# FROM AUTOMATION TO ROLE-PLAY: A POPULATION-LEVEL EVALUATION OF JEV

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can a model designed for fast automation support the role-conditioned decisions needed in a simulated population? We investigate this question in Jev, a newly released System One model that returns structured decisions rather than generated text. Our evaluation proceeds from individual capabilities to population-level behavior, using controlled testbeds to make black-box responses interpretable. PersonaEval and synthetic policy tasks first test role identification and action selection against source-derived or program-computed answers. We then adapt FlockVote into a fixed-population experiment that crosses role representation with shared information. This design distinguishes a change in an initial answer from a change in how the same roles respond to new information. Jev achieves approximately 97% modal-action accuracy on matched policy tasks. Yet, in one historical voting window, converting direct role queries into member-addressed requests changes the aggregate information response by 11.1 percentage points on matched complete records. The corresponding vector interaction remains above a specified tolerance under conservative missing-response bounds. Explicit role instructions and native object inputs leave residual representation effects. Member-level analysis further shows how aligned changes survive aggregation, while cancellation can conceal individual sensitivity. These findings identify a gap between individual decision competence and population-level role-play. They establish concrete validation requirements for using Jev as a shared decision backend before introducing memory, communication, and feedback; the reported probabilities are model responses, not validated human behavioral propensities.

## 1 INTRODUCTION

Jev's release on September 15, 2026 brings a concrete new possibility to agent-based modeling: replacing repeated text generation with a shared model that directly returns decisions. TypeSafe introduces Jev as a System One model for automation, describing a training approach based on Reinforcement Learning for Calibrated Decisions (RLCD) and an interface built around structured judgments rather than arbitrary text (Almeida, 2026). Its native Choice interface returns a selected alternative and a probability vector over a supplied option set (TypeSafe AI, 2026). The reported efficiency makes Jev an attractive candidate for simulations that require many decisions from many roles. But adopting it involves a transfer that the interface itself does not validate: can a model built to automate actions also make decisions on behalf of heterogeneous agents? Establishing the conditions for this reuse is important at the model's introduction, before assumptions about its behavior become embedded in larger systems.

Automation and role-play can share an output format without sharing an objective. Choosing the next browser action asks what will advance a task; choosing for a simulated person asks what follows from that person's information, priorities, and constraints. A role-playing model must allow different roles to make different choices and to respond differently to the same evidence. These decisions underlie the individual and collective behavior represented by generative-agent systems (Park et al., 2023). Maintaining another perspective and reconsidering a choice under changed circumstances can involve the hypothetical reasoning associated with System 2 cognition (Evans & Stanovich, 2013). This motivates our central transfer question: whether a model optimized for rapid System 1 judgments can support role-conditioned decisions with such demands. We investigate the behavioral requirements of that use, rather than infer a cognitive mechanism from Jev's architecture or response speed.



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

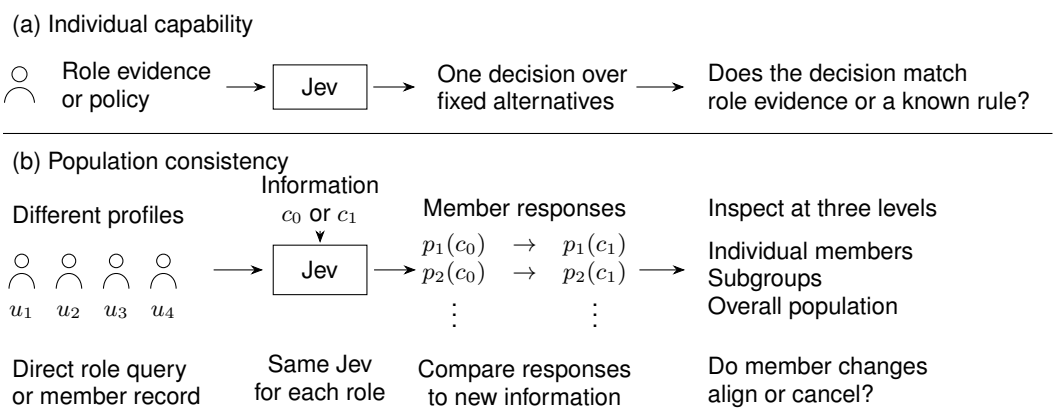



Figure 1: **Two requirements for a role-conditioned decision backend.** An individual test checks whether a decision follows from role evidence or a specified policy. Population use also requires testing how those decisions change with representation and shared information. We query fixed profiles under two information conditions and compare member, subgroup and overall responses before introducing memory or communication.

Individual benchmark performance is a necessary starting point, but it cannot settle this question. Even perfect answers to isolated questions would not establish that the same role retains its behavior when addressed through a member record, placed alongside other profiles, or evaluated under new information. For a black-box model, training exposure also cannot be inspected to determine how much a successful answer depends on familiarity with the tested material. We therefore treat individual scores as preliminary evidence and turn to targeted behavioral tests, following the principle that accuracy alone is insufficient to characterize a model's capabilities (Ribeiro et al., 2020). The relevant evidence is not only whether Jev can make a plausible decision once, but whether its responses remain attributable to the intended role under the conditions in which a simulation would use it.

Population evaluation makes this additional requirement observable. Reusing one backend across many profiles does not make its sensitivities independent. Averaging can attenuate idiosyncratic fluctuations while preserving changes that share a direction; conversely, opposing member changes can leave an aggregate nearly unchanged. A population summary can therefore appear stable without establishing stability for its members or subgroups. Existing work on synthetic respondents already shows that aggregate agreement can coexist with distorted variation and relationships (Bisbee et al., 2024). Our concern is the decision process that produces those aggregates: how the same profiles respond when their representation or surrounding context changes. We study this process before introducing memory and inter-agent communication, so that response formation can be examined separately from the propagation of generated actions.

Our evaluation rests on two complementary experimental foundations. PersonaEval (Zhou et al., 2025c) provides human-authored dialogue and source-derived speaker labels, allowing role identification to be assessed without asking another model to judge generated prose. We retain its released material and labels, and supplement it with synthetic policy tasks in which a program independently computes the correct action. These tasks distinguish recognizing a role from acting under its stated constraints, and distinguish selecting the correct endpoint action from changing that action appropriately when circumstances change. They form the individual-level preliminary study in Section 3. Their purpose is to establish which capabilities make population use plausible and which weaknesses warrant closer examination.

For the population study, we adapt FlockVote's profile-conditioned voting framework (Zhou et al., 2025a). Voting provides a useful middle ground between a narrowly specified rule task and a complete interacting simulation: heterogeneous roles encounter common information, while their decisions remain comparable within a small, interpretable action space. The task concentrates attention on how profile information and contextual evidence enter a choice, rather than on the quality of a generated explanation. We construct a fixed panel of 448 synthetic member occurrences across seven states

and evaluate two historical information windows. FlockVote supplies the experimental setting, but our target is different from election prediction: we ask whether the response attributed to the same population depends on how its members are represented.

The key comparison crosses representation with information. Each target profile is queried directly and through a member-addressed record, with its profile and choice alternatives retained, at both ends of an information window. Comparing the resulting changes separates a static representation offset from a representation-dependent response to new information. This contrast does not require treating an election result as the correct answer for an individual profile, or assuming that the historical material was absent from training. Its interpretability comes from explicit task construction and matched comparisons. Additional controls vary role-binding instructions, serialized versus native object inputs, and co-presented profiles. We examine responses at member, subgroup, and population levels, using repeated calls, matched complete records, and conservative bounds for missing responses. DeepSeek Flash provides a generative reference on corresponding tasks, rather than an architecture-controlled model comparison.

The results reveal a separation between individual competence and population response stability. Jev achieves approximately 97% modal-action accuracy on matched explicit-policy tasks, yet a transition to member-addressed requests reduces accuracy even before other profiles enter the request. In the later historical window, the aggregate Harris-coordinate response differs by 11.1 percentage points between direct and member-addressed queries on 269 complete members. The full-vector interaction remains above the specified tolerance under conservative missing-response bounds. Across the population analyses, we observe both aligned member changes that survive aggregation and opposing changes that conceal individual sensitivity. Clearer role instructions reduce some representation effects, but neither explicit binding nor native object inputs remove the residual dependence. These findings show that the organization of a simulation can affect the information responses it attributes to its roles.

This study provides an early, controlled assessment of Jev's transfer from automation to population-level role-play. Its contribution is to connect individual capability tests to identifiable population contrasts and show where successful action selection fails to guarantee stable role-conditioned behavior. The historical results measure model-response dependence, not fidelity to human voting behavior. For a Jev-based multi-agent system, the immediate requirement is therefore to specify which representational changes should preserve a role, what the returned probabilities are intended to mean, and whether agreement is needed for individual agents, subgroups, or the overall population. These are prerequisites for interpreting subsequent collective behavior: without them, patterns attributed to a simulated society may instead reflect the shared decision backend and the way the system presents its members.

## 2 Related Work

Role-playing research evaluates several capabilities that need not coincide. RoleLLM studies character knowledge and speaking style; CharacterEval assesses role-playing conversations; InCharacter examines personality fidelity through interviews (Wang et al., 2024a; Tu et al., 2024; Wang et al., 2024b). Our focus is the decision attributed to a role, rather than the fluency or persuasiveness of its utterance. Generative Agents and Concordia place such decisions within systems with memory and interaction (Park et al., 2023; Vezhnevets et al., 2023). The present study examines an earlier dependency: whether a shared backend produces stable role-conditioned responses before generated actions affect other agents.

Synthetic respondents offer a means of studying heterogeneous model behavior (Argyle et al., 2023; Aher et al., 2023). Their validity cannot be inferred from an aggregate match alone. Bisbee et al. (2024) find that synthetic survey means can resemble human means while variation, subgroup patterns and relationships differ. We address a complementary source of uncertainty: the representation of a fixed profile may change the response that enters the aggregate. Since our population experiment has no individual human response labels, we evaluate this internal dependence separately from behavioral fidelity.

Machine behaviour and targeted behavioral testing motivate evaluation beyond a single accuracy score (Rahwan et al., 2019; Ribeiro et al., 2020). Sensitivity to ostensibly incidental prompt features

is already documented for language models (Sclar et al., 2024). Building on these observations, we examine the interaction of representation with shared information in Jev's native decision interface, and measure how member-level changes combine into population responses.

## 3 Preliminary: Individual Role-Conditioned Decisions

We assess two distinct capabilities: inferring a role from behavioral evidence and selecting an action under a supplied role policy. The former has labels derived from human-authored material; the latter has program-computed answers. Neither requires a model to judge generated prose. Both backends receive corresponding task information and alternatives. Unless stated otherwise, accuracy uses the argmax of the returned probability vector, not Jev's separately returned choice.

### 3.1 Role identification with PersonaEval

We use PersonaEval (Zhou et al., 2025c) to test role recognition from human-authored dialogue. Each item contains a dialogue, a target response and candidate role profiles; the model identifies the target speaker. We retain the released material, labels and candidate order, adapting only the answer-format instruction to each backend. Literary and Drama each present four candidate profiles; Expertise presents five role categories. All scores follow the released target-speaker labels. Appendix C specifies the Expertise labeling and dataset revision.

Table 1: **Role identification on matched valid items.** Accuracy is reported in percent; Brier loss is the multiclass sum of squared errors against the role label (lower is better). Both models are scored on the same first-response items.

| | | Accuracy | | Brier loss | |
|---|---|---|---|---|---|
| Track | Matched / total | Jev | DeepSeek Flash | Jev | DeepSeek Flash |
| Literary | 23,352 / 26,208 | 65.5 | 75.4 | .4590 | .3571 |
| Drama | 1,594 / 1,658 | 58.5 | 70.8 | .5465 | .4307 |
| Expertise | 669 / 699 | 62.3 | 67.7 | .4852 | .4797 |

Table 1 shows informative but uneven recognition. Class imbalance matters: always choosing Expert attains 67.4% accuracy on the 669 matched Expertise items, above Jev's 62.3%. However, that predictor has 20.0% macro accuracy and .6517 Brier loss, compared with Jev's 51.1% and .4852. Jev therefore discriminates between classes beyond the majority label. Recognition remains weaker than the generative reference on these items, and source-grounded recognition does not by itself establish successful role enactment. We do not repeat PersonaEval's human evaluation or transfer its human scores to this collection.

### 3.2 Action selection under explicit policies

We next supply the role and ask for its action. Four task families combine feasibility constraints with ordered decision priorities. Events implement a change, propose a change without implementing it, or withdraw it; a program independently computes the resulting action. For example, a delivery role must reject infeasible carriers and choose the fastest remaining one. An unimplemented price proposal should not alter the effective alternatives. These are synthetic policies with known answers, not models fitted to human preferences.

Across 494 of 576 role–event units complete for both backends across all three calls, Jev's mean-vector argmax accuracy is 96.96%, compared with 100% for DeepSeek Flash. The corresponding total-variation errors are .07393 and .00135. Jev is less reliable on changes between conditions: its mean deviation from the prescribed event response is .15590 when the action should change and .03493 when it should remain fixed. Appendix C provides the matched-pair definitions. Thus, strong endpoint action accuracy coexists with imperfect conditional responses.

A second comparison retains the same policy and world but represents the actor as a member record. On 405 paired Jev member instances, accuracy falls from 96.54% to 87.90%, while mean total-variation error rises from .07548 to .13861. These instances reuse 58 parent-role units; they are

not 405 independent policies. The transition changes identifiers, request structure and instructions together, so the result concerns that interface transition, not an isolated identifier effect. It shows that individual capability can be lost before any other role is placed in the request.

The evidence supports proceeding to a population test, particularly for explicit-policy action selection. It also identifies a separate issue with probability interpretation. When an action depends on an unrevealed random signal, choosing its mode and returning its future frequency are different tasks. In the controlled example of Appendix F, Jev returns $(1, 0, 0)$ for a prescribed distribution $(.75, .25, 0)$ despite answering the conditional action questions correctly. We therefore analyze the following population vectors as model responses, without equating them to human behavioral probabilities.

# 4 Population Evaluation under Shared Information

## 4.1 From profile-based simulation to a matched information contrast

We adapt FlockVote's profile-conditioned framework for the 2024 U.S. presidential election (Zhou et al., 2025a). It conditions synthetic agents on demographic profiles and election information, elicits probabilities over voting alternatives, and aggregates the outputs. Its experiments examine predictions, individual explanations, profile and population ablations, and contextual sensitivity. We retain its profile-based decision setting, but change the evaluation target: does representing the same roles differently alter their response to new information?

A static comparison cannot resolve this question. A member-addressed request might shift all answers by a constant amount while leaving information responses unchanged. We therefore query both representations under two information conditions and compare their changes. The profile and choice space remain fixed within this contrast. Voting is useful here because heterogeneous roles face the same interpretable alternatives; member and population responses can be compared without a free-text evaluator. The observed election result is not a ground-truth distribution for any individual profile.

The study uses 448 synthetic member occurrences, 64 in each of Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania and Wisconsin. We retain their recorded identities from an earlier 896-occurrence panel, including repeated profiles. Profiles describe state, race, sex, age, occupation, industry, education and religion. These records form a fixed experimental panel, not a representative sample of people recruited for this study. Equal state weights define our overall readout; it is not a national vote-share estimator.

We evaluate two historical information windows. The earlier window moves from pre-debate campaign material to material covering the assassination attempt and Republican convention, with Biden/other/Trump alternatives. The later window moves from the Harris campaign reset to the economic-policy rollout and Democratic convention, with Harris/other/Trump alternatives. These are stages 3–5 and 6–7 of the supplied materials. "Other" includes another candidate or abstention. The intervening candidate replacement is excluded from the information contrast. Each condition is queried independently: previous generated answers, memory and messages from other agents are absent. These are interventions on supplied text, not causal estimates of real events, and the historical material may have appeared in pretraining.

## 4.2 Role representation and context controls

The primary contrast uses an explicit instruction to decide for the target role in both representations. The direct form presents its profile and asks for its decision. The member form places the profile in a one-member container and addresses the decision by member identifier. The alternatives and role evidence are retained, but the wrapper and question structure change. This is a compound representation contrast rather than a claim of perfectly equivalent token sequences.

We test two possible explanations for the resulting difference. A binding control compares inherited task wording with explicit target-role instructions in both representations. An input-type control replaces the serialized member state with a native object while retaining the member task. Comparing either member form with the direct form tests the whole representation; comparing the two member forms tests serialization more specifically.

A separate co-member experiment varies which profiles are displayed together. Within each block of sixteen members, two groups of eight exchange selected same-position pairs. Every target retains its profile and its position within a group, while its contextual partners change. The request asks for one decision per member. This operation tests contextual dependence without conflating it with a target-position change or the propagation of generated answers. Appendix A.1 specifies the request variants and the additional one-question, multiple-question and order controls.

Jev returns native Choice probabilities. DeepSeek Flash generates probability JSON for corresponding profiles, information and alternatives. We use it as an efficiency-oriented generative reference, while retaining the collected configuration with thinking enabled and high reasoning effort. The comparison is neither cost-matched nor an ablation of architecture: models, interfaces and hosted services differ. Basic settings are given in Appendix A.

### 4.3 Response estimands and incomplete observations

Let $\bar{p}^r_{b,i}(c)$ be backend $b$'s mean of three probability vectors for member $i$, representation $r$ and information $c$. For each window $(c_0, c_1)$, define

$$\Delta^r_{b,i} = \bar{p}^r_{b,i}(c_1) - \bar{p}^r_{b,i}(c_0), \qquad d_{b,i} = \Delta^{\text{member}}_{b,i} - \Delta^{\text{direct}}_{b,i}. \tag{1}$$

The interaction $d_{b,i}$ is zero when the representation effect is the same at both information endpoints. Overall responses average equally across states and equally across the included members within each state. Writing these weights as $w_i$, we compare the signed coordinate mean $\sum_i w_i d_{b,i}(a)$ with the mean absolute member change $\sum_i w_i |d_{b,i}(a)|$. Cancellation is measured by the latter minus the absolute value of the former. For a full vector contrast $v$, we also report $\frac{1}{2}\|v\|_1$; for an interaction, this magnitude can exceed one.

Complete-case comparisons retain the same members across every required condition and all three calls. A backend comparison additionally intersects the backend supports. We also report conservative missing-response bounds for the full planned panel, retaining each available response at its original weight and allowing unavailable vectors to vary over the simplex. These bounds concern the finite panel, not sampling uncertainty, and use different weights from the complete-case estimates. The specified half-$L_1$ tolerance is .02. Appendix B gives the construction and the constraints on grouped inference. Missing responses are neither imputed as zero nor scored as incorrect decisions.

## 5 Results

### 5.1 Representation changes the response to information

On the 269 Jev members complete across the later window, both representations and all three calls, the mean probability assigned to Harris decreases by 5.43 percentage points under the direct query and 16.50 under the member query. The difference in responses is approximately 11.1 points. Its full-vector half-$L_1$ magnitude is .11066, with a full-panel missing-response range of $[.02583, .19176]$. Even the conservative lower bound exceeds .02. A constant representation offset therefore cannot account for the observed information response on this panel.

Figure 2 examines the corresponding comparison on the 239 members complete for both backends. This smaller common support is used throughout the figure. Jev's direct and member responses are $-5.76$ and $-17.04$ percentage points; the reference model's are $-.62$ and $-.82$. The difference between Jev's two lines is not simply a different initial probability: their slopes diverge. Matching both information endpoints is what makes that distinction observable.

The earlier window prevents a general directional interpretation. On its 210-member cross-backend support, Jev's Biden-coordinate response is $-18.23$ points directly and $-11.71$ in member form: packaging attenuates rather than accentuates the decline. The corresponding interaction is $+6.52$ points, compared with $-.11$ for the reference. Candidate identities, materials and supported profiles differ between windows, so these signs do not identify a causal difference between events. Moreover, the earlier within-Jev full-panel range includes zero. The evidence establishes representation-dependent responses in the later window, not a candidate-independent direction of the representation effect.

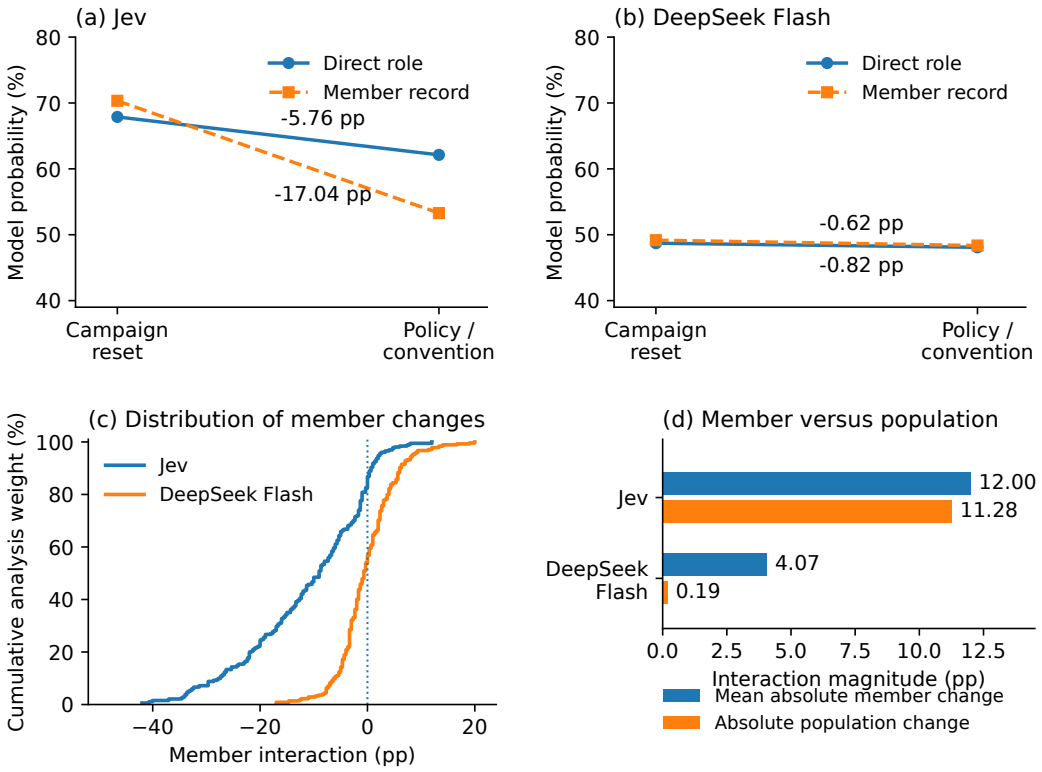



Figure 2: **Representation affects information responses, with different degrees of cancellation.** All panels use the same 239 complete members in the later window and equal state weights. (a,b) Mean model probability for the Harris alternative under direct and member queries; numbers mark each information response. (c) Weighted distribution of the Harris-coordinate interaction $d_{b,i}(\text{Harris})$ in Equation 1, in percentage points. (d) Absolute population interaction versus mean absolute member interaction on that same coordinate. These are descriptive complete-case estimates, not confidence intervals. The 269-member Jev-only estimate in the text has a separate missing-response bound.

### 5.2 Aligned member changes survive aggregation

The 239-member comparison distinguishes two explanations for an aggregate shift. Jev's mean signed interaction on the Harris coordinate is $-11.28$ points, close in magnitude to its mean absolute member interaction of 12.00 points. Negative interactions account for 82.53% of the analysis weight. Its seven state means are all negative on this support. By contrast, DeepSeek Flash has a signed mean of $-.19$ points but a mean absolute member interaction of 4.07 points; 53.89% of its weight is negative. The near-zero aggregate therefore conceals opposing changes rather than unchanged member responses.

This difference is consequential for population use. Averaging can suppress idiosyncratic changes, but it need not suppress a shared component. The familiar inequality $\|\sum_i w_i d_i\|_1 \le \sum_i w_i \|d_i\|_1$ permits cancellation; the measurements show how much occurs here. They do not establish how the effect scales to a larger population. All three coordinates are retained in the analysis: Jev's mean interaction is $(-11.28, +2.86, +8.42)$ points in Harris/other/Trump order. The decrease in one coordinate is therefore not transferred entirely to the other major candidate.

The backend contrast remains conditional on coverage. Its full-vector magnitude is .11086, but the full-panel range $[.00389, .21818]$ crosses .02. Thus the member structures differ descriptively on the matched records, while a full-panel backend separation at that tolerance remains unresolved. Individual-call estimates preserve Jev's negative interaction in the later window ($-11.75$ to $-10.97$ points), but three calls on the same profiles are not independent population replications.

### 5.3 Explicit binding and native inputs leave residual effects

Clearer role instructions might reduce ambiguity about whose decision is requested. To test this explanation, we match direct and member queries under both inherited and explicit binding. Jev's representation magnitude decreases from .1274 to .0663 at the first later-window endpoint and from .1636 to .1088 at the second (Table 2). The decline is a within-backend, four-condition comparison; it is not estimated by subtracting unrelated complete subsets.

Table 2: **Explicit role instructions reduce, but do not remove, Jev's representation effect.** Entries are half-$L_1$ distances between population means. The binding comparison matches all four conditions within each row. Native-object comparisons use their own direct/object matched support and include the member wrapper.

| Information | Matched members | Inherited binding | Explicit binding |
|---|---|---|---|
| Campaign reset | 277 | .1274 | .0663 |
| Policy / convention | 256 | .1636 | .1088 |
| | | Native object member versus direct query | |
| | Matched members | Point estimate | Full-panel range |
| Campaign reset | 360 | .0690 | [.0283, .0989] |
| Policy / convention | 328 | .1003 | [.0430, .1427] |

Residual dependence persists under explicit binding. For Jev's bound direct/member comparison, the full-panel ranges are $[.0276, .0976]$ and $[.0548, .1508]$, both above .02. Native object member inputs also remain separated from the direct query (Table 2). Serialization alone is therefore insufficient to explain the whole representation effect. This does not show that input type is irrelevant: the object-versus-serialized member comparison changes only that type, whereas comparison with the direct request includes the wrapper. The controls narrow two explanations without identifying a unique internal cause.

### 5.4 Aggregate agreement can conceal member and subgroup uncertainty

The co-member comparison changes contextual partners while preserving each target's profile and slot. Figure 3 separates two quantities often conflated in a population evaluation: the average distance between a member's responses, and the distance between population means. For Jev at the campaign-reset condition, these are .06709 and .01528 on 130 complete members. The apparent aggregate stability is conditional: its full-panel upper bound is .19345. Substantial missing coverage prevents concluding that Jev is insensitive to the operation.

The reference model provides a better-covered example. Its 432 complete members change by .04239 on average, while the aggregate changes by .00294. The full-panel upper bound, .01507, is below .02. Yet Wisconsin's upper bound is .08615, so overall agreement does not establish agreement for every state. At the next information condition, even the overall upper bound rises to .02976. These are limits of the agreement claim, not proof of a subgroup error or a paired backend ranking.

The relevant validation level consequently depends on the application. A simulation that tracks named characters needs member-level stability; a study reporting subgroup responses needs subgroup-level evidence. An overall average cannot establish either. This remains true before communication is introduced, because co-presented role descriptions already influence the inputs from which individual decisions are formed.

## 6 Discussion and Conclusion

Jev's individual performance makes it a plausible component for role-conditioned decision systems, especially when roles follow explicit policies. The population experiments reveal a separate requirement: the responses attributed to those roles must be evaluated under the representations in which they will actually be used. In the tested deployment, member-addressed requests alter the response to shared information, and similarly directed changes persist in the aggregate. Clearer role instructions

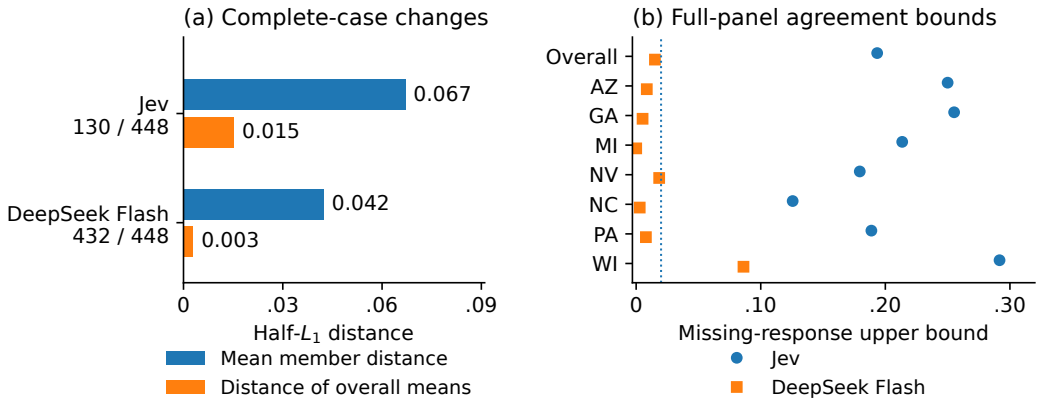



Figure 3: **A small overall difference is not a member- or subgroup-level guarantee.** Same-slot exchanges change the profiles surrounding each target. (a) Complete-case mean member distances and distances between overall means at the campaign reset. Jev uses 130 complete members; DeepSeek Flash uses 432. These are separate within-backend samples. (b) Full-panel missing-response upper bounds for overall and state readouts of the same operation; each row retains all planned members. The line at .02 marks the tolerance, not statistical significance. A bound above it leaves agreement unresolved.

and native inputs do not remove all of this dependence. Conversely, a small aggregate change may reflect cancellation rather than stable individuals.

Representation invariance is application dependent: another actor's profile may legitimately change a socially situated decision. The historical contrasts measure response dependence, not errors against human labels. Demographic descriptions underdetermine preferences, and a consistent model could still reproduce stereotypes. Our identification test supplies evidence about recognition, while explicit-policy tasks provide narrower correctness criteria. Neither establishes that the historical probability outputs are calibrated human propensities. A simulation must therefore specify which changes should preserve its roles' behavior before aggregate agreement can be interpreted as success. For known exogenous randomness, Appendix F shows how a program can combine conditional decisions using supplied signal weights. This restricted construction neither calibrates historical voting probabilities nor corrects the observed representation effects.

For deployment, these distinctions suggest a concrete validation sequence. Fix the profiles, alternatives and information conditions, then evaluate the actual request forms that the system will use. Compare both endpoint decisions and changes across information conditions: agreement at one endpoint does not establish agreement about how a role responds to new information. Inspect signed and absolute member changes together, and report subgroup behavior alongside the overall mean when subgroup conclusions matter. Retain planned-panel weights when bounding missing responses, rather than treating a small complete-case difference as evidence of agreement. Replacing the decision backend or changing the request wrapper should trigger the same checks; matching the output schema alone does not establish that the resulting simulation measures the same behavior.

The findings are limited to one hosted Jev deployment, the collected representations and a fixed synthetic panel. Missing responses weaken several full-panel comparisons, and compound request changes prevent attributing the effects to a particular architectural mechanism. The experiment also stops before memory and inter-agent feedback. Future studies could retain fixed initial roles and matched request forms while introducing these components, testing whether feedback amplifies, attenuates or redirects representation sensitivity. Such comparisons would separate the formation of role responses from the propagation of generated decisions, rather than attributing all observed changes to interaction. Throughout these extensions, the distinction established here remains essential: an accurate individual choice, a stable member response and an acceptable population aggregate are different empirical claims. Fast structured decisions can support role-play, but their behavioral composition must be tested rather than inferred from their output type.

## AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, assisted with research design, implementation, analysis, literature review, visualization and manuscript revision. Jev and DeepSeek Flash were the experimental backends. Reported measurements were computed from recorded responses; AI-assisted writing did not replace missing observations. The authors are responsible for the experiments, analysis and final manuscript.

## ETHICS STATEMENT

This study introduces no new human-participant experiment. PersonaEval supplies published human-authored material and source-derived labels; the policy tasks and demographic profiles are synthetic. Historical voting scenarios are used to evaluate model-response dependence, not to estimate individual intentions or forecast elections. Demographic conditioning may reproduce stereotypes. Neither internal consistency nor aggregate agreement justifies treating a synthetic population as a substitute for a real community in consequential decisions.

## REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the tasks and estimands. The appendices specify input representations, probability validation, missing-response bounds and additional results. Appendix C identifies the fixed PersonaEval dataset revision. Hosted model aliases are mutable, so repeated API calls are not guaranteed to return identical responses.

## REFERENCES

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

Diogo Almeida. Introducing System One models & Jev. TypeSafe AI Blog, September 2026. URL https://typesafe.ai/blog/introducing-system-one-models-and-jev.

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

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

Jonathan St. B. T. Evans and Keith E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013. doi: 10.1177/1745691612460685.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://arxiv.org/abs/2304.03442v2.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486,

2019. doi: 10.1038/s41586-019-1138-y. URL https://www.nature.com/articles/s41586-019-1138-y.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324v2.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.638. URL https://aclanthology.org/2024.acl-long.638/.

TypeSafe AI. Choice. Official documentation, 2026. URL https://docs.typesafe.ai/primitives/choice.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia, 2023. URL https://arxiv.org/abs/2312.03664v2.

Noah Wang, Z. Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

Lingfeng Zhou, Yi Xu, Zhenyu Wang, and Dequan Wang. FlockVote: LLM-empowered agent-based modeling for simulating U.S. presidential elections. In *ICAIS*, 2025a. URL https://arxiv.org/abs/2512.05982v1.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. PersonaEval dataset. Hugging Face dataset, 2025b. URL https://huggingface.co/datasets/lingfengzhou/PersonaEval.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. PersonaEval: Are LLM evaluators human enough to judge role-play? In *Conference on Language Modeling*, 2025c. URL https://arxiv.org/abs/2508.10014v1.

# A Models and Input Construction

The experiments use the hosted `typesafe-ai/jev` model through the native TypeSafe interface and the official `deepseek-flash` chat-completions model. Jev receives a state, a question and named Choice alternatives. DeepSeek Flash receives corresponding content and an instruction to return a JSON probability vector; thinking is enabled with high reasoning effort and a maximum output length of 16,384 tokens. These settings describe the collected comparison, not a claim of equal computational effort. The model aliases do not identify immutable weight snapshots.

For PersonaEval, the common material consists of the released dialogue, target-speaker instruction and candidate profiles. Jev receives this material as its state and the candidate names as Choice alternatives. DeepSeek Flash receives the same material with a probability-JSON adapter in a user message. Policy tasks use explicit role and world descriptions; the generative adapter separates system and user instructions. Reference answers are held out of both requests. Probability coordinates are aligned by alternative identity before scoring.

The population panel has 448 fixed occurrences, selected by recorded identity from an earlier 896-occurrence panel. Repeated demographic profiles remain separate occurrences, but do not supply independent evidence about new kinds of people. No output-based filtering is used to select the panel. The two information windows use the same options internally; the earlier window has Biden, other and Trump, and the later window Harris, other and Trump. We refer to their endpoints as stages 3/5 and 6/7 to identify the supplied materials, not to imply that generated decisions propagate along a trajectory.

## A.1 Request contrasts

Table 3 summarizes the representation changes. A direct request and its member-addressed counterpart contain the same target profile and choice meanings but differ in their request structure. The explicit-binding variants instruct the backend to make the target role's decision rather than answer from an unspecified perspective. The object control changes the serialized member state to a native object, retaining the member question. It exists only for Jev.

Table 3: **Input changes in the population experiment.** Profiles and option meanings are fixed within each comparison. Multiple profiles share input context, not generated answers.

| Contrast | First condition | Second condition |
|---|---|---|
| Representation | One profile, directly addressed | One-member record, addressed by member identifier |
| Role binding | Inherited task instruction | Explicit instruction to decide for the target profile |
| Input type | Serialized member state | Native object member state (Jev only) |
| Co-presentation | Only the target profile | Eight profiles, one target question |
| Question count | Eight profiles, one target question | Same profiles, one question per member |
| Order | Original profile order | Reversed order of the same profiles |
| Co-members | Original groups of eight | Same-slot exchanges between paired groups |

For the co-member operation, label two original groups in a sixteen-member block as $(A_1, \ldots, A_8)$ and $(B_1, \ldots, B_8)$. A fixed mask selects positions at which $A_j$ and $B_j$ are exchanged. Each member stays in position $j$ of its new group. Masks differ across blocks. This preserves target identity and slot while changing contextual partners. Group responses are analyzed for the members actually present after exchange; a group index alone does not imply unchanged membership.

The known-rule actor-to-member comparison is a separate collection. It preserves the parent world's event, alternatives and role policy, but changes the actor identifier to a member identifier, introduces a one-member list, changes instruction emphasis and omits the actor-form prediction-target field. DeepSeek Flash's output keys also change. These differences make it a composite interface transition.

We do not interpret it as a randomized single-field intervention or merge it with the historical population contrast.

# B Scoring and Missing-Response Bounds

A valid response assigns a finite probability in $[0, 1]$ to every requested alternative, with total probability within $10^{-5}$ of one. We retain accepted values without renormalization. Invalid or unavailable responses remain missing. For grouped requests, individually valid member responses are retained even if other members are unavailable. Correctness and availability are therefore separate measurements.

For a one-hot reference $e_y$, multiclass Brier loss is $\sum_a (p(a) - e_y(a))^2$. In a prescribed stochastic task, squared distribution error is instead $\sum_a (p(a) - q(a))^2$, where $q$ is the known action law. Total variation is $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$. A change between two response vectors is not an accuracy measure unless a corresponding reference change is available. Policy-response deviation compares $(p_1 - p_0)$ with $(q_1 - q_0)$ using the same half-$L_1$ norm.

## B.1 Finite-panel bounds

Consider a vector contrast $v = \sum_j c_j p_j$ with its original planned weights included in $c_j$. Combine repeated terms first. Let $k$ be the contribution of available responses, $M$ the sum of absolute coefficient weights for unavailable responses, and $\epsilon$ the accepted probability-sum residual budget. Unavailable vectors can take any value in the probability simplex. With $C = \frac{1}{2}\sum_j |c_j|$, a conservative interval for $\frac{1}{2}\|v\|_1$ is

$$\left[\max\left\{0, \frac{1}{2}\|k\|_1 - \frac{M+\epsilon}{2}\right\}, \quad \min\left\{C + \frac{\epsilon}{2}, \frac{1}{2}\|k\|_1 + \frac{M+\epsilon}{2}\right\}\right]. \tag{2}$$

A difference between two distributions has $C = 1$; a difference of such effects can have $C = 2$. This triangle bound need not be the sharp identified set. For error relative to a fixed reference $q$, the known contribution is adjusted by $q$ and the probability-simplex cap is $1 - \min_a q(a)$, with the numerical residual budget retained.

Complete-case estimates and these bounds have different targets. The former reweight complete members within each state; the latter retain all planned members and every available repetition at the planned weight. A complete-case point need not lie in the interval for the full-panel quantity. We consequently do not draw the missing-response range as a sampling error bar around that point. A lower bound above .02 establishes a finite-panel separation at the chosen tolerance; an upper bound below .02 establishes agreement at that tolerance. A range crossing .02 leaves the comparison unresolved.

For rule-based groups, we first construct the distribution or reference error at the specified level—member, eight-member cohort or sixteen-member world—and then average over planned units. We do not average unrelated action worlds before measuring their error. Error-change bounds are $[L_1 - U_0, U_1 - L_0]$. Reused parent-role predictions in the actor-to-member comparison receive the appropriate occurrence weights but are not treated as independently observed copies.

## B.2 Dependence and uncertainty

The main figures describe a fixed panel. Their member distributions and binding summaries are descriptive analyses of collected responses. They do not provide confidence intervals for a population of human respondents. Repeated calls evaluate the same inputs, and shared profiles or contexts can induce dependence even without communication.

Where grouped bootstrap intervals are available in the underlying analysis, policy tasks retain scenario clusters and historical tasks retain state-stratified sixteen-member blocks. Only four such blocks occur per state. Missing responses can make a state's complete support empty in a bootstrap resample, preventing the intended grouped interval. We do not replace these unavailable intervals with an independent-member bootstrap. The missing-response bounds answer a different question and remain reportable without a sampling model.

# C Additional Individual-Level Results

## C.1 Source-grounded role identification

We use the fixed PersonaEval release (Zhou et al., 2025b) at Hugging Face revision `f65c77a992414af9de49debdc6f607b4a2e472d8`. Literary, Drama and Expertise contain 26,208, 1,658 and 699 items. The released Expertise instruction names Character2 as the target responder. We preserve that instruction and its label, despite the broader role–audience framing of the task. All stated scores refer to this executed classification. We neither relabel the release nor infer human-audience understanding from the result.

On the 669 common Expertise items, a constant Expert prediction has accuracy .67414, macro accuracy .200 and Brier loss .6517. Jev and DeepSeek Flash have macro accuracies .511 and .656. The models were not queried in a strictly contemporaneous randomized comparison; Table 1 is a matched valid-item comparison, not a missing-free full-test ranking. Some valid vectors assign zero mass to the true class, so log loss can be infinite; the reported Brier loss remains finite.

A separate repeated-query sample contains 256 planned items per track. Table 4 retains only items complete for all five calls from each model and evaluates the mean of the first $K$ vectors. This holds the scored questions fixed as $K$ changes. Repeated responses do not increase the number of independent questions.

Table 4: **Brier loss after averaging repeated role-identification responses.** Each track uses the same items across models and all values of $K$. These supports differ from Table 1.

| Track | Model | Complete / 256 | $K = 1$ | $K = 3$ | $K = 5$ |
|---|---|---|---|---|---|
| Literary | Jev | 150 | .4047 | .4064 | .4071 |
| | DeepSeek Flash | 150 | .3246 | .3006 | .2925 |
| Drama | Jev | 220 | .5178 | .5152 | .5169 |
| | DeepSeek Flash | 220 | .3967 | .3486 | .3435 |
| Expertise | Jev | 206 | .4570 | .4583 | .4576 |
| | DeepSeek Flash | 206 | .4403 | .4069 | .3925 |

## C.2 Rule execution and conditional response

The four narrative families instantiate feasibility filtering followed by ordered priorities. One retained delivery example has alternatives A/B/C with fees 140/150/170 and travel times 140/170/150. A accepts documents only, while the object is a fragile crate. The role rejects infeasible services, minimizes travel time and breaks ties by fee. C is therefore correct; its proposed fee reduction is not implemented. Jev returns $(0, .01, .99)$ and DeepSeek Flash returns $(0, 0, 1)$. The program's answer is not supplied to either model.

Table 5 distinguishes endpoint performance from conditional changes. The response rows use pairs matched on representation and both endpoints, pooled across the collected representations. They are not additional independent scenarios. Retaining endpoint error is important because a constant incorrect response could otherwise appear invariant to irrelevant events.

Table 5: **Prescribed actions and their changes.** Accuracy uses the argmax of the three-response mean vector. All other entries are mean half-$L_1$ errors, lower being better.

| Measurement | Matched units/pairs | Jev | DeepSeek Flash |
|---|---|---|---|
| Endpoint accuracy | 494 | .96964 | 1.00000 |
| Endpoint total variation | 494 | .07393 | .00135 |
| Event response: action should change | 222 | .15590 | .00000 |
| Event response: action should stay fixed | 215 | .03493 | .00155 |
| Role switch: action should change | 433 | .11893 | .00154 |

### C.3 Actor-to-member transition

Thirty-two policy worlds supply two parent roles each and eight member instances per role, for 512 planned members. Only the proposed-but-unimplemented event is used in this transition. The original policy is unchanged. Table 6 uses each model's paired three-response support; the rows have different members and are not a matched backend contrast.

Table 6: **Oracle error under the actor-to-member transition.** The final column bounds the mean error increment over all 512 planned members.

| Model | Paired / 512 | Actor TV | Member TV | Full-panel error change |
|---|---|---|---|---|
| Jev | 405 | .07548 | .13861 | [.03330, .10817] |
| DeepSeek Flash | 496 | .00538 | .00605 | [−.00977, .00065] |

The 405 Jev member instances use 58 parent-role units in 31 worlds; the 496 reference instances use 62 parent roles in 31 worlds. Jev's accuracy changes from 96.54% to 87.90% on the paired instances. At the sixteen-member-world level, only six Jev worlds have complete paired responses, while the all-32-world mean error increment is bounded by $[.01301, .08788]$. This bound uses available contributions beyond the six complete worlds and does not treat repeated parent roles as new observations.

## D Population Responses and Control Results

### D.1 Information windows and response coordinates

Table 7 reports every option coordinate on the common backend support for each window. The earlier and later windows have 210 and 239 complete members, respectively; within a window both backends use exactly the same occurrences. Coordinate changes sum to zero up to accepted numerical residuals. Different windows are not a matched comparison between events or candidates.

Table 7: **Representation-by-information interactions.** Values are signed percentage points, not total-variation distances. The first coordinate is Biden in the earlier window and Harris in the later one.

| Window | Model | Matched / 448 | Biden/Harris | Other | Trump |
|---|---|---|---|---|---|
| Earlier | Jev | 210 | +6.519 | +2.309 | −8.828 |
| | DeepSeek Flash | 210 | −.112 | +.370 | −.259 |
| Later | Jev | 239 | −11.278 | +2.856 | +8.422 |
| | DeepSeek Flash | 239 | −.192 | +.010 | +.182 |

Table 8 separates within-backend interactions from the difference between them. Each point uses its indicated complete support, while each range retains all 448 planned members. A full-vector norm bound is not an interval for a signed Harris or Biden coordinate. In particular, the later within-Jev separation is stronger than the all-panel comparison between backends.

Table 8: **Full-vector interactions and missing-response bounds.** Half-$L_1$ magnitudes and bounds use the same contrast but different complete-case and full-panel weighting.

| Window | Contrast | Complete / 448 | Half-$L_1$ | Full-panel range |
|---|---|---|---|---|
| Earlier | Jev | 255 | .08720 | [.00000, .17115] |
| | DeepSeek Flash | 384 | .00191 | [.00000, .04881] |
| | Between backends | 210 | .08569 | [.00000, .21334] |
| Later | Jev | 269 | .11066 | [.02583, .19176] |
| | DeepSeek Flash | 400 | .00222 | [.00000, .02770] |
| | Between backends | 239 | .11086 | [.00389, .21818] |

### D.2 Repeated responses

Table 9 evaluates each call on the same complete support used in Table 7. Jev's interaction retains its sign within each window. The signed aggregate can be much smaller than the mean absolute member interaction, especially for the generative reference. The repeated-query comparison does not subtract a noise floor from the representation effect.

Table 9: **Repeated-call sensitivity on fixed common members.** The three call columns are signed interactions for coordinate $a$ (Biden/Harris), in percentage points. The final column is the weighted mean absolute interaction on that coordinate, using each member's three-response mean.

| Window | Model | Call 1 | Call 2 | Call 3 | Mean $\lvert d_i(a)\rvert$ |
|---|---|---|---|---|---|
| Earlier | Jev | $+6.829$ | $+6.210$ | $+6.518$ | 8.936 |
| | DeepSeek Flash | $+.117$ | $+.245$ | $-.697$ | 3.993 |
| Later | Jev | $-11.109$ | $-10.971$ | $-11.753$ | 12.003 |
| | DeepSeek Flash | $-.309$ | $-.491$ | $+.224$ | 4.066 |

Across the twelve same-input repetition pairs in the later window, mean member absolute differences on the Harris coordinate range from 2.206 to 2.766 points for Jev and 4.214 to 5.037 for DeepSeek Flash. The corresponding signed aggregate differences range from $-.578$ to .226 and $-.422$ to $.575$. These descriptive ranges concern repeated answers at a fixed endpoint, whereas Equation 1 combines changes across representations and information.

### D.3 Binding and native input

The full four-condition binding comparison for DeepSeek Flash has 417 complete members at the campaign reset and 423 at the policy/convention condition. Its inherited/explicit packaging magnitudes are .01086/.01228 and .01067/.01360. Unlike Jev's descriptive decrease in Table 2, those smaller magnitudes do not shrink under explicit binding. Supports differ across backends, so their changes are not a paired comparison of improvements.

Table 10 reports the explicitly bound representation and co-member endpoint effects for both models. The direct/native-object Jev comparison additionally has earlier-window magnitudes .10169 and .08838, with full-panel ranges $[.05068, .14145]$ and $[.03709, .13084]$ on 344 and 336 complete members. All four object-versus-direct comparisons include the wrapper. None alone identifies a pure input-type effect.

Table 10: **Endpoint effects under explicit binding and co-member exchange.** Each row uses its own three-response complete support. Reset and policy denote the later window's two information conditions.

| Operation | Model | Information | $n$ / 448 | Half-$L_1$ | Full-panel range |
|---|---|---|---|---|---|
| Representation | Jev | Reset | 362 | .06645 | [.02763,.09757] |
| | DeepSeek Flash | Reset | 421 | .01202 | [.00000,.02620] |
| | Jev | Policy | 336 | .11329 | [.05484,.15083] |
| | DeepSeek Flash | Policy | 426 | .01359 | [.00554,.02190] |
| Co-members | Jev | Reset | 130 | .01528 | [.00000,.19345] |
| | DeepSeek Flash | Reset | 432 | .00294 | [.00000,.01507] |
| | Jev | Policy | 145 | .00963 | [.00000,.21346] |
| | DeepSeek Flash | Policy | 416 | .00261 | [.00000,.02976] |

At the later endpoint, the paired between-backend difference in the representation effect has 321 complete members, magnitude .10965 and full-panel range $[.04042, .15278]$. This is an endpoint comparison. It does not resolve the less precise representation-by-information backend interaction in Table 8.

## E Additional Organization and Composition Tests

The rule-based population tasks extend the actor-to-member transition to eight co-presented profiles, multiple questions, reversed order and co-member exchange. The two original groups have role mixtures 6:2 and 2:6; same-slot exchanges preserve the sixteen-member total at 8:8 while changing group composition. A group's reference distribution follows the actual roles in that group. Fixed-original-member averages are a separate analysis, not the same group after exchange.

Table 11: **Paired complete coverage for Jev's rule-based organization tests.** The columns refer to different units and cannot substitute for one another. Each condition requires all three responses.

| Transition | Members / 512 | Fixed cohorts / 64 | Worlds / 32 |
|---|---|---|---|
| Actor to single-member request | 405 | 22 | 6 |
| Single profile to eight profiles, one question | 362 | 5 | 1 |
| One question to eight questions | 234 | 6 | 0 |
| Original to reversed order | 104 | 13 | 1 |
| Original to exchanged co-members | 156 | 10 | 3 |

For Jev, the one-question/eight-question comparison has no complete sixteen-member world. Finer-level complete pairs therefore cannot establish a world-level accuracy comparison. Full-panel bounds constrain some distribution changes while permitting either direction of the oracle-error change. We retain this unresolved comparison rather than using a different task to fill its missing observations.

A separate composition analysis forms nested subsets of the historical panel using identities alone, without consulting outputs or validity. Two hundred orderings select 1, 4, 16 or 64 members per state. These are alternative compositions of a single fixed panel. For the reference model's campaign-reset co-member contrast, subsets of 112 members have median aggregate magnitude .00471 and median mean-member magnitude .04243; 152 of 200 have full-panel upper bounds below .02. The worst upper bound is .03598. At 448 members every ordering is the same full panel. These results show composition dependence, not a universal minimum reliable population size. The corresponding Jev representation comparison uses a different perturbation and retains an aligned shift; it is not an experiment that changes only the backend.

## F Choice Confidence and Future-Action Distributions

A probability over possible correct choices is not automatically the probability of a future random action. We distinguish three targets in controlled finite worlds: the action distribution before a random signal is revealed, the identity of the most likely action, and the action after the signal is specified. The signal law and role-to-action mapping are inputs; the reference vector is computed independently. On 77 role–world units complete across both models, all three targets and three responses, Jev's mean total-variation errors are .35000, .00000 and .00000, respectively. DeepSeek Flash's errors are .00000, .03788 and .00000. Identifying a mode and recovering the future distribution are empirically distinct here.

A conditional-composition experiment tests whether an external program can supply the known randomization while the model handles each deterministic branch. Let $z$ be the signal, with supplied probabilities $P(z)$. The program combines conditional decisions as

$$q_b^{\mathrm{cond}}(a) = \sum_z P(z)\, p_b(a \mid z). \tag{3}$$

This is compared with one direct future-distribution query and with the mean of $K$ direct queries, where $K$ is the number of signals. The repeated-direct comparison matches calls, not tokens, information or difficulty. All fully conditioned reference actions are deterministic.

In one recorded example, three amber tickets prescribe action A and one cyan ticket prescribes B. Both direct Jev calls return $(1, 0, 0)$, as does the mode query. The conditional amber and cyan answers are $(1, 0, 0)$ and $(0, 1, 0)$. Their weighted combination is $(.75, .25, 0)$, which is also the direct reference-model response.

The base panel contains 48 worlds, half with two signals and half with three, and two roles per world. A second panel tests positional permutations of ticket tuples $(6, 3, 1)$, $(5, 3, 2)$, $(8, 7, 5)$ and $(34, 33, 33)$: 24 instances but 21 unique count vectors. Three repetitions yield 288 and 144 planned role–world–repeat units. Table 12 matches both models and all three future-distribution methods. Jev's direct errors are often concentrated on the mode, but seven of the 134 matched second-panel units have a wrong mode; overconcentration alone is not a complete account of error.

Table 12: **Future-distribution error on common complete units.** All entries are mean total-variation error. Each model and method uses the same 274 base-panel and 134 second-panel units.

| Panel | Model | Direct once | Direct $K$ times | Conditional mixture |
|---|---|---|---|---|
| Base | Jev | .35354 | .35344 | .00000 |
| | DeepSeek Flash | .00000 | .00000 | .00000 |
| Second | Jev | .37731 | .37532 | .00000 |
| | DeepSeek Flash | .00000 | .00000 | .00000 |

All 1,138 valid Jev conditional vectors match their deterministic references, but missing branches prevent claiming exactness over the full plan. Missing conditional responses remain unknown at their original signal weights rather than being discarded; the conditional-mixture error upper bounds are .01215 and .01410. The full-panel improvement ranges over repeated direct querying are $[.33541, .35914]$ and $[.35060, .37627]$. This construction handles known exogenous randomness outside the model. It does not establish general probabilistic reasoning, calibration to human behavior, or a correction for the historical representation effects.