Is It the Audience or the Question? A Neutral-Paraphrase Control for Adjudicating Claimed Moderators of LLM Persona Effects

Anonymous authors
Paper under double-blind review

Abstract

We set out to confirm an appealing hypothesis: that a large language model's confidence shifts more for some audiences than others because the underlying evidence is contested. A control we built to verify this instead dismantled it. Using a comparable-magnitude, identity-free neutral-paraphrase baseline, we show that the apparent audience-specific moderation is largely an artifact: content-neutral rephrasings of the same questions move confidence as much as personas do, and a biased spread statistic (max−min range) had manufactured an apparent equivalence that an unbiased one (ω^2) overturns. What genuinely survives is narrower, and we report it as such: audience personas induce more systematic confidence variation than neutral paraphrases across all seven biomedical models tested, and on a current frontier model (GPT-5.1), with 90% CIs excluding zero, but this effect is biomedicine-specific and is not organized by contestedness. Where a contestedness relationship appears (four of seven models), it is carried by the audience-free condition as often as the persona; in climate the audience-free condition carries it outright (ρ_neutral = +0.44, p = 0.005; persona n.s.), and in biomedicine the persona-vs-neutral gap is inconclusive. The contribution is therefore a control protocol for adjudicating any claimed moderator of LLM behavior, and a cautionary worked example in which the protocol overturned our own headline result, the paper's warning enacted on its own data. We do not claim the trap is pervasive; we show it is real, easy to fall into, and cheap to test for.

1. Introduction

Large language models change their answers depending on who appears to be asking. Recent work documents this in clinical settings: assigning a model a physician decision-style or professional role shifts its action thresholds and expressed confidence under fixed clinical evidence (Klang et al., 2026; Abdullahi et al., 2026). That personas move behavior is established. The open and more consequential question is interpretive: when persona-induced variation is *larger on some inputs than others*, what licenses attributing that structure to the model reasoning about the audience?

This paper is about that question, a control that answers it, and the discipline of reporting only what the control actually shows. The instance we study is an attractive hypothesis we set out to confirm: that audience-persona effects on biomedical confidence are **gated by evidence contestedness** — near-zero where clinicians agree, largest where they disagree. The hypothesis is plausible, has a clean mechanistic story (audience conditioning has "room to act" only where the model represents multiple interpretations), and produces a striking correlation. It is the kind of result that gets published.

A single control complicates it. To test the hypothesis we built a comparable-magnitude, identity-free neutral baseline: a content-neutral rephrasing of the same questions, with no audience invoked. The variation we attributed to audience-gating is reproduced — and in one domain *exceeded* — by that rephrasing, so the apparent contestedness-gating does not survive as an audience-specific effect. But the picture is not uniform, and the contribution of this paper is partly in *not* overclaiming the uniformity: audience effects do not vanish,

and what changes is their interpretation rather than their existence. We therefore make a scoped claim, and we make the protocol — not the dissolution — the headline.

Concretely, the contribution is:

- **A protocol.** A neutral-paraphrase control for testing a claimed moderator of prompt-conditioned behavior: rephrase the same inputs with no adopted identity, then compare persona-induced variation against the neutral baseline, range-against-range and ρ-against-ρ on matched claims. We position it as a refinement of existing single-point controls (ğ2), not a new idea.
- **A demonstration.** Applied to contestedness-gating, an intuitive and publishable moderator, the protocol overturns it: the gating is not audience-specific in climate and not established as audience-specific in biomedicine.
- **An honest positive.** Audience effects survive the control, but their interpretation changes: a magnitude effect persists while its organization by contestedness does not.

In detail, the worked example splits by domain. In **climate**, the neutral condition tracks contestedness *more* strongly than the persona condition, so the moderator is demonstrably not audience-specific. In **biomedicine**, the contestedness signal appears on four of seven models (mostly audience-favoring); what biomedicine licenses, once the biased max−min is replaced by ω^2, is that personas produce 2–15× more systematic variation than neutral (no equivalence) while the persona-vs-neutral *correlation* gap stays inconclusive — so biomedicine settles neither direction.

We frame the central claim associationally. We do not show personas have no effect; we show that an observed correlation between a semantic property (contestedness) and behavioral variation is not, by itself, evidence the model processes that property, because a comparable-magnitude neutral perturbation can produce the same correlation. Our results suggest that claims of semantic moderation should routinely be checked against comparable neutral perturbations. We validate this protocol on a single worked example, audience-conditioning gated by contestedness, and do not claim the finding itself generalizes to other moderators.

2. Related work and positioning

Persona conditioning in clinical LLMs. The two closest concurrent studies inject persona as *role-adoption* via a system-level "You are a {persona}" declaration. Klang et al. (2026) define physician decision-style personas, evaluate twenty open-weight models on 5,000 emergency and discharge instances, and report persona prompts shift affirmative decision rates under fixed evidence. Abdullahi et al. (2026) study professional roles and bold/cautious styles across triage and patient-safety tasks, finding systematic, non-monotonic, context-dependent effects. Our design differs on the conditioning axis: rather than instructing the model to *be* a role, we vary *who is asking* ("a researcher / clinician / patient / regulator is asking"). **This is a deliberate scope restriction, not a coverage claim:** our results speak to *audience-conditioning*, and we do not generalize them to role-adoption, which may engage different mechanisms.

On the novelty of controlling for generic prompting. Abdullahi et al. already include generic-prompt controls — a "Helpful Assistant" persona and a "No Persona" condition — to distinguish clinically grounded persona effects from generic assistant behavior. We therefore do *not* claim to be first to control for generic prompting. Our null differs in two respects, both necessary for our test. First, prior baselines are *single points*; a single neutral point can show personas differ from a generic baseline but cannot estimate how much variation *any* surface rephrasing induces, which requires a *distribution* of neutral rephrasings. Our test is range-vs-range and is impossible with a point baseline. Second, prior neutral conditions still adopt an *identity* ("You are a helpful assistant"), whereas our null invokes no "you are" at all and varies only question framing, isolating audience identity specifically. We position the null as a refinement of existing controls, and locate novelty in (i) the moderator tested and (ii) the adjudication result.

Moderators of persona effects. Klang et al. report no moderator analysis; Abdullahi et al.'s moderator is *task acuity*. Neither tests whether persona effects track *evidence contestedness*. The contestedness moderator is novel against both.

Consistency and calibration. Abdullahi et al.'s "consistency rate" compares logit-implied to generated labels within a single prompt — latent-vs-surface alignment, not output spread across surface variations. Our neutral-range is a different construct: variation of the generated answer *across* matched rephrasings. Calibration work establishes that verbalized confidence is unstable (Lin et al., 2022; Kadavath et al., 2022; Tian et al., 2023; Omar et al., 2025); we treat confidence as a behavioral readout, and ğ4.5 adds a logit-convergence check precisely because verbalized confidence is noisy.

Persona steering and representations. A separate line manipulates or monitors persona *directions* in activation space (Chen et al., 2025). Those are control/steering methods; our question is the inverse — whether an *observed* moderation is causally about the persona.

Prompt sensitivity — the literature our baseline comes from, and our relation to it. Sclar et al. (2024) introduce FormatSpread and recommend reporting a *range* across plausible prompt formats. Mizrahi et al. (2024) call for multi-prompt evaluation over large paraphrase pools. Most directly, ProSA (Zhuo et al., 2024) shows prompt sensitivity is an outward appearance of decoding confidence, greatest on harder, lower-confidence inputs. We are explicit that **our conceptual core is close to a corollary of ProSA**: if sensitivity tracks input difficulty, then any difficulty-correlated semantic property (contestedness among them) will correlate with paraphrase-induced range whether or not an audience is present. Our addition is operational — turning that corollary into an *adjudication test for a named moderator*, and demonstrating a plausible moderator that the test catches. We frame the contribution accordingly (protocol + worked example), not as a discovery independent of ProSA.

Closest framing precedents. Hu et al. (2025) study how disclosed asker attributes shift factual answers; their DV is answer *correctness* robustness, ours is confidence-*range*, and they apply no neutral-paraphrase baseline. Vaugrante et al. (2024) document a "looming replication crisis" and share our skeptical spirit, but use *replication* as the instrument; we use a comparable-magnitude neutral baseline that separates property-specific effects from generic instability *within a single experiment*.

3. Methods

3.1 The neutral-paraphrase null and the two tests

For a (claim, model) cell, **persona-range** is max−min of persona-mean confidence across the four persona conditions, on a 0–1 scale; **neutral-range** is the identical statistic across four content-neutral paraphrases that carry no audience identity but vary in surface form by a comparable (slightly smaller) amount. Both are max−min over four values, directly comparable.

> **Range statistic — known limitation and fix.** Max−min of four 3-sample means is upward-biased by within-condition sampling noise — contested claims plausibly have higher sampling variance, which can inflate their max−min independent of any systematic condition effect, conflating systematic between-condition structure with noise. We therefore also report an unbiased variance-component statistic — per-cell ω^2 (omega-squared) for the between-condition factor from a one-way model over the 4 conditions $\times$ 3 samples — and re-run both tests on it (results in ğ4.5 and Appendix C).

The null supports two tests.

- **Magnitude (range-equivalence) test.** Does persona-range exceed neutral-range? Paired Wilcoxon across claims, plus a two-one-sided-test (TOST) equivalence test within a ± 0.10 margin, per model.
- **Moderator-specificity test.** Is a putative moderator's correlation with persona-range *audience-specific*? We compute ρ(contestedness, persona-range) = ρ_persona and ρ(contestedness, neutral-range) = ρ_neutral and compare. If ρ_neutral ≈ 0 while ρ_persona is large, the moderation is audience-specific; if ρ_neutral matches or exceeds ρ_persona, the apparent moderation is generic stimulus instability.

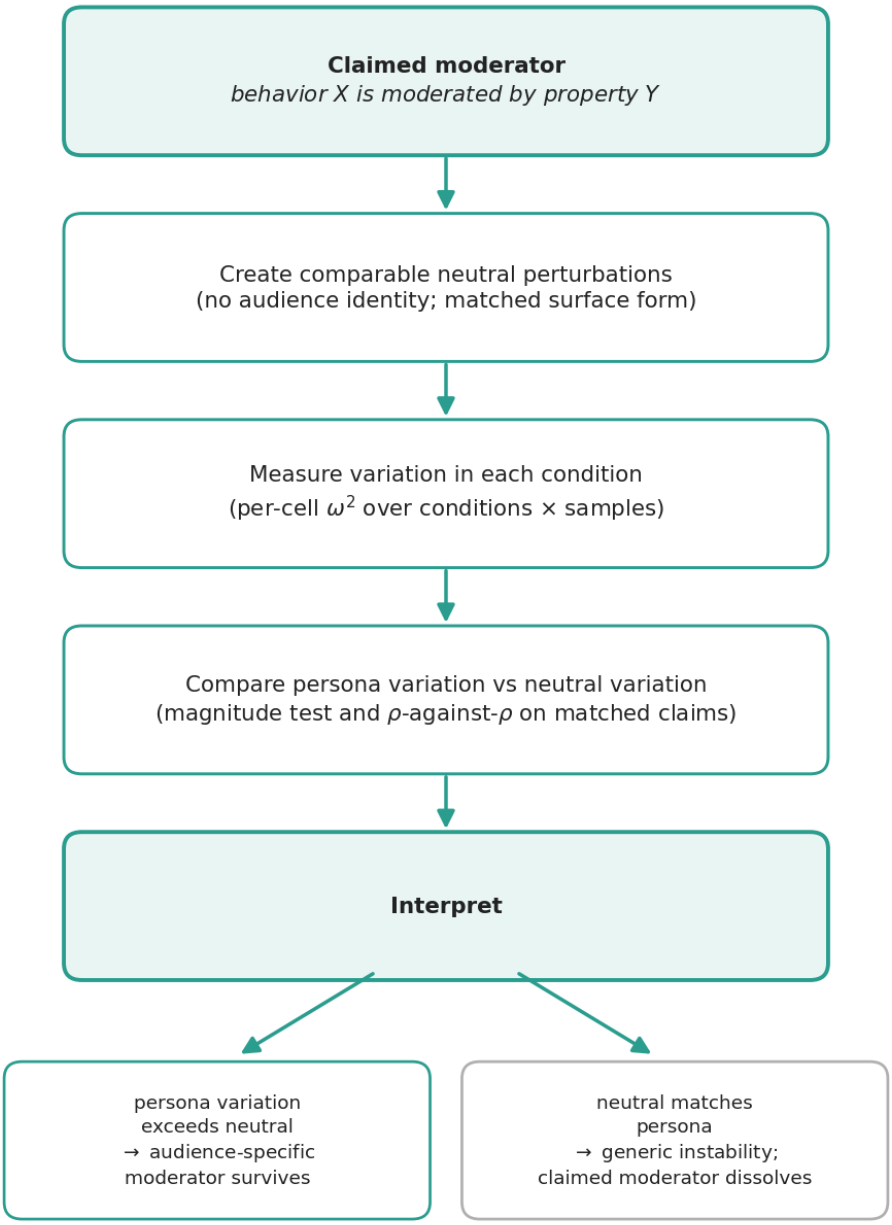

Figure 1: **The neutral-perturbation control protocol: adjudicate a claimed moderator by comparing persona-induced variation against an identity-free neutral baseline of comparable surface form.**

Two cautions on the moderator-specificity test, both load-bearing. (i) Establishing *equivalence* of two correlations requires more power than equivalence of two ranges; our conclusion rests on ρ_persona *not exceeding* ρ_neutral, never on a claim that the correlations are equal. (ii) These are **signed** correlations against *certainty* (so more-negative = stronger tracking). The comparison that matters is in **magnitude**: $|\rho$_persona$|$ vs $|\rho$_neutral$|$. We flag this explicitly because in biomedicine $|\rho$_persona$| > |\rho$_neutral$|$ (ğ4.2), i.e. biomedicine does **not** pass the moderator-specificity test at the correlation level; only climate does.

3.2 "Comparable magnitude" — a lexical proxy, stated as such

The four neutral preambles are slightly smaller surface perturbations than the four persona preambles (mean 10.5 vs ~13 tokens; ~1.5× less character-edit distance and char-3-gram cosine shift from the bare question). We call this **comparable *lexical* surface**, and we concede it is a proxy: a persona preamble injects a *pragmatic frame* (who is asking) that may move the model along a different representational pathway than a syntactic reshuffle of equal edit distance, so lexical equivalence does not guarantee equivalence in the perturbation dimension that matters. The lexical asymmetry is *a fortiori* favorable to the moderator-specificity conclusion (a milder neutral perturbation that nonetheless tracks contestedness at least as strongly indicates the relationship is not driven by richer persona content), but it cuts the *other* way for the residual magnitude effect (ğ4.3): a gentler neutral condition biases the magnitude test toward persona > neutral, so the small surviving persona excess is an **upper bound**, not a point estimate.

3.3 Claim corpora

Biomedicine (40 claims). Ten settled-consensus anchors, 25 contested-middle claims, and 5 high-effect claims carried from prior work (Anonymous, 2026) that are excluded from inferential correlations to avoid selection on the dependent variable. The contested subset skews toward supplements/complementary medicine — a topic confound we do not fully dissolve but only *relocate*, and address via cross-domain replication (ğ5.5).

Climate (40 claims, Climate-FEVER). From Diggelmann et al. (2020), annotating claims as supports/refutes/not-enough-info/disputed with per-claim evidence-label annotations. We sampled 10 low-contestedness (unanimous) and 30 contested claims, stratifying so the categorical label and continuous entropy score are not rank-identical (Spearman +0.58 across the 40). Each declarative claim became a natural polar question where the verb allowed (36/40), with an "is it true that…" wrapper otherwise (4/40); ğ4.5 reports a drop-the-fallbacks sensitivity check.

3.4 Persona and neutral conditions

Biomedical personas: researcher, clinician, patient, regulator. Climate personas: climate-scientist, policy-maker, journalist, concerned-public. Each preamble named the audience asking. The four neutral preambles varied surface form (formality, syntax) while naming no audience and carrying no epistemic stance, were fixed in a version-controlled config *before* the dry run and any inspection of induced range, and were reused byte-identical across both domains. They were not iterated against observed range; this pre-specification is what licenses the neutral condition as a control rather than a quantity tuned to match the persona spread. Every condition used the same structured prompt (yes/no/uncertain/it-depends; 0–4 confidence; implied action; brief explanation), the same regex parser, and the same /4→[0,1] normalization.

3.5 Models

Seven-model panel. Three frontier models — Claude Opus 4.1, GPT-4o, Gemini 2.5 Flash — on all 40 claims per domain; three open-weight models added in a panel extension — DeepSeek-V3.2, Kimi-K2, Qwen3-Max — on all 40 claims per domain (Gemini 2.5 Pro deliberately excluded on cost grounds; full per-model results in Appendix E); and Llama-3.1-70B-Instruct on a 14-claim subset (reported in Appendix D as uninformative, not as evidence) — seven models in total. All via OpenRouter. Each (claim × condition × model) cell sampled three times at temperature 0.7 (baseline throughout); biomedical persona-range additionally verified across a {0.0, 0.3, 0.7, 1.0} sweep (ğ4.5). To test whether the audience-magnitude effect is specific to these model generations, we additionally replicated the biomedicine magnitude test on a current frontier model, GPT-5.1 (mid-2026), via the identical pipeline (magnitude test only; reasoning-capped; see ğ4.3 and its footnote).

3.6 Contestedness measures

Biomedical contestedness from clinician ratings: four clinicians (of seven invited) rated each claim's evidence-certainty (1–5) and reported confidence in their own rating, yielding a weighted-mean certainty and inter-

clinician disagreement (rating SD). Thirty of forty claims received ≥3 ratings; the primary analysis uses these 30, with an all-40 sensitivity check. Climate contestedness measured two ways from the dataset (no new annotation): a categorical label (disputed vs settled) and a continuous Shannon entropy over per-claim evidence labels.

3.7 Analysis

The unit is the claim. Correlations use Spearman ρ with **exact permutation p-values** (100,000 resamples) and bootstrap 95% CIs on ρ; exact permutation is used because the asymptotic normal (z) reference, applied at n = 30, produces p-values that are systematically too small (this corrected an earlier-draft error; ğ4.5). Group comparisons use Mann-Whitney U with tie correction. Range equivalence uses TOST. We pre-specify the equivalence margin at ±0.10 on the 0–1 scale on measurement-granularity grounds: one step on the underlying 0–4 elicitation scale equals 0.25, so ±0.10 is below a single elicitation step.

> **On the margin and "negligible."** We *do not* call sub-margin effects "behaviorally negligible." The literature we build on (Klang et al., 2026) concerns *action thresholds*, where a sub-step confidence shift near a decision boundary can flip a recommendation. "Below measurement granularity" is therefore a statement about resolution, not about downstream stakes. We report full 90% CIs so readers can apply their own margin, and we flag every CI that excludes zero as a real, if sub-margin, effect.

Across the inferential family we report Benjamini-Hochberg FDR-adjusted q alongside raw p (ğ4.5). We state plainly that this family is a *curated subset* of comparisons run, that the biomedical analysis is **exploratory** (the Gemini −0.63 is the comparison that "popped" in a wide garden of forking paths), and that FDR control here is indicative, not exhaustive. The climate analysis is the confirmatory arm.

4. Results

4.1 An apparent contestedness pattern — robust across the seven-model panel (biomedicine, exploratory)

Pooling models, persona-range correlates negatively with weighted clinician certainty ($\rho = -0.54$, exact-permutation p = 0.0027, n = 30) and positively with inter-clinician disagreement ($\rho = +0.60$, p = 0.0006). Both survive dropping the five prior-effect claims (certainty $-0.52 \rightarrow -0.57$). The disagreement correlation is fragile (collapses at n = 40, p = 0.16, because under-rated claims cannot exhibit high disagreement); weighted certainty is the robust predictor.

Decomposed by model (max−min, T = 0.7, certainty, n = 30; the unbiased ω^2 seven-model panel is in Appendix E):

Model	ρ_persona	p (exact perm)
Gemini	**−0.627**	0.0003
GPT-4o	−0.198	0.29 (ns)
Claude	−0.208	0.27 (ns)
Llama (n = 7)	+0.018	0.99 (uninformative)

Across the seven-model panel, the contestedness→variation relationship is not a single-model artifact: it reaches significance in four of seven models in biomedicine (Gemini, DeepSeek-V3.2, Kimi-K2, Qwen3-Max) and two of seven in climate (Gemini, Qwen3-Max). The relationship the paper sets out to interrogate is therefore robustly real, not a curiosity of one system. (Full per-model ρ and ω^2 for all seven models in Appendix E.)

4.2 The moderator-specificity test (biomedicine): audience-favoring but inconclusive under ω^2 (range-equivalence does not survive)

Under ω^2, the biomedical picture shifts on both axes. Personas produce 2–15× the systematic between-condition variation of neutral paraphrases (Gemini 2.1×, Claude 2.8×, GPT-4o 15×; all 90% CIs exclude zero), and they track contestedness numerically more strongly ($|\rho_\text{persona}| = 0.593$, $p = 0.0007$ vs $|\rho_\text{neutral}| = 0.396$). The Gemini contestedness correlation that anchored ğ4.1 survives the unbiased statistic (-0.627 max−min → -0.593 ω^2, still significant), so the headline correlation is not a max−min artifact. But the persona-vs-neutral gap in that correlation — 0.197 — does not reach significance at n = 30 (bootstrap 95% CI [−0.05, +0.47], includes zero). We therefore report biomedicine as numerically audience-favoring but inconclusive on moderator-specificity: it does not show that the contestedness relationship is audience-free (the v1/v2 reading), nor does it establish that the relationship is audience-specific. What biomedicine no longer supports is the claim it supported under max−min — equivalence — which ω^2 removes entirely.

4.3 What survives: a residual audience-magnitude effect, and the GPT-4o dissociation

Under ω^2 the paired magnitude test shows a persona excess on **all seven** biomedical models (Figure 3): persona ω^2 exceeds neutral ω^2 by +0.07 to +0.17 (original panel — Gemini +0.136, GPT-4o +0.174, Claude +0.167, Llama +0.117; added panel — DeepSeek +0.072, Kimi +0.081, Qwen3-Max +0.072; all 90% CIs exclude zero; per-model in Appendix E), i.e. personas induce 2–15× the systematic variation of neutral rephrasing. This is a genuine *magnitude* effect, **not** a contestedness-gated one — audiences move confidence more than arbitrary rephrasing, not selectively where evidence is contested. Per ğ3.2 the gentler neutral perturbation makes this an upper bound.

The effect is not confined to these model generations: on a current frontier model (GPT-5.1, mid-2026), persona ω^2 exceeded neutral by +0.064 (90% CI [+0.021, +0.107]; Wilcoxon $p = 0.012$; TOST ±0.10 $p = 0.084$), the same small-but-real character seen across the panel.[1]

The cleanest dissociation in the dataset is now GPT-4o, not Claude: its neutral ω^2 is essentially zero (0.012) while its persona ω^2 is 0.186 — a 15× ratio, a model that all but ignores neutral rephrasing yet moves systematically on audience identity. Claude shows the same direction more weakly (0.091 vs 0.258, 2.8×). We attach a caveat that tempers this: median ω^2 is zero across conditions for these models, so the dissociation lives in the upper tail of claims rather than the typical claim, and the GPT-4o ratio in particular is sensitive to its near-zero neutral floor. We report it as the strongest single instance of audience-as-signal in the data and as the principled limit of any "audience is incidental" reading — not as a general law, and not on the typical claim.

Within the persona conditions, a descriptive ordering recurs (researcher lowest confidence, patient highest, clinician and regulator between). We report this gradient as **descriptive only**; given ğ4.2 it does not by itself establish audience-specific gating by contestedness. (We flag it as a candidate audience-specific signal worth a neutral-baselined test in future work — ğ5.6.)

4.4 Cross-domain replication (climate, confirmatory): the magnitude effect does not generalize; the moderator-specificity result holds and is unambiguous here

We treat biomedicine as exploratory and climate as confirmatory. For the confirmatory arm we fixed a decision rule before climate results were seen (Appendix A; committed ~7 hours before result generation — documentation of analysis order, not formal pre-registration).

The magnitude effect does not generalize to climate. Under ω^2 it is mixed and mostly reversed: neutral exceeds persona on Gemini (-0.116 [-0.190, -0.041]) and Claude (-0.023 [-0.122, $+0.076$]), with

[1] GPT-5.1 does not accept a temperature parameter via OpenRouter and ran at provider default; the persona-vs-neutral contrast is internally temperature-matched so the magnitude test is unaffected, but cross-model level comparison with the T=0.7 panel carries this caveat. The model was run with reasoning capped (≤128 tokens); a capped-vs-uncapped spot-check confirmed no material effect on elicited confidence.

only GPT-4o persona-favoring (+0.055 [+0.007, +0.103]); Gemini's persona ω^2 is ≈ 0.006. The biomedical audience-magnitude effect is domain-specific.

The moderator-specificity result holds — and here it is unambiguous. Climate ranges are small (8–17 of 40 claims move; settled claims are flat), so the domain is low-powered; we read it as *consistent-with* rather than *strengthening*. On Gemini (again the only model with any contestedness signal), the audience-free neutral condition tracks contestedness *more* strongly than the persona condition, in both operationalizations, and crucially the signs are positive so the magnitude comparison is direct (ρ on ω^2, n = 40):

Arm	ρ_persona (p)	ρ_neutral (p)	gap (ρ_neutral − ρ_persona)
Categorical (disputed vs settled)	+0.151 (0.45, ns)	+0.329 (0.064, marginal)	+0.178
Continuous (label entropy)	+0.209 (0.20, ns)	+0.437 (0.005, sig)	+0.228

In both arms the persona correlation is non-significant while the neutral correlation is significant: contestedness predicts confidence-range on climate, and the audience-free condition carries it (Figure 2). **The single load-bearing fact is ρ_neutral = +0.44 (p = 0.005) significant where ρ_persona is not** — this is a real signal, not a power artifact (a power deficit would weaken both arms, not selectively spare the neutral one). This is the moderator-specificity failure in its cleanest form, and it is a *climate* result. We do not describe climate as "strengthening" the biomedical finding; biomedicine did not establish audience-non-specificity to strengthen.

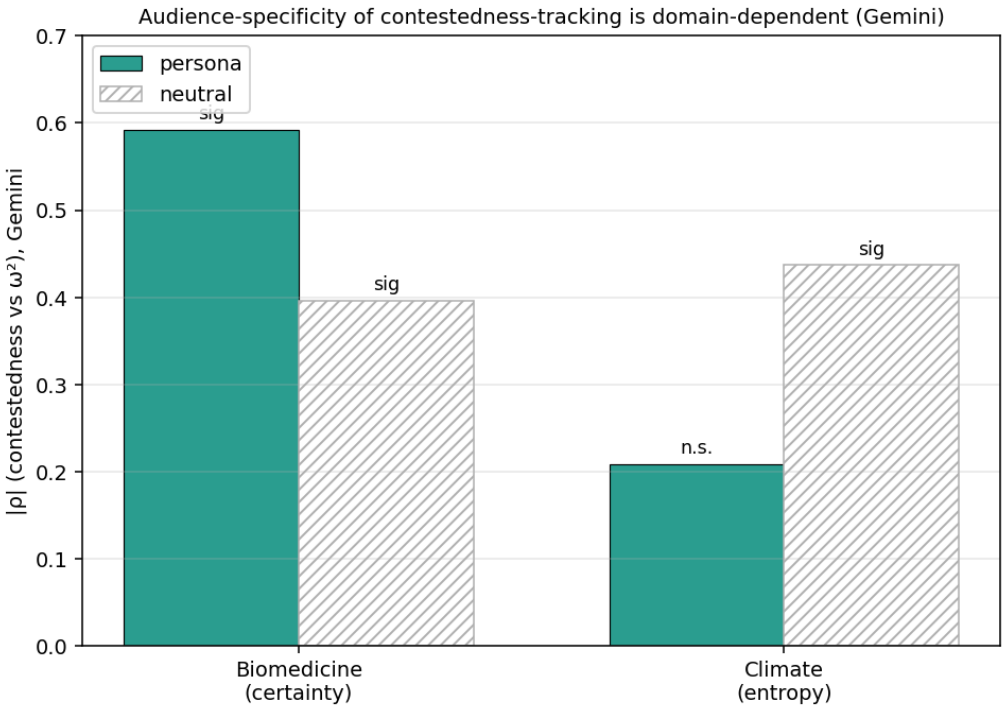

Figure 2: **Whether contestedness-tracking is audience-specific is domain-dependent.** Magnitude of the contestedness–ω^2 correlation under persona vs. neutral conditions on Gemini (the only model with a contestedness signal), by domain. In biomedicine the persona bar exceeds the neutral bar ($|\rho$_persona$|$ = 0.59 vs. $|\rho$_neutral$|$ = 0.40), but the gap is not significant (95% CI [−0.05, +0.47] includes zero), so biomedicine is inconclusive on audience-specificity. In climate the neutral bar exceeds the persona bar and only the neutral correlation is significant (ρ_neutral = +0.44, p = 0.005; persona n.s.), so the moderator is not audience-specific there. The two domains point in opposite directions; the audience-specific reading is not supported.

Range-equivalence holds uniformly (TOST, ±0.10):

Domain	Model	mean diff (persona − neutral)	90% CI	TOST p	Verdict
Biomedical	Gemini	+0.022	[−0.002, +0.045]	<0.001	equivalent
Biomedical	GPT-4o	+0.048	[+0.024, +0.072]	<0.001	equivalent (CI excludes 0)
Biomedical	Claude	+0.042	[+0.016, +0.068]	<0.001	equivalent (CI excludes 0)
Climate	Gemini	−0.033	[−0.058, −0.009]	<0.001	equivalent
Climate	GPT-4o	+0.000	[−0.025, +0.025]	<0.001	equivalent
Climate	Claude	−0.010	[−0.041, +0.020]	<0.001	equivalent

This equivalence does not survive the ω^2 statistic of ğ4.5 and is retained only to document the artifact; see Appendix C.

The two biomedical models with the residual magnitude effect (GPT-4o, Claude) have CIs excluding zero: a small, real, sub-margin persona excess. No climate model shows even a sub-margin positive excess; two are negative.

4.5 Robustness

Statistics reconciliation. Earlier-draft Spearman p-values used an asymptotic normal reference where the Student-t reference (df = n − 2) is appropriate, inflating significance ~3×. All reported p-values are exact-permutation. Qualitative conclusions unchanged.

Multiple comparisons. Over the 13-test inferential family, 11/13 survive Benjamini-Hochberg FDR at q = 0.05. The two affected: a Gemini researcher-asymmetry test (raw p = 0.052) does not survive (q = 0.067); a tie-corrected Claude settled-vs-contested comparison (raw p = 0.023) does (q = 0.033). This family is curated and biomedicine is exploratory.

Temperature and scale. Biomedical persona-range is temperature-invariant across {0.0, 0.3, 0.7, 1.0} (mixed-effects coefficient ≈ 0, p = 0.33), and the persona ordering is preserved on a continuous 0–100 re-elicitation. *Limitation:* the 0–100 re-elicitation verified the persona *ordering* but did not re-test the persona-vs-neutral *magnitude* comparison on the continuous scale; since both max−min and ω^2 here rest on four coarse means, this is left to a focused follow-up (ğ5.5).

Variance-component (ω^2) — run; full tables in Appendix C. Replacing max−min with the pre-committed unbiased ω^2 (N_PERM=100,000): **(i) the Gemini biomedical correlation survives** — $\rho(\omega^2$_persona, certainty) = −0.593, p = 0.0007 (matched-N max−min −0.627); ω^2↔max−min rank agreement is high (bio 0.87, climate 0.89), so the ğ4.1–4.2 *correlation* conclusions stand. **(ii) Range-equivalence does NOT survive.** Persona ω^2 exceeds neutral ω^2 on all seven biomedical models (mean diff +0.07 to +0.17; 90% CIs exclude 0; TOST non-equivalent; per-model in Appendix E), mean ω^2 persona 0.10–0.26 vs neutral 0.01–0.13 (2–15×). The max−min equivalence was a shared sampling-noise-floor artifact; per ğ3.1 the unbiased statistic governs, so the magnitude finding is now *persona > neutral* (general across biomedical models), not equivalence. **(iii) Biomedicine moderator-specificity** unchanged in direction, larger in size: |ρ_persona| 0.593 > |ρ_neutral| 0.396 (gap 0.197; bootstrap 95% CI [−0.05, +0.47] — includes 0, so *numerically audience-favoring but inconclusive*). **(iv) Climate** entropy arm holds (ρ_neutral +0.44, p = 0.005; persona ns); categorical arm weakens to marginal (ρ_neutral +0.33, p = 0.064).

Logit-convergence check — could not be run. No provider in our OpenRouter routing returned token-logprobs for these runs, so the specified logit-derived check (`logit_convergence.py`, Appendix C) could not be computed. The construct-validity implications — and why the unbiased ω^2 statistic already rules out a shared elicitation-noise artifact — are addressed in ğ5.2.

Fallback question form (climate). Dropping the four "is-it-true" fallbacks changes nothing (the climate magnitude pattern is unchanged and correlation p-values move within 0.01).

Decoupling check (climate). Categorical and continuous contestedness correlate at +0.58 — related but separable, not the disputed label twice.

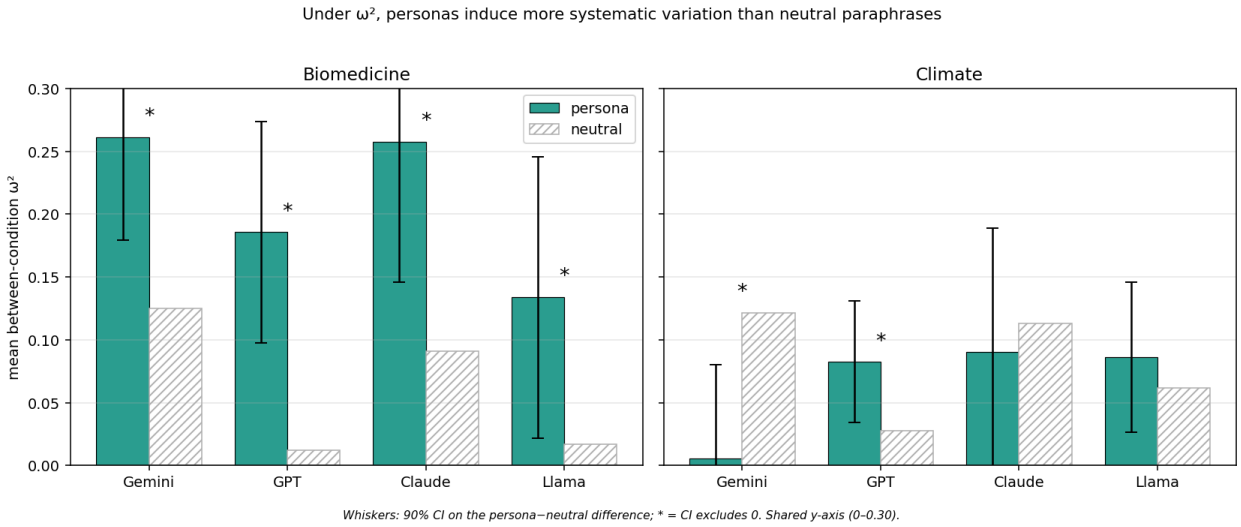

Figure 3: **Under the unbiased ω^2 statistic, audience personas induce more systematic variation than neutral paraphrases — in biomedicine, on every model.** Mean between-condition ω^2 per arm, by model and domain. In biomedicine persona ω^2 exceeds neutral on every frontier model shown here (2–15×; all paired 90% CIs exclude zero), and on all three added open-weight models as well — 7/7 (Appendix E) — so the audience-magnitude effect is general within the domain. In climate it does not generalize: neutral exceeds persona on Gemini (≈20×) and Claude, with only GPT-4o persona-favoring. The biased max−min statistic (now superseded) had shown these arms as equivalent; ω^2 removes the shared sampling-noise floor and reveals the systematic persona excess.

5. Discussion

5.1 Apparent semantic moderation is not model sensitivity — scoped to what we showed

The sharpest lesson is one we did not plan to deliver. We pre-specified, before seeing results, that ω^2 rather than max−min was the defensible spread statistic — precisely because max−min is upward-biased by within-condition sampling noise. That pre-commitment then overturned one of our own headline claims: the range-equivalence that held cleanly under max−min dissolves under ω^2, because the noise floor that inflated both arms equally was doing the work the "equivalence" was credited to. The paper's central warning — that a biased spread statistic can manufacture an apparent null that a property-specific effect does not actually support — is therefore not only argued but instantiated on our own data, against our own prior framing. We regard this as the strongest evidence we can offer that the control matters: it caught us.

The structural lesson outlives our specific hypothesis: when persona-induced variation is larger on some inputs than others, the instinct is to attribute the pattern to the model processing what distinguishes those inputs. Our result shows that inference needs a comparable-magnitude neutral baseline before it is licensed.

Our evidence is no longer carried by a single model. The contestedness signal replicates across four models in biomedicine and two in climate, spanning three independent open-weight families in addition to Gemini, so the premise, that a semantic property co-varies with where models are unstable, is robust. What remains scoped is the audience-specificity question, and here the panel reveals a split the neutral control is built to expose: in climate, where two models show the signal, both are neutral-carried (the relationship survives removing the audience); in biomedicine, of the four models showing the signal, one (DeepSeek) is neutral-carried while three lean audience-favoring, so biomedicine does not cleanly support the dissolution. The dissolution is thus climate-robust and biomedicine-mixed, now observed across models rather than asserted from one. We therefore present the neutral-perturbation control as a general protocol and contestedness-gated audience-conditioning as a single worked example, not as evidence that semantic moderators are

generally confounded. We restrict even that example to audience-conditioning: we did not test role-adoption, emotional tone, demographic signals, or political framing, and we do not claim our result transfers to them.

5.2 Construct validity of the verbalized-confidence dependent variable

Our dependent variable throughout is verbalized confidence, and the calibration literature establishes that verbalized confidence is noisy and elicitation-sensitive (Lin et al., 2022; Kadavath et al., 2022; Tian et al., 2023). A natural objection is therefore that the effects we report — the audience-magnitude excess and its domain-dependent organization — could be artifacts of a noisy readout rather than properties of model behavior. The objection has a precise answer, supplied by the choice of statistic.

ω^2 measures systematic between-condition variance: within-condition sampling noise enters the residual term and is partialled out by construction. Verbalized-confidence noise that is shared across conditions — the elicitation floor the objection invokes — therefore cannot inflate ω^2, because it contributes to within-condition variance, not between. This is not an assumption but the mechanism we observed: it is exactly why the apparent range-equivalence under max−min dissolved under ω^2 (ğ4.4). A shared noise floor made the two arms look equal under the biased statistic; the unbiased one removed it. The same removal that overturned our prior claim protects the surviving ones: a noise floor common to both arms cannot produce a 2–15× persona excess in ω^2, because the floor is precisely what ω^2 discards.

What ω^2 does not establish is calibration: it tells us that confidence varies systematically by condition, not that the confidence values track ground-truth correctness. Our claims are scoped to variation, not calibration — we treat confidence as a behavioral readout (ğ3.5), and we make no claim that the elicited values are well-calibrated. The guarantee is therefore partial: systematic between-condition structure is real and not a noise artifact; whether it reflects calibrated uncertainty is a separate question we do not answer. The clean test of the stronger property would correlate verbalized confidence against a logit-derived decisiveness on the answer token, which we specify and release as `logit_convergence.py` (Appendix C); it could not be run here because the OpenRouter chat-completions routes for our panel returned no token-logprobs. We flag this as the single most valuable extension: a logprob-exposing route would convert the partial guarantee into a full one. Until then, our results should be read as claims about systematic, audience-conditioned variation in expressed confidence — which is what the protocol adjudicates — and not as claims about the calibration of that confidence.

5.3 Why a distribution, and why no identity

Two design choices distinguish this null from prior single-point baselines. A *distribution* of rephrasings makes range-against-range possible: the question is not "do personas differ from a generic baseline" (they can, trivially) but "do personas differ from each other *more than arbitrary rephrasings differ from each other*." A null that invokes *no identity* isolates audience-conditioning from identity-adoption. We recommend both as defaults when validating a claimed moderator of prompt-conditioned behavior.

5.4 The surviving effect and the dissociation as the claim's limit

Personas are not inert. The audience-magnitude effect is the most robust audience-specific finding in the study: general within biomedicine, replicated on a current frontier model, and biomedicine-specific (ğ4.3). What fails to generalize is its organization by contestedness, not its existence. The GPT-4o dissociation (ğ4.3) is a clean demonstration, confined to one model, that audience identity can be treated as signal independent of surface form; we read it as the limit of the dissolution, not as support for it.

5.5 Limitations

Sample sizes are modest (40 claims/domain; 30 with ≥3 raters in biomedicine; a 14-claim open-weights subset). Biomedical contestedness is **entangled with topic** (contested skews to supplements/CAM), so "contested" partly indexes a *training-corpus region*; the cross-domain recurrence on climate shows the relationship is not unique to that region but does **not** dissolve the confound — it *relocates* it, since climate

carries its own topical correlates of disputedness. We are therefore claiming "instability tracks something correlated with contestedness," not "contestedness per se." Climate ranges are small and power is limited (though the neutral correlations remain significant). We test a single prompt template. The open-weights arm is underpowered and reported as uninformative. The (now-superseded) max−min equivalence verdict depended on the ±0.10 margin and the discrete 0–4-derived statistic; under the unbiased ω^2 that equivalence does not hold (ğ4.5), and neither statistic was re-confirmed on the continuous 0–100 scale. The two contestedness constructs differ across domains (expert certainty vs annotator entropy). We provide no mechanistic evidence beyond connecting the result to ProSA's decoding-confidence account; the dissolution is behavioral. The DV is verbalized confidence; its construct validity — and why the unbiased ω^2 already excludes a shared elicitation-noise artifact — is addressed in ğ5.2. Our seven-model panel is too small for model-family claims; per-model differences are descriptive.

5.6 Implications and future work

The immediate methodological implication restates the protocol: without a comparable-magnitude neutral-perturbation baseline, a stimulus artifact and a model property are not distinguishable. Concrete next steps: (i) scale the control into a field audit across many claimed moderators, domains, and models, to estimate how often published moderation survives — *this* is what would convert our "the trap is real" into "the trap is pervasive (or not)"; (ii) neutral-baseline the expertise gradient (researcher→patient), which is a candidate audience-specific signal we did not adjudicate; (iii) re-confirm the ω^2 magnitude finding on a continuous elicitation scale (the variance-component statistic itself is now applied, ğ4.5); (iv) extend the null from audience-conditioning to role-adoption.

6. Conclusion

The lesson outlived the hypothesis. We set out to show that audience-persona effects on LLM confidence are gated by evidence contestedness, and a control we built to confirm it overturned it instead, on our own data. That is the paper's contribution stated plainly: an apparent semantic moderator of model behavior is not evidence of model sensitivity to that semantics until a comparable-magnitude, identity-free neutral baseline has separated it from generic prompt-instability, and a biased spread statistic can manufacture a moderator that an unbiased one dissolves.

Three findings survive, scoped. Magnitude: audience personas induce more systematic confidence variation than neutral paraphrases, general across the biomedical panel, but biomedicine-specific. Specificity: that variation is no more organized by contestedness than neutral variation is — in climate the audience-free condition carries the signal outright (ρ_neutral = +0.44, p = 0.005), while in biomedicine the gap is inconclusive. Dissolution: the moderator's audience-specificity is therefore domain-dependent, clean in climate and unproven in biomedicine. Building on the prompt-sensitivity literature (Sclar et al., 2024; Mizrahi et al., 2024; Zhuo et al., 2024) rather than apart from it, the contribution is a protocol, turning paraphrase-spread into an adjudication test for a named moderator, and an honest worked example showing the test catches a plausible, publishable-looking moderator. We do not claim the trap is pervasive; we claim it is real, easy to fall into, and cheap to test for.

Acknowledgements

Omitted for double-blind review.

Data and code availability

Claim corpora, neutral and persona templates (with pre-run commit timestamps), reconciliation and neutral-null analysis scripts, the variance-component and logit-convergence scripts, per-claim range audit tables for both domains, and the decision-rule file are released as anonymized supplementary material for review (anonymized repository; author-identifying commit metadata stripped).

References

Abdullahi, T., Ghosh, S., Fraser, H. S., León Tramontini, D., Abbasi, A., Bourjeily, G., Eickhoff, C., & Singh, R. (2026). The Persona Paradox: Medical Personas as Behavioral Priors in Clinical Language Models. arXiv:2601.05376.

Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2020). Climate-FEVER: A Dataset for Verification of Real-World Climate Claims. arXiv:2012.00614.

Hu, et al. (2025). Who's Asking? Evaluating LLM Robustness to Inquiry Personas in Factual Question Answering. arXiv:2510.12925.

Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221.

Klang, E., Gorenstein, A., Omar, M., & Nadkarni, G. N. (2026). Personas Shift Clinical Action Thresholds in Large Language Models. medRxiv. DOI:10.64898/2026.01.01.26343302.

Li, C., Wang, J., Zhang, Y., et al. (2023). Large Language Models Understand and Can Be Enhanced by Emotional Stimuli. arXiv:2307.11760.

Lin, S., Hilton, J., & Evans, O. (2022). Teaching Models to Express Their Uncertainty in Words. Transactions on Machine Learning Research.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of What Art? A Call for Multi-Prompt LLM Evaluation. TACL. arXiv:2401.00595.

Omar, M., Agbareia, R., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Benchmarking the Confidence of Large Language Models in Answering Clinical Questions. JMIR Medical Informatics. DOI:10.2196/66917.

Anonymous (2026). Structural vs sampling-induced behaviors in LLMs: evidence from persona-conditioned confidence. Prior non-archival venue; citation anonymized for double-blind review.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design. ICLR. arXiv:2310.11324.

Tian, K., Mitchell, E., Zhou, A., et al. (2023). Just Ask for Calibration. EMNLP.

Vaugrante, L., et al. (2024). A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions. arXiv:2409.20303.

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. Findings of EMNLP 2024, 1950–1976. arXiv:2410.12405.

Appendix A. Decision rule (climate, confirmatory arm)

Fixed and committed to version control before climate results were generated. Documentation of analysis order and transparency, not formal pre-registration; the biomedical dissolution that motivated it was exploratory.

- *Effect A (audience magnitude) reproduces* if persona-range > neutral-range (paired Wilcoxon $p < 0.05$) on ≥ 2 of 3 frontier models.

- *Effect B (moderator-specificity) reproduces* if ρ_persona does not exceed ρ_neutral **in magnitude** (point gap small or favoring neutral; ρ_neutral significant where ρ_persona is not) on the model(s) where ρ_persona is significant. (We do not require statistical equivalence of the correlations.)
- *Neither* → effect is biomedical-specific.

Outcome: Effect A (audience magnitude) did not generalize to climate (under the pre-committed ω^2, only GPT-4o is persona-favoring; Gemini and Claude are neutral-led); Effect B reproduced in climate (neutral steeper than persona on Gemini in both arms, positive signs, unambiguous). *Note added in v2:* in biomedicine, by the same magnitude criterion |ρ_persona| (ω^2 0.593) exceeds |ρ_neutral| (0.396), so Effect B did **not** reproduce in biomedicine — the confirmatory result is climate-carried. v1 obscured this by reading a signed gap as if it favored neutral.

Commit timeline (UTC): dry run 00:50–00:59; full run starts 01:00:32; decision rule committed 01:01:33; full run finishes 07:14:43; results generated 08:04:19. The rule predates result generation by ~7 hours; the commit followed run launch by ~61 s (zero results analyzed in between). We claim only that the rule was fixed before results were seen or analyzed.

Appendix B. Note on earlier framings

An earlier, non-archival version of this work reported contestedness-gating of persona effects as a positive finding. Two corrections changed that conclusion. First, a statistics error: Spearman p-values used an asymptotic-normal reference where the Student-t (exact-permutation) reference is appropriate, which had overstated significance at our sample sizes. Second, a missing control: the earlier version lacked a comparable-magnitude, identity-free neutral-paraphrase baseline; adding it shows the apparent moderation is not audience-specific in the way originally claimed, and under the unbiased ω^2 statistic the apparent range-equivalence does not hold. This version corrects both and supersedes that earlier report.

Appendix C. Variance-component statistic (specification + results)

For each (claim, model) and each arm (persona / neutral), fit a one-way model with `condition` (4 levels) over the 3 samples per condition and compute ω^2 (unbiased between-condition variance fraction). Use ω^2 in place of max−min for: (a) the moderator-specificity correlations $\rho(\omega^2$, contestedness), persona vs neutral, per model, both domains; (b) the magnitude test, mean ω^2_persona vs ω^2_neutral (paired Wilcoxon + TOST). Insert results here and reconcile against ğ4.1–4.4. Script: `variance_component.py`.

Results (ω^2, N_PERM=100,000; matched-N 3 samples/condition).

Moderator-specificity — $\rho(\omega^2$, contestedness), persona vs neutral:

Domain / measure	Model	ρ_persona (p)	ρ_neutral (p)	verdict
Biomed certainty (n=30)	Gemini	−0.593 (0.0007)	−0.396 (0.031)	persona-favoring, inconclusive
Biomed certainty	GPT-4o	−0.146 (ns)	−0.249 (ns)	both ns
Biomed certainty	Claude	−0.208 (ns)	−0.205 (ns)	both ns
Climate entropy (n=40)	Gemini	+0.209 (ns)	**+0.437 (0.005)**	neutral carries signal
Climate entropy	GPT-4o	+0.180 (ns)	+0.001 (ns)	both ns
Climate entropy	Claude	+0.164 (ns)	+0.122 (ns)	both ns
Climate categorical (n=40)	Gemini	+0.151 (ns)	+0.329 (0.064)	neutral marginal

Magnitude — mean ω^2_persona − ω^2_neutral (90% CI):

ω^2↔max−min rank agreement: biomed 0.866, climate 0.890 (>0.8 → correlation/rank conclusions stable).

Reconciliation: (a) Gemini biomed correlation survives ω^2 — ğ4.1–4.2 not softened to max−min-only. (b) Biomed fails moderator-specificity under ω^2 too (|ρ_p|>|ρ_n|, gap 0.197, 95% CI [−0.05,+0.47], inconclusive). (c) Magnitude excess CIs exclude 0 on the four original-panel biomedical models (was GPT-4o/Claude only on max−min); the three added models also exclude 0 (Appendix E) — 7/7. (d) Climate

Domain	Model	mean diff	90% CI	TOST(±0.10)	levels (persona / neutral)
Biomed	Gemini	+0.136	[+0.055,+0.218]	not equiv	0.261 / 0.125
Biomed	GPT-4o	+0.174	[+0.086,+0.262]	not equiv	0.186 / 0.012
Biomed	Claude	+0.167	[+0.055,+0.278]	not equiv	0.258 / 0.091
Climate	Gemini	−0.116	[−0.190,−0.041]	n/a (neutral>persona)	0.006 / 0.121
Climate	GPT-4o	+0.055	[+0.007,+0.103]	not equiv	0.082 / 0.028
Climate	Claude	−0.023	[−0.122,+0.076]	incl 0	0.090 / 0.113

entropy ρ_neutral significant (+0.44) where persona is not; categorical now marginal (p=0.064). (e) Rank agreement >0.8 both domains. **Range-equivalence is the one claim that does not survive** (persona ω^2 2–15× neutral, biomed) — abstract/ğ4.2/ğ4.4 reframed.

The specified-but-unrun logit-convergence construct-validity check (`logit_convergence.py`) is released alongside; see ğ5.2 for why it could not be run and what ω^2 establishes in its place.

Appendix D. Open-weights arm (Llama, n = 7–14)

Reported as uninformative (small effective sample, unstable signs), not as evidence. Llama-3.1-70B-Instruct was run on a 14-claim subset (4 personas + 4 neutral, 3 samples, T = 0.7); 7 of the 14 claims have ≥ 3 clinician ratings and enter the certainty correlation.

Quantity	Llama-3.1-70B (14-claim subset)
Claims run / with ≥3 ratings	14 / 7
$\rho(\omega^2$_persona, certainty), n = 7	+0.401 (p = 0.43, ns)
$\rho(\omega^2$_neutral, certainty), n = 7	−0.225 (p = 0.63, ns)
Mean ω^2 persona / neutral	0.134 / 0.017
Magnitude (persona − neutral ω^2)	+0.117, 90% CI [+0.005, +0.229]
Verdict	uninformative (n too small; signs unstable)

Both correlation arms are non-significant and sign-unstable at n = 7; we exclude the open-weights arm from all inferential claims and report it here only for completeness.

Appendix E. Seven-model panel (ω^2, per model, both domains)

Added in the panel extension (DeepSeek-V3.2, Kimi-K2, Qwen3-Max via OpenRouter, standard routes, T=0.7, 40 claims × 4 persona + 4 neutral × 3 samples; Gemini 2.5 Pro deliberately excluded on cost grounds). ρ on ω^2; permutation p (100k). "Signal" = at least one arm significant at p<0.05. Magnitude = persona ω^2 > neutral ω^2 (paired, 90% CI excludes 0).

Biomedicine (clinician certainty, n=30):

Model	ρ_persona (p)	ρ_neutral (p)	ω^2 persona / neutral	signal	direction	magnitude p>n
Gemini 2.5 Flash	−0.593 (.0007)	−0.396 (.031)	0.261 / 0.125	yes	persona-favoring	yes
GPT-4o	−0.146 (ns)	−0.249 (ns)	0.186 / 0.012	no	—	yes
Claude Opus 4.1	−0.208 (ns)	−0.205 (ns)	0.258 / 0.091	no	—	yes
Llama-3.1-70B (n=7)	+0.401 (ns)	−0.225 (ns)	0.134 / 0.017	uninformative	—	yes
DeepSeek-V3.2	−0.308 (.098)	−0.506 (.004)	0.125 / 0.054	yes	**neutral-carried**	yes
Kimi-K2	−0.443 (.014)	−0.318 (.088)	0.100 / 0.019	yes	persona-favoring	yes
Qwen3-Max	−0.409 (.025)	−0.310 (.097)	0.158 / 0.086	yes	persona-favoring	yes
GPT-5.1 (current frontier)	— (not run)	— (not run)	0.105 / 0.041	not run	—	yes (+0.064 [+0.021, +0.107])

Biomedicine signal: **4 of 7** (Gemini, DeepSeek, Kimi, Qwen). Magnitude (persona > neutral): **7 of 7** (all CIs exclude 0). Direction split among signal models: 1 neutral-carried (DeepSeek), 3 audience-favoring.

GPT-5.1 (current frontier, mid-2026) is a separate magnitude-only recency replication (ğ4.3), not part of the seven-model panel and not counted in the totals above; the contestedness correlation (ρ, signal, direction) was not run for it. It was run reasoning-capped at provider-default temperature (see the ğ4.3 note).

Climate (label entropy, n=40):

Model	ρ_persona (p)	ρ_neutral (p)	ω^2 persona / neutral	signal	direction	magnitude p>n
Gemini 2.5 Flash	+0.209 (ns)	+0.437 (.005)	0.006 / 0.121	yes	**neutral-carried**	no
GPT-4o	+0.180 (ns)	+0.001 (ns)	0.082 / 0.028	no	—	no
Claude Opus 4.1	+0.164 (ns)	+0.122 (ns)	0.090 / 0.113	no	—	no
DeepSeek-V3.2	+0.224 (ns)	−0.042 (ns)	0.078 / 0.096	no	—	no
Kimi-K2	+0.130 (ns)	+0.244 (ns)	0.028 / 0.048	no	—	no
Qwen3-Max	+0.156 (ns)	+0.352 (.026)	0.125 / 0.100	yes	**neutral-carried**	no

Climate signal: **2 of 7** (Gemini, Qwen3-Max), both neutral-carried. Magnitude: 0 of 7 (does not generalize). Categorical operationalization: no model reaches significance.

Panel API cost (OpenRouter, three added models): $1.51 total (DeepSeek $0.13, Qwen3-Max $0.85, Kimi $0.53). Scripts: `experiments/runners/panel_runner.py`, `panel_adapt.py`.