526)
		right	0.006 (−0.059, 0.076)	0.498 (0.485, 0.512)	0.073 (0.000, 0.196)	1.017 (0.976, 1.053)	0.500 (0.481, 0.520)
	Rubric score	no persona	0.080† (0.000, 0.111)	0.501 (0.500, 0.503)	0.200 (0.000, 0.500)	1.045 (1.042, 1.049)	0.503 (0.500, 0.508)
		left	0.049† (−0.012, 0.107)	0.502 (0.500, 0.506)	0.300 (0.000, 0.600)	1.050 (1.046, 1.055)	0.506 (0.500, 0.514)
		right	0.049† (−0.012, 0.107)	0.502 (0.500, 0.506)	0.300 (0.000, 0.600)	1.050 (1.046, 1.055)	0.506 (0.500, 0.514)
	CSPE	no persona	0.319 (0.197, 0.445)	0.641 (0.588, 0.693)	0.984 (0.980, 0.988)	1.271 (0.903, 1.674)	0.722 (0.651, 0.797)
		left	0.320 (0.207, 0.429)	0.642 (0.594, 0.689)	0.986 (0.982, 0.990)	1.361 (0.972, 1.785)	0.723 (0.655, 0.796)
		right	0.315 (0.206, 0.426)	0.641 (0.594, 0.689)	0.985 (0.980, 0.989)	1.285 (0.910, 1.694)	0.722 (0.657, 0.794)
Llama-3.1	Likert	no persona	0.156 (0.103, 0.203)	0.557 (0.534, 0.579)	0.349 (0.281, 0.410)	0.862 (0.765, 0.945)	0.601 (0.567, 0.630)
		left	0.064 (−0.042, 0.190)	0.519 (0.474, 0.574)	0.397 (0.318, 0.464)	0.826 (0.644, 1.036)	0.551 (0.484, 0.636)
		right	0.214 (0.120, 0.296)	0.592 (0.546, 0.629)	0.478 (0.396, 0.547)	1.160 (0.858, 1.468)	0.645 (0.572, 0.707)
	Rubric score	no persona	0.270 (0.188, 0.348)	0.602 (0.567, 0.634)	0.606 (0.543, 0.659)	1.146 (0.955, 1.337)	0.674 (0.626, 0.723)
		left	0.139 (0.015, 0.271)	0.545 (0.496, 0.600)	0.699 (0.655, 0.748)	0.903 (0.629, 1.156)	0.591 (0.511, 0.682)
		right	0.273 (0.188, 0.352)	0.613 (0.572, 0.649)	0.599 (0.539, 0.654)	1.107 (0.866, 1.283)	0.688 (0.633, 0.736)
	CSPE	no persona	0.340 (0.251, 0.426)	0.648 (0.609, 0.685)	0.978 (0.968, 0.985)	1.062 (0.750, 1.368)	0.759 (0.696, 0.819)
		left	0.289 (0.203, 0.370)	0.626 (0.589, 0.664)	0.976 (0.969, 0.982)	1.118 (0.833, 1.424)	0.721 (0.660, 0.786)
		right	0.292 (0.216, 0.363)	0.628 (0.595, 0.659)	0.979 (0.974, 0.984)	1.181 (0.889, 1.472)	0.726 (0.672, 0.781)
GPT-OSS	Likert	no persona	0.184 (0.107, 0.251)	0.575 (0.540, 0.606)	0.255 (0.214, 0.295)	1.230 (0.930, 1.489)	0.618 (0.554, 0.680)
		left	0.206 (0.111, 0.281)	0.589 (0.543, 0.624)	0.300 (0.258, 0.344)	1.195 (0.918, 1.417)	0.667 (0.597, 0.729)
		right	0.248 (0.153, 0.330)	0.607 (0.563, 0.646)	0.313 (0.275, 0.351)	1.196 (0.972, 1.422)	0.688 (0.623, 0.749)
	Rubric score	no persona	0.166 (0.103, 0.228)	0.569 (0.543, 0.598)	0.264 (0.218, 0.320)	1.158 (0.972, 1.331)	0.622 (0.568, 0.671)
		left	0.163 (0.131, 0.189)	0.575 (0.557, 0.589)	0.371 (0.322, 0.426)	1.152 (0.907, 1.360)	0.640 (0.607, 0.669)
		right	0.148 (0.099, 0.195)	0.566 (0.542, 0.586)	0.380 (0.356, 0.402)	1.085 (0.830, 1.313)	0.619 (0.567, 0.665)
	CSPE	no persona	0.332 (0.181, 0.468)	0.645 (0.579, 0.702)	0.879 (0.851, 0.905)	1.410 (1.014, 1.771)	0.708 (0.608, 0.795)
		left	0.300 (0.156, 0.418)	0.632 (0.567, 0.683)	0.835 (0.772, 0.879)	1.236 (0.868, 1.632)	0.685 (0.570, 0.788)
		right	0.290 (0.146, 0.407)	0.628 (0.565, 0.680)	0.844 (0.790, 0.888)	1.264 (0.882, 1.625)	0.680 (0.559, 0.785)
DeBERTa-v3-large	NLI	invariant	0.209 (0.097, 0.329)	0.592 (0.539, 0.644)	1.000‡ (1.000, 1.000)	0.986 (0.632, 1.299)	0.658 (0.583, 0.736)
all-mpnet-base-v2	Semantic similarity	invariant	0.129 (0.002, 0.258)	0.558 (0.501, 0.615)	1.000‡ (1.000, 1.000)	0.958 (0.757, 1.160)	0.577 (0.475, 0.682)

Continued on the next page

Table 10: Supplementary joint configuration summaries. Human-correspondence metrics use the nine-cell composite in Equation 16; joint configuration agreement averages all 36 persona–wording configuration pairs. Entries are point estimates with 95% issue-bootstrap confidence intervals. Bold marks the best point estimate within an evaluator. Chance is 0.5 for WPC and AUROC, and 1 for top-K enrichment.

Method	Spearman ρ	WPC	Joint configuration agreement	Top-K enrichment	AUROC
Llama-2-7B-Chat					
Likert	-0.002 (-0.093, 0.078)	0.492 (0.462, 0.515)	0.116 (0.038, 0.258)	0.989 (0.879, 1.067)	0.496 (0.450, 0.529)
Rubric score	0.049 (-0.012, 0.107)	0.502 (0.500, 0.506)	0.250 (0.017, 0.506)	1.050 (1.046, 1.055)	0.506 (0.500, 0.514)
Token length	0.140 (0.056, 0.215)	0.559 (0.518, 0.596)	**1.000** (**1.000**, **1.000**)	0.983 (0.625, 1.361)	0.592 (0.531, 0.656)
CSPE	**0.320** (**0.206**, **0.434**)	**0.643** (**0.593**, **0.691**)	0.975 (0.969, 0.980)	**1.236** (**0.847**, **1.681**)	**0.724** (**0.656**, **0.799**)
Llama-3.1-8B					
Likert	0.160 (0.084, 0.240)	0.567 (0.529, 0.604)	0.316 (0.258, 0.369)	0.979 (0.813, 1.156)	0.613 (0.554, 0.677)
Rubric score	0.231 (0.144, 0.322)	0.595 (0.555, 0.635)	0.532 (0.461, 0.600)	0.921 (0.708, 1.097)	0.676 (0.614, 0.736)
Token length	0.132 (0.049, 0.206)	0.556 (0.514, 0.592)	**1.000** (**1.000**, **1.000**)	1.066 (0.778, 1.382)	0.587 (0.527, 0.650)
CSPE	**0.310** (**0.230**, **0.389**)	**0.636** (**0.600**, **0.671**)	0.953 (0.935, 0.967)	**1.125** (**0.819**, **1.438**)	**0.737** (**0.680**, **0.798**)
GPT-OSS-120B					
Likert	0.255 (0.147, 0.339)	0.609 (0.559, 0.648)	0.288 (0.257, 0.319)	**1.410** (**1.132**, **1.674**)	0.691 (0.612, 0.758)
Rubric score	0.202 (0.160, 0.244)	0.592 (0.573, 0.611)	0.344 (0.315, 0.377)	1.212 (0.934, 1.458)	0.667 (0.625, 0.707)
Token length	0.135 (0.051, 0.211)	0.557 (0.515, 0.593)	**1.000** (**1.000**, **1.000**)	1.067 (0.799, 1.377)	0.589 (0.528, 0.652)
CSPE	**0.317** (**0.169**, **0.443**)	**0.639** (**0.574**, **0.693**)	0.814 (0.757, 0.861)	1.285 (0.965, 1.563)	**0.695** (**0.583**, **0.794**)

Table 9: Complete persona-specific results (continued).

Evaluator	Method	Persona	Spearman ρ	WPC	Wording stability ρ_{cfg}	Top-K enrichment	AUROC

† Some issue-level Spearman correlations are undefined because the Rubric scores are constant; these correlations are assigned zero before pooling. ‡ The value 1.000 is structural: the same scorer output is used in all nine configuration slots.

Table 9 expands the primary results in Table 2. It reports all five metrics and their 95% bootstrap confidence intervals for each evaluator, method, and persona. As in the main analysis, Spearman correlation, WPC, top-K enrichment, and AUROC use the three-wording composite within a fixed persona; wording stability holds the evaluator and persona fixed and compares the three wording variants.

F Supplementary Summaries Across the Evaluator Grid

The main results hold persona fixed when measuring wording stability and computing human correspondence. The robustness analyses in Appendix D and supplementary experiments use the full nine-cell composite in Equation 16 and joint configuration agreement across all 36 configuration pairs. Table 10 records these summaries, together with token length, top-K enrichment, and AUROC. Token length is invariant to instruction wording and has weak human correspondence, illustrating the complementary roles of stability and human-outcome validation.

Paired differences against the strongest elicited-response baseline. Table 11 reports paired CSPE-minus-baseline differences and their 95% issue-clustered bootstrap intervals for each evaluator.

Table 11: Paired differences in human correspondence between CSPE and the strongest elicited-response baseline for each evaluator. Differences use the full nine-cell composite in Equation 16. Confidence intervals are obtained from the paired issue-clustered bootstrap; positive values favor CSPE. The strongest baseline is the Rubric score for Llama-2 and Llama-3.1, and Likert for GPT-OSS.

Evaluator	Δ Spearman ρ [95% CI]	Δ WPC [95% CI]
Llama-2-7B-Chat	0.272 [0.115, 0.429]	0.140 [0.083, 0.191]
Llama-3.1-8B	0.079 [−0.026, 0.185]	0.041 [−0.010, 0.088]
GPT-OSS-120B	0.062 [−0.060, 0.169]	0.030 [−0.025, 0.075]

Table 12: Resolution-controlled comparison of native argmax, continuous soft-label, and CSPE readouts. Stability is the joint configuration agreement averaged over all 36 persona–wording configuration pairs; human correspondence is measured by within-issue Spearman ρ. Brackets give 95% issue-bootstrap confidence intervals. Bold highlights CSPE.

Evaluator	Method	Stability	Human correspondence ρ
Llama-2-7B-Chat	Likert	0.116 [0.036, 0.256]	−0.002 [−0.096, 0.077]
	Soft Likert	0.517 [0.486, 0.551]	0.028 [−0.144, 0.132]
	Rubric score	0.250 [0.014, 0.508]	0.049 [−0.011, 0.106]
	Soft Rubric score	0.699 [0.653, 0.740]	0.030 [−0.119, 0.132]
	CSPE	**0.975 [0.969, 0.980]**	**0.320 [0.217, 0.441]**
Llama-3.1-8B	Likert	0.316 [0.257, 0.369]	0.160 [0.084, 0.239]
	Soft Likert	0.584 [0.500, 0.661]	0.237 [0.139, 0.330]
	Rubric score	0.532 [0.463, 0.595]	0.231 [0.146, 0.324]
	Soft Rubric score	0.769 [0.714, 0.819]	0.273 [0.169, 0.376]
	CSPE	**0.953 [0.934, 0.966]**	**0.310 [0.233, 0.391]**
GPT-OSS-120B	Likert	0.288 [0.255, 0.318]	0.255 [0.148, 0.347]
	Soft Likert	0.586 [0.559, 0.612]	0.267 [0.170, 0.359]
	Rubric score	0.344 [0.316, 0.378]	0.202 [0.160, 0.244]
	Soft Rubric score	0.562 [0.535, 0.590]	0.188 [0.138, 0.245]
	CSPE	**0.814 [0.762, 0.861]**	**0.317 [0.175, 0.451]**

G Resolution-Controlled Elicited-Response Baselines

We examine how retaining the five-label likelihood distribution affects the stability and human correspondence of Likert and Rubric scores. We replace each baseline's hard label selection with a continuous readout from this distribution.

For normalized sequence log likelihoods $\ell_1, \ldots, \ell_5$, we define

$$p_k = \frac{\exp(\ell_k)}{\sum_{j=1}^{5} \exp(\ell_j)}, \qquad \widetilde{A} = \sum_{k=1}^{5} k p_k. \tag{22}$$

The soft Rubric score uses $\widetilde{A}$ directly. Soft Likert likewise replaces the native Likert baseline's argmax rating with its corresponding probability-weighted mean. All messages, evaluator checkpoints, prompts, persona and wording configurations, and evaluation metrics are otherwise unchanged. We compute joint configuration agreement over all 36 pairs of the nine persona–wording configurations, and difference intervals use paired 2,000-repetition issue bootstraps.

Table 12 shows that continuous soft-label readouts increase joint configuration agreement for both baselines. Across the three evaluators, CSPE exceeds soft Likert by 0.228–0.458 and soft Rubric by 0.184–0.276, with all paired 95% confidence intervals above zero.

H Matched-Quartet Construct-Validation Details

We construct a 2×2 matched-quartet dataset about renewable-energy subsidies. Fifty sets cross stance strength (high versus low) with persuasion level (high versus low), yielding 200 texts in four

Table 13: Condition means for the matched-quartet construct validation. CSPE and NLI use different numerical scales.

Method	HH	HL	LH	LL
CSPE	0.8981	−1.4905	−0.3363	−1.9118
NLI	0.9431	0.9742	0.5102	−0.2683

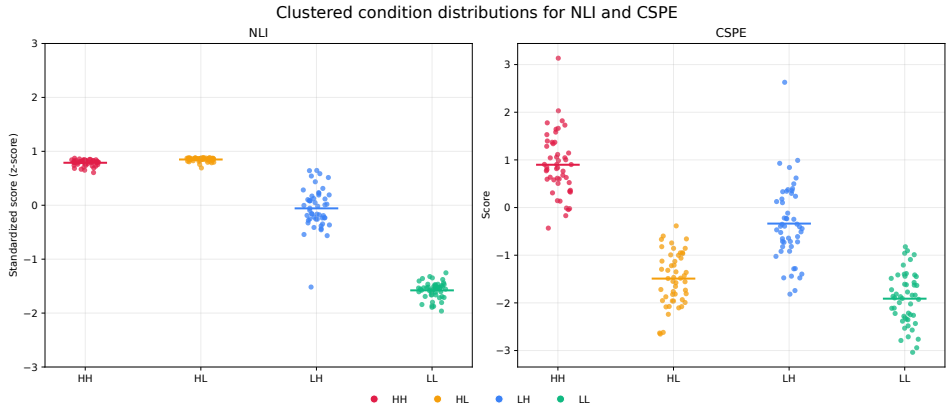

Figure 5: Comparison of standardized NLI and CSPE scores across condition clusters.

conditions: high stance/high persuasion (HH), high stance/low persuasion (HL), low stance/high persuasion (LH), and low stance/low persuasion (LL). Within each set, HH and LH share the same high-persuasion core and differ only in their stance prefix; HL and LL analogously share the same low-persuasion core. All four texts target the same claim and therefore use the same pro probes, con probes, and neutral reference.

We evaluate CSPE with Llama-3.1-8B under the same nine persona–wording configurations. Each configuration is standardized against the 720-message reference distribution and the nine standardized scores are averaged. As a semantic-entailment reference, we use `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` (Laurer et al., 2023): each text is the premise, each stance probe is a hypothesis, and the NLI score is mean pro entailment probability minus mean con entailment probability. All comparisons remain within matched sets. We define the stance main effect as $[(\mathrm{HH} - \mathrm{LH}) + (\mathrm{HL} - \mathrm{LL})]/2$, the persuasion main effect as $[(\mathrm{HH} - \mathrm{HL}) + (\mathrm{LH} - \mathrm{LL})]/2$, and the interaction as $\mathrm{HH} - \mathrm{HL} - \mathrm{LH} + \mathrm{LL}$. Confidence intervals use 10,000 matched-set bootstrap replicates.

Figure 5 shows that the NLI reference is driven mainly by stance strength: it does not distinguish high- from low-persuasion texts under strong stance and even scores HL above LH. In contrast, CSPE clearly separates the four conditions, assigning higher scores to more persuasive texts at both stance levels.

Table 13 gives the condition means underlying the factorial effects in Table 4. For CSPE, the simple persuasion contrast is $\mathrm{HH} - \mathrm{HL} = 2.3886$ at high stance and $\mathrm{LH} - \mathrm{LL} = 1.5755$ at low stance. For NLI, the corresponding contrasts are -0.0311 and 0.7785, explaining the opposite interaction signs reported in the main text.

I Complete Component-Ablation Results

The component analysis evaluates persuasiveness scores over the pooled message set. Each benchmark issue contains one target claim, whose neutral-reference term is a constant additive offset across all messages in that issue. Consequently, the primary within-issue rank-based metrics, including Spearman correlation and WPC, are invariant to neutral-reference subtraction.

Pooling all 720 messages places multiple target claims in the same comparison, so their claim-specific neutral terms no longer cancel. This diagnostic therefore assesses cross-claim offset correction.

For the supplementary global AUROC diagnostic, macro-averages across the nine evaluator–persona combinations are 0.620 for Full, 0.548 for Pro only, 0.514 for Con only, and 0.569 without neutral reference. The respective changes from Full are -0.072, -0.106, and -0.052.

For every ablation variant, Spearman, WPC, wording stability, top-K enrichment, and AUROC are computed over the pooled 720-message set. The pooled set replaces the issue-specific message set in each definition: Spearman is computed once without Fisher-z pooling; wording stability correlates each pair of wording-specific score vectors over all 720 messages; top-K uses $K = 144$; AUROC uses the global top and bottom quartiles; and WPC includes all pooled message pairs. Issue-clustered bootstrap replicates continue to resample the ten complete issues. The full, Pro-only, Con-only, and without-neutral-reference variants share this pooled evaluation scope, which differs from the primary within-issue analysis.

Table 14 reports the evaluator- and persona-specific results underlying the macro-averages in Table 5. All entries use the global 720-message analysis defined above. Human ρ, WPC, AUROC, and top-K enrichment measure global correspondence with the Hackenburg human outcomes; wording stability measures global agreement across wording variants.

Table 14: Complete global component-ablation results on the Hackenburg benchmark. Each metric is computed over the pooled 720-message set and is reported as a point estimate followed by its 95% confidence interval in brackets.

Model	Persona	Variant	Human ρ	WPC	Global wording stability	AUROC	Top-K enrichment
Llama-3.1-8B	no persona	Full	0.391 [0.266, 0.489]	0.674 [0.619, 0.714]	0.977 [0.953, 0.990]	0.631 [0.588, 0.666]	1.528 [0.972, 2.257]
		Pro only	0.170 [−0.011, 0.326]	0.572 [0.489, 0.644]	0.989 [0.978, 0.994]	0.556 [0.494, 0.609]	1.146 [0.521, 1.702]
		Con only	0.013 [−0.163, 0.196]	0.510 [0.428, 0.594]	0.988 [0.976, 0.993]	0.505 [0.446, 0.567]	0.938 [0.382, 1.458]
		Without neutral reference	0.186 [0.020, 0.384]	0.581 [0.502, 0.671]	0.997 [0.990, 0.998]	0.563 [0.506, 0.631]	1.562 [0.660, 2.396]
Llama-3.1-8B	left	Full	0.346 [0.232, 0.426]	0.653 [0.605, 0.685]	0.979 [0.970, 0.987]	0.615 [0.577, 0.644]	1.389 [0.868, 1.840]
		Pro only	0.121 [−0.043, 0.280]	0.551 [0.473, 0.624]	0.991 [0.981, 0.994]	0.540 [0.484, 0.593]	1.076 [0.486, 1.597]
		Con only	0.040 [−0.131, 0.203]	0.521 [0.442, 0.595]	0.987 [0.976, 0.993]	0.513 [0.455, 0.569]	1.042 [0.451, 1.493]
		Without neutral reference	0.150 [−0.002, 0.340]	0.565 [0.494, 0.651]	0.997 [0.990, 0.998]	0.551 [0.499, 0.615]	1.250 [0.624, 2.014]
Llama-3.1-8B	right	Full	0.317 [0.194, 0.420]	0.640 [0.587, 0.687]	0.981 [0.971, 0.988]	0.605 [0.564, 0.642]	1.458 [0.799, 1.876]
		Pro only	0.112 [−0.047, 0.280]	0.546 [0.472, 0.624]	0.990 [0.979, 0.994]	0.537 [0.483, 0.593]	1.007 [0.486, 1.632]
		Con only	0.035 [−0.124, 0.194]	0.519 [0.446, 0.590]	0.983 [0.969, 0.992]	0.511 [0.458, 0.564]	0.938 [0.451, 1.354]
		Without neutral reference	0.171 [0.003, 0.349]	0.574 [0.496, 0.656]	0.996 [0.985, 0.998]	0.557 [0.500, 0.618]	1.424 [0.659, 2.083]
Llama-2-7B-Chat	no persona	Full	0.393 [0.211, 0.541]	0.679 [0.596, 0.743]	0.993 [0.986, 0.995]	0.635 [0.571, 0.687]	1.875 [1.181, 2.604]
		Pro only	0.181 [−0.004, 0.329]	0.580 [0.495, 0.645]	0.986 [0.978, 0.991]	0.560 [0.498, 0.611]	1.354 [0.694, 1.944]
		Con only	0.072 [−0.029, 0.182]	0.534 [0.488, 0.586]	0.986 [0.977, 0.991]	0.524 [0.491, 0.561]	1.007 [0.556, 1.458]
		Without neutral reference	0.184 [0.004, 0.393]	0.582 [0.498, 0.672]	0.997 [0.992, 0.998]	0.562 [0.501, 0.633]	1.562 [0.660, 2.326]
Llama-2-7B-Chat	left	Full	0.418 [0.243, 0.539]	0.692 [0.613, 0.743]	0.993 [0.988, 0.995]	0.644 [0.582, 0.686]	1.910 [1.319, 2.604]
		Pro only	0.081 [−0.083, 0.233]	0.536 [0.458, 0.603]	0.986 [0.978, 0.991]	0.527 [0.471, 0.579]	1.215 [0.590, 1.667]
		Con only	0.201 [0.116, 0.276]	0.591 [0.552, 0.628]	0.988 [0.981, 0.993]	0.567 [0.539, 0.593]	1.319 [0.938, 1.701]
		Without neutral reference	0.181 [0.002, 0.388]	0.581 [0.498, 0.674]	0.998 [0.994, 0.999]	0.561 [0.500, 0.633]	1.562 [0.660, 2.326]
Llama-2-7B-Chat	right	Full	0.407 [0.234, 0.535]	0.685 [0.607, 0.740]	0.992 [0.986, 0.995]	0.639 [0.579, 0.684]	1.667 [1.076, 2.500]
		Pro only	0.078 [−0.087, 0.225]	0.534 [0.455, 0.600]	0.984 [0.975, 0.991]	0.526 [0.470, 0.576]	1.111 [0.556, 1.667]
		Con only	0.199 [0.109, 0.272]	0.590 [0.551, 0.625]	0.987 [0.981, 0.992]	0.567 [0.537, 0.591]	1.285 [0.868, 1.633]
		Without neutral reference	0.186 [0.005, 0.390]	0.583 [0.498, 0.673]	0.998 [0.993, 0.998]	0.563 [0.501, 0.632]	1.528 [0.694, 2.326]
GPT-OSS-120B	no persona	Full	0.274 [0.114, 0.418]	0.621 [0.549, 0.684]	0.916 [0.870, 0.943]	0.591 [0.538, 0.641]	1.146 [0.799, 1.736]
		Pro only	0.076 [−0.036, 0.173]	0.533 [0.484, 0.577]	0.877 [0.814, 0.913]	0.525 [0.488, 0.557]	0.833 [0.590, 1.389]

Continued on the next page

Table 14: Complete component-ablation results on Hackenburg (continued).

Model	Persona	Variant	Human ρ	WPC	Global wording stability	AUROC	Top-K enrichment
			0.049	0.522	0.850	0.516	1.007
		Con only	$[-0.037, 0.132]$	$[0.482, 0.559]$	$[0.789, 0.888]$	$[0.487, 0.543]$	$[0.521, 1.389]$
			0.243	0.608	0.959	0.582	1.285
		Without neutral reference	$[0.053, 0.422]$	$[0.519, 0.686]$	$[0.924, 0.969]$	$[0.517, 0.643]$	$[0.625, 2.118]$
			0.329	0.647	0.894	0.611	1.319
GPT-OSS-120B	left	Full	$[0.136, 0.467]$	$[0.562, 0.705]$	$[0.810, 0.927]$	$[0.546, 0.658]$	$[0.903, 2.014]$
			0.242	0.607	0.892	0.579	1.285
		Pro only	$[0.098, 0.355]$	$[0.545, 0.657]$	$[0.854, 0.915]$	$[0.532, 0.616]$	$[0.833, 1.944]$
			−0.113	0.449	0.800	0.463	0.590
		Con only	$[-0.219, -0.014]$	$[0.400, 0.492]$	$[0.740, 0.863]$	$[0.427, 0.496]$	$[0.312, 0.903]$
			0.263	0.616	0.940	0.588	1.319
		Without neutral reference	$[0.061, 0.427]$	$[0.526, 0.687]$	$[0.892, 0.953]$	$[0.520, 0.644]$	$[0.764, 2.049]$
			0.329	0.647	0.893	0.610	1.319
GPT-OSS-120B	right	Full	$[0.127, 0.471]$	$[0.559, 0.706]$	$[0.820, 0.926]$	$[0.542, 0.658]$	$[0.868, 1.945]$
			0.241	0.607	0.886	0.579	1.181
		Pro only	$[0.096, 0.352]$	$[0.545, 0.655]$	$[0.849, 0.907]$	$[0.531, 0.617]$	$[0.833, 1.875]$
			−0.108	0.451	0.807	0.464	0.694
		Con only	$[-0.208, -0.018]$	$[0.405, 0.491]$	$[0.764, 0.855]$	$[0.431, 0.494]$	$[0.347, 0.903]$
			0.271	0.620	0.946	0.590	1.319
		Without neutral reference	$[0.064, 0.435]$	$[0.528, 0.690]$	$[0.906, 0.956]$	$[0.521, 0.645]$	$[0.729, 2.083]$

J SINGLE-CONFIGURATION PERFORMANCE AND AVERAGING

This analysis compares individual configurations with the nine-cell composite reported in Table 10. We compute WPC separately for each configuration, pooling comparisons across the same ten issues, and compare these values with the WPC obtained after averaging all nine standardized configuration scores.

Figure 3 shows that individual Llama-3.1-8B Likert configurations span WPC 0.497–0.579, while configurations of the Rubric score span 0.528–0.602. Score averaging raises WPC relative to the mean of the nine individual configurations by 0.034 [95% CI: 0.011, 0.058] for Likert and 0.030 [0.013, 0.048] for the Rubric score. For GPT-OSS-120B, the corresponding gains are 0.056 [0.030, 0.080] and 0.047 [0.038, 0.057]. All four intervals exclude zero. For Llama-2-7B-Chat, both baselines are at chance, and averaging produces no gain.

CSPE's nine configurations span WPC 0.639–0.645, 0.624–0.647, and 0.609–0.650 for Llama-2-7B-Chat, Llama-3.1-8B, and GPT-OSS-120B, respectively. Its averaging gains are 0.001 and 0.002 for the two Llama evaluators, with both gain intervals including zero, and 0.009 [95% CI: 0.002, 0.017] for GPT-OSS-120B. Even the lowest observed CSPE WPC exceeds the highest observed Likert or Rubric WPC for every evaluator: 0.639 versus 0.502 for Llama-2, 0.624 versus 0.602 for Llama-3.1, and 0.609 versus 0.590 for GPT-OSS.

Token length's nine points coincide, averaging changes nothing, and its WPC remains 0.556–0.559.

Figure 3 measures each configuration's correspondence with aggregate human outcomes. Across the tested grid, individual CSPE configurations have WPC values close to the nine-cell composite. Table 2 provides the complementary comparison of wording stability and human correspondence separately within each persona.

K EVALUATION PROMPTS

The complete input is constructed as follows:

```
[Persona prefix] + [Prompt template] + [MESSAGE]
```

K.1 PERSONA PREFIXES

None: `""`

Left: `Reader context: The reader generally holds left-leaning political values.`

Right: `Reader context: The reader generally holds right-leaning political values.`

K.2 LIKERT PROMPTS

W1: `After reading the message, rate how much the reader would agree with the position it advocates. Use 1 for strongly disagree, 2 for disagree, 3 for neither agree nor disagree, 4 for agree, and 5 for strongly agree. Return only the final number.`

W2: `Read the message and indicate the reader's degree of agreement with its advocated position. Choose 1 (strongly disagree), 2 (disagree), 3 (neither agree nor disagree), 4 (agree), or 5 (strongly agree). Give only the final number.`

W3: `Using the message below, judge how strongly the reader would agree with the position being promoted. The scale is 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, and 5 = strongly agree. Output only the final number.`

K.3 RUBRIC PROMPTS

W1: `Judge the message's overall persuasiveness for the reader. Consider clarity, relevance, quality of support, and likely impact as four criteria. Use 1 for not at all persuasive, 2 for slightly persuasive, 3 for moderately persuasive, 4 for very persuasive, and 5 for extremely persuasive. Return only the final number.`

W2: `Evaluate how persuasive the message would be to the reader, using a holistic judgment of clarity, relevance, quality of support, and likely impact. Choose 1 (not at all persuasive), 2 (slightly persuasive), 3 (moderately persuasive), 4 (very persuasive), or 5 (extremely persuasive). Give only the final number.`

W3: `Rate the message's overall persuasive strength for the reader by considering clarity, relevance, quality of support, and likely impact together. The scale is 1 = not at all persuasive, 2 = slightly persuasive, 3 = moderately persuasive, 4 = very persuasive, and 5 = extremely persuasive. Output only the final number.`

K.4 CSPE PROMPTS

W1: `The following is a passage provided for reading.`

```
Passage: [MESSAGE]
```

W2: `Below is a text presented for reading.`

```
Text: [MESSAGE]
```

W3: `A written passage is supplied below to be read.`

```
Passage: [MESSAGE]
```

Direct: `Reading approach: Read the text directly as presented.`

Stepwise: `Reading approach: Read the text carefully, considering its statements step by step.`