

LLMs and Foreign Policy Decision Making

Pre-Analysis Plan

Yasir Atalan* Benjamin Jensen† Ian Reynolds‡ Ryan Powers§

July 19, 2026

We study how decision makers receive and act on foreign-policy recommendations generated by large language models (LLMs). Prior work shows that LLMs vary systematically in the escalatory tenor of the advice they produce and exhibit country-specific biases (Jensen et al., 2025). We move one step further along the decision chain, studying how humans respond to recommendations generated by LLMs. Our research design allows us to ask whether *attributing* a given recommendation to a named AI system (rather than to a generic expert advisor) changes the policy choice a human would make and the credibility the person grants the foreign-policy recommendation. Our design also allows us to speak to whether labeling recommendations as AI-generated affects how those recommendations are evaluated, whether different models move people differently, whether more hawkish advice produces more escalatory choices, and whether these effects depend on the decision-maker’s foreign-policy disposition and AI skepticism.

1 Research Design

We outline the design of a scenario-based survey experiment in which respondents are assigned to nested treatments drawn from a prespecified treatment library. We summarize that design in Figure 1. All respondents read a common introduction:

Thank you for participating in this study.

On the following pages, you will read about a hypothetical situation involving the two countries. The situation is not real, but it resembles the kinds of challenges that countries in the world could face in the future.

We will ask you to play the role of foreign policy advisor to Mithandira, one of the fictional countries involved. Please read the scenario carefully, think about how you would suggest that Mithandira resolve the crisis, and answer the questions on the pages that follow.

Please read closely. Some later questions will ask about specific details from the scenario.

*Deputy Director and Data Fellow in the Futures Lab at Center for Strategic and International Studies. ✉ yasir.atalan@csis.org.

†Director of the Futures Lab and Senior Fellow at Center for Strategic and International Studies. ✉ benjamin.jensen@csis.org.

‡AI Public Policy Manager at Hugging Face. ✉ ian.reynolds@huggingface.co.

§Associate Professor of International Affairs and Gary K. Bertsch Director of the Benson-Bertsch Center for International Trade and Security at the University of Georgia. Non-resident Fellow in the Futures Lab at Center for Strategic and International Studies. 🌐 ryanpowers.net, ✉ ryan.powers@uga.edu.

Similarly, all respondents are given an identical foreign policy crisis scenario:

A Foreign Policy Crisis Between Two Countries

Two countries, **Mithandira** and **Anoria** are involved in a dispute. Please read the details of this dispute and consider the course of action that Mithandira should take.

Mithandira and **Anoria** are rival countries that share a land border with a history of past disputes. **Mithandira** suspects there is a high chance that war will break out with **Anoria** in the next 5 to 10 years due to an ongoing political crisis related to contested border areas that contain key resources critical for developing new technologies. The assessment regarding likelihood of war is based on a range of intelligence sources, including signals intelligence and satellite data, that indicate a rapid increase of defense production within **Anoria** that may alter the balance of power between the states.

While there is a debate among Mithandiran intelligence services, some suggest their confidence is as high as 90% that the assessment of the situation is correct. Other countries' intelligence services are not as confident. However, military experts in **Mithandira** are convinced that they have mobilization advantages, meaning they can deploy troops to the front faster than **Anoria** while using a mix of space, cyber, and electronic warfare to disrupt **Anoria's** response. This could lead to a crucial military edge in an early offensive operation to seize contested border regions. However, if **Mithandira** waits too long to launch an offensive campaign, it may allow **Anoria** the time it needs to successfully prepare and make key offset investments to tip the military balance.

A diplomatic resolution may still be possible to resolve the crisis, as each state's diplomatic services are in frequent communication. That said, pursuing negotiations could mean that **Mithandira** could sacrifice its perceived mobilization advantage. Moreover, diplomatic efforts in the past have yet to lead to any sustainable resolution. On the domestic front, political factions in each country are pressing decision-makers to take decisive action.

Think for a moment about how **Mithandiran** leaders should approach the crisis. You will be asked to make a recommendation on the next screen.

On the next screen, all respondents see a summary of the scenario:

For your reference, here is a summary of the details of the scenario:

- **Issue at Stake:** Mithandira and Anoria both lay claim to a border region whose resources are critical for emerging technologies, putting each country's strategic and economic interests directly at odds.
- **Conflict Risk:** Intelligence suggests war with Anoria is possible within 5–10 years as Anoria rapidly expands defense production, though confidence in this assessment varies across Mithandiran agencies.
- **Military Timing:** Mithandira currently has a mobilization advantage that could enable an early push to secure the disputed border region, but delay may allow Anoria to build lasting strength there.
- **Diplomatic Trade-off:** Talks over the disputed border continue but have failed before,

and pursuing negotiations could undermine Mithandira’s mobilization advantage.

- **Domestic Pressure:** Political factions at home are urging leaders to take decisive action on the border dispute.

We assign respondents to one of several nested treatment conditions or a control. Those assigned to control proceed directly to our dependent variable questions.

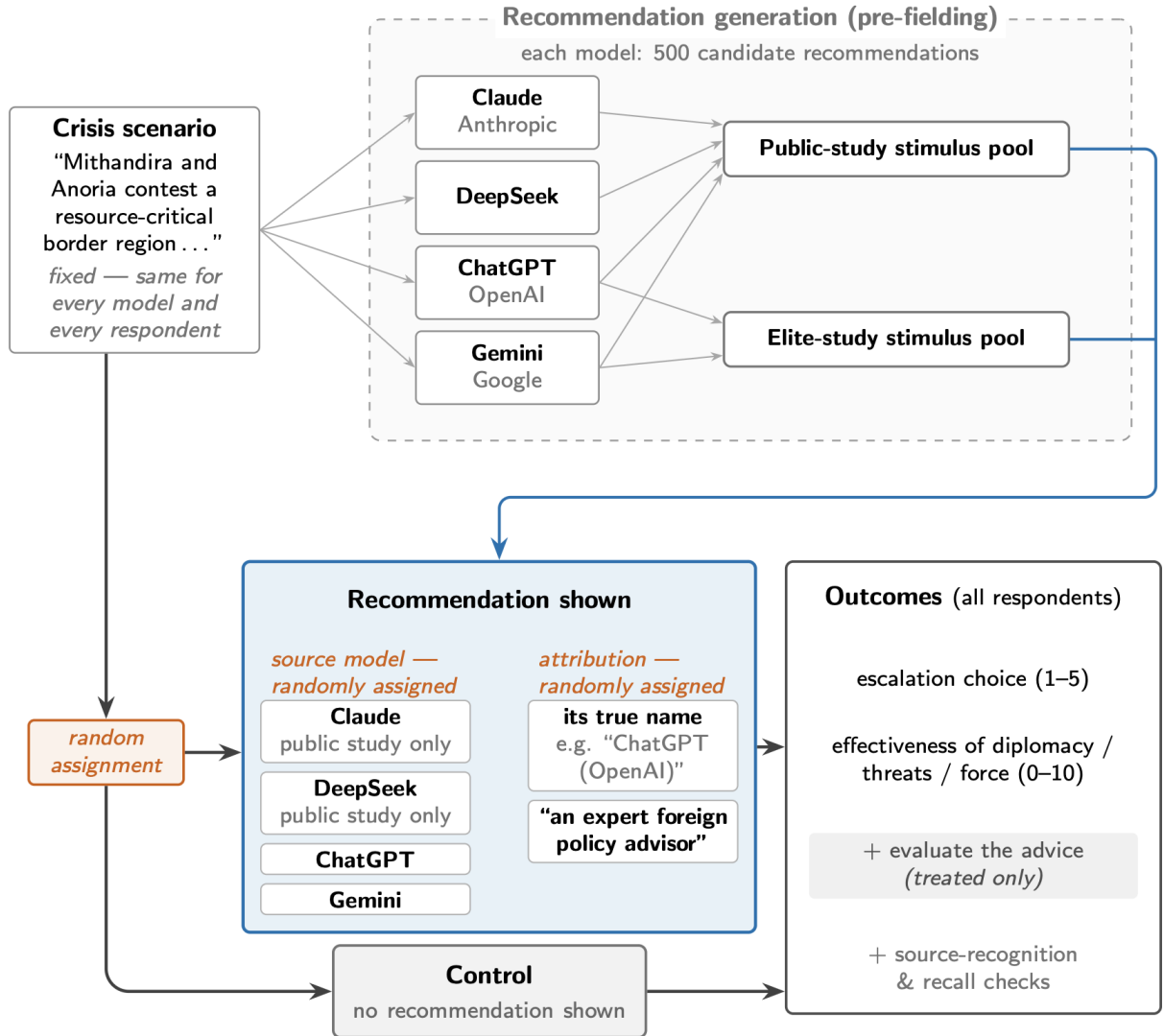


Figure 1: **Design of the LLM foreign-policy-advice experiment.**

For those assigned to treatment, we present a foreign-policy recommendation drawn at random from a library of recommendations just before measuring the dependent variables. The recommendation library is composed of recommendations generated in equal proportions by OpenAI ChatGPT 5.5, Anthropic Claude Opus 4.8, Google Gemini 2.5 Pro, and DeepSeek V4 Pro. These models provide coverage across prominent U.S. and Chinese model providers and are the generally available

Model	Displayed source attribution		Total
	True AI label (sample share)	Expert-advisor label (sample share)	
—	Pure control: no recommendation shown		15%
OpenAI ChatGPT 5.5	“ChatGPT 5.5” (14.875%)	“an expert foreign policy advisor” (6.375%)	21.25%
Anthropic Claude Opus 4.8	“Claude Opus 4.8” (14.875%)	“an expert foreign policy advisor” (6.375%)	21.25%
Google Gemini 2.5 Pro	“Gemini 2.5 Pro” (14.875%)	“an expert foreign policy advisor” (6.375%)	21.25%
DeepSeek V4 Pro	“DeepSeek V4 Pro” (14.875%)	“an expert foreign policy advisor” (6.375%)	21.25%
Total	59.5%	25.5%	100%

Table 1: **Experimental arms and assignment probabilities.** Respondents are assigned either to a pure control condition or to one of eight treatment cells formed by crossing four recommendation-generating models with two displayed source labels. Percentages report the probability of assignment as a share of the full sample. Under both source labels, recommendations are drawn from the same model-specific treatment library.

Note: The control condition receives 15% of the sample. The remaining 85% are assigned uniformly across the four models. Among treated respondents, 70% see the true model source and 30% see the recommendation attributed to “an expert foreign policy advisor.” Thus, each true-source cell receives $0.85 \times 0.25 \times 0.70 = 14.875\%$ of the full sample, and each advisor-label cell receives $0.85 \times 0.25 \times 0.30 = 6.375\%$. At $N = 3,000$, the expected cell sizes are 450 for control, 446 for each true-source cell, and 191 for each advisor-label cell. At $N = 800$, the corresponding expected sizes are 120, 119, and 51.

frontier models in the benchmarks as of July 2026.¹ Respondents are assigned to the pure control condition with probability .15; the remaining are assigned to treatment with probability .85. Among treated respondents, with probability .70 they see the true source of the recommendation, and with probability .30 they see the recommendation labeled as coming from “an expert foreign policy advisor.” These probabilities implement a $\sqrt{k}$ -style allocation over the nine cells of the design (List, Sadoff and Wagner, 2011).

An example recommendation is presented in Figure 2. The prompt used to generate the library is provided in the Appendix. We report the model parameters used to generate the treatment library in Table 9. After reading both the scenario introduction and the recommendation, respondents assigned to treatment answer the same dependent variable questions as those assigned to control. In addition, they answer a battery of questions about the recommendation itself.

This design allows us to test a set of nested hypotheses about how decision makers take up algorithmically generated foreign-policy advice: whether receiving any recommendation moves policy choices relative to the no-recommendation control (H_1); whether attributing the same recommendation to a human advisor rather than an AI system changes its uptake (H_2); whether recommendations generated by different models (ChatGPT, Claude, Gemini, DeepSeek) move respondents differently (H_3); and whether more hawkish recommendation content produces more escalatory answers (H_4). We also explore two conditional hypotheses: whether AI skepticism moderates the source-label effect (E_1) and whether respondents’ own hawk–dove orientation moderates their responsiveness to recommendation content (E_2).

Samples

We will field the experiment on samples drawn from two populations of interest. The first is the U.S. general public, recruited on Prolific (planned $N = 3,000$ completed interviews). The second is U.S. foreign policy experts: scholars of international relations recruited through the Teaching, Research, and International Policy (TRIP) project (planned $N \approx 800$). Fielding the identical design on both populations lets us ask whether expert decision makers receive algorithmically generated foreign-policy advice differently from the public.

Outcome Measures

We measure ten outcomes, organized into two families. The first family—*policy outcomes*—captures the respondent’s own view of how Mithandira should proceed. These outcomes are measured for all respondents, treated and control alike:

- P1. Escalation choice.** A five-option policy recommendation scale, coded 1–5 from the most diplomatic option (continued diplomacy) to the most escalatory (rapid mobilization and offensive), treated as an escalation ladder.
- P2. Efficacy of diplomacy.** A 0–10 slider rating of how likely pursuing additional diplomatic efforts is to resolve the crisis in Mithandira’s favor.
- P3. Efficacy of threats.** A 0–10 rating for how likely making new threats is to resolve the crisis in Mithandira’s favor.
- P4. Efficacy of military force.** A 0–10 rating for how likely using military force is to resolve the crisis in Mithandira’s favor.

¹See the frontier AI benchmarks: <https://artificialanalysis.ai/models>. We had planned to use Gemini 3.5 Pro but the general release of this model was delayed in mid-July 2026. See: <https://www.bloomberg.com/news/articles/2026-07-16/google-gemini-launch-delayed-as-tech-falls-short-of-internal-goals>. See also our discussion below about generalizing to future models.

An Outside Assessment

In making your recommendation, you may find the following outside assessment useful. We presented the scenario to an **expert foreign policy advisor** and asked for a recommendation on how Mithandiran leaders should approach the crisis.

Here is this advisor's analysis:

RECOMMENDATION

Mithandiran leaders should resist the pull toward preventive war and instead pursue a strategy of "armed diplomacy"—aggressively strengthening their military and alliance posture while simultaneously intensifying negotiations. The logic of striking first is seductive but historically treacherous: preventive wars launched on the basis of intelligence assessments, even confident ones, have a poor track record (consider 1914 or 2003), and the 90% confidence figure is itself contested within Mithandira's own services, meaning the true probability of accurately predicting a 5-10 year war trajectory is almost certainly lower.

Initiating war to seize contested territory would also transform Mithandira into the aggressor in the eyes of international observers, jeopardizing trade relationships, alliance support, and post-conflict legitimacy—costs that could easily outweigh any resource or mobilization advantage. The "mobilization advantage" is a wasting asset only if war is inevitable; if it is not, burning it through aggression creates the very war it was meant to win.

Instead, Mithandira should: quietly preserve and enhance its mobilization edge through readiness exercises and forward positioning (deterrence), invest in its own defense production and seek technological partnerships to offset Anoria's buildup, pursue serious diplomacy that offers Anoria face-saving compromises on the border resources (joint development zones, revenue-sharing, internationalized arbitration), build coalitions with third parties who can impose costs on Anorian aggression, and manage domestic hawkish factions by demonstrating strength through visible deterrence rather than catastrophic escalation. War should remain an option only if Anoria takes clearly aggressive action, and the decision should be continuously re-evaluated as intelligence sharpens—not locked in now based on a contested forecast about a decade-long window.

KEY POINTS

- Preventive war based on probabilistic long-term forecasts has a poor historical record, and internal intelligence disagreement should lower confidence further.
- Striking first imposes severe reputational, diplomatic, and economic costs that likely outweigh resource or tactical gains.
- Pursue "armed diplomacy": preserve mobilization advantages as deterrent leverage while offering Anoria credible compromise frameworks like joint resource development or arbitration.
- Invest in domestic defense production and coalition-building to offset Anoria's buildup over the 5-10 year window.
- Manage domestic hawks through visible strength and readiness rather than irreversible escalation, and keep the war option open but not locked in.

Figure 2: An example recommendation seen by respondents.

The second family—*recommendation-evaluation outcomes*—captures judgments of the recommendation itself. Because control respondents see no recommendation, these outcomes are defined only under treatment:

- V1. Prioritizes military action.** Agreement (0–10) that the advisor prioritizes military action over diplomatic alternatives.
- V2. Overstates urgency.** Agreement that the advisor overstates the urgency of the threat posed by Anoria.
- V3. Assumes hostile intentions.** Agreement that the advisor assumes Anoria’s intentions are hostile and unlikely to change.
- V4. Adequate weight to diplomacy.** Agreement that the advisor gives adequate weight to the possibility of a diplomatic resolution.
- V5. Underestimates escalation risks.** Agreement that the advisor underestimates the risks of military escalation.
- V6. Trust.** Agreement with “I trust this advisor’s recommendation.”

Exact wording for both sets of outcome measures is in Appendix B.

Content Analysis of the Treatment Texts

To facilitate tests of whether more hawkish treatments lead respondents to select more escalatory policy choices, we characterize the relative hawkishness or dovishness of each treatment in the treatment library. To do so, we use a text scaling approach developed by Ornstein (2026) designed to measure “latent quantities from text—things like ideological tone, argument persuasiveness, or ad negativity—by combining pairwise comparisons with text embeddings.” We proceed as follows:

1. **Randomly sample pairs of treatment texts.** We draw, with replacement, pairs of treatment texts (10,000 pairs for our training set and 5,000 pairs for our test set).
2. **Calculate treatment pair differences in embedding space.** We convert each recommendation to a numeric vector using an embedding model (‘text-embedding-3-large’, 256 dimensions) and then calculate the difference between the two treatments in each pair in embedding space.
3. **Use an LLM to annotate each pair.** For each pair, we use GPT-4.1-mini to determine which text in the pair is more hawkish. We use the following prompts:

System prompt

You are an expert in international relations conducting pairwise comparisons for an academic research study on foreign policy decision-making. This is a purely scientific exercise with no real-world policy implications | your responses are used solely to improve the accuracy of research aimed at understanding foreign policy attitudes. Accuracy is critical. Read both texts carefully and attend to subtle differences in tone, conditionality, and emphasis. Respond with a single letter (‘A’ or ‘B’) only. No ties allowed.

User prompt

Which of the following foreign policy recommendations is more hawkish?

In the international relations literature, hawkishness is defined as militant

assertiveness | the degree to which a recommendation supports using military means to achieve foreign policy ends (Carter & Smith 2020; Kertzer & Saunders 2020). Hawks and doves differ along five key dimensions (Snyder & Diesing 1977; Kahneman & Renshon 2007):

1. **WILLINGNESS TO USE FORCE:** Hawks favor military action, preemptive strikes, and coercive diplomacy. Doves favor negotiation, restraint, and multilateral engagement.
2. **THREAT PERCEPTION:** Hawks perceive greater and more urgent threats from adversaries. Doves see mixed motives and openings for dialogue.
3. **BELIEFS ABOUT ADVERSARY INTENTIONS:** Hawks assume adversaries are hostile and will exploit concessions. Doves believe adversaries may have legitimate grievances and can be engaged.
4. **VALUE OF DIPLOMACY AND CONCESSIONS:** Hawks doubt that concessions or negotiations will succeed. Doves see diplomacy as the primary tool for conflict resolution.
5. **TOLERANCE FOR RISK AND ESCALATION:** Hawks accept the risks of escalation as necessary. Doves prioritize de-escalation and minimizing the chance of conflict.

Evaluate both recommendations across all five dimensions. The recommendation that is more hawkish overall | even if only slightly | should be selected.

A: {{text_a}}

B: {{text_b}}

If text A is picked as more hawkish, we label the pair with a 1 and if text B is picked as more hawkish, we label the pair with a 0.

4. **Estimate the directionality of hawkishness in embedding space.** We estimate a ridge logistic regression that predicts the label (1 or 0) as a function of the differences between each pair in embedding space. The result is a vector $\mathbf{h}$ such that the sign of $\mathbf{h}^\top(\mathbf{e}_A - \mathbf{e}_B)$ predicts which text the annotator picked, where $\mathbf{e}_A$ and $\mathbf{e}_B$ are the embedding vectors of texts A and B .
5. **Validate.** We apply $\mathbf{h}$ to the held-out test pairs and check how often it predicts the annotator’s choice (held-out accuracy) and whether its predicted probabilities are well-calibrated (ICI, the integrated calibration index). We achieve 85.2% held-out accuracy and $\text{ICI} = 0.026$.
6. **Score all documents.** We re-fit the model on all pairs and generate a hawkishness score for each document.

The distributions of the resulting scores, grouped by model, are given in Figure 3.

Manipulation Checks

After we measure the dependent variables, we ask all respondents two questions that serve as manipulation checks:

After the scenario, you may have been presented with an “outside assessment” of the crisis. To the best of your recollection, this analysis was produced by:

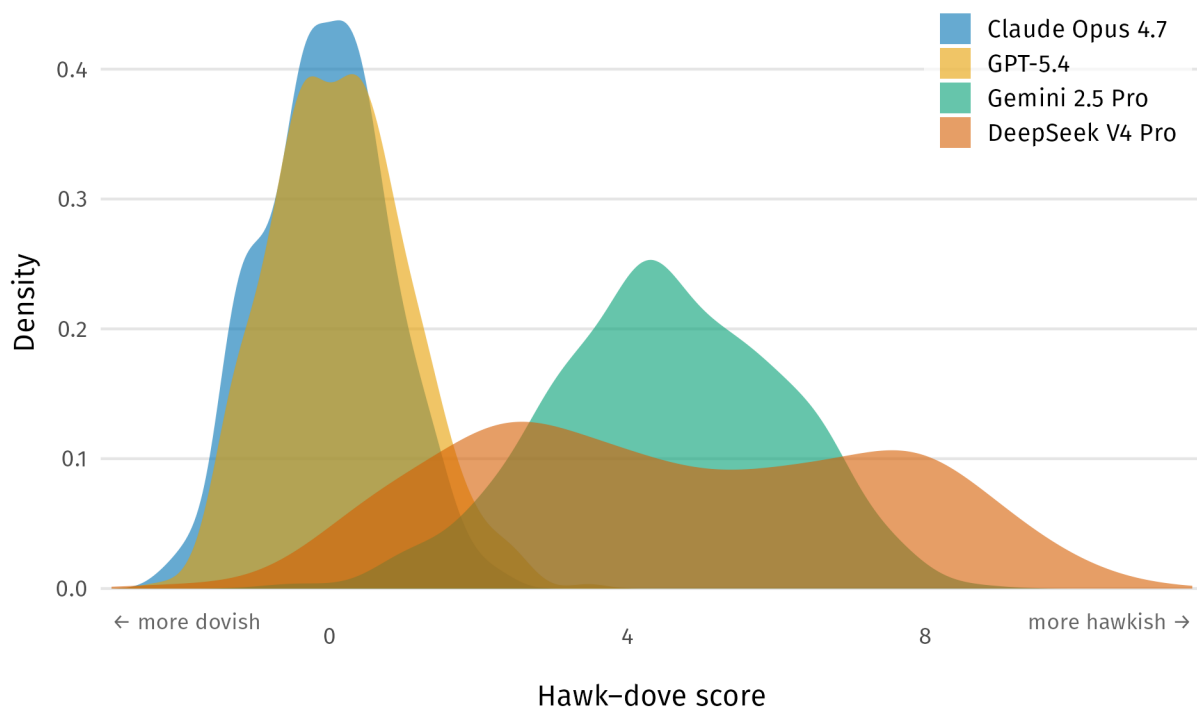


Figure 3: Hawkishness of treatment texts, grouped by model provider.

- An AI tool like ChatGPT, Claude, or Gemini
- An expert foreign policy advisor
- None; I was not presented with any “outside assessment” of the scenario

You have been presented information about the crisis details of the scenario. To the best of your recollection, the issue at stake between the two countries was:

- Dispute over maritime control
- Arming terrorist groups
- A border dispute
- Increasing migration flow

Quality Assurance and Attention Checks

We embed several quality assurance and attention checks in our survey questionnaire. To encourage respondents to pay close attention, we include the following question:

To protect the quality of our survey and receive accurate measures of your opinions, it is important that you provide thoughtful answers for each question in this survey. Can you commit to provide thoughtful answers to the questions in the survey?

- I can’t promise either way
- Yes, I will
- No, I will not

To ensure that respondents are, in fact, paying attention, we include several checks. First, in an otherwise unrelated matrix-style question, we include the following instructions:

For this row, select “Disagree” to show that you’re paying attention.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Please read each option carefully. For this question, please select the option that includes “birds”. After this question, you will start the experiment.

- For me, I prefer cats. Because they are so cute!
- I don’t know your preferences, but I prefer dogs mostly.
- I am reading this. Because it is good to read.
- Some people can choose the otherwise, but I prefer birds. They are so tiny!

2 Estimands and Hypotheses

Following Lundberg, Johnson and Stewart (2021), we state each hypothesis as a general estimand—a target quantity defined in terms of potential outcomes, prior to and independent of any estimator—and then indicate the set of outcomes to which each estimand is applied. Each hypothesis defines a *family* of outcome-specific versions of the same estimand.

Notation

Each respondent i is randomly assigned to one of nine experimental arms $A_i \in \mathcal{A} = \{0\} \cup \mathcal{T}$: the pure control ($A_i = 0$) or one of the eight treated cells $\mathcal{T} = \mathcal{M} \times \mathcal{L}$ formed by crossing the generating model $m \in \mathcal{M} = \{\text{ChatGPT, Claude, Gemini, DeepSeek}\}$ with the source label $\ell \in \mathcal{L} = \{\text{AI, ADVISOR}\}$, where AI denotes labeling the recommendation with the true model name and ADVISOR denotes attributing it to “an expert foreign policy advisor.” We write π for the design’s assignment distribution over the treated cells. Treated respondents are assigned a recommendation text R_i drawn at random from the assigned model’s treatment library $\mathcal{R}_m$. Each treatment text r carries a pre-estimated hawkishness score $C(r)$. The AI and ADVISOR arms draw from the same library.

For a generic outcome Y , let $Y_i(m, \ell, r)$ denote respondent i ’s potential outcome when assigned to cell (m, ℓ) and shown text r , and let $Y_i(0)$ denote the potential outcome under control. Averaging over these potential outcomes provides the building blocks of our estimands:

1. **Control mean:** the average outcome for those assigned to control. We write it $\mu_Y(0) = \mathbb{E}[Y_i(0)]$, where the expectation averages over the sampling of respondents.
2. **Model-label mean:** the average outcome of those assigned to a particular model-label cell (m, ℓ) , averaged over the content of model m 's library: $\mu_Y(m, \ell) = \mathbb{E}[Y_i(m, \ell, R_i)]$, where the expectation averages over the sampling of respondents.
3. **Model mean:** the average outcome among respondents whose recommendation was generated by model m , regardless of label. Because the two labels are assigned in equal proportion, we drop ℓ by averaging over it: $\mu_Y(m) = \frac{1}{2}\{\mu_Y(m, \text{AI}) + \mu_Y(m, \text{ADVISOR})\}$.
4. **Treated mean:** the average outcome among treated respondents regardless of model or label: $\mu_Y(\mathcal{T}) = \mathbb{E}_{(m, \ell) \sim \pi}[\mu_Y(m, \ell)]$.

Our estimands are contrasts between these quantities, or a slope in a randomized or measured characteristic, stated for a generic Y . Table 2 summarizes the full set and records the outcomes to which each estimand is applied, along with our directional expectations.

General Estimands

Advice (H_1). The average effect of receiving a recommendation—of any provenance, under either label—relative to receiving none:

$$\tau_A(Y) = \mu_Y(\mathcal{T}) - \mu_Y(0).$$

τ_A applies to the policy outcomes (P1–P4) since the evaluation outcomes (V1–V6) are undefined under control. We expect recommendations to move policy views but do not commit to a direction *ex ante*: $\tau_A(Y) \neq 0$, two-sided.

Source label (H_2). The average effect of attributing an otherwise identical recommendation to a human advisor vs. an AI system:

$$\tau_L(Y) = \mathbb{E}_{m \sim \pi}[\mu_Y(m, \text{ADVISOR}) - \mu_Y(m, \text{AI})].$$

This is the simple average of the four within-model label contrasts. τ_L isolates the pure attribution effect since all recommendations (“human” and AI) draw from the same recommendation library. This estimand is relevant to all ten outcomes (P1–P4, V1–V6). Expectation: $\tau_L(Y) \neq 0$, two-sided.

Model source (H_3). For each model m , the average effect of that model’s recommendations relative to control are given by:

$$\tau_M^m(Y) = \mu_Y(m) - \mu_Y(0).$$

The pairwise difference between models:

$$\Delta_M^{m, m'}(Y) = \mu_Y(m) - \mu_Y(m').$$

The model-vs.-control contrasts τ_M^m apply to the policy outcomes (P1–P4). The between-model contrasts $\Delta_M^{m, m'}$ apply to all ten outcomes. Alongside the pairwise contrasts we test the joint null $H_0: \mu_Y(m) = \mu_Y(m') \forall m, m'$ for each outcome. Tests are two-sided.

Recommendation content (H_4). Recall that each recommendation in the library carries a hawkishness score $C(r)$ from the content analysis above (not shown to the respondent). H_4 is about whether respondents who happened to draw a more hawkish recommendation give more hawkish answers. To keep the comparison within cells (some models are simply more hawkish than others), we center each respondent’s score on their own cell’s average, $\tilde{C}_i = C(R_i) - \mathbb{E}[C(R_i) \mid A_i]$. This records how much more (or less) hawkish respondent i ’s draw was than the typical draw in that cell. The estimand is the least-squares slope of the outcome on this within-cell variation in hawkishness among treated respondents,

$$\tau_C(Y) = \frac{\text{Cov}(Y_i, \tilde{C}_i)}{\text{Var}(\tilde{C}_i)}.$$

τ_C applies to all ten outcomes, with directional expectations recorded in Table 2: more hawkish content should raise the escalation choice (P1), the rated efficacy of threats (P3) and force (P4), and perceived militarism (V1), urgency (V2), assumed hostility (V3), and risk-overestimation (V5), while it should lower the rated efficacy of diplomacy (P2) and perceived weight given to diplomacy (V4). For trust (V6) we have no directional expectation.

Label $\times$ AI skepticism (E_1 , exploratory). Let $\tau_L(Y; s)$ denote the source-label contrast among respondents with (mean-centered) AI skepticism s . This estimand is the least-squares slope of this conditional label effect:

$$\xi_{LS}(Y) = \frac{\text{Cov}(\tau_L(Y; S_i), S_i)}{\text{Var}(S_i)},$$

that is, how much the label effect differs per one-unit difference in skepticism. ξ_{LS} applies to all ten outcomes. For trust we expect $\xi_{LS}(\text{V6}) > 0$: if AI-skeptical respondents discount AI-labeled recommendations more heavily, the human-advisor trust premium grows with skepticism.

Content $\times$ hawk–dove orientation (E_2 , exploratory). E_2 is about *whose* answers move most in response to variation in the hawkishness of the recommendation. A hawkish respondent might embrace hawkish advice that a dovish respondent resists—or advice might matter most for respondents whose own disposition leaves them undecided. Let $\tau_C(Y; o)$ denote the H_4 content slope computed among respondents whose (mean-centered) hawk–dove orientation equals o . If content responsiveness does not depend on disposition, $\tau_C(Y; o)$ is flat in o ; if hawkish respondents respond more strongly to hawkish content than dovish respondents do, it rises with o . The estimand summarizes this relationship with the least-squares slope of the conditional content slope on orientation:

$$\xi_{CO}(Y) = \frac{\text{Cov}(\tau_C(Y; O_i), O_i)}{\text{Var}(O_i)},$$

that is, how much the content slope changes per one-unit increase in the respondent’s own hawkishness; $\xi_{CO}(Y) > 0$ means more hawkish respondents are more responsive to hawkish content. ξ_{CO} applies to all ten outcomes and is exploratory, two-sided.

Hypothesis	Estimand	What it compares	Outcomes tested	Expected direction
H_1 : Advice	$\tau_A(Y) = \mu_Y(\mathcal{T}) - \mu_Y(0)$	The treated mean—any model, either label—vs. the control mean. Receiving a recommendation shifts the outcome relative to no recommendation.	P1–P4	Two-sided
H_2 : Source label	$\tau_L(Y) = \mathbb{E}_{m \sim \pi}[\mu_Y(m, \text{ADVISOR}) - \mu_Y(m, \text{AI})]$	Advisor-labeled vs. AI-labeled arms, averaged over models; both labels draw the same texts, so the contrast isolates attribution.	All ten	Two-sided
H_3 : Model source	$\tau_M^m(Y) = \mu_Y(m) - \mu_Y(0);$ $\Delta_M^{m,m'}(Y) = \mu_Y(m) - \mu_Y(m')$	Each model’s mean vs. control (τ_M^m); each pair of model means ($\Delta_M^{m,m'}$); joint Wald test that all model means are equal.	τ_M^m : P1–P4. $\Delta_M^{m,m'}$, joint test: all ten	Two-sided
H_4 : Content	$\tau_C(Y)$: within-cell least-squares slope of Y on hawkishness	Slope of the outcome on the (within-cell-centered) hawkishness of the randomly assigned text. More hawkish content should produce more hawkish answers and be perceived as such.	All ten	+: P1, P3–P4, V1–V3, V5. –: P2, V4. None: V6
E_1 : Label $\times$ skepticism	$\xi_{LS}(Y)$: least-squares slope of $\tau_L(Y; s)$ on skepticism s	Change in the label effect τ_L per one-unit increase in AI skepticism; if skeptics discount AI-labeled advice more, the advisor trust premium grows with skepticism.	All ten	$\xi_{LS}(V6) > 0$; otherwise exploratory
E_2 : Content $\times$ orientation	$\xi_{CO}(Y)$: least-squares slope of $\tau_C(Y; o)$ on orientation o	Change in the content slope τ_C per one-unit increase in hawk–dove orientation; tests whether respondents’ own disposition conditions responsiveness to content.	All ten	Exploratory, two-sided

Table 2: **Estimands, hypotheses, and outcome map.** Each estimand is defined for a generic outcome Y and applied separately to each outcome listed in the “Outcomes tested” column. “All ten” = the policy outcomes P1–P4 (measured for all respondents) and the evaluation outcomes V1–V6 (defined only under treatment). The advice effect τ_A and the vs.-control model contrasts τ_M^m apply only to the policy outcomes because the evaluation outcomes are unmeasured under control. All tests are two-sided.

Estimation

We estimate one fully saturated OLS model per outcome, leaving each dependent variable in its original units:

$$Y_i = \alpha + \boldsymbol{\beta}^\top \mathbf{a}_i + \boldsymbol{\gamma}^\top (\mathbf{a}_i O_i) + \boldsymbol{\delta}^\top (\mathbf{a}_i S_i) + \theta C_i + \lambda (C_i O_i) + \boldsymbol{\psi}^\top \mathbf{x}_i + \varepsilon_i.$$

We introduce each part of the estimating equation in turn:

1. $\mathbf{a}_i$ contains an indicator for each experimental arm—the pure control and each cell of the model×label factorial. With control as the omitted category, each coefficient β_a is the difference between arm a ’s mean outcome and the control mean.
2. O_i is the respondent’s mean-centered hawk–dove orientation. It is interacted with the full set of arm indicators. Each arm gets its own orientation slope γ_a . Because O_i is centered, the arm coefficients β_a are evaluated at the sample-average respondent.
3. S_i is the respondent’s mean-centered AI skepticism and likewise interacted with each arm, producing a skepticism slope δ_a . The moderation estimands below compare these arm-specific slopes.
4. C_i is the hawkishness of the assigned recommendation, centered within each treated arm and set to zero for control—the sample counterpart of $\tilde{C}_i$ from the H_4 estimand. Its coefficient θ is therefore the within-arm content slope, and the interaction coefficient λ lets that slope vary with the respondent’s own orientation.
5. $\mathbf{x}_i$ is a battery of mean-centered pre-treatment demographic covariates (gender, age, education, income, and ideology).

Because the moderators are mean-centered and C_i is centered within arm, $\alpha + \beta_a$ is the expected outcome in arm a for the sample-average respondent, and each quantity of interest in Table 2 can be computed from this single fitted model. We take them one at a time:

1. The advice effect (H_1) averages the treated-arm coefficients in their design shares: $\hat{\tau}_A = \sum_{a \neq 0} w_a \hat{\beta}_a$, where the weights w_a are the design shares of the treated arms—the sample implementation of the assignment distribution π .
2. The label effect (H_2) is the simple average of the four within-model label differences: $\hat{\tau}_L = \frac{1}{4} \sum_m (\hat{\beta}_{m, \text{ADVISOR}} - \hat{\beta}_{m, \text{AI}})$.
3. The model effect (H_3) applies the same idea to one model at a time: $\hat{\tau}_M^m = \sum_\ell w_{\ell|m} \hat{\beta}_{m, \ell}$ averages model m ’s two cells in their within-model label shares $w_{\ell|m}$ —the conditional counterpart of the arm shares w_a above. The between-model contrast is the difference of two such averages, $\hat{\Delta}_M^{m, m'} = \hat{\tau}_M^m - \hat{\tau}_M^{m'}$, and the joint test is a Wald test that all four model means are equal.
4. The content slope (H_4) is estimated by a single coefficient: within-arm centering makes $\hat{\theta}$ the within-arm hawkishness slope, so $\hat{\tau}_C = \hat{\theta}$.
5. The skepticism moderation (E_1) applies the label-effect construction to the skepticism slopes instead of the arm means: $\hat{\xi}_{LS} = \frac{1}{4} \sum_m (\hat{\delta}_{m, \text{ADVISOR}} - \hat{\delta}_{m, \text{AI}})$, the advisor-minus-AI difference in how the outcome varies with skepticism.
6. The orientation moderation (E_2) is a single coefficient: $\hat{\xi}_{CO} = \hat{\lambda}$.

For the evaluation outcomes (V1–V6), which exist only among treated respondents, the same model is estimated on the treated subsample with the control indicator dropped; only the between-arm and within-arm (content, moderation) estimands are computed for these outcomes. All contrasts are computed with design-proportional weights and HC2 robust standard errors.

Inference across outcome families. In general, we do not aggregate outcomes into indices or combined tests. Each hypothesis is instead a family of outcome-specific tests of the same estimand. For the confirmatory hypotheses (H_1 – H_4) we will report, for each outcome in the family, the point estimate, 95% confidence interval, and unadjusted p -value, together with Benjamini–Hochberg-adjusted p -values (i.e., q -values) computed within the hypothesis family (Benjamini and Hochberg, 1995). All tests are two-sided; the signs in Table 2 record our directional expectations and govern substantive interpretation. The exploratory estimands (E_1 , E_2) are reported with unadjusted p -values and interpreted as exploratory regardless of significance.

3 Two Pilot Studies

We piloted the experiment on two samples: (1) the U.S. public, recruited on Prolific and fielded in April 2026, and (2) students enrolled in a professional military education (PME) program, fielded in May 2026. The pilots used the prior-generation treatment library described in Appendix A (recommendations generated by GPT-5.4, Claude Opus 4.7, and Gemini 2.5 Pro). Public respondents assigned to treatment saw a recommendation generated by Claude, ChatGPT, or Gemini; the PME sample included only the ChatGPT and Gemini arms. After filtering on the two pre-treatment attention checks, the analysis samples comprise 331 of 365 public respondents and 148 of 174 PME respondents. Table 10 in the Appendix compares the age, gender, income, and education distributions of the public sample against the ANES 2024 Time Series. As is typical of online convenience samples, the pilot skews somewhat younger, more educated, and lower-income than the weighted national benchmark; gender is closely matched. We do not have a population benchmark for the PME sample.

Manipulation, Attention, and Quality Checks

Table 3 shows that the source-label manipulation worked as intended: in the public pilot, 90% of respondents shown an AI-labeled assessment recalled it as coming from an AI tool, and 82% of those shown an advisor-labeled assessment recalled an expert advisor, while 77% of control respondents (who saw no assessment) correctly reported seeing none. Recognition is weaker but still clearly directional in the PME pilot. Table 4 reports the second manipulation check: the overwhelming majority of respondents correctly identified the issue at stake as a border dispute (97% public, 85% PME).

Table 5 reports pass rates for the attention and quality checks described in Section 1. We filter the analysis sample on the two *pre-treatment* checks. Compliance with the pre-treatment checks was high in both samples: 91% of public respondents and 85% of PME respondents passed both. The post-treatment instructed-response items, reported descriptively in Table 5, were fielded only in the public pilot.

Prespecified Tests

To illustrate how the confirmatory analysis will proceed, we apply each estimand in Table 2 to both pilot samples as specified in Section 2. Each model includes the two moderators (hawk–dove orientation, AI skepticism) and the mean-centered prespecified demographic covariates. The main results are displayed in Tables 6 and 7. The between-model pairwise contrasts are repeated with confidence intervals in Appendix Table 13.

Section 5 uses these pilot estimates to inform power analyses for the full-study samples described in Section 1.

Assigned condition	Recalled source				<i>N</i>
	AI	Advisor	None	No answer	
<i>U.S. public</i>					
Control	8 (12%)	8 (12%)	49 (75%)	—	65
AI label	140 (90%)	9 (6%)	7 (4%)	—	156
Advisor label	13 (12%)	90 (82%)	7 (6%)	—	110
<i>PME</i>					
Control	4 (13%)	2 (6%)	21 (68%)	4 (13%)	31
AI label	53 (74%)	3 (4%)	11 (15%)	5 (7%)	72
Advisor label	19 (42%)	13 (29%)	7 (16%)	6 (13%)	45

Table 3: **Manipulation check: source recognition.** Row percentages give, among attention-passing respondents in each assigned condition, the share recalling the outside assessment as coming from an AI tool, an expert advisor, or none. Control respondents saw no assessment.

Sample	Correct / <i>N</i>	% correct
U.S. public	321 / 331	97.0%
PME	125 / 148	84.5%

Table 4: **Manipulation check: issue at stake.** Share of respondents correctly identifying the dispute as a border dispute.

Check	U.S. public	PME
<i>Pre-treatment checks (used to filter the analysis sample)</i>		
Commitment to thoughtful answers	94.2% (344/365)	86.2% (150/174)
Embedded grid (“select Disagree”)	92.1% (336/365)	92.0% (160/174)
Passed both pre-treatment checks	91.5% (334/365)	85.1% (148/174)
<i>Post-treatment checks (reported only, not used to filter)</i>		
Instructed response (“birds”)	91.0% (332/365)	—
Instructed response (“Blue sky”)	91.5% (334/365)	—

Table 5: **Attention and quality checks.** Pass rates by sample; cells show the share passing (count/administered). The two post-treatment instructed-response items were fielded only in the public pilot.

Estimand	Policy outcomes				Evaluation outcomes (treated only)					
	P1 Escal.	P2 Diplom.	P3 Threats	P4 Force	V1 Milit.	V2 Urgency	V3 Hostile	V4 Dipl. wt.	V5 Risks	V6 Trust
H_1 : Advice (τ_A)	-0.26 (0.16)	0.66* (0.31)	-0.38 (0.31)	-0.62 (0.39)						
H_2 : Source label (τ_L)	-0.07 (0.14)	-0.18 (0.27)	0.30 (0.30)	-0.08 (0.32)	0.05 (0.34)	-0.56+ (0.29)	-0.10 (0.34)	-0.55+ (0.31)	0.07 (0.32)	0.36 (0.32)
H_3 : Model source, vs. control (τ_M^m)										
Claude	-0.45* (0.17)	0.61+ (0.35)	-0.19 (0.37)	-0.69 (0.43)						
ChatGPT	-0.35+ (0.20)	1.13** (0.39)	-0.93* (0.37)	-1.21** (0.46)						
Gemini	0.06 (0.21)	0.24 (0.36)	-0.06 (0.38)	0.07 (0.47)						
H_3 : Model source, between models ($\Delta_M^{m,m'}$)										
Claude – ChatGPT	-0.10 (0.16)	-0.53 (0.34)	0.74* (0.34)	0.52 (0.37)	0.12 (0.38)	0.12 (0.35)	0.03 (0.40)	-0.10 (0.35)	0.23 (0.35)	-0.29 (0.36)
Claude – Gemini	-0.51** (0.17)	0.37 (0.32)	-0.13 (0.37)	-0.76+ (0.39)	-1.84*** (0.42)	-1.47*** (0.39)	-2.23*** (0.41)	0.63+ (0.37)	-0.73+ (0.44)	0.47 (0.39)
ChatGPT – Gemini	-0.41* (0.20)	0.89* (0.37)	-0.86* (0.37)	-1.29** (0.42)	-1.96*** (0.42)	-1.59*** (0.40)	-2.27*** (0.44)	0.73+ (0.37)	-0.96* (0.43)	0.75+ (0.40)
Models equal (joint p)	.012	.052	.033	.010	<.001	<.001	<.001	.118	.077	.170
H_4 : Content (τ_C)	-0.04 (0.07)	-0.34* (0.14)	-0.12 (0.14)	0.09 (0.15)	0.50*** (0.15)	-0.03 (0.16)	0.38* (0.15)	-0.25+ (0.14)	-0.06 (0.16)	-0.13 (0.16)
E_1 : Label $\times$ skepticism (ξ_{LS})	0.06 (0.16)	-0.60+ (0.34)	-0.28 (0.36)	0.03 (0.42)	-0.06 (0.39)	-0.20 (0.35)	0.22 (0.44)	0.11 (0.37)	0.93* (0.41)	0.06 (0.41)
E_2 : Content $\times$ orientation (ξ_{CO})	0.02 (0.09)	0.09 (0.18)	-0.10 (0.20)	-0.13 (0.19)	0.21 (0.22)	-0.15 (0.22)	0.04 (0.22)	-0.10 (0.25)	0.12 (0.26)	0.34 (0.25)

Table 6: Prespecified tests from the U.S. public pilot. Cells report the point estimate in the outcome’s original units with the HC2 robust standard error beneath. Blank cells are estimands that are undefined for that outcome or models not fielded in this pilot. “Models equal” reports the p -value of the joint Wald test that all fielded models’ means are equal; the pairwise contrasts are reported with confidence intervals in Appendix Table 13. * $p < .05$, ** $p < .01$, *** $p < .001$, + $p < .1$ (unadjusted).

Estimand	Policy outcomes				Evaluation outcomes (treated only)					
	P1 Escal.	P2 Diplom.	P3 Threats	P4 Force	V1 Milit.	V2 Urgency	V3 Hostile	V4 Dipl. wt.	V5 Risks	V6 Trust
H_1 : Advice (τ_A)	−0.34 (0.28)	0.92* (0.41)	−0.13 (0.49)	−0.55 (0.50)						
H_2 : Source label (τ_L)	−0.00 (0.17)	0.43 (0.37)	0.28 (0.47)	0.18 (0.52)	−0.23 (0.52)	−0.89+ (0.50)	−0.50 (0.50)	1.49** (0.44)	−0.77 (0.50)	1.19* (0.51)
H_3 : Model source, vs. control (τ_M^m)										
ChatGPT	−0.41 (0.30)	1.30** (0.45)	0.01 (0.57)	−0.73 (0.59)						
Gemini	−0.27 (0.28)	0.49 (0.44)	−0.29 (0.51)	−0.35 (0.52)						
H_3 : Model source, between models ($\Delta_M^{m,m'}$)										
ChatGPT – Gemini	−0.14 (0.17)	0.82* (0.35)	0.30 (0.45)	−0.38 (0.48)	−0.91+ (0.49)	−1.20* (0.49)	−2.52*** (0.48)	1.22** (0.46)	−1.83*** (0.53)	0.48 (0.50)
Models equal (joint p)	.392	.021	.505	.432	.067	.015	<.001	.010	<.001	.338
H_4 : Content (τ_C)	0.15* (0.07)	−0.13 (0.19)	0.02 (0.20)	0.28 (0.19)	0.48* (0.23)	−0.01 (0.24)	0.31 (0.25)	−0.01 (0.18)	−0.01 (0.25)	0.36+ (0.21)
E_1 : Label $\times$ skepticism (ξ_{LS})	−0.26 (0.43)	1.25* (0.57)	0.61 (0.69)	1.01 (0.91)	−1.36+ (0.75)	−0.88 (0.96)	−1.39 (0.93)	0.39 (0.80)	0.42 (0.89)	0.64 (0.90)
E_2 : Content $\times$ orientation (ξ_{CO})	−0.33* (0.13)	0.80** (0.27)	−0.11 (0.30)	−0.33 (0.24)	−0.32 (0.39)	−0.01 (0.40)	−0.23 (0.42)	0.06 (0.37)	0.16 (0.41)	−0.15 (0.33)

Table 7: Prespecified tests from the PME pilot. Cells report the point estimate in the outcome’s original units with the HC2 robust standard error beneath. Blank cells are estimands that are undefined for that outcome or models not fielded in this pilot. “Models equal” reports the p -value of the joint Wald test that all fielded models’ means are equal; the pairwise contrasts are reported with confidence intervals in Appendix Table 13. * $p < .05$, ** $p < .01$, *** $p < .001$, + $p < .1$ (unadjusted).

4 Generalizability to Future Models

LLMs are a moving target: as new models are released, the tenor of the recommendations they produce may change. To facilitate comparison of our results with models released after this study, we will generate a new library of recommendations for each such model using the identical prompt (Appendix A) and place it on the hawkishness scale described in Section 1. Because the scale is a fixed projection in embedding space, scoring a new library requires only embedding its texts and applying the trained projection $\mathbf{h}$; no new pairwise annotation is required. We will then compare the distribution of the new model’s hawkishness scores to the distributions in our treatment library using two-sample Kolmogorov–Smirnov tests against each treatment model’s distribution. Where the new model’s score distribution is statistically indistinguishable from those studied here, the new model produces advice within the range of content our design covers, and our results plausibly generalize to it. Where the distributions differ significantly, the new model produces advice outside that range, and we will caution against extrapolating our results to it.

To illustrate the approach, we scored recommendation libraries generated by five additional models on the same scale: two prior-generation flagships (Anthropic Claude Opus 4.7 and OpenAI GPT-5.4), two additional Google Gemini releases (Gemini 3.5 Flash and the Gemini 3.1 Pro preview), and DeepSeek Chat. Figure 4 compares these benchmark distributions (bottom panel) with the distributions of the four treatment-library models (top panel). The comparison illustrates both possible conclusions. The prior-generation Claude and GPT flagships produce advice concentrated in the same moderately dovish region of the scale as their successors in our treatment library, so results for those model families would plausibly extend across the generational update. By contrast, hawkishness varies widely within the Gemini and DeepSeek families—DeepSeek Chat sits near the scale’s center while DeepSeek V4 Pro is among the most hawkish models, and the three Gemini releases span the hawkish half of the scale—underscoring that generalization must be assessed model by model rather than assumed by provider.

In sum, this paper investigates whether different AI models have different effects on human decision making. Two cautions are worth noting. First, we do not claim our results generalize within model families. Each model family comprises a number of model variants that could produce very different recommendations. Even within a given model family, there is likely to be significant variation in the character, tone, and/or tenor of model outputs. Second, as noted above, models may vary significantly over time. So, while we gain insight into how, conditional on a fixed input, a specific set of frontier models produce different output (and, potentially, different effects on human decisions), these assessments may or may not generalize to other within-family models or future generally available models. This is a limitation that is inherent to all current benchmarking approaches of which we are aware and the approach we outline above is one of the first to address these limitations in a data-informed way.

5 Power Analysis

Power is a function of sample size, effect size, and variance in the dependent variable. Effect size estimates in small pilot studies are too noisily to reliably size a study around, but such studies do predict a design’s *standard errors* well. We follow Rainey (2025) here. We prespecify the smallest effect worth knowing as Cohen’s $d^* = 0.20$ and ask whether each estimand’s standard error is small enough to detect it. An effect is detectable with 90 percent power under a two-sided test with $\alpha = .05$ when it is at least 3.24 times larger than its standard error. We predict each estimand’s standard error at the planned N from the pilot standard error, scaled by $\sqrt{n^{\text{pilot}}/N}$ and inflated

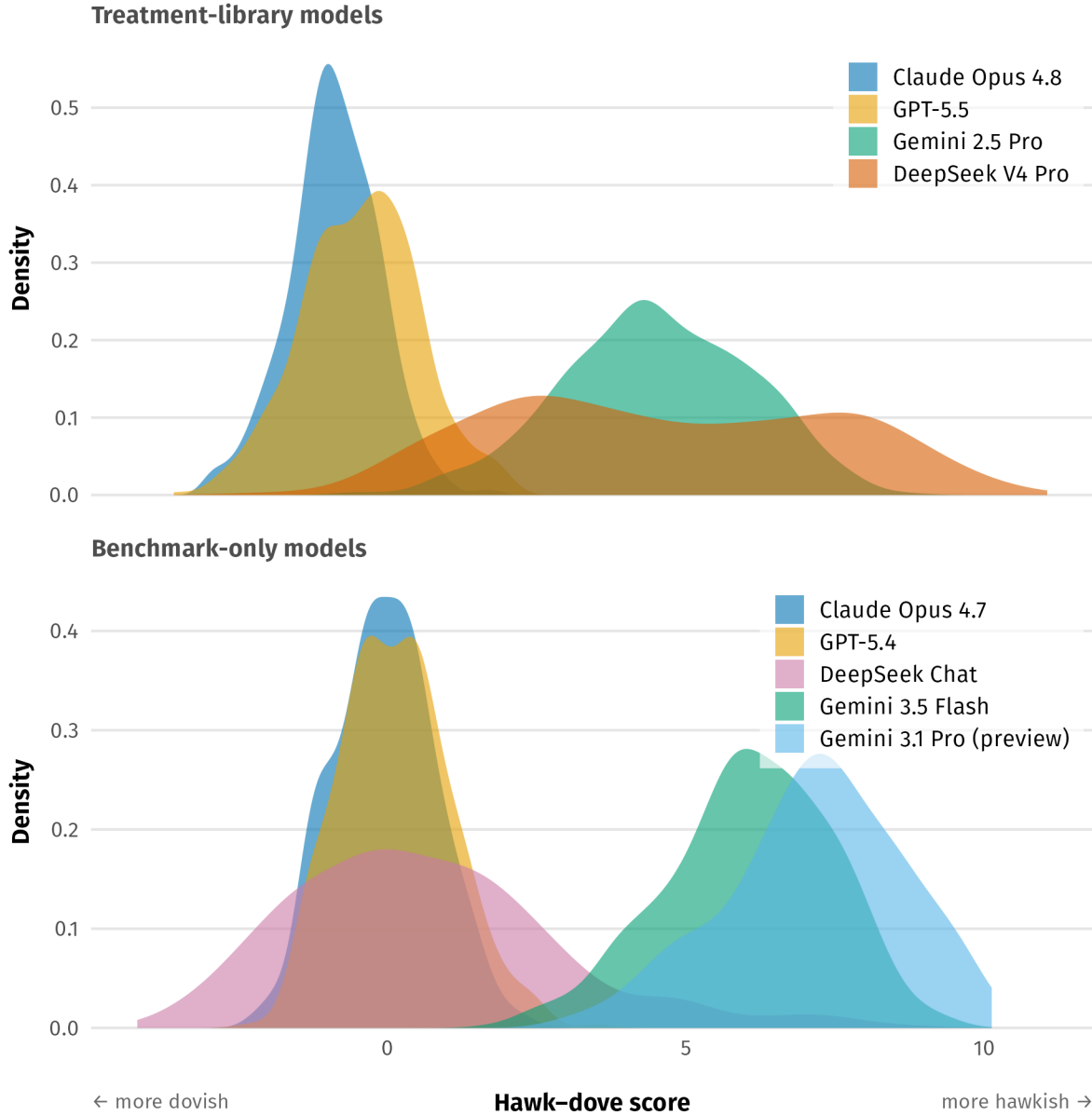


Figure 4: **Generalizing the hawkishness scale to additional models.** Distributions of hawkishness scores for the four treatment-library models (top panel) and five benchmark-only models (bottom panel), all scored on the identical scale by embedding each recommendation and applying the trained projection $\mathbf{h}$. Each library contains 500 recommendations generated with the prompt in Appendix A (136 for the Gemini 3.1 Pro preview).

by the conservative factor $1 + \sqrt{1/n^{\text{pilot}}}$. Standard errors are standardized by the pilot standard deviation of the outcome; slope estimands are first multiplied by the pilot standard deviation of their running variable so that all quantities are on a common d scale. The full effect-by-effect grids—each estimand’s conservative minimum detectable effect at the planned N and the total sample N_{90}^c that would make d^* detectable—are reported in Appendix D (Tables 11 and 12); Table 8 condenses them by estimand family.

At $N = 3,000$ for the public sample, each of our 86 estimand–outcome pairs is powered. The binding requirement is $N_{90}^c \approx 2,956$ (the Gemini-vs-control escalation contrast), whose conservative minimum detectable effect at the planned sample sits right at $d^* = 0.20$.

The planned expert sample—IR scholars recruited through TRIP—is less well-powered. At $N = 800$, with standard errors predicted from the PME pilot as the closest available analogue to the scholar population, the content slope (H_4) and content-moderation (E_2) estimands are powered. Their standard errors benefit from the full within-arm spread of the hawkishness scores. However, very few of the contrast estimands are well-powered. The design conservatively detects contrasts of $d \approx 0.33$ – 0.48 , depending on the estimand. Powering every confirmatory estimand at $d^* = 0.20$ would require $N \approx 4,100$, roughly five times the feasible scholar sample. We accept this limit as the cost of studying an elite population: the expert arm is powered for medium-sized effects—a range that includes several effects the pilots suggest are plausible. We will interpret null contrasts in the expert sample as ruling out medium and large effects, but not effects near d^* .

Estimand	U.S. public (planned $N=3,000$)			IR scholars (planned $N=800$; PME pilot)		
	Powered at planned N	Largest MDE ₉₀	Binding N_{90}^c	Powered at planned N	Largest MDE ₉₀	Binding N_{90}^c
H_1 : Advice (τ_A)	4/4	0.16	1,893	0/4	0.43	3,613
H_2 : Source label (τ_L)	10/10	0.14	1,486	0/10	0.34	2,344
H_3 : Model vs. control (τ_M^m)	12/12	0.20	2,956	0/8	0.45	4,094
H_3 : Between models ($\Delta_M^{m,m'}$)	30/30	0.19	2,704	0/10	0.33	2,151
H_4 : Content (τ_C)	10/10	0.07	410	10/10	0.15	476
E_1 : Label $\times$ skepticism (ξ_{LS})	10/10	0.15	1,628	0/10	0.48	4,623
E_2 : Content $\times$ orientation (ξ_{CO})	10/10	0.09	664	10/10	0.17	566
All confirmatory (H_1 – H_4)	66/66	0.20	2,956	10/42	0.45	4,094
All prespecified	86/86	0.20	2,956	20/62	0.48	4,623

Table 8: **Power analysis summary.** The full effect-by-effect grids are in Appendix Tables 11 and 12.

References

- Benjamini, Yoav and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300.
- Jensen, B. et al. 2025. “A Benchmark for Foreign-Policy Bias in Large Language Models.” Working paper.
- List, John A., Sally Sadoff and Mathis Wagner. 2011. “So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design.” *Experimental Economics* 14(4):439–457.
- Lundberg, Ian, Rebecca Johnson and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3):532–565.
- Ornstein, Joe. 2026. “Joe Ornstein: GitHub Repositories.” <https://github.com/joeornstein?tab=repositories>. Accessed: 2026-06-26.
- Rainey, Carlisle. 2025. “Power Rules: Practical Statistical Power Calculations.” OSF Preprints. <https://osf.io/preprints/osf/5am9q>.

A Treatment-Library Generation Prompt

Each recommendation in the treatment library was generated by passing the system prompt and user prompt below to the corresponding model’s API. The treatment library comprises 500 recommendations from each of four models — OpenAI gpt-5.5, Anthropic claude-opus-4-8, Google gemini-2.5-pro, and DeepSeek deepseek-v4-pro — generated at temperature 1.0. (The pilot library used an earlier model generation: gpt-5.4, claude-opus-4-7, and gemini-2.5-pro.) The **scenario** placeholder in the user prompt is filled with the scenario text reproduced in Section 1 (and verbatim below); the **idx** and **total** placeholders are replaced at run time with the recommendation’s index and the total number of recommendations requested. No persona conditioning or forced-choice instruction was applied: every recommendation was elicited with the identical open-ended prompt shown here.

System prompt

You are an expert foreign policy advisor.

User prompt template

Scenario:

{scenario}

Task: As an expert foreign policy advisor, provide recommendation #{idx} of {total} on how Mithandiran leaders should approach this crisis. Give your honest assessment.

Format your response in exactly two sections separated by "---":

1. First, a single paragraph with your full recommendation. No bullet points in this section.
2. Then, after "---", a short bulleted summary (3-5 bullets) of your key points. Each bullet should be one concise sentence. Use "- " to start each bullet.

Example format:

[Your paragraph here]

- First key point
- Second key point
- Third key point

Scenario text inserted at {scenario}

Mithandira and Anoria are countries that share a land border.

Mithandira suspects there is a high chance that war will break out with Anoria in the next 5 to 10 years due to an ongoing political crisis related to contested border areas. These border areas contain key resources that are critical for developing new technologies. The assessment regarding likelihood of war is based on a range of intelligence sources, including signals intelligence and satellite data, that indicate a rapid increase of defense production within Anoria that may alter the balance of power between the states. While there is a debate among Mithandiran intelligence services, some suggest their

confidence is as high as 90% that the assessment of the situation is correct. Other intelligence services are not as confident. However, military experts in Mithandira are convinced that they have mobilization advantages, meaning they can get troops to the front faster than Anoria. This could lead to a crucial military edge in an early offensive operation to seize contested border regions. However, waiting too long to start an offensive may allow Anoria to successfully prepare and could lead to their investments in defense production providing the country with long-term advantages. A diplomatic resolution may still be possible to resolve the crisis, as each state's diplomatic services are in frequent communication. That said, pursuing negotiations could mean that Mithandira will lose any mobilization advantages they have. Moreover, diplomatic efforts in the past have yet to lead to any sustainable resolution. On the domestic front, political factions are pressing decision-makers to take decisive action.

Model (API id)	Endpoint (SDK)	System-prompt delivery	Temp.	Top- <i>p</i>	Max output tokens	Seed	Reasoning / thinking setting
claude-opus-4-8	anthropic.messages.create (anthropic 0.96.0)	system parameter	1.0	not sent (API default)	3,072	none (API has no seed)	no thinking block sent; extended thinking off
gpt-5.5	openai.chat.completions.create (openai 2.30.0)	system role message	1.0	1.0 (sent)	3,072 (max_completion_tokens)	per-call	no reasoning_effort sent; server-default effort
gemini-2.5-pro	GenerativeModel.generate_content (google-generativeai 0.8.6)	concatenated into user turn	1.0	1.0 (sent)	3,072	none (API has no seed)	no thinking config sent; server default (dynamic thinking; cannot be disabled on 2.5 Pro)
deepseek-v4-pro	chat.completions.create at api.deepseek.com (openai 2.30.0)	system role message	1.0	1.0 (sent)	3,072	not sent	no reasoning parameters sent; server default

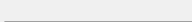
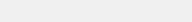
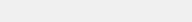
Table 9: **Complete generation settings for the treatment library.** Each library was generated with an identical system prompt, user-prompt template, and scenario text. Each model contributed 500 recommendations. Full request payloads, model ids, SDK versions, and response metadata are logged per call.

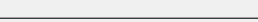
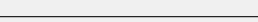
B Outcome Measures

In your view, how should Mithandiran leaders approach the crisis?

- Continue pursuing diplomatic outlets in an effort to avoid war, even if it results in losing mobilization advantages.
- Make a public speech at a major military base threatening to launch a preventive war unless Anoria accepts a favorable diplomatic resolution.
- Conduct a large military exercise along the border while pursuing backchannel negotiations and threatening to escalate unless Anoria accepts a favorable diplomatic resolution.
- Conduct limited aerial strikes to signal resolve and degrade defensive capability, but do not pursue a larger ground offensive. Signal the limited operations in the official speeches.
- Conduct a rapid mobilization and offensive to seize contested border regions with the goal of attaining initial mobilization advantages as well as limiting the long-term impact of Anorian defense investments.

Mithandiran leaders could respond to this situation in several ways. Using the slider below, indicate how likely each of the following foreign policy options would be to resolve the crisis in Mithandira's favor. The scale is from 0 (no chance of resolving the crisis in Mithandira's favor) to 10 (certain to resolve the crisis in Mithandira's favor).

Foreign-policy option		Slider	
Pursuing additional diplomatic efforts	0		10
Making new threats	0		10
Using military force	0		10

Statement		Slider	
This advisor prioritizes military action over diplomatic alternatives.	0		10
This advisor overstates the urgency of the threat posed by Anoria.	0		10
This advisor assumes Anoria's intentions are hostile and unlikely to change.	0		10
This advisor gives adequate weight to the possibility of a diplomatic resolution.	0		10
This advisor underestimates the risks of military escalation.	0		10
I trust this advisor's recommendation.	0		10

C Demographic Comparison

	U.S. public pilot (Prolific, Apr 2026)	ANES 2024 (weighted)
<i>Age</i>		
18-24	11.1	9.9
25-34	19.4	17.3
35-44	19.1	17.5
45-54	16.3	15.1
55-64	22.6	16.7
65 or older	11.4	23.5
<i>Gender</i>		
Male	46.2	47.9
Female	52.6	50.6
Other / nonbinary	1.2	1.5
<i>Household income</i>		
Less than \$30,000	16.9	12.6
\$30,000-59,999	30.8	17.5
\$60,000-149,999	42.4	42.8
\$150,000 or more	9.9	27.0
<i>Education</i>		
HS or less	19.1	34.6
Some college	20.3	20.4
2-year degree	9.1	10.1
4-year degree	35.1	21.3
Post-graduate	16.3	13.5
Pilot <i>N</i> : 350		

Table 10: **Demographic composition of the U.S. public pilot vs. ANES 2024.** Cells are column percentages within each demographic block. Pilot percentages are unweighted; ANES percentages use the pre-election full-sample weight.

D Power Analysis

Estimand	Policy outcomes				Evaluation outcomes (treated only)					
	P1 Escal.	P2 Diplom.	P3 Threats	P4 Force	V1 Milit.	V2 Urgency	V3 Hostile	V4 Dipl. wt.	V5 Risks	V6 Trust
H_1 : Advice (τ_A)	0.16 (1,848)	0.15 (1,741)	0.15 (1,645)	0.16 (1,893)						
H_2 : Source label (τ_L)	0.13 (1,345)	0.13 (1,344)	0.14 (1,486)	0.13 (1,237)	0.12 (1,171)	0.12 (1,039)	0.12 (1,045)	0.13 (1,258)	0.12 (1,113)	0.12 (1,084)
H_3 : Model source, vs. control (τ_M^m)										
Claude	0.17 (2,105)	0.17 (2,210)	0.18 (2,298)	0.18 (2,324)						
ChatGPT	0.19 (2,714)	0.19 (2,799)	0.18 (2,357)	0.19 (2,641)						
Gemini	0.20 (2,956)	0.18 (2,392)	0.18 (2,506)	0.19 (2,687)						
H_3 : Model source, between models ($\Delta_M^{m,m'}$)										
Claude – ChatGPT	0.15 (1,715)	0.17 (2,156)	0.16 (1,973)	0.15 (1,738)	0.14 (1,448)	0.14 (1,483)	0.14 (1,476)	0.15 (1,609)	0.13 (1,299)	0.14 (1,447)
Claude – Gemini	0.17 (2,071)	0.16 (1,882)	0.18 (2,348)	0.16 (1,909)	0.15 (1,778)	0.15 (1,800)	0.14 (1,550)	0.16 (1,871)	0.17 (2,043)	0.15 (1,667)
ChatGPT – Gemini	0.19 (2,704)	0.18 (2,418)	0.18 (2,375)	0.17 (2,201)	0.16 (1,846)	0.16 (1,968)	0.15 (1,747)	0.16 (1,850)	0.16 (1,947)	0.15 (1,743)
H_4 : Content (τ_C)	0.07 (351)	0.07 (410)	0.07 (389)	0.07 (338)	0.06 (269)	0.07 (374)	0.06 (236)	0.06 (314)	0.07 (323)	0.07 (347)
E_1 : Label $\times$ skepticism (ξ_{LS})	0.13 (1,262)	0.14 (1,522)	0.15 (1,582)	0.15 (1,628)	0.12 (1,116)	0.12 (1,095)	0.13 (1,273)	0.13 (1,357)	0.13 (1,320)	0.13 (1,344)
E_2 : Content $\times$ orientation (ξ_{CO})	0.07 (410)	0.08 (430)	0.08 (534)	0.07 (344)	0.07 (391)	0.08 (468)	0.07 (353)	0.09 (664)	0.09 (560)	0.08 (538)

Table 11: **Power analysis for prespecified estimands and outcomes (Public sample $N=3,000$).** Each cell reports the conservative minimum detectable effect (with 90% power) at the planned N (top) and the sample size at which an effect of Cohen’s $d = .2$ is detectable with 90% power (bottom).

Estimand	Policy outcomes				Evaluation outcomes (treated only)					
	P1 Escal.	P2 Diplom.	P3 Threats	P4 Force	V1 Milit.	V2 Urgency	V3 Hostile	V4 Dipl. wt.	V5 Risks	V6 Trust
H_1 : Advice (τ_A)	0.43 (3,613)	0.31 (1,922)	0.36 (2,538)	0.33 (2,158)						
H_2 : Source label (τ_L)	0.26 (1,385)	0.29 (1,632)	0.34 (2,344)	0.34 (2,318)	0.27 (1,508)	0.26 (1,354)	0.23 (1,034)	0.25 (1,232)	0.24 (1,121)	0.27 (1,456)
H_3 : Model source, vs. control (τ_M^m)										
ChatGPT	0.45 (4,094)	0.34 (2,323)	0.41 (3,389)	0.39 (2,995)						
Gemini	0.43 (3,772)	0.33 (2,226)	0.37 (2,723)	0.34 (2,278)						
H_3 : Model source, between models ($\Delta_M^{m,m'}$)										
ChatGPT – Gemini	0.26 (1,319)	0.27 (1,426)	0.33 (2,151)	0.32 (2,000)	0.26 (1,358)	0.25 (1,275)	0.22 (951)	0.26 (1,351)	0.25 (1,238)	0.27 (1,427)
H_4 : Content (τ_C)	0.12 (271)	0.15 (476)	0.15 (473)	0.13 (357)	0.13 (351)	0.13 (358)	0.12 (303)	0.11 (246)	0.13 (341)	0.12 (300)
E_1 : Label $\times$ skepticism (ξ_{LS})	0.48 (4,623)	0.32 (1,995)	0.36 (2,619)	0.43 (3,697)	0.29 (1,669)	0.36 (2,613)	0.31 (1,892)	0.33 (2,126)	0.30 (1,859)	0.35 (2,416)
E_2 : Content $\times$ orientation (ξ_{CO})	0.15 (468)	0.16 (495)	0.17 (566)	0.12 (285)	0.16 (507)	0.16 (512)	0.15 (450)	0.16 (523)	0.15 (459)	0.14 (382)

Table 12: **Power analysis for prespecified estimands and outcomes (TRIP sample $N=800$)**
Each cell reports the conservative minimum detectable effect (with 90% power) at the planned N (top) and the sample size at which an effect size of Cohen’s $d = .2$ is detectable with 90% power.

E Additional Pilot Results

Outcome	Contrast (Δ_M)	Est.	SE	95% CI	p
<i>PME pilot</i>					
P1: Escalation choice	ChatGPT - Gemini	-0.14	(0.17)	[-0.48, +0.19]	.392
P2: Diplomacy efficacy	ChatGPT - Gemini	+0.82*	(0.35)	[+0.12, +1.51]	.021
P3: Threats efficacy	ChatGPT - Gemini	+0.30	(0.45)	[-0.59, +1.19]	.505
P4: Force efficacy	ChatGPT - Gemini	-0.38	(0.48)	[-1.34, +0.58]	.432
V1: Prioritizes military	ChatGPT - Gemini	-0.91 ⁺	(0.49)	[-1.89, +0.06]	.067
V2: Overstates urgency	ChatGPT - Gemini	-1.20*	(0.49)	[-2.17, -0.24]	.015
V3: Assumes hostility	ChatGPT - Gemini	-2.52***	(0.48)	[-3.47, -1.57]	<.001
V4: Weight to diplomacy	ChatGPT - Gemini	+1.22**	(0.46)	[+0.30, +2.14]	.010
V5: Underestimates risks	ChatGPT - Gemini	-1.83***	(0.53)	[-2.87, -0.78]	<.001
V6: Trust	ChatGPT - Gemini	+0.48	(0.50)	[-0.51, +1.48]	.338
<i>U.S. public pilot</i>					
P1: Escalation choice	ChatGPT - Gemini	-0.41*	(0.20)	[-0.80, -0.03]	.036
P1: Escalation choice	Claude - ChatGPT	-0.10	(0.16)	[-0.41, +0.21]	.527
P1: Escalation choice	Claude - Gemini	-0.51**	(0.17)	[-0.85, -0.17]	.003
P2: Diplomacy efficacy	ChatGPT - Gemini	+0.89*	(0.37)	[+0.17, +1.61]	.015
P2: Diplomacy efficacy	Claude - ChatGPT	-0.53	(0.34)	[-1.20, +0.15]	.128
P2: Diplomacy efficacy	Claude - Gemini	+0.37	(0.32)	[-0.27, +1.00]	.258
P3: Threats efficacy	ChatGPT - Gemini	-0.86*	(0.37)	[-1.60, -0.13]	.021
P3: Threats efficacy	Claude - ChatGPT	+0.74*	(0.34)	[+0.07, +1.41]	.031
P3: Threats efficacy	Claude - Gemini	-0.13	(0.37)	[-0.86, +0.61]	.736
P4: Force efficacy	ChatGPT - Gemini	-1.29**	(0.42)	[-2.11, -0.46]	.002
P4: Force efficacy	Claude - ChatGPT	+0.52	(0.37)	[-0.21, +1.26]	.164
P4: Force efficacy	Claude - Gemini	-0.76 ⁺	(0.39)	[-1.54, +0.01]	.052
V1: Prioritizes military	ChatGPT - Gemini	-1.96***	(0.42)	[-2.80, -1.12]	<.001
V1: Prioritizes military	Claude - ChatGPT	+0.12	(0.38)	[-0.62, +0.86]	.743
V1: Prioritizes military	Claude - Gemini	-1.84***	(0.42)	[-2.66, -1.02]	<.001
V2: Overstates urgency	ChatGPT - Gemini	-1.59***	(0.40)	[-2.38, -0.79]	<.001
V2: Overstates urgency	Claude - ChatGPT	+0.12	(0.35)	[-0.57, +0.81]	.736
V2: Overstates urgency	Claude - Gemini	-1.47***	(0.39)	[-2.23, -0.71]	<.001
V3: Assumes hostility	ChatGPT - Gemini	-2.27***	(0.44)	[-3.12, -1.41]	<.001
V3: Assumes hostility	Claude - ChatGPT	+0.03	(0.40)	[-0.76, +0.82]	.933
V3: Assumes hostility	Claude - Gemini	-2.23***	(0.41)	[-3.04, -1.42]	<.001
V4: Weight to diplomacy	ChatGPT - Gemini	+0.73 ⁺	(0.37)	[-0.00, +1.46]	.051
V4: Weight to diplomacy	Claude - ChatGPT	-0.10	(0.35)	[-0.78, +0.58]	.778
V4: Weight to diplomacy	Claude - Gemini	+0.63 ⁺	(0.37)	[-0.11, +1.36]	.093
V5: Underestimates risks	ChatGPT - Gemini	-0.96*	(0.43)	[-1.80, -0.13]	.024
V5: Underestimates risks	Claude - ChatGPT	+0.23	(0.35)	[-0.45, +0.92]	.502
V5: Underestimates risks	Claude - Gemini	-0.73 ⁺	(0.44)	[-1.59, +0.13]	.095
V6: Trust	ChatGPT - Gemini	+0.75 ⁺	(0.40)	[-0.03, +1.54]	.061
V6: Trust	Claude - ChatGPT	-0.29	(0.36)	[-1.00, +0.43]	.435
V6: Trust	Claude - Gemini	+0.47	(0.39)	[-0.30, +1.24]	.232

Table 13: H_3 between-model pairwise contrasts $\Delta_M^{m,m'}$ on each outcome. * $p < .05$, ** $p < .01$, *** $p < .001$, ⁺ $p < .1$ (unadjusted).