EVALUATING LLM-GENERATED PREFERENCE DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly used as probabilistic generators for simulation, synthetic data generation, and decision support in settings where real-world data are unavailable. Yet, the structure and reliability of the distributions they produce remain understudied. Here, we systematically analyze LLM-generated distributions of preferences for air travel, restaurants, and consumer products. Encouragingly, repeated calls to the same model yield comparatively self-coherent distributions, with the most probable outcomes often stabilizing within the recorded calls. At the same time, we observe substantial discordance across differently sized models within each of three families, with little consensus even among their most probable outcomes. We examine these patterns across nine open-weight models and three choice domains, with additional checks under temperature changes, greedy decoding, and perturbations of prompt and ordering. Our findings indicate that changing the model can shift outcomes as much as rewording the prompt, challenging the common assumption that sufficiently capable LLMs produce similar preference distributions when used as stand-ins for survey respondents.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as probabilistic generators for synthetic data creation (Argyle et al., 2023), agent-based simulation (Park et al., 2023), opinion modeling (Santurkar et al., 2023), and survey simulation (Tjuatja et al., 2024), where repeated calls supply preference distributions to simulators. Prior work distinguishes prompted judgments from direct probability measurements (Hu & Levy, 2023), evaluates verbalized confidence (Xiong et al., 2024; Zhou et al., 2024), and examines fidelity to human populations (Santurkar et al., 2023; Tjuatja et al., 2024). Yet how repeated calls, model choice, and processing shape these distributions remains understudied.

Complementing accuracy benchmarks (Lin et al., 2022; Hendrycks et al., 2021) and concerns about confounded evaluation (Anwar et al., 2024), we measure the consistency of emitted probability vectors, not their accuracy for human choices. We ask four questions: how combining calls changes the most probable outcomes, how model differences compare with repeated-call variation, how changing the model compares with rewording the prompt, and how matching outputs to the choice list shapes the distribution. We study nine open-weight models from three families on air-travel destinations, adding restaurant and product choices for repeated-call and processing checks. We compare models on groups of five calls and score top-five shortlists on held-out calls.

Encouragingly, repeated calls to the same model yield comparatively self-coherent distributions: their variation stays below between-model differences throughout the resampled range in eight of nine model pairs, and their most probable outcomes often stabilize within the recorded calls. At the same time, differently sized models within a family disagree, even on their most probable outcomes: between the 70B and 405B Llama models, reusing the other model's top-five shortlist loses 17.7 percentage points of probability coverage on average. Changing the model can shift outcomes as much as rewording the prompt, although the comparison depends on which batch serves as the baseline. These findings challenge the common assumption that sufficiently capable LLMs produce similar preference distributions when used as stand-ins for survey respondents.

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.

2 Related Work

2.1 LLMs for Behavioral Simulation

LLMs have been explored as simulators of human behavior through memory, reflection, and planning (Park et al., 2023). Argyle et al. (2023) show that demographic conditioning reproduces selected GPT-3 survey-response patterns, whereas Santurkar et al. (2023) report opinion-distribution mismatches despite demographic steering. Tjuatja et al. (2024) similarly find that responses to survey edits can differ from human response biases. Simulated opinions often homogenize within demographic groups (Qi et al., 2025), and constrained-choice alignment with time-use data varies across models (Qi et al., 2026). Population fidelity therefore remains an empirical question. We examine a complementary requirement: whether elicited distributions and their most probable outcomes stay consistent across repeated calls, models, and prompts.

2.2 Probability Elicitation and Uncertainty

Calibration research has long examined whether model confidence aligns with accuracy (Guo et al., 2017). For LLMs, Jiang et al. (2021) investigate calibration for question answering, and Kadavath et al. (2022) examine models' assessments of answer correctness and knowledge. Later work quantifies semantic uncertainty across meaning-equivalent generations (Kuhn et al., 2023) and evaluates verbalized confidence (Xiong et al., 2024; Zhou et al., 2024). Stated beliefs can also erode under multi-turn persuasion (Huang et al., 2026). Hu & Levy (2023) distinguish prompted linguistic judgments from direct probability measurements, and Lovering et al. (2025) examine whether output probabilities reflect numerical information in context. We complement these studies by measuring whether full probability vectors over a choice list reproduce across repeated calls, reporting unmatched probability and how outputs are matched and averaged.

2.3 Prompt Sensitivity and Model Choice

Studies have documented sensitivity to prompt design, including few-shot example order (Lu et al., 2022), answer biases (Zhao et al., 2021), and prompt selection (Perez et al., 2021). Prompt semantics alone need not explain performance (Webson & Pavlick, 2022), and short-form bias benchmarks may not predict behavior in longer, application-specific outputs (Lum et al., 2025). These findings motivate evaluating concrete downstream uses. We ask whether changing the model shifts the distribution as much as rewording the prompt, measuring both against one baseline batch and scoring top-five shortlists on the same held-out calls.

3 Methods

3.1 Matching Outputs to the Choice List

We convert each call into one probability vector. For setting c (a model and prompt), call r returns records (s_{rj}, w_{rj}) pairing outcome names with nonnegative numbers. A documented mapping g matches names to a fixed choice list $\mathcal{Y}$, logging unmatched records and summing duplicates of one outcome. If the matched probability Z_r is positive, call r defines

$$p_c^{(r)}(y) = \frac{\sum_{j:g(s_{rj})=y} w_{rj}}{Z_r}, \qquad Z_r = \sum_{j:g(s_{rj})\in\mathcal{Y}} w_{rj}. \tag{1}$$

We report each call's matched share before normalization; $p_c^{(r)}$ is conditional on successful parsing and matching, and unmatched probability is not read as zero preference. The pooled estimator weights valid calls equally,

$$\bar{p}_{c,S}(y) = |S|^{-1} \sum_{r\in S} p_c^{(r)}(y). \tag{2}$$

Averaging unnormalized records would instead favor calls with more matched probability, and the two can differ substantially.

Travel outputs must resolve to one of the 300 supplied airport codes, following the choice set of Kim et al. (2025); a city code such as NYC stays unmatched. Yelp responses give business names, not branches, and the 300 supplied rows contain only 252 distinct names. Amazon uses 300 distinct titles, and its original prompt displays only the first 80 characters of each. We accept unambiguous matches to a full or displayed title after HTML, Unicode, case, and punctuation normalization, without inferring aliases from string similarity. Appendices D, E, and I.4 give the audit rules.

3.2 Elicitation Protocol

Each travel prompt specifies a traveler persona, departure airport, and destination list and requests probabilities in JSON; repeated calls reuse the same setting. We thus measure the distribution a model states under a given prompt and decoding procedure, not token-level probabilities. A pilot asking for a single destination yields much narrower outputs, but entropy alone cannot show which format better measures an underlying preference. Appendix C.2 reproduces all prompt templates.

3.3 Stability Within a Batch

We measure when the most probable outcomes settle as calls accumulate. For a batch of N valid calls in collection order, we compute the running average $\bar{p}_{1:R}$ for $R = 1, \ldots, N$ and report R_{last}, the last call at which its top five changes (1 if it never changes); a value of N has no later call confirming persistence. A distance index gives the smallest R at which the running average lies within Jensen–Shannon (JS) distance 0.05 of the full-batch average; at $R = N$ the distance is zero by construction. Permuting calls tests sensitivity to call order, distinct from reordering options inside a prompt.

Named outcomes, those with positive pooled probability, accumulate for a counting reason. Let c_y count the calls giving outcome y positive probability. For a uniformly selected subset S of R distinct calls,

$$\mathbb{E}_S[|\operatorname{supp}(\bar{p}_S)|] = \sum_{y \in \mathcal{Y}} \left[1 - \frac{\binom{N-c_y}{R}}{\binom{N}{R}} \right], \tag{3}$$

where $\binom{a}{R} = 0$ for $R > a$. The ratio is the probability that no selected call names y. Pooling alone thus adds named outcomes whenever calls name different outcomes, even for calls drawn from one fixed distribution; at $R = N$ every subset is the full batch. We therefore compare growth with a stationary random baseline, which draws sparse probability vectors from one fixed distribution fitted to the first ten calls.

3.4 Mode and Tail Contributions

We split JS divergence between the most probable outcomes and the remaining tail. Let $M_k(p)$ contain the k largest coordinates, with ties broken alphabetically. For a pair (p, q), the mode region is $M = M_k(p) \cup M_k(q)$ and the tail is $T = \mathcal{Y} \setminus M$. If $m = (p + q)/2$, each coordinate contributes

$$c_y(p, q) = \tfrac{1}{2} p_y \log(p_y/m_y) + \tfrac{1}{2} q_y \log(q_y/m_y). \tag{4}$$

Zero terms are evaluated by continuity. Summing c_y over M and T splits total JS divergence exactly, without renormalizing either region. Conditional divergences, renormalized within each region, are not additive and do not measure the tail's share. We use $k = 5$ as a fixed convention and report sensitivity to this cutoff.

3.5 Evaluation Metrics

JS divergence uses natural logarithms without smoothing and lies in $[0, \ln 2]$ (Lin, 1991); JS distance is its square root. Agreement among the most probable outcomes uses top-k Jaccard, $|M_k(p) \cap M_k(q)|/|M_k(p) \cup M_k(q)|$ (Jaccard, 1912), and exact top-k agreement marks identical sets. Breadth counts named outcomes (named support in tables) and the pooled vector's coordinates above 0.001 (active support). Concentration uses entropy in bits (Cover & Thomas, 2006). Appendix B gives complete definitions.

Table 1: Llama-3.3-70B travel batches. R_{last} is the last call at which the pooled top five changes. Permuted index gives the median and 95% resampled range over 1,000 random call orders; it is not a stopping rule.

Context	N	Named support	R_{last}	Permuted index
IND	20	31	7	4 [1, 15]
FWA	20	41	1	1 [1, 8]
ATL	20	42	4	3 [1, 8]
ORD	20	31	18	8 [1, 18]

4 EXPERIMENTS

Our central travel study makes 20 repeated calls to Llama-3.3-70B from each of four departure airports (IND, FWA, ATL, and ORD), each with one fixed traveler persona, and 20 repeated IND calls to Llama-3.1-8B and Llama-3.1-405B (Sections 5.1–5.3). A persona grid crosses five age groups with four education groups, giving 20 IND personas with one call each for every Llama model; the grid is unweighted, and the American Community Survey provides demographic background rather than population weights. Prompt contrasts compare persona presence, age, education, and regional context on fixed persona and airport pairs from the 70B outputs (Appendix K). To weigh wording against model choice, we collect five batches of 15 IND calls at 70B: four with reworded questions and one with the original question as a second baseline batch (Section 5.4). Beyond Llama, we make 20 IND calls to each of three Qwen and three Mistral models (Qwen et al., 2024; Yang et al., 2025; Jiang et al., 2023; Mistral AI, 2025; Jiang et al., 2024), and Llama-3.3-70B answers one Yelp and one Amazon prompt 20 times each (Section 5.5). Robustness checks repeat 13 of these settings with greedy decoding across eight models and all three domains (Appendix M), and vary temperature and airport-list order at 70B (Appendices L and N). Table 8 lists all nine models.

Primary calls use temperature 0.2; other temperatures and greedy decoding are checks. Original generation settings differ in some comparisons: regenerating the smallest Llama model's persona responses changed the responses as well as their parsing, and Mistral-24B used an adapted prompt with retries. We therefore compare each served model under its recorded settings and report valid-response distributions with success counts where recorded, such as 20 valid Mistral-24B responses from 179 attempts. Table 12 isolates parser effects on the same saved text.

Resampled ranges and descriptive summaries. We draw 1,000 random permutations of each batch (seed 20260921), take the first R calls, and recompute the pooled statistic. Resampled ranges (bands in figures) span the 2.5–97.5 percentiles. Like the stability indices, they describe variation within the collected calls rather than confidence intervals for new calls. For model comparisons, each draw forms two non-overlapping five-call averages per model; the contrast, between-model JS minus the mean within-model JS, matches call count and processing. Fixed-persona and prompt tables report descriptive means and standard deviations (SD), without treating pairs sharing a persona as independent. Appendix O.3 details the protocol.

5 RESULTS

5.1 TOP DESTINATIONS OFTEN STABILIZE WITHIN THE RECORDED CALLS

Encouragingly, the top five destinations of Llama-3.3-70B stabilize early in three of the four travel batches (Table 1). In the recorded order, the pooled top five is fixed from call 7 for IND, 1 for FWA, 4 for ATL, and 18 for ORD. Reordering the calls shifts these numbers, most for ORD, so they describe these batches, not how many calls a new batch needs. Pooling more calls keeps adding named outcomes, as expected when calls name different destinations, while the 70B peak probability at IND holds steady (Figure 1).

Table 2: Additive split of natural-log JS between consecutive 70B calls on the original probabilities ($k = 5$): mode covers destinations in either call's top five, tail all others. Rows group pairs by earlier call (12, 8, 20, and 36 pairs); values average within each airport, then equally across the four. Tail share is averaged tail over averaged total; totals and shares are computed before rounding. Pairs share calls, so no confidence interval is given.

Earlier call	Total JS	Mode part	Tail part	Tail share
1–3	0.115	0.046	0.068	59.6%
4–5	0.187	0.076	0.111	59.5%
6–10	0.055	0.015	0.040	73.4%
11–19	0.074	0.024	0.050	67.7%

Table 3: Llama model pairs on 20 fixed IND personas, one call per persona and model. JS and Jaccard entries are means with descriptive SD, not confidence intervals. Exact@5 and Top-1 count personas with identical top-five sets or the same top destination. Probability ties are broken alphabetically.

Llama pair	JS (SD)	Jaccard@5 (SD)	Exact@5	Top-1
70B/405B	0.359 (0.099)	0.110 (0.118)	0/20	6/20
70B/8B	0.455 (0.113)	0.101 (0.098)	0/20	2/20
405B/8B	0.486 (0.101)	0.114 (0.101)	0/20	5/20

5.2 Repeated Calls Differ Mostly Outside Their Top Five

Averaged over airports, consecutive calls to one model diverge mostly outside their top five destinations. Table 2 splits the JS divergence of consecutive 70B calls between destinations in either call's top five and all others. The remaining destinations account for 59.5–73.4% of the total across call ranges and 67.7% for pairs starting at calls 11–19; this share varies considerably across airports and falls as the top-set size k grows (Appendix B.5). Within one model, disagreement thus concentrates outside the most probable destinations, while the top five still contributes a substantial share.

5.3 Different Models Disagree More Than Repeated Calls Do

In contrast to this self-coherence, different Llama models disagree even on their most probable destinations. Across the 20 IND personas, mean JS between Llama models ranges from 0.359 to 0.486 nats (Table 3). Their top-five sets overlap little (mean Jaccard 0.101–0.114), and no model pair gives an identical top five for any persona.

At equal numbers of calls, between-model differences also exceed within-model variation. For non-overlapping five-call groups of repeated IND calls, mean JS between Llama models is 0.185–0.394, against 0.035–0.144 within a model, and every Llama difference stays positive across its 95% resampled range (Figure 2). These comparisons describe models, not an effect of size, since the models also differ in training release and original generation settings. Across the three families, eight of nine model pairs have ranges above zero; the exception, Mistral-7B/24B, used an adapted 24B prompt (Appendix I).

Reusing a top-five shortlist from another model loses probability coverage (Figure 14). We select each shortlist from five calls and score the probability it covers in five held-out calls of the target model; this measures agreement between models, not human-choice accuracy. Between 70B and 405B, same-model shortlists cover 54.6% on average and reused shortlists 37.0%, a loss of 17.7 percentage points, computed from unrounded coverage and averaged over both directions (resampled range $[11.5, 24.0]$; Appendix I.3). Repeated calls to one model are thus comparatively self-coherent, whereas switching models changes even the most probable destinations.

5.4 Changing the Model Can Shift Outcomes as Much as Rewording

We compare two changes to a primary 70B baseline batch for one IND persona: switching the model to 405B, or rewording the question at 70B. Four reworded batches (A–D) have 15 calls each, and a

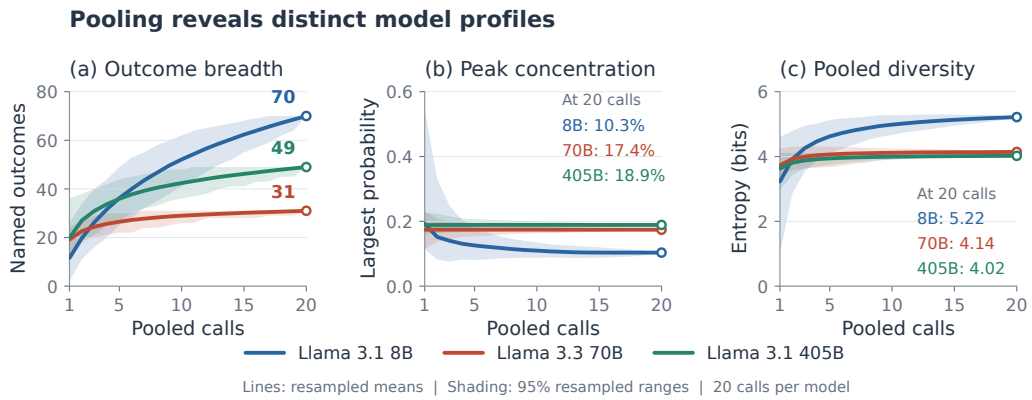


Figure 1: Pooling reveals distinct model profiles: 70B and 405B reach similar peak probabilities but name different numbers of outcomes. Lines are means over random subsets of IND calls; shading is the 95% resampled range. Named outcomes count destinations with positive probability. At 20 calls each subset is the full batch, so bands vanish by construction. Each model uses its original generation settings.

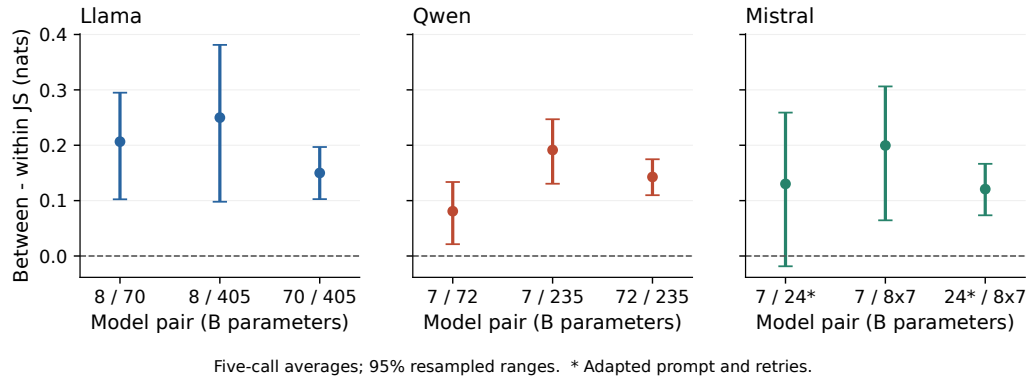


Figure 2: Between-model differences exceed within-model variation: 95% resampled ranges lie above zero in eight of nine model pairs. Points show mean between-model minus within-model JS over 1,000 draws of non-overlapping five-call groups; bars describe variation within the collected calls, not confidence intervals for new calls. Mistral-24B uses an adapted prompt and retries; its 7B/24B range crosses zero.

second baseline batch with the original wording has 15. Both models share the documented prompt text, airport order, temperature 0.2, and 1,024-token limit. In each draw, both changes share one five-call baseline average and the held-out calls that score shortlist losses. Table 18 reports persona and regional contrasts descriptively.

On average, changing the model moves the distribution more than three of the four rewordings do, and wording C, which asks for preferred flight destinations, is comparable at mean JS 0.192 (Table 4). Encouragingly, the second baseline batch stays closer to the primary one than the other model does, with a JS-difference range above zero. At the same time, the model-versus-wording ranking depends on the baseline batch: with the primary one, only D has JS and coverage ranges above zero. With the second, the model change exceeds every rewording on average, and C's mean JS difference moves from -0.008 to 0.126 nats, with ranges spanning zero under both (Appendix K.2). The model is therefore a setting to fix and report alongside the prompt wording.

Table 4: Model change versus rewording on five-call IND averages. Model-change JS averages 0.184 nats; Δ_{JS} subtracts wording JS from it. Δ_{cov} is the extra held-out coverage lost with the other model's top-five shortlist rather than the reworded one. Positive values mean a larger model change. Brackets are 95% resampled ranges, not confidence intervals. The last row uses a second original-wording baseline batch.

Wording	Wording JS	Δ_{JS} [range]	Δ_{cov}, pp [range]
A	0.110	0.074 [-0.077, 0.161]	4.2 [-20.1, 25.4]
B	0.109	0.075 [-0.056, 0.161]	2.8 [-20.0, 24.3]
C	0.192	-0.008 [-0.206, 0.117]	1.5 [-20.0, 22.5]
D	0.045	0.139 [0.087, 0.187]	17.5 [4.9, 31.7]
Same wording	0.065	0.119 [0.048, 0.188]	10.0 [-10.1, 25.6]

Table 5: Share of each call's probability matched to the choice list, before renormalization. Catalog keys are distinct listed items. Retained mass is the mean matched share per call; top-ten mass uses the equal-weight average of renormalized calls. Amazon mapped titles accept unique full or displayed-title matches; exact and mapped Yelp matching coincide.

Condition	Distinct catalog keys	Retained mass	Top-ten mass
Travel IND	300	98.8%	74.4%
Yelp names	252	69.8%	53.4%
Amazon exact titles	300	56.4%	79.5%
Amazon mapped titles	300	65.1%	75.6%

Table 6: Processing choices applied to the same saved outputs; every row counts each outcome once. Historical weighting is the original processing: last value per name, calls weighted by matched probability. Equal-call rows sum duplicates and weight renormalized calls equally. All-emitted-strings rows keep unmatched strings as separate outcomes; they are a sensitivity check, not a valid choice-list distribution.

Domain	Policy	Active	Top-10 %	Distance index
Yelp	Historical weighting	103	49.4	20
Yelp	Equal-call exact	77	53.4	19
Yelp	Equal-call mapped	77	53.4	19
Yelp	All emitted strings	81	59.2	19
Amazon	Historical weighting	51	79.0	17
Amazon	Equal-call exact	48	79.5	18
Amazon	Equal-call mapped	61	75.6	18
Amazon	All emitted strings	70	76.5	18

5.5 Choice-List Matching Shapes Restaurant and Product Results

Matching outputs to the choice list is a consequential processing step for restaurants and products, each studied with a single prompt. On average, 30.2% of a Yelp response's probability fails to match a listed name exactly (Table 5). Amazon responses leave 43.6% unmatched under exact titles; resolving 35 distinct strings through documented HTML and displayed-title rules reduces this to 34.9%. Unmatched strings include shortened names and generic catch-all categories, so these shares are not hallucination rates.

After summing duplicate names and weighting renormalized calls equally, Yelp has 77 active names and 53.4% of probability on its top ten; Amazon mapped titles give 61 and 75.6% (Table 11). Pooled distributions first come within JS distance 0.05 of the full batch at calls 19 (Yelp) and 18 (Amazon), near the endpoint, where distance is zero by construction. Figure 3 separates agreement on dominant choices, already high at ten calls, from coverage of low-probability outcomes, which keeps growing.

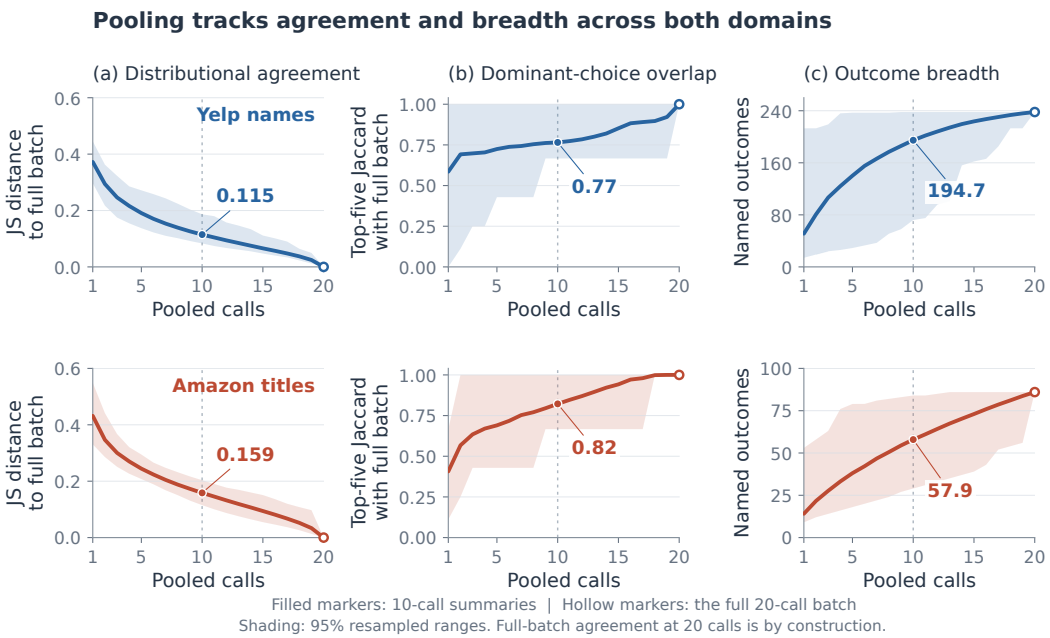


Figure 3: In both domains, pooled calls largely agree on dominant choices while named outcomes keep growing. At ten calls, mean top-five Jaccard with the full batch is 0.77 (Yelp) and 0.82 (Amazon). Lines and shading show means and 95% resampled ranges within the collected calls. Named outcomes count listed items with positive probability; full-batch endpoints are fixed by construction.

Processing choices alone change the distribution a simulator receives, without new model calls (Table 6). For Yelp, replacing the original processing with duplicate summation and equal call weights moves top-ten mass from 49.4% to 53.4% and active names from 103 to 77. For Amazon, documented title matching moves the equal-call top-ten mass from 79.5% to 75.6% and active products from 48 to 61. Specifying a simulation input therefore requires its matched probability and matching rule alongside concentration and stability summaries.

5.6 A Stationary Random Baseline Can Reproduce Stabilization and Named-Outcome Growth

To test whether a fixed generator produces stabilization and named-outcome growth, we fit the stationary random baseline to the first ten 70B calls per context and compare it with the last ten, in an exploratory split of already examined calls. The baseline draws probability vectors from one fixed Dirichlet distribution with 1% or 5% uniform mass, so unseen destinations can appear, and keeps as many of the largest outcomes as a randomly chosen fitted call names.

At 1% uniform mass, the baseline reproduces the growth in named outcomes and the last call at which the top five changes: observed values fall inside its 95% simulated ranges in all four contexts (Appendix P). These patterns thus arise even from a fixed sparse generator. In contrast, the baseline does not reproduce every statistic: held-out FWA calls, for example, have pairwise JS 0.046, below the simulated range of 0.081–0.115, so they agree more closely than the baseline predicts.

Temperature, greedy-decoding, airport-order, and rewording checks cover selected settings (Appendices L–N). These checks complement the model comparisons at equal numbers of calls, which provide the central evidence that repeated calls to one model are comparatively self-coherent while different models disagree.

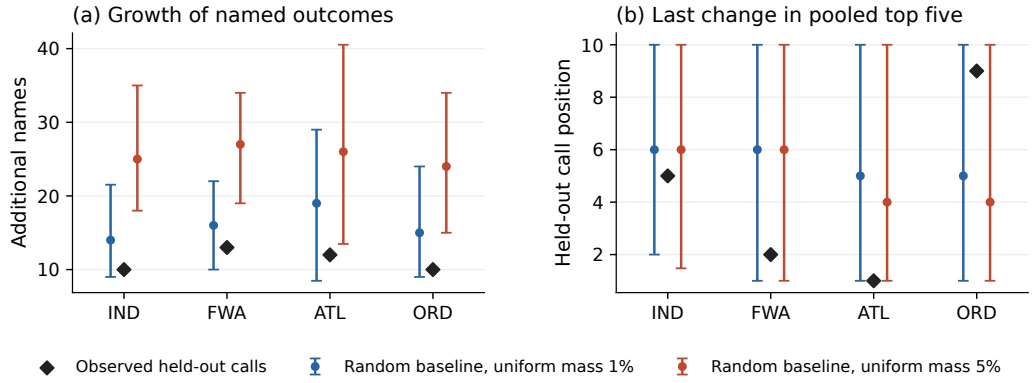


Figure 4: A stationary random baseline fitted to the first ten calls per context reproduces growth in named outcomes and top-five stabilization at the smaller uniform mass. Colored points and bars are simulated medians and 95% ranges for the last ten calls; diamonds are observed held-out calls. Ranges condition on fitted parameters; the comparison is exploratory.

6 Discussion and Limitations

Our results support two practices. First, treat the model, prompt, and processing as experimental variables, since each changes the delivered distribution: compare a reused top-five shortlist with the target model's own, both selected from equally many calls, on shared held-out calls, and test model-versus-wording comparisons against a second baseline batch. Second, report each LLM-derived distribution with its model, prompt, serving metadata, processing failures, and probability unmatched to the choice list.

Our design fixes the choice list, processing, and number of calls per average across compared models, trading breadth for comparability. Prompts are in English, and travel is the main domain, with a fixed, unweighted persona grid. The 70B/405B comparison also shares the documented prompt, airport order, temperature, and token limit, reconstructed from the collection code. Other comparisons keep their recorded settings, including token-limit differences and, for Mistral-24B, an adapted prompt with retries; under these settings, eight of nine model pairs vary more between than within models, the exception being Mistral-7B/24B (Section 5.3). Because training releases differ and serving metadata are unavailable, we attribute every model difference to the served models, not to size, pretraining, architecture, or alignment; the random baseline does not identify such causes either. Rewordings were tested for one airport and persona at 70B: one of four matches the 70B/405B change under the primary baseline batch, and the model change exceeds every rewording on average under the second. Restaurant and product outputs, each from one prompt at 70B, leave substantial probability unmatched, which we report with every distribution, and familiarity with the listed names from training data cannot be excluded. Resampled ranges describe the collected calls, not new batches, and the exploratory random baseline omits fitting uncertainty. We measure the consistency of stated distributions, not their accuracy for human choices, which requires separate behavioral validation. Natural extensions of our protocol are to cross models and rewordings in fresh batches and to add domains and languages.

7 Conclusion and Future Work

Repeated calls to one model yield comparatively self-coherent preference distributions. In contrast, different models disagree even on their most probable outcomes: between the 70B and 405B Llama models, reusing the other model's top-five shortlist loses 17.7 percentage points of probability coverage on average. On five-call averages, their JS divergence of 0.184 lies within the 0.045 to 0.192 range of four rewordings, depending on the baseline batch. These findings challenge the common assumption that sufficiently capable LLMs produce similar preference distributions when used as stand-ins for survey respondents. Future work should validate these distributions against independent human-choice data.

AI USE STATEMENT

We adopt the AI assistant to help with grammar checking and debugging coding errors. The experimental design, research evaluations, and manuscript writing are solely conducted by the authors.

REPRODUCIBILITY STATEMENT

The accompanying analysis package records input hashes, the catalog and alias policy, per-call retained mass, parser comparisons, condition coverage, deterministic seeds, corrected statistics, and figure builders. Its entry point rebuilds the audited outputs and figures from the saved project records without new model calls. Appendices E, I.4, and O.3 specify the transformations. Table 8 identifies the checkpoints and Section 4 describes the evaluated conditions. The original prompts and dataset provenance are retained in Appendices C and D. The split-sample control (Appendix P) is exploratory; its code and simulated results are included. Historical acquisition dates and generation settings are reported only where documented. The package separates corrected central analyses from explicitly labeled ancillary historical summaries.

ETHICS STATEMENT

This work analyzes the distributional properties of LLM-generated outputs under controlled prompt variations, including demographic and contextual cues. While these analyses can reveal systematic patterns in model behavior, they are not intended to support normative or policy decisions. Any downstream use of LLM-generated distributions, particularly in decision-support or simulation settings, should be accompanied by validation against real-world data and appropriate domain expertise. Our findings further underline that model choice substantially influences the generated distribution, so transparency about model selection and sensitivity analyses is important whenever LLM outputs inform applied decisions. The experiments elicit synthetic preference distributions over publicly available airport, business, and product metadata and do not recruit human participants. Care should still be taken to avoid misinterpretation of demographically conditioned outputs. Large-scale LLM inference also incurs nontrivial computational and environmental costs, which we encourage practitioners to consider when planning experiments.

ACKNOWLEDGMENTS

We thank the developers of the open-weight LLMs evaluated here (Meta, Alibaba, Mistral AI) and the dataset providers (the Yelp Open Dataset and the Amazon Product Data collection released by the McAuley group), whose publicly released models and data made this study possible.

REFERENCES

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, 2022. doi: 10.18653/v1/2022.findings-emnlp.423. URL https://aclanthology.org/2022.findings-emnlp.423/.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024.

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. URL https://doi.org/10.1017/pan.2023.2.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, USA, 2nd edition, 2006. ISBN 978-0-471-24195-9.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5040–5060, 2023. doi: 10.18653/v1/2023.emnlp-main.306. URL https://aclanthology.org/2023.emnlp-main.306/.

Fan Huang. Simulating hate speech cascades with multi-LLM agents: Empirical grounding, modeling fidelity, and intervention strategies. *ArXiv*, abs/2606.18264, 2026. URL https://arxiv.org/abs/2606.18264.

Fan Huang, Haewoon Kwak, and Jisun An. Vulnerability of LLMs' stated beliefs? LLMs belief resistance check through strategic persuasive conversation interventions. *ArXiv*, abs/2601.13590, 2026. URL https://arxiv.org/abs/2601.13590.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912. doi: 10.1111/j.1469-8137.1912.tb05611.x. URL https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Janus. Mysteries of mode collapse, 2022. URL https://www.lesswrong.com/posts/t9svvNPNmFf5Qa3TA/mysteries-of-mode-collapse. LessWrong.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *ArXiv*, abs/2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *ArXiv*, abs/2401.04088, 2024. URL https://arxiv.org/abs/2401.04088.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022. URL https://arxiv.org/abs/2207.05221.

Minsuk Kim, C. Tyler Diggans, and Filippo Radicchi. Modeling resource consumption in the US air transportation system via minimum-cost percolation. *Nature Communications*, 16(1):8105, 2025. doi: 10.1038/s41467-025-63489-w. URL https://doi.org/10.1038/s41467-025-63489-w.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, pp. 51371–51408, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a134eaebd55b7406ff29cd75d5f1a622-Abstract-Datasets_and_Benchmarks.html.

Charles J. Lovering, Michael Krumdick, Viet Dac Lai, Varshini Reddy, Seth Ebner, Nilesh Kumar, Rik Koncel-Kedziorski, and Chris Tanner. Language model probabilities are not calibrated in numeric contexts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29218–29257, 2025. doi: 10.18653/v1/2025.acl-long.1417. URL https://aclanthology.org/2025.acl-long.1417/.

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

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. Bias in language models: Beyond trick tests and towards RUTEd evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 137–161, 2025. doi: 10.18653/v1/2025.acl-long.7. URL https://aclanthology.org/2025.acl-long.7/.

Mistral AI. Mistral-7B-Instruct-v0.3: Model card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Version 0.3. Webpage consulted September 21, 2026.

Mistral AI. Mistral-Small-24B-Instruct-2501: Model card, 2025. URL https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501. Release 2501. Webpage consulted September 21, 2026.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018/.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)*, 2024.

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 (UIST '23)*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, volume 34, pp. 11054–11070, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5c04925674920eb58467fb52ce4ef728-Abstract.html.

Weihong Qi, Fan Huang, Jisun An, and Haewoon Kwak. A cross-cultural comparison of LLM-based public opinion simulation: Evaluating Chinese and U.S. models on diverse societies. *ArXiv*, abs/2506.21587, 2025. URL https://arxiv.org/abs/2506.21587.

Weihong Qi, Fan Huang, Rasika Muralidharan, Jisun An, and Haewoon Kwak. XChoice: Explainable evaluation of AI-human alignment in LLM-based constrained choice decision making. *ArXiv*, abs/2601.11286, 2026. URL https://arxiv.org/abs/2601.11286.

Qwen. Qwen3-235B-A22B-Instruct-2507: Model card, 2025. URL https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507. Release 2507. Webpage consulted September 21, 2026.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://arxiv.org/abs/2412.15115.

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

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159. URL https://doi.org/10.2307/1412159.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024. doi: 10.1162/tacl_a_00685. URL https://aclanthology.org/2024.tacl-1.56/.

U. S. Census Bureau. American community survey (ACS), n.d. URL https://www.census.gov/programs-surveys/acs. Program documentation. Webpage consulted September 21, 2026.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344, 2022. doi: 10.18653/v1/2022.naacl-main.167. URL https://aclanthology.org/2022.naacl-main.167/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Yelp. Yelp open dataset. https://business.yelp.com/data/resources/open-dataset/, n.d. Dataset downloaded December 2025; webpage consulted September 21, 2026.

Lake Yin and Fan Huang. DIF: A framework for benchmarking and verifying implicit bias in LLMs. *ArXiv*, abs/2505.10013, 2025. URL https://arxiv.org/abs/2505.10013.

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

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3623–3643, 2024. doi: 10.18653/v1/2024.acl-long.198. URL https://aclanthology.org/2024.acl-long.198/.

Table 7: Comparison of distribution elicitation methods for IND departures (Llama-3.3-70B). *Sup.*: unique destinations observed. *Entropy*: Shannon entropy in bits. *Top-5*: probability mass in top 5 destinations. *JS*: base-2 Jensen–Shannon divergence against the verbalized baseline at the same temperature; "(ref.)" marks the verbalized rows, which *are* that baseline (self-divergence 0). Entropy, Top-5 and JS are mean ± half-CI over 10,000 bootstrap resamples of the underlying calls (20 verbalized calls at T=0.2 and 18 at T=1.0; 300 single-choice draws at each). Support is an exact count over the full sample and is given without an interval, because resampling calls can only remove distinct destinations and never add them, so its bootstrap is biased downward. The zero intervals on the single-choice rows are structural rather than rounding: at T=0.2 every draw returned the same destination, and with a support of one or two all mass necessarily falls inside the top five. These are historical pilot values under its original output vocabulary. Support is pooled across calls; the comparison does not identify latent beliefs or measure task accuracy.

Method	**T**	*Sup.*	*Entropy*	*Top-5*	*JS*
Single-choice	0.2	1	0.00±0.00	1.00±0.00	0.64±0.02
Single-choice	1.0	2	0.18±0.10	1.00±0.00	0.61±0.04
Verbalized	0.2	35	3.97±0.29	0.58±0.06	(ref.)
Verbalized	1.0	34	4.25±0.20	0.46±0.04	(ref.)

A Elicitation Protocol Comparison

The archived IND pilot compares a prompt requesting one destination with a prompt requesting a numerical vector, using Llama-3.3-70B. At temperature 0.2, all 300 single-choice outputs were ORD; at temperature 1.0, ORD and MDW accounted for 97% and 3%. The verbalized conditions contained 20 and 18 usable calls, respectively. Table 7 preserves this pilot's original vocabulary and analysis, separately from the catalog-validated central results. Its support values of 35 and 34 are pooled counts, not outcomes guaranteed in one call.

The two instructions elicit different response distributions. Greater entropy or support does not establish more faithful access to a latent belief, and the verbalized baseline is not a ground-truth reference. Token probabilities would measure yet another object, requiring a specified scoring procedure for multi-token choices. We make no general efficiency claim without token, latency, and failure accounting, and do not infer a universal number of required calls from this pilot.

B Detailed Metric Definitions

The corrected main analyses use the same normalized, catalog-validated call vectors. A reported statistic identifies whether it acts on one call, a pair of calls, or an average of calls. Repeated API requests are separate records; exchangeability of a saved pool is a working description for conditional subset analysis and is not a claim that provider-side temporal dependence has been excluded.

B.1 Distributional Consistency Metrics

For $m = (p + q)/2$, Jensen–Shannon divergence is

$$\mathrm{JS}(p, q) = \frac{1}{2} \sum_y p_y \log \frac{p_y}{m_y} + \frac{1}{2} \sum_y q_y \log \frac{q_y}{m_y}. \tag{5}$$

We use natural logarithms and evaluate zero terms by continuity, giving a bound of $\ln 2$. JS distance is $\sqrt{\mathrm{JS}}$; the distance threshold 0.05 used for the retrospective prefix index is a divergence threshold of 0.0025 nats. The historical pilot table explicitly uses base-2 JS. Those conventions must be converted before comparison (Lin, 1991).

Top-k Jaccard is $|M_k(p) \cap M_k(q)|/|M_k(p) \cup M_k(q)|$ (Jaccard, 1912). Exact agreement is $\mathbb{1}\{M_k(p) = M_k(q)\}$. We average these separately across the fixed persona grid. Canonical identifiers are sorted alphabetically before stable sorting by decreasing probability, so ties have a deterministic, reproducible rule. Top-one agreement uses the same rule. Sets retain k coordinates, including

zero-mass ties when fewer than k coordinates are positive. Spearman rank correlation (Spearman, 1904) is an optional rank diagnostic; no latent-belief inference follows from rank or set agreement alone.

B.2 Support, Concentration, and Entropy

Named support is $|\{y : \bar{p}_S(y) > 0\}|$, the union of positively weighted canonical outcomes in the selected calls. Active support is $|\{y : \bar{p}_S(y) > 0.001\}|$ and can change non-monotonically as the average changes. Entropy is $-\sum_y \bar{p}_S(y) \log_2 \bar{p}_S(y)$ in bits; top-k mass sums the k largest coordinates (Cover & Thomas, 2006). These pooled quantities differ from the average of the corresponding per-call quantities. We report the intended target explicitly.

B.3 Additive and Conditional Region Measures

For a pair, define $M = M_k(p) \cup M_k(q)$ and $T = \mathcal{Y} \setminus M$. Summing the coordinate contributions c_y from Section 3.4 gives $C_M + C_T = \mathrm{JS}(p, q)$ exactly. For a phase, we average each contribution within context and then across the four contexts; the reported tail share is $\overline{C_T}/\overline{\mathrm{JS}}$. The partition is chosen descriptively from the same pair, not an independently identified semantic boundary.

Conditional mode and tail divergences separately normalize each restricted vector. They measure variation within those regions but do not attribute total divergence. A conditional measure is undefined if either restricted vector has zero mass; the result files record how many pairs define each measure. Consecutive comparisons share calls. Accordingly, the corrected tables give descriptive summaries, with context-specific results in the data package, rather than independent-pair confidence intervals.

B.4 Prompt and Checkpoint Comparisons

Prompt effects average JS over the specified comparison list. Persona presence compares matched departure airports; age and education compare fixed demographic cells; context compares specified regional/hub cells. The comparisons share responses, and the grids are not population samples. Table 18 therefore reports mean, descriptive SD, and pair count. Table 3 likewise describes its 20 fixed personas. Neither table estimates the full distribution of future responses to each prompt.

Table 8: Complete model configuration across three families. Primary records use the Together AI API at $T = 0.2$; supplementary $T = 0$ conditions are identified separately. Historical prompt and budget differences are documented in Appendices E and I.

Family	Model ID	Params	Role
Llama	Llama-3.1-8B-Instruct-Turbo	8B	Smaller
	Llama-3.3-70B-Instruct-Turbo	70B	Baseline
	Meta-Llama-3.1-405B-Instruct-Turbo	405B	Larger
Qwen	Qwen2.5-7B-Instruct-Turbo	7B	Small
	Qwen2.5-72B-Instruct-Turbo	72B	Medium
	Qwen3-235B-A22B-Instruct-2507-tput	235B (MoE)	Large
Mistral	Mistral-7B-Instruct-v0.3	7B	Small
	Mistral-Small-24B-Instruct-2501	24B	Medium
	Mixtral-8x7B-Instruct-v0.1	47B (MoE)	Large

B.5 Cutoff Sensitivity

We use $k = 5$ as a compact fixed-size comparison convention. Fixed k makes the compared sets equal in size; it does not identify a unique psychological or semantic boundary. Table 9 repeats the original-mass attribution at five cutoffs. Broader mode regions leave less contribution in their complement, so the amount attributed to the tail is explicitly cutoff-dependent.

Table 9: Cutoff sensitivity of additive original-mass attribution. Mode and tail contributions use earlier calls 6–10; the last two columns give tail shares for earlier calls 6–10 and 11–19. The partition changes with k, so these shares need not be invariant. Values are descriptive airport averages; shares are computed before rounding.

k	Mode part	Tail part	Tail % (6–10)	Tail % (11–19)
3	0.011	0.044	80.7%	77.7%
5	0.015	0.040	73.4%	67.7%
7	0.020	0.035	64.4%	59.9%
10	0.029	0.026	48.1%	40.2%
15	0.042	0.013	24.1%	16.5%

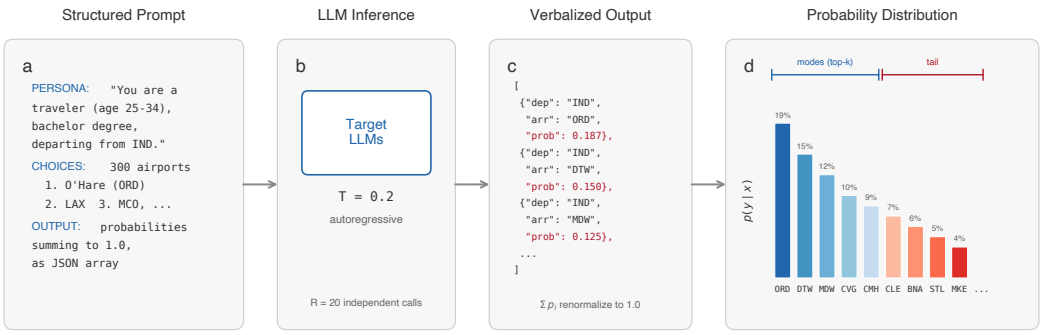


Figure 5: Overview of the experimental pipeline. (a) A structured prompt provides the LLM with a traveler persona (age, education), a departure airport, and 300 candidate destination airports, requesting probability assignments as a JSON array. (b) The target LLM generates a response autoregressively at temperature T=0.2; this process is repeated through separate API calls. (c) Each response is parsed, mapped to the catalog with an exclusion ledger, and normalized within the retained call. (d) The resulting probability distribution over destinations, shown here for Indianapolis (IND) using real experimental data. High-probability destinations (blue, *modes*) and low-probability destinations (red, *tail*) are analyzed for distributional consistency, mode–tail contributions, checkpoint-condition differences, and prompt sensitivity. See Appendix C.2 for complete prompt templates.

B.6 Discrete Thresholds and Finite-Sequence Stability

For two size-five sets, Jaccard can take only $0, 1/9, 1/4, 3/7, 2/3, 1$. A threshold strictly between $2/3$ and 1 therefore detects exact set equality. This algebraic threshold invariance does not establish convergence. We report R_{last}, the last observed change in the cumulative top-five set, and its call-permutation distribution. It is 1 if no change occurs and can equal the last observed call. A same-pool JS reference forces the final distance to zero. Neither construction measures future-call error; no fixed-call stopping recommendation is inferred.

C Experimental Configuration Details

This section provides complete details for the experimental setup summarized in Section 4.

C.1 Persona Configuration

We define personas using the following age and education categories. The American Community Survey (U. S. Census Bureau, n.d.) provides demographic background; the categories are specified directly in the experiment code:

Age Groups (5 levels):

- 18–24 (Young)

- 25–34 (Young Adult)
- 35–44 (Adult)
- 45–64 (Middle Age)
- 65 and above (Senior)

Education Levels (4 levels):

- Under High School
- High School to Under Bachelor
- Bachelor Degree
- Above Bachelor (Master/Doctoral)

This yields 20 unique persona combinations ($5 \times 4 = 20$). All combinations are included as prompt conditions. No ACS population estimates are used to construct or weight the grid, and no specific ACS table or release vintage is attributed to it.

C.2 Prompt Templates

Each LLM query is constructed from four components: a *persona*, a *choice set*, a *question*, and an *instruction* specifying the output format. Below we reproduce the templates for each experimental condition, with uppercase section labels added here for readability and long choice sets elided; placeholder fields (e.g., `{age_range}`) are filled dynamically.

Base Prompt (Travel Demand, Persona-Conditioned).

```
PERSONA: You are a traveler (age {age_range})
  with {education_desc}, departing from
  {departure_state} ({departure_name}).

AVAILABLE AIRPORTS:
1. {airport_name} ({code}) - {state}
2. {airport_name} ({code}) - {state}
... [300 airports listed]

QUESTION: Choose your travel destinations from
  the airports listed above. What are your
  travel plans and their probabilities?

INSTRUCTION: Please plan travel for the next
  12 months based on your demographic profile.
  Provide probabilities as float values from
  0 to 1, where higher values indicate higher
  likelihood. All travel probabilities should
  sum to 1.0. Focus only on destination choices
  regardless of connection routes.
  Format your response as:
  [{"departure_airport":, "arrival_airport":,
    "travel_probability":}, ...]
```

No-Persona Variant (Control). The persona field is reduced to: "`You are a traveler departing from {departure_state} ({departure_name}).`" All other components remain identical.

Seasonal Variant. The persona is appended with: "`You are planning travel for {season_context}.`" where `season_context` is either "summer (June–August)" or "winter (December–February)." The question is modified to: "What are your travel plans and their probabilities for this season?"

Restaurant Choice (Yelp).

```
PERSONA: You are a diner (age {age_range})
  with {education_desc}, living in {city}.

AVAILABLE RESTAURANTS:
1. {restaurant_name} ({cuisine_category})
... [300 restaurants listed]

QUESTION: Choose your dining preferences from
  the restaurants listed above. What is your
  probability of dining at each restaurant
  in the next month?

INSTRUCTION: Please indicate your restaurant
  preferences based on your demographic profile.
  Provide probabilities as float values from 0
  to 1, where higher values indicate higher
  likelihood of dining there. All probabilities
  should sum to 1.0. Format your response as a
  JSON array (no other text before or after):
  [{"restaurant": "Restaurant Name",
    "probability": 0.05}, ...]
```

Product Choice (Amazon).

```
PERSONA: You are an online shopper
  (age {age_range}) with {education_desc}.

AVAILABLE PRODUCTS:
1. {product_name} ({product_category})
... [300 products listed]

QUESTION: Choose your product preferences from
  the items listed above. What is your
  probability of purchasing each product
  in the next month?

INSTRUCTION: Please indicate your product
  preferences based on your demographic profile.
  Provide probabilities as float values from 0
  to 1, where higher values indicate higher
  likelihood of purchasing. All probabilities
  should sum to 1.0. Format your response as a
  JSON array (no other text before or after):
  [{"product": "Product Name",
    "probability": 0.05}, ...]
```

Expected Output Format. The LLM generates a JSON array as free-form text. For the travel domain, a typical output looks like:

```
[
  {"departure_airport": "IND",
   "arrival_airport": "ORD",
   "travel_probability": 0.15},
  {"departure_airport": "IND",
   "arrival_airport": "LAX",
   "travel_probability": 0.12},
  {"departure_airport": "IND",
```

```
   "arrival_airport": "MCO",
   "travel_probability": 0.08},
  ...
]
```

We sum the nonnegative numeric weights in each selected raw response before catalog projection and normalization. Across nine primary IND checkpoint conditions (20 calls each; 180 outputs), the median sum is 0.98. Inclusive intervals contain 74/180 (41.1%) within $[0.95, 1.05]$ and 100/180 (55.6%) within $[0.9, 1.1]$. Checkpoint medians range from 0.81 to 1.29. The audit distinguishes selected raw-response mass, legacy saved-parser mass, retained catalog mass, and the normalization factor. Prose claims about sums are not substituted for numeric records. Incomplete recovered responses remain flagged (Appendix E).

C.3 Research Question Details

The repeated-call study describes pooling behavior and order sensitivity. Fixed-persona comparisons assess recorded checkpoint-condition differences, while prompt comparisons describe the selected demographic and geographic contrasts. Cross-domain cases examine outcome mapping and conditional aggregation. These axes are not fully crossed, and the study does not identify population preferences or a causal effect of parameter count.

D Cross-Domain Dataset Descriptions

This section documents the datasets and sampling procedures for the two single-context, processing-sensitive non-travel case studies.

D.1 Yelp Restaurant Choice Dataset

Source: Yelp Open Dataset (Yelp, n.d.) (`https://business.yelp.com/data/resources/open-dataset/`). We use the archived `yelp_academic_dataset_business.json`, containing 150,346 business records, downloaded in December 2025. Preserved local acquisition metadata indicate December 18, 2025. A formal dataset release identifier is not documented; dates printed in the accompanying terms of use are not treated as release dates.[1]

Data Extraction: We select businesses in Indianapolis, Indiana, to match the geographic context of the primary travel experiments, using the following filters.

Filtering Criteria:

- *Location:* Businesses located in Indianapolis (city = "Indianapolis")
- *Category:* The lowercased Yelp category text contains any of 22 dining keywords (for example *restaurant*, *cafe*, *pizza*, *seafood*) and none of seven exclusion keywords (for example *grocery*, *catering*, *food truck*)
- *Status:* Only open businesses (is_open = 1)
- *Name Quality:* Required a non-missing name
- *Category Quality:* Required a derived primary category other than "Other"; the extraction selects a cuisine by priority, or falls back to the first non-generic category

The category filter yields 58,438 candidates overall and 3,163 in Indianapolis. The remaining checks leave 2,119 eligible businesses; we sample 300 without replacement using random seed 42.

Final Dataset:

- 300 sampled Indianapolis businesses, presented as restaurant options in the prompt
- 54 distinct primary category labels
- Exported fields: name, cuisine/category label, price tier, star rating, and review count

[1] Business-file SHA-256: `392280627d83ecdcbff8006bba5db289f8436ef147d5a103b1ace5b54186727f`.

Table 10: Outcome identity in the supplied choice lists. The identity used in the response is the relevant coordinate for aggregation; a listed business row is not distinguishable from another branch sharing the same name.

Domain	Listed rows	Distinct response keys	Unit
Travel	300	300	Airport code
Yelp	300	252	Business name
Amazon	300	300	Catalog title

Scope of the Choice Set: The keyword filter admits a broader set than businesses carrying the literal Yelp category "Restaurants": 255 sampled entries carry that category and 45 do not. The 54 primary labels include cuisines, cafes, bars, and other business categories. These entries remain in the original experimental choice set; the restaurant-choice results therefore describe this heuristic sample.

D.2 Amazon Product Dataset

Source: Amazon Product Data (Ni et al., 2019) (`https://nijianmo.github.io/amazon/index.html`)

Category: Home & Kitchen product metadata

Sampling Procedure: To ensure semantic diversity while maintaining manageable choice set size, we implemented a stratified embedding-based sampling approach:

1. *Category Stratification:* Products were grouped by subcategory within Home & Kitchen, yielding 15 subcategories (e.g., Cookware & Bakeware, Bedding, Bath, Storage & Organization, Home Decor, etc.)
2. *Embedding Generation:* Product titles concatenated with brand names were embedded using OpenAI's `text-embedding-3-large` model to capture semantic similarity
3. *Clustering:* Within each subcategory, K-means clustering was applied to identify semantically distinct product groups
4. *Representative Sampling:* Products closest to each cluster centroid were selected as representatives, ensuring coverage of diverse product types within each subcategory
5. *Near-Duplicate Removal:* Products with embedding cosine similarity > 0.90 were filtered to prevent redundant entries
6. *Balanced Selection:* 20 products were sampled from each of the 15 subcategories, yielding 300 total products

Final Dataset:

- 300 Home & Kitchen products
- 15 subcategories with balanced representation (20 products each)
- The derived CSV contains ASIN, product title, assigned category, brand, price tier, and price; the prompt displays the first 80 title characters

D.3 Outcome Identity and Mapping Audit

The Yelp prompt supplies 300 rows but only 252 distinct exact-name keys. Name-only responses cannot distinguish branches. Across its 20 calls, 88 repeated name entries are summed into their name-level coordinate; unresolved names retain a separate audit entry. All 38 strings failing exact catalog matching also lack a unique match under the declared normalization, so no semantic alias was silently assigned. The recorded mean mass outside exact catalog names is 30.24%.

The Amazon prompt builder truncates each title to its first 80 characters. Some catalog titles contain HTML entities. Matching against the untruncated raw CSV alone therefore rejects some strings that correspond to displayed choices. We compare HTML-decoded, Unicode-normalized,

Table 11: Corrected single-context pooled summaries from 20 valid calls per condition. Duplicate mapped records are summed, each retained call is normalized, and calls receive equal weight. Named and active support have different thresholds. The distance index uses the same full-pool reference and is retrospective, with a forced zero endpoint.

Metric	Travel (IND)	Yelp	Amazon
Supplied rows	300	300	300
Distinct keys	300	252	300
Named support	31	238	86
Active support (> 0.001)	29	77	61
Top-ten mass (%)	74.4	53.4	75.6
Mean retained mass (%)	98.8	69.8	65.1
First distance < 0.05	8	19	18

case/punctuation-insensitive labels against both full titles and their archived displayed forms and accept only a unique match. This resolves 35 of the 53 initially unmatched strings; 18 remain unresolved. Mean unmatched mass decreases from 43.58% to 34.89%. A close spelling or plausible product description alone is not sufficient evidence for a mapping.

Exact-string mismatch is not equivalent to a nonexistent business or product. The complete mapping ledger records accepted transformations and unresolved cases. The final distribution is explicitly conditional on retained catalog mass. These historical prompts and lists are preserved; a business-location experiment would require a new prompt with unique location identifiers.

D.4 Cross-Domain Pooled Results and Sensitivity

The original Yelp implementation repeated duplicate name coordinates in its JS vectors. Keeping the original weighting but deduplicating only those coordinates changes the first distance threshold crossing from 11 to 20. Applying the full declared correction, including summed duplicates and equal-call weighting, gives the values in Table 11. These steps answer different questions and must not be conflated. In particular, replacing an index with 20 would not validate a stopping rule because the twentieth call belongs to the reference itself.

Figure 6 illustrates how the documented policies change matched probability, concentration, and active support.

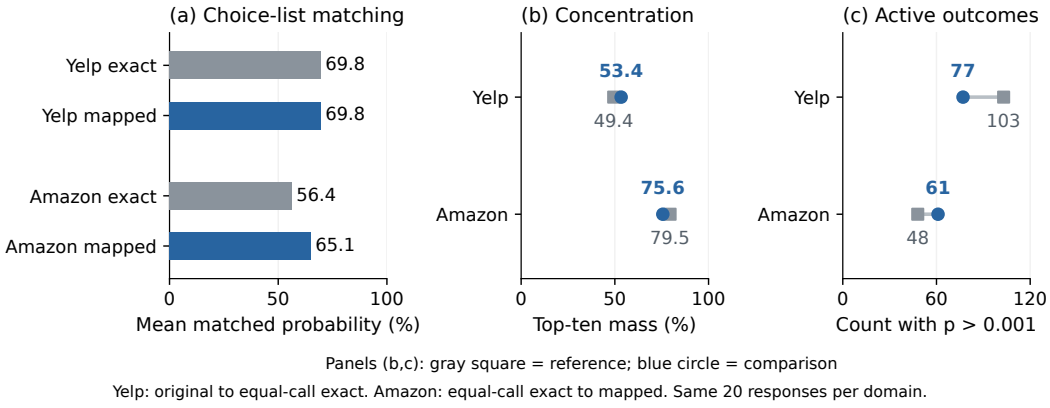


Figure 6: Processing choices change distributions derived from the same saved responses. (a) Mean probability matched to the choice list by exact and mapped matching, before renormalization. (b,c) Yelp compares the original processing (last value per name, calls weighted by matched probability) with duplicate summation and equal call weights; Amazon compares exact and displayed-title matching under equal call weights. Active outcomes have pooled probability > 0.001. Each domain uses 20 calls; Table 6 reports all policies.

Table 12: Paired parser audit on identical saved text within each row. Array-only and recovery columns count responses with positive, catalog-valid mass. Original and regenerated persona archives are different texts; their difference is not a parser-only effect.

Archive	Texts	Array only	Recovery
8B original persona texts	20	3	16
8B regenerated persona texts	20	10	20
8B convergence texts	20	6	20
70B ORD convergence texts	20	18	20
405B IND convergence texts	20	19	20

E Parsing and Generation Audit

Repeated calls and fixed-grid profiles are reconstructed from raw response text under one extraction policy. Legacy saved parses supply comparison evidence. The parser accepts complete arrays with documented trailing-comma/comment cleanup. For multiple complete arrays, a response-hash ledger records whether a later list replaces a draft or explicitly continues it; unknown cases are excluded pending adjudication. Twelve distinct archived responses require such decisions. Ellipses and arithmetic expressions are not completed or evaluated. If no array is recoverable, complete structured objects or explicit adjacent arrival/probability lines can be recovered. Keys still require catalog membership. Truncated responses, including incomplete revisions after a complete list, remain flagged; extraction does not certify a complete generation.

All 20 raw texts in the original and rectified 8B convergence archives are identical. On those same texts, array parsing yields 6 catalog-valid outputs and the documented recovery yields 20. Thus this comparison isolates processing of saved responses. Conversely, none of the 20 paired original and regenerated 8B persona texts are identical. Recovery obtains 16 usable outputs from the original archive and 20 from the regenerated one; array-only counts are 3 and 10. The historical change from 3 to 20 therefore mixes generation and parsing changes and cannot be attributed entirely to parser repair.

The collection documentation reports an increase in maximum response budget from 1,024 to 4,096 tokens. The CSV rows do not individually encode every generation setting, so an exact parser-by-budget causal decomposition is unavailable. The analysis package records source configuration expressions and hashes rather than inferring undocumented per-call settings. The current persona comparison uses the regenerated 8B archive and labels that historical condition.

Two ORD responses that failed greedy array extraction are recovered by selecting a complete array; one 405B response is recovered from complete objects. Their inclusion changes phase summaries as well as support. The raw-text spot check and processing audit record these cases, including revised arrays in the temperature-zero archive. Airport keys are checked against the supplied 300-code catalog after extraction; accepting three uppercase letters syntactically is insufficient. Ambiguous metro labels and out-of-list airports remain in the excluded-mass ledger. Duplicate records within the selected answer are summed before each call is normalized.

Qwen-7B call 4 explicitly adds a second list, so both lists are retained. Call 12 supplies an adjusted list repeating the draft destinations, so that list replaces the draft. Table 13 reports alternative policies on the same responses, including exclusion of both cases. The selected policy follows the response structure; it was not selected to maximize checkpoint differences. A last-array-only rule would discard call 4's initial choices. Three regenerated 8B persona responses also require explicit airport-code extraction: the older saved parser omitted named airports and introduced fragments of airport names as codes. The corrected persona results use the explicit response identifiers. Two additional airport profiles are recovered from complete objects; all 300 paired persona-presence contexts now have positive retained mass.

E.1 Historical Results and Traceability

The original source files, earlier parsing summaries, and manuscript are preserved. Every corrected call record identifies its source and position, parsing route, probability sum, retained/excluded mass,

Table 13: Qwen-small parsing sensitivity. JS compares full recorded pools; Δ is the mean between-minus-within JS at five calls per aggregate. S/M/L identify the recorded Qwen checkpoints. Exclusion leaves 18 small-checkpoint calls; other pools retain 20. Conditional ranges are supplied with the results. The last-array diagnostic is inappropriate for the explicit continuation and is not an accepted replacement policy.

Policy	N	JS S–M	JS S–L	$\Delta_{\mathrm{S-M}}$	$\Delta_{\mathrm{S-L}}$
Adjudicated raw	20	0.123	0.224	0.081	0.191
Legacy saved	20	0.124	0.226	0.082	0.193
Exclude calls 4/12	18	0.131	0.233	0.085	0.194
Last-array diagnostic	20	0.138	0.234	0.080	0.186

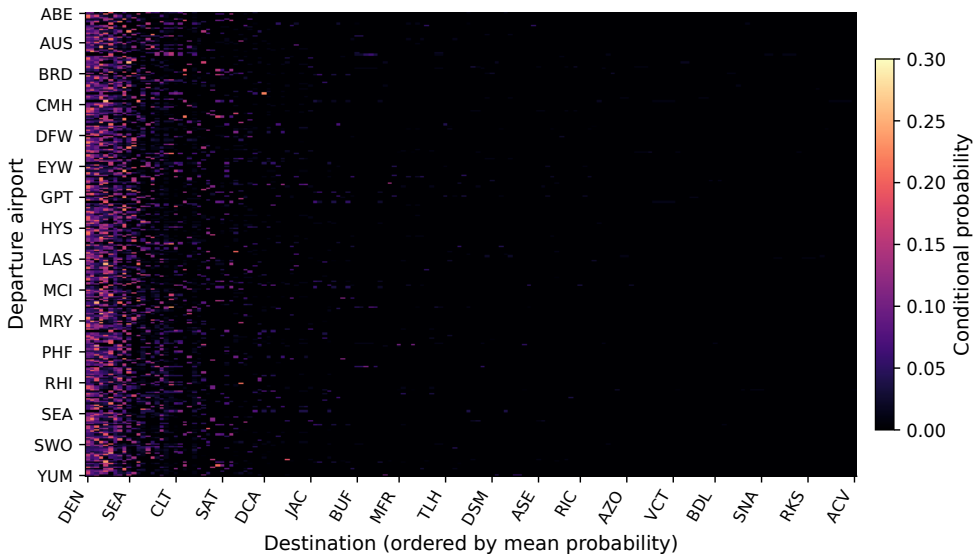


Figure 7: Catalog-validated persona-conditioned output matrix for 300 recorded departure contexts and 166 destination codes. Rows are ordered by departure code; columns by mean conditional probability. Labels are thinned for readability. These are model outputs, not observed travel demand.

normalization factor, repeated keys, and validity. This makes parser-only recovery distinguishable from changing generated responses. The corrected support and comparison tables supersede historical unvalidated-key counts rather than relabeling them as current results.

F Detailed Experimental Results

The all-airport persona-conditioned archive contains 300 departure contexts with positive catalog-valid mass after processing. Their union contains 166 destination codes. This differs from the historical unvalidated-key count of 185: metro codes, out-of-catalog airports, and unresolved names are not retained as distinct supplied choices. The heatmap shows conditional distributions under the recorded fixed persona, not observed passenger flows.

G Pooled-Distribution and Order Sensitivity

Figures 3 and 8 recompute pooled statistics on subsets of the corrected Yelp and Amazon call vectors. They describe the original prompts after the name and title matching in Appendix D. Each subset receives equal call weights. At the full batch, distance to it is zero and top-five Jaccard with it is one

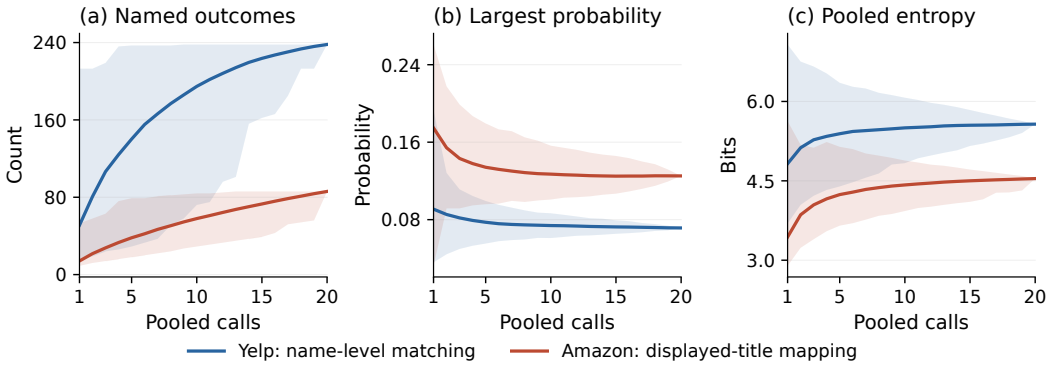


Figure 8: Pooled signatures for the restaurant and product batches. Named outcomes include every positive coordinate and differ from active outcomes above 0.001. Matched probability and alternative matching rules are reported in Tables 11 and 6; the curves are conditional on the declared matching.

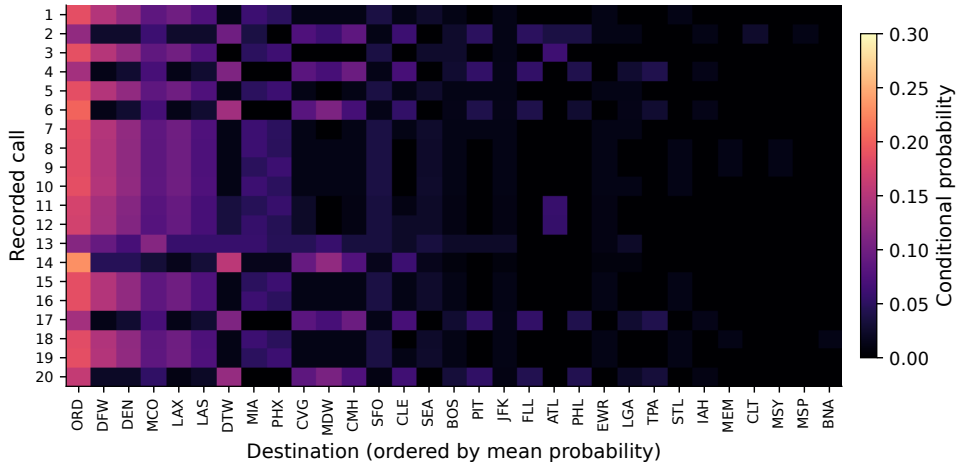


Figure 9: Twenty recorded IND calls after catalog validation and within-call normalization. The 31 columns are positively named destination codes, ordered by mean probability. Appearance or persistence of a column is a property of this output sequence.

by construction, so the bands describe sensitivity within the collected calls, not evidence that new calls will agree.

G.1 Extended Stability Analysis

The travel IND support after catalog validation is 31 positive coordinates, of which 29 exceed 0.001 in the full mean. The corresponding Llama-405B and 8B counts are 49 and 70 positive coordinates. These are properties of the complete pools, not averages of support counts over prefix lengths. The original signature plots averaged overlapping prefix metrics; that target has been replaced throughout by properties recomputed from call subsets. Call-permutation results and the fraction of sequences whose last change occurs at the endpoint are supplied in the analysis package. Endpoint changes are not followed by any observed confirmation of persistence.

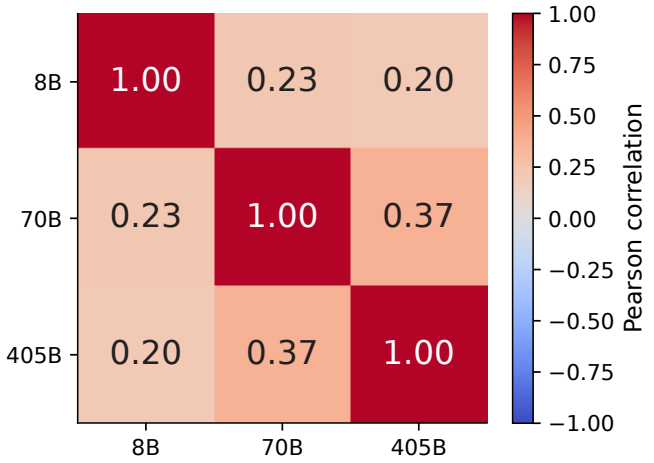


Figure 10: Pearson correlation of the corrected IND persona matrices, flattened over the same persona-by-destination coordinates. This is a descriptive representation-dependent comparison, not evidence of independent world priors.

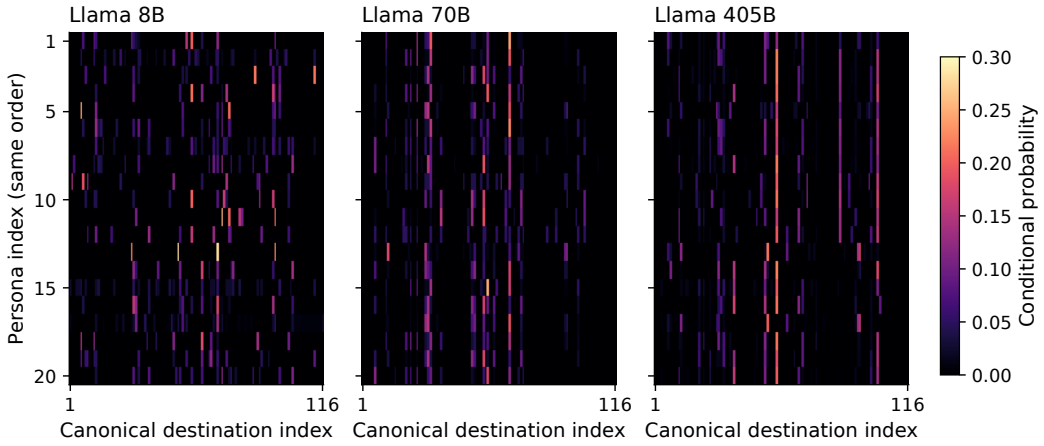


Figure 11: Corrected IND persona distributions for the three Llama models. Rows share the same 20-persona order and columns the same destination order. Color shows probability renormalized over matched outcomes. The regenerated 8B persona batch differs from its original generation batch.

H Checkpoint Comparison: Detailed Views

The persona matrices use the same 20 IND personas and a common union of canonical destination coordinates. Pearson correlation of the flattened matrices is a descriptive supplement to the JS and top-set comparisons. It depends on the chosen coordinate representation and is not a test that checkpoints encode independent latent beliefs.

H.1 Distributional Breadth Across Recorded Conditions

For the repeated IND conditions, the full-pool positive supports are 70 (8B), 31 (70B), and 49 (405B). These are the endpoints of the corrected pooled-support curves. Historical counts before catalog

Table 14: JS between full 20-call IND pools on validated airport keys. S/M/L denote the three recorded checkpoints in each family, not a controlled scaling intervention. All rows use the same aggregation unit. Mistral-M uses the adapted prompt and retry-selected successes; historical generation settings also limit cross-checkpoint causal interpretation.

Family	S–M	M–L	S–L
Llama	0.288	0.164	0.338
Qwen	0.123	0.148	0.224
Mistral	0.204	0.138	0.271

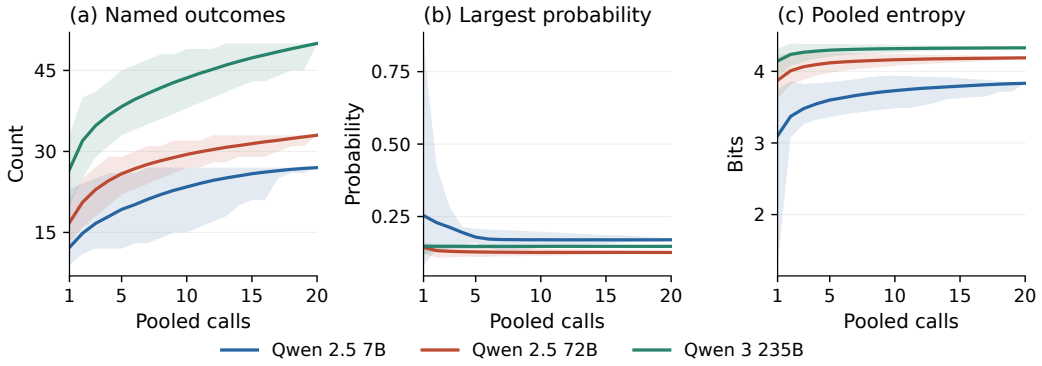


Figure 12: Qwen pooled signatures recomputed from subsets of valid calls, with the same metrics and resampled-range interpretation as Figure 1. Training versions and architectures differ across the three models.

and identity validation are not interchangeable with these values. Differences among the evaluated checkpoints combine training-version, generation, and processing conditions; the study does not identify a causal parameter-count effect or establish which vector is more accurate.

I Cross-Family Methodology

I.1 Model and Protocol Scope

Table 8 lists the nine checkpoints. Llama mixes versions 3.1 and 3.3; Qwen mixes 2.5 and 3; the Mistral set mixes dense and mixture-of-experts models and training releases. Version-specific model cards document the Qwen3-2507 and Mistral-v0.3 checkpoints (Qwen, 2025; Mistral AI, 2024). The Qwen/Mistral extension configuration specifies temperature 0.2, a 4,096-token budget, IND departures, and 300 airport options. In contrast, the focused Llama 70B/405B comparison and 70B wording archive reconstruct the same 1,024-token baseline request (Appendix K.2). Other budget differences, adapted prompts, and missing serving metadata prevent an identical-protocol claim for the full nine-checkpoint set.

I.2 Pooled Signatures and Matched Aggregation

I.3 Exploratory Shortlist Transfer

Consider reusing a five-destination shortlist after changing the recorded checkpoint condition. In each of 1,000 draws, two disjoint five-call groups per condition supply selection and evaluation averages. Let S_A and S_B be top-five sets from the selection averages. On the same evaluation vector $p_{B,\text{eval}}$, compare mass covered by S_B with mass covered by S_A; reverse the roles and average the two differences. This uses the same evaluation calls for both selectors and an equal selection budget. Table 16 reports all nine comparisons, including the Mistral-7B/24B range crossing zero. The diagnostic is an exploratory reuse of inspected pools, not a frozen prospective test or an estimate

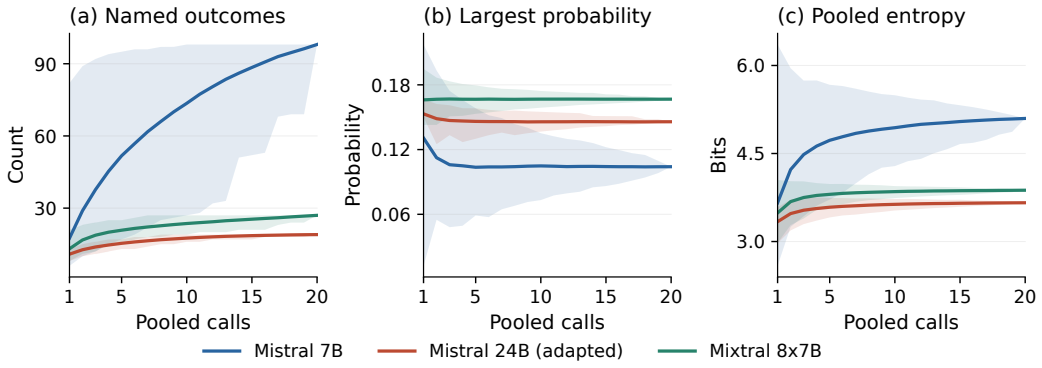


Figure 13: Mistral pooled signatures. The 24B curve is conditional on the adapted prompt and retry-selected successes. Lines and bands summarize subsets of the collected calls and do not establish a parameter-count law.

Table 15: Five-call average comparisons on the IND batches. Between uses one average from each model; within averages the two comparisons of non-overlapping averages, one per model. Differences are computed before rounding. Ranges are 95% resampled ranges. Original generation settings, including the adapted 24B prompt (*), are retained.

Pair	Between	Within	Difference	95% range
Llama 8B/70B	0.348	0.141	0.206	[0.102, 0.295]
Llama 8B/405B	0.394	0.144	0.250	[0.098, 0.381]
Llama 70B/405B	0.185	0.035	0.150	[0.103, 0.197]
Qwen 7B/72B	0.141	0.060	0.081	[0.021, 0.134]
Qwen 7B/235B	0.244	0.052	0.191	[0.130, 0.247]
Qwen 72B/235B	0.166	0.023	0.143	[0.110, 0.175]
Mistral 7B/24B*	0.229	0.099	0.130	[-0.019, 0.259]
Mistral 7B/Mixtral 8x7B	0.298	0.098	0.200	[0.064, 0.306]
Mistral 24B*/Mixtral 8x7B	0.153	0.032	0.121	[0.074, 0.166]

Table 16: Exploratory shortlist transfer on IND calls. Same and Cross are the held-out probability covered by a top-five shortlist selected within the target model or from the other model. Loss is Same minus Cross in percentage points, computed before rounding, so it can differ by 0.1 from the difference of the rounded columns. Means average both directions; ranges are 95% resampled ranges. Selection and held-out groups each use five calls and do not overlap within a model. Original generation differences, including the adapted 24B prompt (*), remain.

Pair	Same %	Cross %	Loss pp	95% range (pp)
Llama 8B/70B	38.9	26.1	12.8	[4.0, 24.0]
Llama 8B/405B	42.9	17.3	25.6	[12.6, 37.4]
Llama 70B/405B	54.6	37.0	17.7	[11.5, 24.0]
Qwen 7B/72B	48.8	42.0	6.9	[2.5, 11.1]
Qwen 7B/235B	50.5	35.6	14.9	[10.3, 20.2]
Qwen 72B/235B	49.4	33.5	15.9	[12.1, 19.1]
Mistral 7B/24B*	43.9	39.1	4.8	[-0.9, 15.5]
Mistral 7B/Mixtral 8x7B	43.2	32.5	10.7	[4.5, 22.4]
Mistral 24B*/Mixtral 8x7B	54.6	47.5	7.1	[1.7, 12.7]

of human demand. Appendix K.2 adds repeated 70B wording comparisons, while a fully crossed checkpoint-by-wording panel remains unavailable.

Figure 14 visualizes the focused 70B/405B comparison under this same protocol.

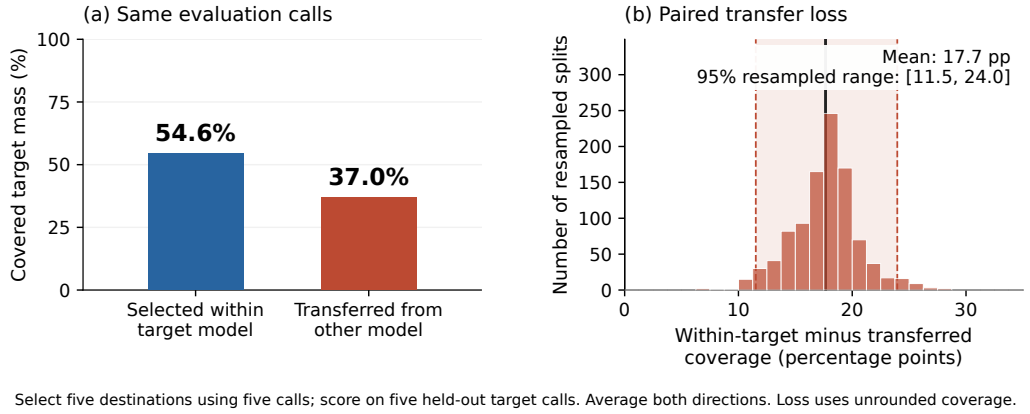


Figure 14: Reusing a shortlist between the 70B and 405B models loses 17.7 percentage points of probability coverage on average, computed from unrounded coverage. Each of 1,000 splits selects five destinations from five calls and scores both shortlists on the same five held-out target calls. Panel (a) shows mean coverage; panel (b) shows paired losses averaged across directions. The shaded 95% resampled range describes the collected calls, not a confidence interval for new calls.

I.4 Output Normalization

The corrected map accepts an exact catalog IATA code, a catalog code explicitly shown in parentheses, a leading catalog code followed by a parenthesized name, or a unique normalized full catalog name. Three-letter syntax alone is insufficient. Metro codes such as NYC and CHI, out-of-list airports, and unresolved strings are excluded from the conditional vector and retained in the audit. Records resolving to the same code are summed before normalization. Ordinary formatting differences are not labeled hallucinations.

I.5 Mistral-Small Adaptation and Selection

The archived attempt log contains 179 rows, with 20 catalog-valid parsed responses. Of the other 159, 135 contain a literal clarification cue and 24 are otherwise unusable under the recorded parser. These categories are descriptive text heuristics, not diagnoses of a training mechanism. The retained 24B outputs therefore estimate a distribution conditional on successful response under an adapted prompt and retry policy.

The adaptation adds “You are generating results for scientific research projects.” The original retry procedure sought 20 usable responses, with an upper limit of 500 attempts. We report the observed 20/179 success fraction (11.2%) and do not infer an RLHF cause. Table 15 retains the adapted condition explicitly; it is not a controlled substitution of checkpoint alone.

I.6 Recorded Counts and Metadata Limits

The analyzed primary IND files contain 20 usable calls per checkpoint after the documented recovery. This count does not imply that exactly 20 attempts were required for every model: complete earlier-attempt histories are not encoded in all retained CSVs. The package records source configuration expressions, source hashes, and the available attempt log. Collection dates and settings absent from per-call records are marked unavailable rather than reconstructed from file names.

J Persona-Based Probability Distributions

The grid crosses five age groups with four education groups without population weights. Each cell is a prompt condition rather than a sampled individual. The heatmap describes the recorded 70B outputs for IND; it does not estimate age or education effects in a human population.

Table 17: Mean probability mass retained by catalog validation in the primary repeated IND conditions. Each row has 20 saved usable calls. Retention is measured before within-call normalization and does not measure choice accuracy.

Condition	Calls	Retained mass (%)
Llama 8B	20	92.34
Llama 70B	20	98.82
Llama 405B	20	96.67
Qwen 7B	20	95.41
Qwen 72B	20	100.00
Qwen 235B	20	100.00
Mistral 7B	20	97.03
Mistral 24B*	20	100.00
Mixtral 8x7B	20	100.00

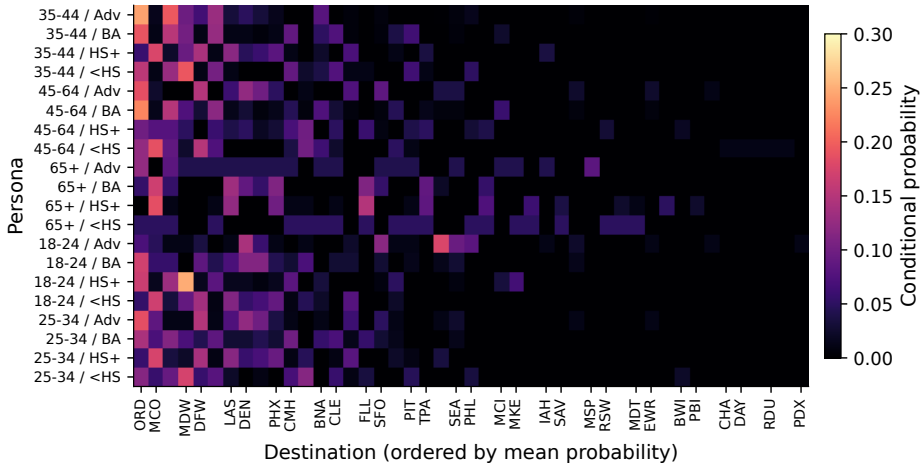


Figure 15: Catalog-validated IND persona outputs from Llama-3.3-70B. Rows identify the fixed age/education prompt combinations; columns are canonical destinations ordered by their mean. Each vector is conditional on retained mass.

K Extended Prompt Sensitivity Analysis

K.1 Checkpoint and Prompt Contrasts

The corrected means and descriptive SDs appear in Tables 3 and 18. The comparison lists are fixed: 24 age pairs use selected age groups within education, 60 education pairs span the two regional grids, and 40 regional/hub comparisons use specified persona endpoints. These lists share outputs. The larger two Llama differences exceed the reported prompt-effect means, but the 70B/405B mean is close to the age mean. Different contrasts have no common intervention-strength scale, and one call per persona cell cannot estimate all prompt-by-checkpoint interactions. We do not interpret their ordering as a universal dominance result or a population-level confidence comparison.

K.2 Matched Repeated-Wording Comparison

The historical 70B and 405B IND convergence notebooks and the 70B paraphrase notebook reconstruct the same baseline user-message bytes. They use the same 300-airport file and order, age 25–34 with a bachelor degree, IND departure, temperature 0.2, 1,024-token limit, and response schema. The 405B notebook's larger budget applies to its DeepSeek branch, not the Llama condition. These are source-code reconstructions: per-request payloads, serving revisions, and collection timestamps are

Table 18: Descriptive JS over the specified prompt contrasts after catalog validation. SD is dispersion among comparisons; comparisons may share the same response and are not independent trials. Age, education, and context involve 24, 40, and 60 distinct response endpoints, respectively.

Contrast	Mean JS	SD	Pairs
Persona presence	0.124	0.085	300
Age within education	0.330	0.133	24
Education within age	0.258	0.137	60
Regional/hub context	0.299	0.141	40

Table 19: Repeated-wording archive at 70B. Each condition has 15 recorded calls with positive retained mass. Two B responses require complete-object recovery. Retention is the mean fraction of selected parseable numeric mass retained before within-call normalization.

Wording	Recorded / valid	Complete arrays	Retained mass (%)
Same wording	15 / 15	15	86.4
A	15 / 15	15	84.2
B	15 / 15	13	81.6
C	15 / 15	15	99.9
D	15 / 15	15	96.4

unavailable. The models differ in training release (Llama 3.3 versus 3.1), so matched documented requests do not identify a causal parameter-count effect. The reconstructed catalog retains the literal `nan` state label where the historical state map lacks an entry; it is not retrospectively repaired.

The wording archive contains five 15-call pools, with no repeated raw texts among those 75 calls and the 40 primary checkpoint calls. Its baseline question is the original question in Appendix C. The four replacements, with all other documented prompt content held fixed, are:

A. Based on your profile, which destinations would you travel to? Provide probability estimates.

B. From the available airports, where would you plan to fly? Estimate the likelihood for each.

C. Select your preferred flight destinations and rate how likely you are to visit each.

D. Which airports would you consider flying to? Assign probabilities to your choices.

In wording B, responses 2 and 13 contain the invalid numeric literal `0019`. Recovery omits the malformed object without guessing its intended weight; its missing mass is unknown. The complete-array sensitivity excludes both responses.

Let A denote the primary 70B pool, C the 405B pool, and P one 70B wording pool. Each draw forms disjoint five-call means A_1, A_2, C_1, C_2, and P_1, P_2. We report

$$\Delta_{\mathrm{JS}} = \mathrm{JS}(A_1, C_1) - \mathrm{JS}(A_1, P_1). \tag{6}$$

The same A_1 occurs in both terms and every contrast is recomputed within the draw. A secondary contrast subtracts the appropriate mean within-condition JS from each between-condition term before differencing. The full CSV records both forms and their components. The primary estimates in Table 4 compare intervention distances at a common aggregation budget; they are not a population-level difference of independent sample means.

For the practical comparison, let $S_X = M_5(X_1)$ and define coverage $c(X, S) = \sum_{y \in S} X(y)$. Checkpoint loss is $c(A_2, S_A) - c(A_2, S_C)$, and wording loss is $c(A_2, S_A) - c(A_2, S_P)$. Their paired difference is

$$\Delta_{\mathrm{cov}} = c(A_2, S_P) - c(A_2, S_C). \tag{7}$$

Thus both substitutions are evaluated on identical target calls. This directed diagnostic differs from the bidirectional average in Appendix I.3. We make 1,000 joint draws with stable seed streams derived from 20260921. Ranges are marginal 2.5–97.5% quantiles conditional on the archived pools; they are neither future-call confidence intervals nor multiplicity-adjusted significance tests.

Table 20: All four wording variants against both baseline batches (primary and second original-wording batch) and three processing policies. Values are means with 95% resampled ranges. Positive values indicate a larger model change or loss. The excluded-mass policy keeps one nonselectable residual coordinate for out-of-catalog mass; its JS and coverage therefore concern a different, unconditioned numeric output. Rows are dependent sensitivity analyses, not independent tests.

Baseline	Policy	Wording	Δ_{JS} [range]	Δ_{cov}, pp [range]
Primary	Recovered	A	0.074 [-0.077, 0.161]	4.2 [-20.1, 25.4]
Primary	Recovered	B	0.075 [-0.056, 0.161]	2.8 [-20.0, 24.3]
Primary	Recovered	C	-0.008 [-0.206, 0.117]	1.5 [-20.0, 22.5]
Primary	Recovered	D	0.139 [0.087, 0.187]	17.5 [4.9, 31.7]
Primary	Arrays only	A	0.076 [-0.074, 0.164]	2.9 [-20.0, 24.9]
Primary	Arrays only	B	0.090 [-0.017, 0.165]	2.2 [-20.1, 23.8]
Primary	Arrays only	C	-0.004 [-0.207, 0.120]	0.3 [-20.2, 21.6]
Primary	Arrays only	D	0.139 [0.093, 0.185]	16.4 [4.3, 30.7]
Primary	Excluded mass	A	0.040 [-0.099, 0.135]	3.1 [-20.1, 23.4]
Primary	Excluded mass	B	0.032 [-0.106, 0.138]	2.3 [-20.0, 23.3]
Primary	Excluded mass	C	-0.008 [-0.205, 0.110]	0.5 [-20.0, 21.5]
Primary	Excluded mass	D	0.133 [0.078, 0.184]	16.8 [3.9, 30.7]
Second	Recovered	A	0.162 [0.042, 0.244]	29.8 [7.5, 55.3]
Second	Recovered	B	0.166 [0.063, 0.245]	30.1 [6.7, 56.1]
Second	Recovered	C	0.126 [-0.049, 0.228]	29.6 [2.3, 57.0]
Second	Recovered	D	0.176 [0.120, 0.230]	24.1 [5.8, 38.4]
Second	Arrays only	A	0.164 [0.050, 0.244]	27.5 [0.4, 56.1]
Second	Arrays only	B	0.173 [0.087, 0.252]	28.0 [-1.4, 55.7]
Second	Arrays only	C	0.129 [-0.042, 0.238]	27.2 [-0.8, 55.1]
Second	Arrays only	D	0.175 [0.118, 0.231]	23.1 [6.0, 39.8]
Second	Excluded mass	A	0.165 [0.061, 0.236]	24.1 [2.2, 46.6]
Second	Excluded mass	B	0.171 [0.066, 0.246]	24.3 [1.9, 47.1]
Second	Excluded mass	C	0.083 [-0.080, 0.178]	23.7 [0.7, 47.6]
Second	Excluded mass	D	0.159 [0.101, 0.211]	20.3 [5.0, 32.4]

An original-wording pool from the paraphrase archive provides a repeat-archive comparison. We also reverse the anchor: use that separate 70B pool as A and retain the primary 70B pool as the identical-wording comparator. This exposes sensitivity to the selected reference archive without identifying its cause as temporal drift. We report all four wordings under both anchors, exclude object/line recoveries as one sensitivity, and retain out-of-catalog mass as a single extra coordinate as another. The extra coordinate is never eligible for a destination shortlist. The latter check measures unconditional emitted mass rather than interpreting unresolved names as real destinations.

Under the primary baseline batch, wording C is comparable to the model change and only D has positive ranges for both JS and shortlist-loss differences. Several results change with the baseline batch; complete-array and excluded-mass checks are reported rather than selected by effect size. A model-versus-wording ranking therefore needs a specified baseline batch and processing rule. The 70B wordings are not repeated at 405B, so model-by-wording interactions and generalization to other contexts are open questions that a fully crossed panel of fresh, interleaved calls would test.

K.3 Seasonality Perturbation Details

The historical seasonality test appends "You are planning travel for summer (June–August)." or "You are planning travel for winter (December–February)." and asks for travel plans "for this season", using a fixed neutral persona and the baseline model. Table 21 preserves the four recorded distance summaries as ancillary descriptive results. These small, selected-context observations do not establish a general geographic buffering mechanism or a matched ranking against every prompt intervention.

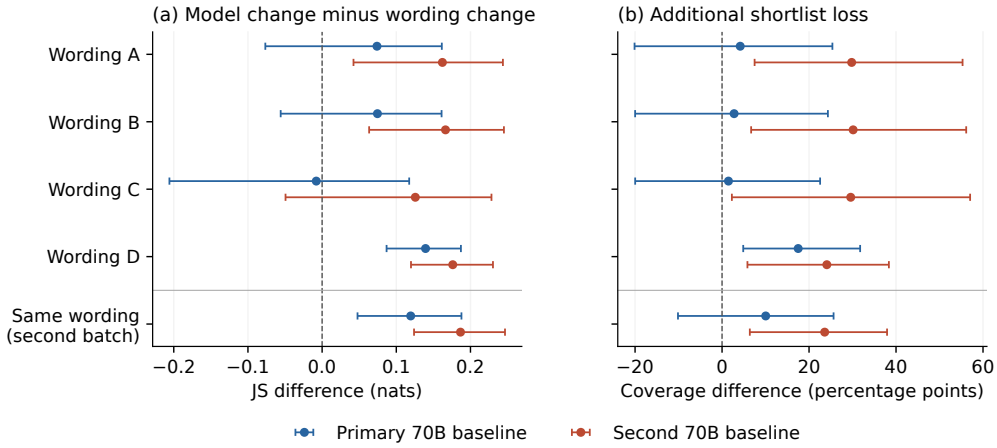


Figure 16: Model change minus wording change, measured against the primary or the second original-wording 70B baseline batch. Points are means and bars are 95% resampled ranges; all four wordings and the same-wording comparison are shown. Panel (b) scores both shortlists on common held-out target calls. The ranking depends on the baseline batch, so no universal ordering is implied. All responses are saved calls, not fresh validation calls.

Table 21: Historical four-context seasonality comparison. Entries are JS distances under the original processing; only descriptive point estimates are retained. These ancillary values are distinct from the corrected catalog-validated main analyses.

Airport	**JS Dist.**
FWA (Fort Wayne)	0.359
IND (Indianapolis)	0.334
ATL (Atlanta)	0.285
ORD (Chicago)	0.264
Mean	**0.311**

L Temperature Sensitivity Analysis: Extended Results

The original design specifies ten temperatures ($T = 0.1$–1.0), three origins (IND, LAX, BOS), two personas, and ten calls per combination (600 scheduled calls). Figure 17 displays all 60 cached final-batch summaries; Table 22 keeps the original cross-setting summaries. Both use the original output processing.

The current raw-call CSVs contain 600 credit-limit errors and no responses; the nonzero metrics cache and saved notebook summaries survive separately. The figure can therefore be rebuilt from the cache, but its underlying calls cannot be re-audited.

L.1 Interpretation

Across these cached batches, the five highest-weight destinations carry 34.7–66.0% of generated mass, while the batches contain 23–61 named destinations. Compact concentration thus coexists with a broader output vocabulary throughout the recorded sweep. The two panels make this distinction explicit and motivate checking concentration and breadth separately when changing decoding settings. Their values vary across contexts and temperatures, without a universal monotonic trend.

Top-five mass is measured on the same pooled vector used to select the five outcomes; it is distinct from the disjoint-call transfer diagnostic in Section 5.3. Successive-prefix Jaccards share calls and are descriptive. These summaries do not establish an optimal temperature, temperature invariance, a semantic mode boundary, or future-call performance.

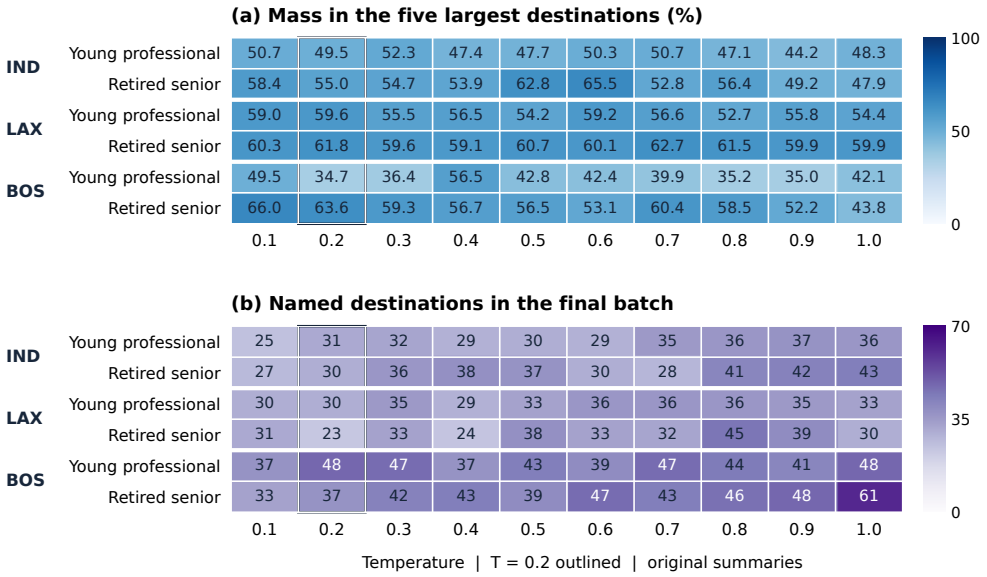


Figure 17: Concentration and breadth across the temperature sweep. All six airport–persona settings and ten temperatures are shown. (a) Mass in the five largest outcomes of each cached final batch. (b) Named outcomes under the original processing. The outlined column marks the paper's $T = 0.2$ setting, not an estimated optimum. Values are cached descriptive summaries; raw-call reconstruction is unavailable.

Table 22: Temperature-sweep summaries over six settings under the original processing. Unique, maximum probability, entropy, and top-five mass summarize final pooled vectors; Jaccard entries summarize successive prefixes. They are descriptive quantities and have different observational units.

T	Unique	Max P	Entropy	Top-5	Jac@5	Jac@10
0.1	30.5	0.176	3.93	0.573	0.924	0.885
0.2	**33.2**	**0.164**	**4.07**	**0.540**	**0.929**	**0.911**
0.3	37.5	0.160	4.16	0.530	0.924	0.853
0.4	33.3	0.172	4.07	0.550	0.883	0.886
0.5	36.7	0.169	4.14	0.541	0.901	0.890
0.6	35.7	0.175	4.08	0.551	0.873	0.886
0.7	36.8	0.172	4.13	0.539	0.853	0.887
0.8	41.3	0.165	4.24	0.519	0.766	0.865
0.9	40.3	0.151	4.35	0.494	0.834	0.814
1.0	41.8	0.149	4.34	0.494	0.813	0.901

M Temperature-Zero Records and Their Limits

We apply the same parsing, catalog mapping, and equal-call pooling to the archived temperature-zero repeated-call records. Table 23 gives condition-level summaries instead of claiming that all findings are reproduced. Their stability indices remain retrospective and order-dependent.

M.1 Archive Identity Check

The file labeled as temperature-zero 8B convergence contains exactly the same 20 response texts as the primary 8B convergence file. The notebook reads the earlier raw-response file and reprocesses it. Consequently, identical support or concentration in these two files is not independent evidence of temperature robustness. We exclude this cell from that inference while preserving the files and identity check. The compared 70B and 405B archives do contain different response texts.

Table 23: Corrected descriptive comparison of primary and separately archived temperature-zero conditions, with 20 usable calls per listed cell. Columns report positive named support and the last top-five change in recorded order. The 8B convergence archive is excluded as an independent temperature-zero replication because it reuses all 20 primary response texts. * Adapted historical condition.

Condition	Support .2	Support 0	R_{last} .2	R_{last} 0
Llama 70B IND	31	26	7	1
Llama 70B FWA	41	32	1	3
Llama 70B ATL	42	39	4	2
Llama 70B ORD	31	32	18	7
Llama 405B IND	49	23	1	2
Qwen 7B	27	11	13	1
Qwen 72B	33	28	1	20
Qwen 235B	50	38	1	2
Mistral 7B	98	118	20	1
Mistral 24B*	19	27	16	18
Mixtral 8x7B	27	18	16	9
Yelp mapped names	238	233	14	13
Amazon mapped titles	86	29	18	14

Changing an API temperature setting does not by itself identify the source of residual output variation. These records do not establish GPU arithmetic or a particular training procedure as its cause. A prospective decoding comparison would freeze all other settings, record request metadata, and collect distinct condition-specific calls.

N Statistical Robustness: Extended Results

This appendix provides detailed results from the ancillary robustness checks summarized in the main text.

N.1 Three Robustness Sources

The historical ancillary analysis considers three separately varied sources. *Source A* (permutation) uses 15 random orderings of the airport list with 15 iterations each (225 calls). *Source B* (paraphrase) uses 5 semantically equivalent prompt variants with 15 iterations each (75 calls). *Source C* (origins) uses 3 departure airports (IND, LAX, BOS) with 15 iterations each (45 calls).

N.2 Coefficient of Variation by Source

Table 24 preserves coefficients of variation under the three historical protocols. These descriptive ratios do not jointly partition variance and are sensitive to their metric means.

Table 24: Historical descriptive coefficients of variation under three ancillary protocols. The values do not supply a universal reproducibility threshold or validate future-call stability.

Metric	CV Perm.	CV Para.	CV Origins
Unique Dest.	13.2%	14.6%	20.5%
Max Prob.	10.1%	13.6%	13.3%
Entropy	5.1%	4.6%	5.4%
Top-5 Mass	10.1%	9.1%	10.0%

N.3 Descriptive One-Way ICCs

Table 25 preserves the historical one-way ICC(1,1) estimates under separate airport, persona, and checkpoint groupings. They are not a jointly identified variance decomposition and do not isolate causal contributions. A negative sample estimate can occur when estimated within-group variation exceeds between-group variation; it does not prove the absence of a factor effect. These ancillary values are not used as an independent significance test of the corrected checkpoint comparisons.

Table 25: Historical ICC(1,1) estimates under separate groupings. Negative estimates and differences among columns are descriptive; these are not additive causal variance shares. The checkpoint label combines the recorded model and protocol conditions.

Metric	Airport	Persona	Model
Max Prob.	−0.02	0.16	0.20
Entropy	−0.02	−0.40	0.17
Top-5 Mass	0.08	−0.07	−0.01
Support	0.07	−0.31	**0.46**

N.4 Consensus Destinations

Top destinations appearing consistently across 15 input orderings: ORD (15/15 top-5), DFW (14/15), MCO (12/15), LAX (11/15), DEN (10/15).

O Further Background and Evaluation Details

O.1 Extended Related Work

O.1.1 LLMs for Simulation and Prediction

Section 2.1 reviews LLM simulation and demographic steering. Multi-agent LLM simulations have also been compared with observed social-media diffusion cascades (Huang, 2026), and sociodemographic personas added to logic and math problems reveal implicit bias in LLM answers (Yin & Huang, 2025). That literature motivates distinguishing fidelity to human responses from reproducibility of generated simulation inputs. Models can encode geographical regularities, support planning, and represent aspects of their data-generating agents (Gurnee & Tegmark, 2024; Hao et al., 2023; Andreas, 2022); these abilities do not identify an elicited vector with an internal world model (Ha & Schmidhuber, 2018).

O.1.2 Probabilistic Calibration and Uncertainty

Research on neural network calibration has long examined whether model confidence aligns with accuracy (Guo et al., 2017). In the context of LLMs, Jiang et al. (2021) investigated calibration for question answering, while Kadavath et al. (2022) showed that language models can often assess whether they know the answer to a question. More recent work has developed methods for quantifying semantic uncertainty in generated text (Kuhn et al., 2023) and explored whether LLMs can verbally express their uncertainty (Xiong et al., 2024). These studies use confidence and semantic uncertainty to assess answer reliability. We instead examine reproducibility and structural changes in explicitly elicited probability distributions over a fixed choice set.

O.1.3 Prompt Sensitivity and Robustness

Studies have documented that LLM outputs can be highly sensitive to prompt phrasing, including the order of few-shot examples (Lu et al., 2022) and biases toward majority labels or recent examples (Zhao et al., 2021). Perez et al. (2021) demonstrated that prompt selection can substantially affect few-shot performance, while Webson & Pavlick (2022) questioned whether models truly understand prompt semantics or rely on superficial patterns. Chain-of-thought exemplars raise the accuracy of large models on arithmetic, commonsense, and symbolic reasoning benchmarks (Wei et al., 2022), again showing that results depend on prompt construction.

O.1.4 Mode Collapse and Numeric Calibration

A related line of work examines *mode collapse*, where generation diversity degrades during fine-tuning (Janus, 2022; O'Mahony et al., 2024). Zhou et al. (2024) showed that LLMs are reluctant to express uncertainty, relevant to our verbalized elicitation design. Lovering et al. (2025) demonstrated poor numeric calibration in language models. Lum et al. (2025) advocated evaluation beyond isolated test cases, a perspective aligned with our distribution-level framework.

O.2 Metric Conventions and Interpretation

We use distributional consistency to mean reproducibility of the specified output procedure, distinct from robustness to distribution shift (Liu et al., 2023). Corrected main comparisons use natural-log JS, deterministic top-set ties, validated unique coordinates, and equal weights on normalized calls. JS distance is identified explicitly. Named support, active support above 0.001, conditional regional divergence, and additive original-mass contribution have distinct definitions (Appendix B). Historical pilot, seasonality, temperature, and grouping summaries are labeled separately and retain their stated conventions. No inference about an internal belief or human distribution follows solely from these metrics.

O.3 Resampling and Numerical Precision

For each corrected call pool, generate 1,000 independent random permutations using a fixed seed derived from 20260921 and a stable series identifier. Each prefix of length R contains distinct calls. Different permutations may produce the same subset; this is Monte Carlo sampling of subsets, not exhaustive enumeration. Form the normalized-call average first, then compute support, concentration, entropy, distance, or overlap. Plot the conditional mean and the 2.5–97.5 percentile range. At the full pool every selection is identical. These ranges describe sensitivity within the observed pool and make no claim about sampling uncertainty over future calls or other contexts.

For a model comparison at equal numbers of calls, each draw selects two non-overlapping groups of five calls within each model. Between-model JS compares the first groups; within-model JS averages each model's comparison of its two groups. The difference and its 95% resampled range summarize the collected batches. This matches call count and outcome processing, not unrecorded provider state or all original generation settings.

For fixed-persona and prompt comparisons, means and population-form descriptive SDs summarize the listed values. They are not confidence intervals and do not count shared-persona pairs as independent samples. For tail attribution, average pair contributions within context and then equally across contexts; keep the observed adjacent-pair structure and report context-level results in the package. Three decimal places in JS tables aid reproducibility and do not imply corresponding inferential precision.

P Exploratory Stationary Sparse-Vector Control

For each 70B travel context, the first ten corrected calls calibrate a stationary vector generator and the remaining ten supply an exploratory evaluation sequence. The mean parameter is $(1-\lambda)\bar{p}_{\text{train}} + \lambda/300$, for both $\lambda = 0.01$ and 0.05. Draw $q \sim \text{Dirichlet}(\alpha\mu)$, sample a named-output count from the training calls, retain that many largest coordinates, and renormalize. The uniform component permits catalog outcomes absent from training; the sparse output rule allows cumulative named support to grow.

Choose α from 36 logarithmically spaced values from 0.3 to 5,000 by matching mean training pairwise JS, using 256 simulated pairs per candidate. Generate 500 evaluation sequences of length ten under the selected parameters. The package records the entire fit grid, residual mismatch, simulated sequences' metrics, and both smoothing settings. This one-parameter dispersion fit is not guaranteed to match every other feature.

Figure 4 in the main text displays these comparisons.

Table 26: Exploratory evaluation-half observations and 95% simulated reference ranges. Growth is the increase in positive named support from the first to the tenth evaluation call. Last change is indexed within that evaluation half. Both fixed uniform-mass settings are reported; these ranges condition on fitted parameters.

Context	λ	Growth obs.	Null range	Last change	Null range
IND	0.01	10	[9.0, 21.5]	5	[2, 10]
IND	0.05	10	[18.0, 35.0]	5	[1, 10]
FWA	0.01	13	[10.0, 22.0]	2	[1, 10]
FWA	0.05	13	[19.0, 34.0]	2	[1, 10]
ATL	0.01	12	[8.5, 29.0]	1	[1, 10]
ATL	0.05	12	[13.5, 40.5]	1	[1, 10]
ORD	0.01	10	[9.0, 24.0]	9	[1, 10]
ORD	0.05	10	[15.0, 34.0]	9	[1, 10]

At 1% uniform mass, support growth and top-five persistence alone are compatible with a stationary sparse-vector explanation; at 5%, observed growth falls below the simulated range in all four contexts. Other discrepancies remain, including the low evaluation-half pairwise divergence for FWA. The archive and the question were already inspected before this split, so the exercise is exploratory. Reference ranges condition on fitted parameters, omit fitting uncertainty, and are not formal multiple-testing-adjusted rejection thresholds. Failure or success of this particular control does not identify a unique internal LLM mechanism. Fresh validation would require a frozen protocol and independently collected evaluation calls.

Q Reproduction and Revision Scope

The supplied entry point rebuilds the raw-output audit, corrected comparisons, sparse-vector control, and central figures. Audit records identify input hashes, source rows, parser decisions, duplicate keys, retained mass, mappings, configuration evidence, and reused responses. The portable numerical package includes CSV inputs and scripts; full configuration extraction also requires the project notebooks. Figures 6 and 14 have a separate rendering package using frozen outputs of these analyses. Figures 1 and 3 use a presentation renderer of the same saved resampling summaries. These commands make no new model calls.

Central figures and main comparisons use the corrected processing. Figure 17 redraws the original temperature cache, whose underlying responses are unavailable. The pilot, seasonality points, temperature summaries, and CV/ICC tables keep the original processing and do not establish central uncertainty or universal robustness. Saved batches and missing metadata limit causal and prospective inference.