LLMs as Implicit Probabilistic World Models: Distributional Stability, Convergence, and Scale Discordance

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used to generate structured predictions and simulate human decision-making, yet the distributional behavior of their outputs remains underexplored. We treat LLMs as implicit probabilistic world models and introduce a systematic framework for evaluating their output distributions under repeated sampling, prompt perturbations, and model scaling. Across models and domains, we observe a consistent *modes-first, tails-later* dynamics: high-probability outcomes stabilize within a few iterations, while low-probability alternatives continue to evolve, with post-convergence variability dominated by the distribution tail rather than shifts in dominant predictions. We further find pronounced *scale discordance*, with low agreement between models of different sizes, suggesting that scaling alters underlying probabilistic representations rather than merely refining them. Prompt-based perturbations, including persona conditioning and seasonal context, induce systematic but bounded distributional shifts that are smaller in magnitude than differences across model scales. Experiments on air travel destination choice, with cross-domain verification in restaurant and product selection, demonstrate the robustness of these behaviors. Together, our results provide a distribution-level perspective on LLM stability and sensitivity, with implications for probabilistic simulation and decision-support applications.

1 Introduction

A *world model* is an internal representation that enables an agent to simulate, predict, and reason about its environment (Ha and Schmidhuber, 2018). Recent evidence suggests that Large Language Models (LLMs) develop such representations implicitly through training on text: they encode spatial and temporal structure (Gurnee and Tegmark, 2024), support planning through internal simulation (Hao et al., 2023), and can be understood as models of the agents that produced their training data (Li et al., 2023). This has motivated a growing body of work treating LLMs as *implicit probabilistic world models*: systems that encode beliefs about real-world phenomena and can generate structured predictions that approximate human judgment.

Building on this view, researchers have employed LLMs for probabilistic applications across diverse domains. Argyle et al. (2023) demonstrated that LLMs can simulate survey responses that replicate demographic subgroup patterns. Park et al. (2023) used LLM-based agents to simulate human social behavior in interactive environments. Santurkar et al. (2023) examined whose opinions LLMs reflect when prompted with demographic information. These applications share a common premise that LLM outputs can be treated as samples from a meaningful probability distribution over possible responses, conditioned on contextual prompts.

Despite this growing reliance on LLMs as probabilistic generators, the *distributional behavior* of their outputs remains underexplored. Most existing evaluations focus on factual accuracy of individual outputs (Lin et al., 2022), calibration of confidence estimates (Kadavath et al., 2022), or benchmark performance on standardized tasks (Hendrycks et al., 2021), leaving unanswered fundamental questions about how LLM-generated probability distributions behave under repeated sampling. Prior work has documented prompt sensitivity at the output level (Lu et al., 2022), but systematic analysis of full distributional behavior remains underexplored. Specifically: Do LLM-generated distributions converge under repeated sampling? Which parts of the distribution stabilize first? How do models with different scales compare in their distributional priors? And how do prompt perturbations reshape entire distributions rather than individual outputs?

These questions constitute a critical gap in our understanding of LLMs as probabilistic systems.

Many downstream applications (e.g., simulation, synthetic data generation, and decision support) implicitly assume that repeated LLM outputs approximate samples from a coherent and stable probability distribution. In this work, we address this gap by empirically characterizing the structure, stability, and sensitivity of LLM-generated output distributions. Rather than evaluating whether an LLM produces a correct answer, we study how the entire output distribution evolves under repeated sampling, controlled prompt perturbations, and changes in model scale.

Across models, prompts, and domains, LLM-generated distributions exhibit a consistent *modes-first, tails-later* convergence pattern: high-probability outcomes stabilize within a few iterations, while low-probability alternatives continue to evolve, with post-convergence variability dominated by the distribution tail. We further observe pronounced *scale discordance*, with low agreement between models of different sizes under identical prompts, suggesting that scaling alters underlying probabilistic representations. Our investigation proposes a new perspective on LLM evaluation. We argue that understanding distributional stability, sensitivity, and relativity is essential for responsibly using LLMs in settings that rely on probabilistic reasoning or simulation applications.

2 Related Work

2.1 LLMs for Simulation and Prediction

Recent work has explored LLMs as simulators of human behavior (Park et al., 2023; Argyle et al., 2023). Studies have shown that LLMs can reproduce survey response patterns and exhibit demographic-consistent preferences when prompted with persona information (Santurkar et al., 2023). However, the distributional properties of LLM outputs (stability, convergence, and model sensitivity) remain largely uncharacterized.

2.2 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). While these studies focus on calibration of single-point predictions or confidence scores, our work analyzes full probability distributions over outcome spaces, providing a more complete characterization of LLM distributional behavior.

2.3 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 and Pavlick (2022) questioned whether models truly understand prompt semantics or rely on superficial patterns. Prompting strategies such as chain-of-thought (Wei et al., 2022) have been shown to elicit more reliable reasoning, yet the sensitivity of outputs to prompt variations remains a fundamental challenge. Our work extends this analysis to the distributional level, measuring how prompt perturbations systematically reshape entire output distributions rather than individual predictions.

3 Method

To enable this analysis, we adopt experimental settings in which LLMs are prompted to generate explicit probability distributions over large discrete choice sets. This formulation allows us to repeatedly query models under identical conditions and directly compare distributions across iterations. Our primary experiments focus on air travel destination choice, where we prompt LLMs to generate probability distributions over 300 U.S. airports given departure context and traveler persona information. While we instantiate our experiments in domains involving human preferences (travel demand, restaurant choice, and product selection), the evaluation framework and metrics we introduce are domain-agnostic and applicable to any task involving structured probabilistic generation.

3.1 Problem Formulation

We treat LLMs as *implicit probabilistic world models* that generate distributions over discrete outcome spaces conditioned on natural language prompts. Formally, given a prompt x describing a contextual scenario (e.g., departure airport, traveler persona), an LLM induces a probability dis-

tribution over outcomes $y \in \mathcal{Y}$: $p(y \mid x)$. In our setting, $\mathcal{Y}$ corresponds to a finite set of candidate destinations. Rather than treating LLM outputs as single-point predictions, we explicitly analyze the distributional structure, stability, and sensitivity of $p(y \mid x)$ under repeated sampling and controlled prompt perturbations.

Our analysis focuses on three fundamental questions: (1) *Distributional Stability*: how consistent are LLM-generated distributions across repeated sampling under identical prompts? (2) *Mode–Tail Dynamics*: how do high-probability outcomes ("modes") and low-probability outcomes ("tails") evolve over repeated generations? (3) *Sensitivity and Scale Discordance*: how do distributions change across prompt perturbations (e.g., persona conditioning), and do models of different scales exhibit discordant priors?

3.2 Iterative Sampling Protocol

For each prompt x, we perform repeated sampling runs indexed by $r = 1, \ldots, R$, producing a sequence of distributions:

$$\{p^{(r)}(y \mid x)\}_{r=1}^{R} \tag{1}$$

Each run corresponds to an independent model invocation with identical generation settings (e.g., temperature) unless otherwise specified. This protocol allows us to empirically examine whether LLM outputs behave as independent noisy samples or as structured stochastic realizations of a stable latent distribution. Following our empirical observations, we set $R = 20$, which is sufficient to observe convergence behavior across all model sizes.

3.3 Mode–Tail Decomposition

To characterize internal distributional dynamics, we decompose each distribution into high-probability modes and low-probability tails. **Mode Definition.** For a fixed k, we define the mode set: $\mathcal{M}_k(p) =$ Top-k outcomes ranked by $p(y \mid x)$. **Tail Definition.** The tail is defined as the complement: $\mathcal{T}_k(p) = \mathcal{Y} \setminus \mathcal{M}_k(p)$. This decomposition enables us to separately analyze the stability of dominant semantic beliefs (modes) and exploratory variability in low-probability outcomes (tails). In our experiments, we set $k = 5$ for mode–tail decomposition and agreement metrics, as this threshold effectively captures the primary semantic beliefs while remaining interpretable across domains. We also report $k = 10$ results in our tables; the findings are consistent across both thresholds, indicating robustness to the choice of k.

3.4 Evaluation Metrics

We employ a suite of complementary metrics to characterize distributional behavior across three dimensions: stability, tail dynamics, and cross-condition comparison.

Distributional Stability. We measure overall distributional similarity using Jensen–Shannon (JS) divergence (Lin, 1991), which is symmetric and bounded in $[0, 1]$. Mode stability is quantified via Jaccard@k, measuring the overlap of top-k outcome sets between runs. Rank consistency is assessed using Spearman correlation over mode orderings. **Tail Dynamics.** To characterize behavior in low-probability regions, we track support growth (number of outcomes above threshold ϵ), Shannon entropy (distributional uncertainty), and tail mass (probability allocated outside the top-k modes). We further decompose JS divergence into mode and tail contributions to localize sources of instability. **Cross-Condition Analysis.** For prompt sensitivity, we compute JS divergence between distributions under different prompt conditions (persona, demographics, location).

For scale discordance analysis, we compare distributions across model sizes using both JS divergence and agreement@k (proportion of contexts with identical top-k sets). The joint interpretation of these metrics distinguishes scale refinement (consistent beliefs across sizes) from scale discordance (fundamentally different priors). Complete mathematical definitions are provided in Appendix K.

4 Experiments

We employ structured prompts providing LLMs with a traveler persona (age and education), a departure airport, 300 possible destination airports, and instructions to output preferences as probabilities. Note that these airports are in the contiguous US considered in the work (Kim et al., 2025). Personas are defined across 5 age groups and 4 education levels, yielding 20 unique combinations. Experiments primarily use Llama-3.3-70B via the Together AI API[1], with extension experiments spanning 9 models across three families (Llama, Qwen (Qwen Team, 2024), Mistral (Jiang et al., 2023,

[1] https://api.together.ai/

2024)) at small, medium, and large scales. All experiments use the temperature $T = 0.2$. See Appendix L for complete configuration details.

Our experiments examine three aspects of LLM distributional behavior: (1) stability and convergence under repeated sampling across 20 iterations; (2) sensitivity to prompt perturbations, including persona demographics, season, and geographic context; and (3) cross-scale agreement and cross-domain generalization using restaurant choice (Yelp) and product shopping (Amazon) datasets. See Appendix I for dataset descriptions.

To quantify convergence, we define two metrics. Let $|\mathcal{S}_r|$ denote the support size (number of unique outcomes with non-zero probability) at iteration r. The *growth percentage* measures the relative increase in support: Growth $= (|\mathcal{S}_{20}| - |\mathcal{S}_1|)/|\mathcal{S}_1| \times 100\%$. The *stabilization point* is the first iteration r^* such that the top-k mode set remains unchanged for all subsequent iterations: $\mathcal{M}_k(p^{(r^*)}) = \mathcal{M}_k(p^{(r)})$ for all $r > r^*$.

To analyze the distributional variability, we rely on the JS divergence:

$$\mathrm{JS}(P, Q) = \frac{1}{2}\big[D_{\mathrm{KL}}(P\|M) + D_{\mathrm{KL}}(Q\|M)\big], \quad (2)$$

where P and Q are probability distributions, $M = (P + Q)/2$ is a mixture distribution of P and Q, and $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler divergence. Based on this measure, we construct three metrics: (a) $\mathrm{JS}_{\mathrm{total}}$ computed using $P = p^{(r)}$ and $Q = p^{(r')}$; (b) $\mathrm{JS}_{\mathrm{mode}}$ computed using renormalized mode distributions $P = \tilde{p}^{(r)}_{\mathcal{M}}$ and $Q = \tilde{p}^{(r')}_{\mathcal{M}}$; and (c) $\mathrm{JS}_{\mathrm{tail}}$ computed analogously using renormalized tail distributions $P = \tilde{p}^{(r)}_{\mathcal{T}}$ and $Q = \tilde{p}^{(r')}_{\mathcal{T}}$. The mode and tail distributions are obtained by partitioning the original probabilities according to the mode–tail criterion and renormalizing each subset to sum to one. In all cases, the divergence is evaluated over the reference mode set defined at the later iteration $r' > r$.

These metrics provide complementary views: $\mathrm{JS}_{\mathrm{total}}$ captures aggregate distributional change, while $\mathrm{JS}_{\mathrm{mode}}$ and $\mathrm{JS}_{\mathrm{tail}}$ isolate whether variability arises from changes in core semantic beliefs (modes) or peripheral exploration (tails). Note that they do *not* form an additive decomposition (i.e., $\mathrm{JS}_{\mathrm{total}} \neq \mathrm{JS}_{\mathrm{mode}} + \mathrm{JS}_{\mathrm{tail}}$). Further details on all evaluation metrics can be found in Appendix K.

Context/Model	Iter 1	Iter 20	Growth	Stable
IND	19	32	+68.4%	Iter 3
FWA	16	42	+162.5%	Iter 5
ATL	15	43	+186.7%	Iter 3
ORD	18	34	+88.9%	Iter 3
Model Size Comparison (IND)				
Baseline (70B)	19	32	+68.4%	Iter 3
Larger (405B)	30	54	+80.0%	Iter 4
Smaller (8B)	29	100	+244.8%	Iter 17

Table 1: Convergence results for **travel demand** (US airports) showing outcome counts at iterations 1 and 20, growth percentage (relative increase in unique outcomes), and stabilization point (iteration at which top-k modes stop changing). Airports: IND (Indianapolis), FWA (Fort Wayne), ATL (Atlanta), ORD (Chicago).

Iteration	Jacc.@5	Jacc.@10	Spearman	JS Div.
$1 \to 2$	0.60	0.50	0.75	0.120
$2 \to 3$	0.72	0.62	0.82	0.085
$3 \to 4$	0.80	0.70	0.88	0.055
$5 \to 10$	0.85	0.75	0.92	0.035
$10 \to 20$	0.92	0.85	0.95	0.020

Table 2: Mode stability metrics for **travel demand** (US airports) across iterations. High Jaccard@5 (>0.90) and Spearman correlation (>0.95) after iteration 10 confirm that core distributional structure stabilizes rapidly.

5 Results

We organize our findings around five empirical claims that characterize how LLMs function as implicit probabilistic world models.

5.1 Rapid Stabilization of High-Probability Mass

Our primary finding is that LLM-generated distributions exhibit a consistent *modes-first* convergence pattern: high-probability outcomes emerge and stabilize within the first few sampling iterations, while subsequent sampling primarily affects low-probability alternatives. Across contexts and model sizes, convergence metrics in Table 1 show that top-k modes stabilize by iterations 3–5. This early stabilization is further confirmed by high agreement among dominant outcomes, with Jaccard@5 exceeding 0.9 and Spearman rank correlation exceeding 0.95 by later iterations (Table 2). Figure 1 illustrates this behavior across model scales, where both baseline and larger models rapidly stabilize their core distributional structure, and smaller models exhibit delayed but qualitatively similar convergence.

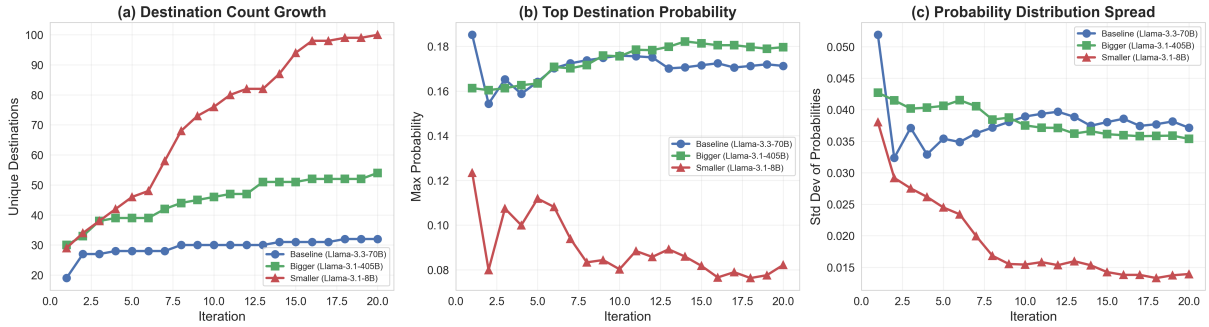

Figure 1: Convergence comparison for **travel demand** (US airports) across three LLM sizes over 20 iterations. (a) Unique outcome count growth; (b) maximum probability evolution; (c) probability distribution spread.

Phase	JS_{total}	JS_{mode}	JS_{tail}
Iter 1–3	0.121	0.087	0.204
Iter 3–5	0.072	0.025	0.189
Iter 5–10	0.045	0.011	0.169
Iter 10–20	0.028	0.004	0.152

Table 3: Mode-tail JS divergence for **travel demand** (US airports) across convergence phases. JS_{total}, JS_{mode}, and JS_{tail} are computed independently on the full distribution, renormalized mode region, and renormalized tail region, respectively; they are not additive. After early iterations, mode divergence approaches zero while tail divergence remains substantial, demonstrating that post-convergence variability arises from low-probability outcomes.

Iter.	Support	Entropy	TM@5	TM@10
1	21.75	3.888	0.435	0.285
5	20.50	3.812	0.415	0.268
10	19.75	3.756	0.402	0.255
20	19.50	3.732	0.397	0.248

Table 4: Tail dynamics metrics for **travel demand** (US airports) over iterations. Gradual decrease in support and entropy indicates convergence, while substantial remaining tail mass (∼40% at TM@5) reflects inherent distributional uncertainty beyond the dominant modes.

5.2 Post-Convergence Variability is Tail-Dominated

Having established rapid mode stabilization, we next examine the source of post-convergence variability. Our analysis reveals that nearly all remaining divergence arises from tail outcomes rather than from instability in core semantic beliefs.

Table 3 reports JS divergence computed separately over the full distribution, the renormalized mode region, and the renormalized tail region across consecutive iterations. In early iterations (1–3), mode divergence is substantial (JS_{mode} = 0.087), indicating ongoing formation of core beliefs. After convergence, however, mode divergence drops to near zero (JS_{mode} = 0.004), while tail divergence remains elevated (JS_{tail} = 0.152), showing that residual variability is driven almost entirely by low-probability outcomes rather than changes in dominant predictions. Complementing this analysis, Table 4 shows that although tail support size contracts slightly over iterations (21.75 → 19.50), a substantial fraction of probability mass persists in the tail (approximately 40% at TM@5), indicating sustained uncertainty over non-modal alternatives even after convergence.

Together, these results support a characterization of LLMs as encoding deterministic core beliefs (modes) with stochastic peripheral exploration (tails). This modes-first, tails-later dynamic has important implications for how LLM-generated distributions should be interpreted and used in downstream applications.

5.3 Model Scale Induces Distributional Discordance

Our most surprising finding concerns *scale discordance*: LLMs of different sizes exhibit extremely low distributional agreement, even when prompts and decoding parameters are held constant. This indicates that larger models do not simply refine smaller ones; rather, they encode fundamentally different latent world priors.

Table 5 presents pairwise agreement metrics across three model sizes (8B, 70B, 405B). JS divergences are substantial (0.365–0.501), agreement@5 is extremely low (0.081–0.124), and top-1 match rates range from only 5% to 30%. These values indicate near-chance agreement on which outcomes should dominate. Further analysis is in Appendix C.

Model Pair	JS Div.	Agr.@5	Top1
70B vs 405B	0.365	0.124	0.30
70B vs 8B	0.474	0.081	0.05
405B vs 8B	0.501	0.081	0.20

Table 5: Scale discordance metrics for **travel demand** (US airports) across three LLM sizes. High JS divergence and low agreement@k indicate that different model scales encode fundamentally different world priors rather than scaled refinements of similar distributions.

Family	S-M	M-L	S-L
Llama	0.474	0.365	0.501
Qwen	0.138	0.693*	0.693*
Mistral	0.693*	0.693*	0.281

Table 6: Cross-family scale discordance for **travel demand** (US airports), measured by JS divergence. *Maximum JS values indicate output format variations requiring normalization; see Appendix J for details.

Cross-Family Validation. To assess whether scale discordance is architecture-specific, we replicate our analysis across two additional model families: Qwen (Qwen Team, 2024) and Mistral (Jiang et al., 2023, 2024). This extension spans nine models covering three training origins, three size tiers, and both dense and mixture-of-experts architectures. As shown in Table 6, substantial scale discordance persists across all families, with JS divergence ranging from 0.138 to 0.281 between comparable model sizes. While the degree of inter-scale agreement varies—for example, Qwen exhibits higher correlation at smaller scales than Llama—no family eliminates the phenomenon. The asterisked JS values (0.693) indicate cases of complete distributional non-overlap. In addition, all families exhibit the same modes-first convergence behavior: Qwen models stabilize by iterations 1–2 and Mistral models by iterations 2–4, consistent with the convergence dynamics observed in Llama, confirming that this pattern is architecture-independent.

5.4 Prompt Perturbations are Bounded and Structured

We next examine the sensitivity of LLM-generated distributions to controlled prompt perturbations. Table 7 reports JS divergence with bootstrap confidence intervals for different perturbation types. Persona conditioning induces bounded shifts (JS = 0.134), while demographic attributes produce larger effects, including age (JS = 0.330) and education (JS = 0.261). Context specification (regional vs. hub) also substantially reshapes distributions (JS = 0.310). Figure 4 (Appendix B) illustrates how persona information systematically redistributes probability mass across outcome types.

Perturbation Type	Mean JS	95% CI	N
Persona (with vs without)	0.134	[0.124, 0.144]	298
Age (within education)	0.330	[0.276, 0.385]	24
Education (within age)	0.261	[0.229, 0.294]	60
Context (regional vs hub)	0.310	[0.266, 0.354]	40

Table 7: Prompt sensitivity effect sizes for **travel demand** (US airports) with 95% bootstrap confidence intervals. N indicates the number of pairwise distribution comparisons: persona comparisons across 298 departure contexts; age comparisons within each education level; education comparisons within each age group; and context comparisons across 20 personas. Age perturbations produce the largest distributional shifts among prompt-level factors, followed by context and education effects.

Airport	JS Div.	95% CI
FWA (Fort Wayne)	0.359	[0.316, 0.412]
IND (Indianapolis)	0.334	[0.298, 0.375]
ATL (Atlanta)	0.285	[0.255, 0.317]
ORD (Chicago)	0.264	[0.230, 0.306]
Mean	**0.311**	—

Table 8: Seasonality perturbation effect sizes for **travel demand** (US airports). JS divergence between summer (June–August) and winter (December–February) distributions. Smaller airports show larger seasonal effects than major hubs.

Seasonality Perturbation Test. To further contextualize perturbation strength, we compare summer and winter travel settings using a minimal prompt modification. As shown in Table 8, the average seasonal effect ($\bar{\text{JS}} = 0.311$) is comparable to age-based perturbations and larger than education-based ones. Seasonal sensitivity varies by departure airport, with large hubs (ATL, ORD) exhibiting smaller shifts than smaller airports (FWA, IND), suggesting that strong geographic anchors partially attenuate temporal context effects.

5.5 Distributional Dynamics Generalize Across Domains

To assess whether our findings generalize beyond the primary domain, we replicate our convergence analysis in two additional domains using real-world datasets: restaurant choice (Yelp) and online product shopping (Amazon).

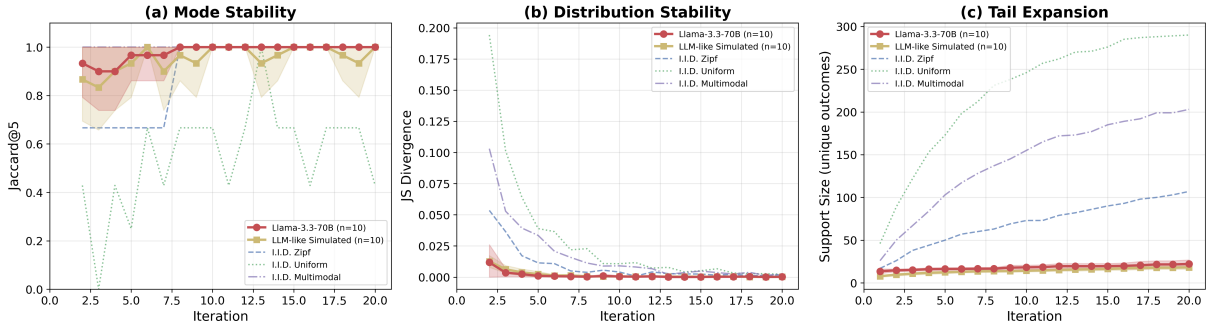

Figure 2: Control experiment comparing real LLM generation against baselines for **travel demand**. Red with shaded band: Llama-3.3-70B (n=10 runs, mean ± std). Gold with shaded band: LLM-like staged simulation (n=10 parameter variants). Dashed lines: i.i.d. sampling from fixed distributions (Zipf, Uniform, Multimodal). The real LLM exhibits distinct convergence dynamics that cannot be fully replicated by either i.i.d. sampling or deliberately staged simulations.

Metric	Travel	Yelp	Amazon
Options	300	300	300
Mode stable (iter)	3–5	4	2
JS converge (iter)	5–7	11	17
Final support	185	103	51
Mode mass (%)	75.2%	49.4%	79.0%

Table 9: Cross-domain comparison of convergence metrics across **travel demand**, **Yelp restaurants**, and **Amazon products**. All domains provide 300 options; final support indicates unique items with probability >0.1% in the converged distribution. Travel metrics represent ranges or averages across 300 departure airport contexts, whereas Yelp and Amazon values are from single-context experiments.

Yelp Restaurant Domain. Using 300 restaurants from Indianapolis extracted from the Yelp Open Dataset, we observe rapid mode stabilization consistent with travel domain findings. Table 9 summarizes convergence metrics.

Amazon Product Domain. Using 300 Home & Kitchen products sampled via stratified embedding-based selection from Amazon metadata, we observe the same fundamental pattern with domain-specific variations.

Support Size Calculation. We compute final support as the number of unique items with non-negligible probability ($> 0.1\%$) in the converged distribution. For travel, this yields 185 unique destinations aggregated across 300 departure contexts, while Yelp and Amazon produce 103 and 51 unique items, respectively, after 20 iterations. Across all domains, the modes-first, tails-later pattern persists: Jaccard@5 reaches 1.0 by iterations 2–4, indicating early mode stabilization, while JS divergence continues to decrease as tail exploration proceeds (Figure 12, Appendix D).

Key Cross-Domain Findings. The modes-first, tails-later convergence pattern holds consistently across travel demand, restaurant selection, and product choice, suggesting it reflects a general property of LLM distributional generation rather than domain-specific behavior. The strength of mode dominance varies by domain, with Amazon exhibiting the highest concentration (79% in top-10), followed by travel (75%) and Yelp (49%), consistent with differences in utility structure. Support diversity likewise varies, but all domains converge within a similar number of iterations (mode stabilization by iterations 2–5), underscoring the robustness of the convergence dynamics. Additional visualizations and analyses are provided in Appendix G.

5.6 Control Experiment: Fixed-Distribution Sampling

A potential concern is that the modes-first, tails-later pattern may arise trivially from sampling any fixed distribution or from staged sampling in general. To address this, we compare LLM iterative sampling against (i) i.i.d. sampling from fixed synthetic distributions (Zipf, Uniform, Multimodal) and (ii) a calibrated "LLM-like" staged sampling process designed to mimic early mode formation followed by gradual tail expansion. Figure 2 summarizes these comparisons with confidence bands across independent runs. We observe three consistent differences: real LLM sampling exhibits more rapid and abrupt mode stabilization (iterations 3–5) than either baseline; its JS divergence

trajectories remain distinct from both i.i.d. and simulated processes; and its tail support expands along a characteristic path that differs qualitatively from both fixed-distribution sampling and tuned staged simulations. Together, these results indicate that the observed convergence dynamics are not a trivial consequence of generic sampling procedures.

These results confirm that LLM iterative generation exhibits asymmetric stabilization dynamics distinct from both classical i.i.d. sampling and synthetic staged processes. The modes-first, tails-later pattern reflects structured, non-stationary generation behavior intrinsic to LLM inference rather than a sampling artifact or trivial consequence of any two-stage process.

6 Discussion

LLMs as Structured Stochastic Generators. Our results support viewing LLMs as *structured stochastic generators*: high-probability semantic beliefs are represented stably, while low-probability alternatives are explored stochastically. The observed modes-first, tails-later dynamic indicates that common decoding strategies, such as nucleus sampling and temperature scaling, primarily affect tail exploration rather than altering core semantic content. Persona and contextual conditioning induce systematic but bounded redistribution of probability mass, constrained by strong semantic anchors in the model's internal world representation.

Scale Discordance and Model Choice. Low agreement between distributions generated by models of different sizes challenges the assumption that scaling merely refines existing beliefs. Instead, different model scales encode distinct probabilistic world priors, a pattern we observe consistently across three model families (Llama, Qwen, and Mistral) with diverse training origins and architectures. These results suggest that scale discordance is intrinsic to LLM scaling rather than an artifact of specific designs. Practically, model size should be treated as a primary experimental variable, as changing scale can induce larger distributional shifts than prompt perturbations, and results may not transfer reliably across models or families.

Cross-Domain Generalization. The modes-first, tails-later convergence pattern generalizes across three choice domains—travel demand, restaurant selection, and product choice—despite substantial differences in domain semantics and utility structure. While the qualitative convergence behavior is consistent, its magnitude varies by domain: settings with clearer utility hierarchies exhibit stronger mode concentration, whereas more diffuse preference spaces yield flatter distributions. This suggests that convergence dynamics are domain-agnostic in form but domain-dependent in strength.

Practitioner Takeaways. For applications relying on LLM-generated probability distributions, a small number of samples ($\approx$5) is typically sufficient to identify dominant outcomes, while additional sampling primarily refines tail uncertainty rather than altering decisions. Given pronounced scale discordance, model selection should not be viewed solely as a cost–performance tradeoff, as different scales may yield qualitatively different distributions. Together, these findings highlight the importance of distribution-level evaluation when deploying LLMs in simulation, preference modeling, and decision-support settings.

7 Conclusion and Future Work

This work introduces a systematic framework for evaluating the *distributional behavior* of large language models, focusing on stability, convergence, and scale discordance. Through iterative sampling, we show that LLM-generated distributions consistently exhibit a *modes-first, tails-later* dynamic: high-probability outcomes stabilize rapidly, while residual variability is concentrated in low-probability alternatives. We further demonstrate that distributional differences across model scales substantially exceed those induced by prompt perturbations, revealing pronounced scale discordance in LLM probabilistic representations. These behaviors generalize across three choice domains—travel demand, restaurant selection, and product choice—and across three model families (Llama, Qwen, and Mistral), suggesting that they are intrinsic to contemporary LLMs rather than artifacts of specific domains or architectures.

Several directions for future work remain. An important extension is to assess the alignment between LLM-generated distributions and real-world human preferences. Further research could also develop theoretical explanations linking distributional dynamics to training objectives and optimization processes, as well as extend the analysis to additional domains, closed-source models, and alternative architectural designs.

Limitations

This study focuses on discrete choice settings in which the outcome space can be explicitly enumerated and normalized into a probability distribution. While we demonstrate consistent distributional behaviors across travel demand, restaurant selection, and product choice, our analysis is not designed to address open-ended generation tasks (e.g., summarization or creative writing), where outcome spaces are unbounded and distributional structure is less clearly defined. In addition, all experiments are conducted under a fixed decoding temperature ($T = 0.2$); exploring how alternative decoding strategies and temperature settings influence convergence dynamics and stability remains an important direction for future work.

Ethical Considerations

This work analyzes the distributional properties of LLM-generated outputs under controlled prompt variations, including demographic and contextual cues. While such 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 also highlight that model choice substantially influences generated distributions, underscoring the importance of transparency about model selection and sensitivity analyses when LLM outputs are used in applied settings. Although our experiments rely solely on synthetic data and do not involve personal information, care should be taken to avoid misinterpretation or inappropriate use of demographically conditioned outputs. Finally, as large-scale inference incurs nontrivial computational costs, practitioners should consider efficiency and sustainability alongside reliability when deploying LLM-based systems.

Acknowledgments

This project is supported by institutional funds. Additional details will be provided in the camera-ready version.

References

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

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. John Wiley & Sons.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

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. 2023. Mistral 7b. *arXiv preprint arXiv: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, and 7 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2021. 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.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Minsuk Kim, C Tyler Diggans, and Filippo Radicchi. 2025. Modeling resource consumption in the us air transportation system via minimum-cost percolation. *Nature Communications*, 16(1):8105.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Proceedings of the International Conference on Learning Representations*.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2023. Language models as agent models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12379–12396.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

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

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2300–2344.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Proceedings of the International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706.

A Detailed Experimental Results

Figure 3 shows the complete probability matrix across 300 contexts, revealing that LLM-generated distributions encode clear geographic structure with major hub locations appearing as consistently high-probability modes.

Distributional Stability and Convergence: Iteration analysis demonstrated that LLM distributions converge rapidly, with modes stabilizing by iterations 3–5 while tails continue to expand. Hub extensions confirmed consistent convergence patterns across context types. Model comparison showed convergence holds across scales, though smaller models require more iterations.

Sensitivity to Prompt Perturbations: Comparative analysis revealed perturbation effect sizes: model choice (JS=0.37–0.50) > age (JS=0.33) $\approx$ season (JS=0.31) > context (JS=0.31) > education (JS=0.26) > persona presence (JS=0.13). Seasonality experiments across 4 airports showed large hubs exhibit smaller seasonal effects (ORD: JS=0.26, ATL: JS=0.28) compared to smaller airports (IND: JS=0.33, FWA: JS=0.36), indicating stronger geographic anchoring at major hubs.

Scale Discordance and Generalization: Cross-model analysis revealed extremely low agreement: pairwise correlations near zero ($|r| < 0.21$) and agreement@5 below 13%. Different model

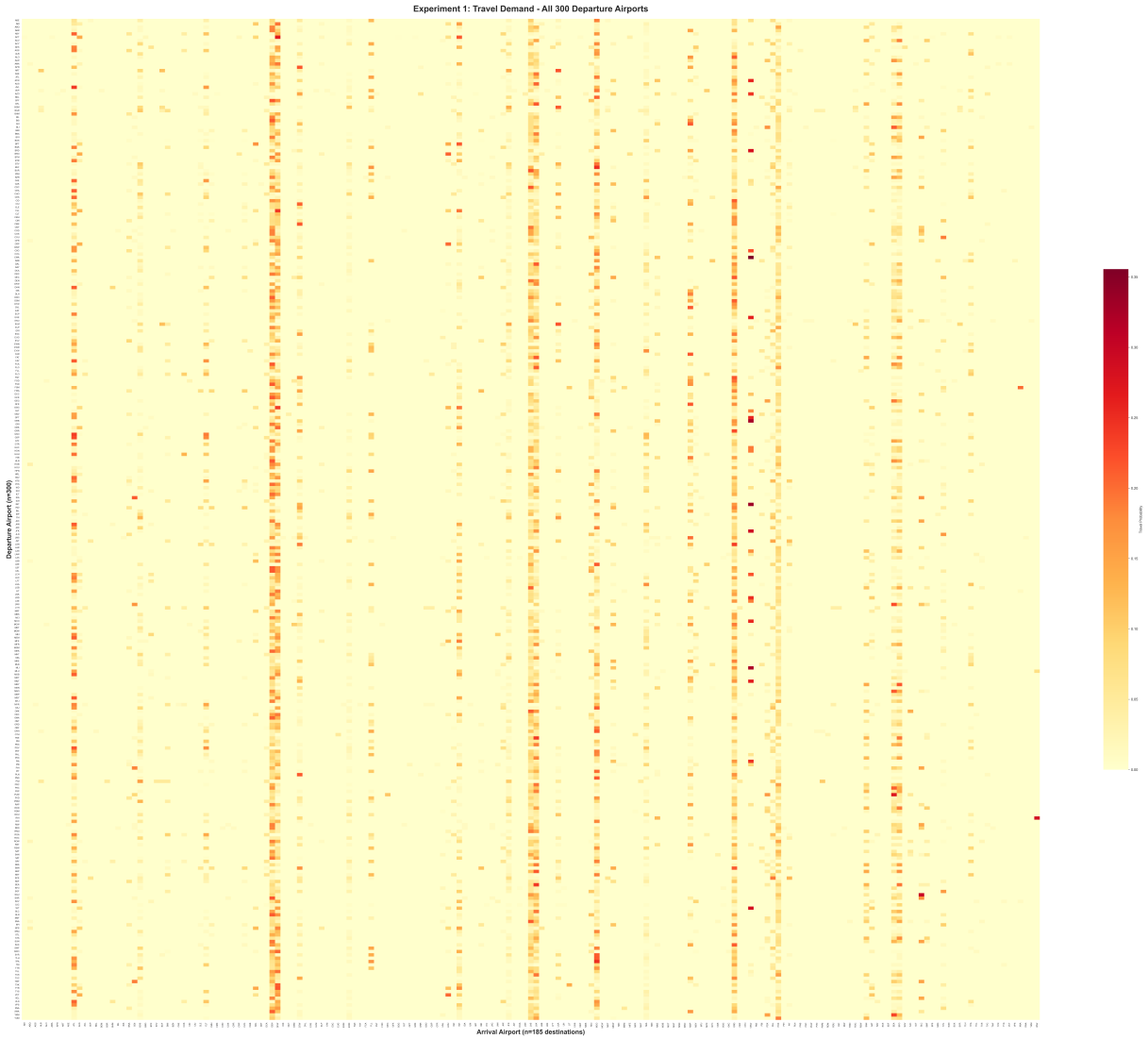

Figure 3: Probability distribution heatmap showing 300 departure airports (rows) × 185 unique destination airports (columns) that appeared across all contexts. Of the 300 airports provided as options, only 185 received non-zero probability in at least one context. Bright vertical bands reveal consistently high-probability modes.

scales encode fundamentally different world priors. Cross-domain replication in restaurant choice (Yelp, 300 restaurants from Indianapolis) and product shopping (Amazon, 300 Home & Kitchen products) confirmed the modes-first, tails-later pattern generalizes across domains. Mode stabilization occurred by iterations 2–4 in all domains, with domain-specific variations in mode dominance: Amazon (79%), Travel (∼75%), Yelp (49%).

B Persona-based Probability Distributions

Figure 4 shows the effect of persona conditioning on distributions. The difference heatmap reveals systematic patterns: persona information increases probabilities for certain outcome types (red) while decreasing others (blue), confirming that prompt perturbations induce meaningful, interpretable shifts.

Figures 5–6 show the complete persona-destination probability matrices for regional contexts, revealing how demographic factors influence the full distributional structure.

C Model Size Comparison: Full Heatmaps

Figure 7 visualizes the scale discordance phenomenon discussed in Section 5.3.

Figures 8–11 provide detailed comparisons of probability distributions across model sizes.

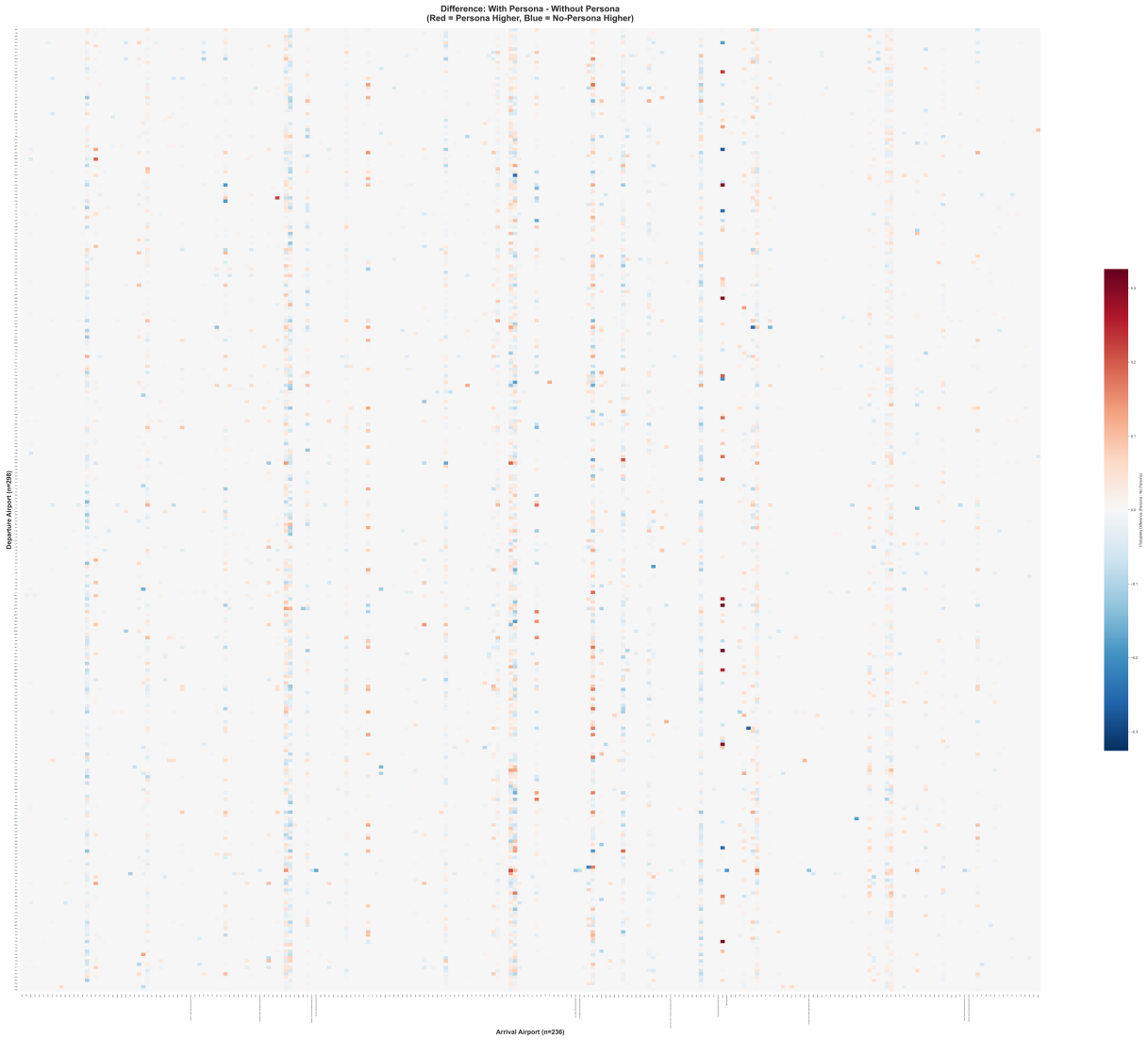

Figure 4: Difference heatmap showing $\Delta_{\text{persona}} = p_{\text{with}} - p_{\text{without}}$. Red: persona increases probability; Blue: persona decreases probability. The pattern reveals systematic but bounded distributional shifts induced by persona conditioning.

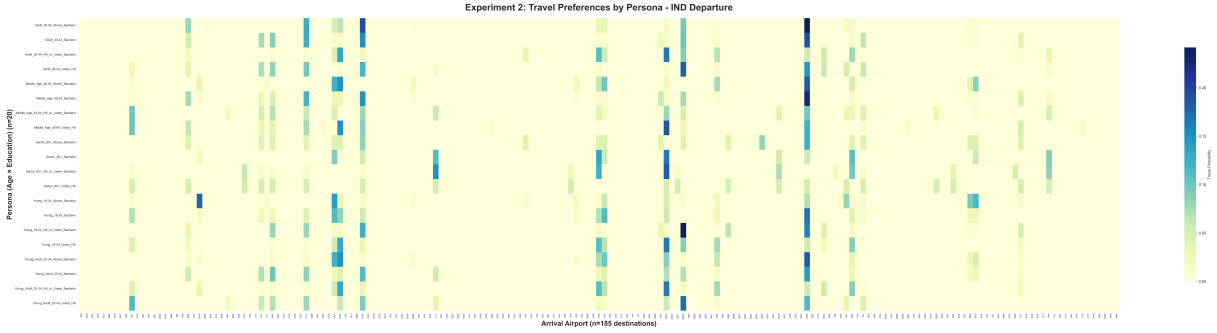

Figure 5: Persona-based probability distributions from Regional Context A (IND). Rows: 20 demographic personas (Age $\times$ Education). Columns: destination outcomes. Senior personas show concentrated preferences; younger personas exhibit more diverse distributions.

D Convergence Analysis: Full Iteration Heatmaps

Figure 12 visualizes cross-domain convergence patterns discussed in Section 5.5.

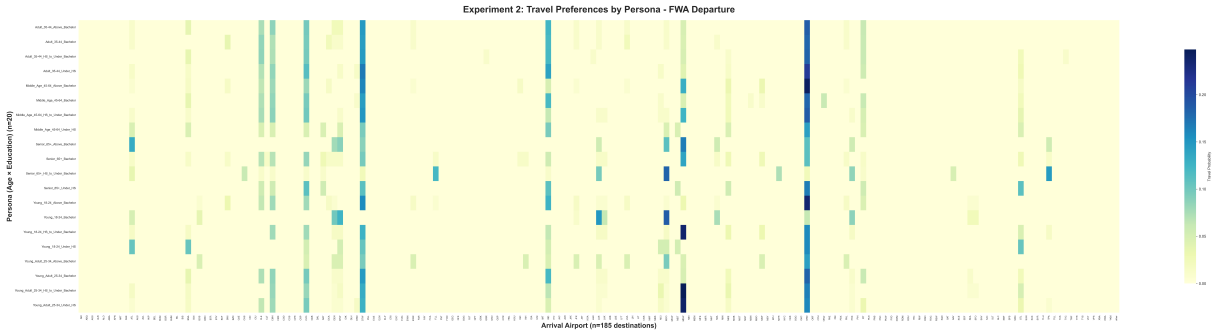

Figure 6: Persona-based probability distributions from Regional Context B (FWA). Comparison with Figure 5 reveals both consistency (hub dominance) and variation (persona-specific differences) across contexts.

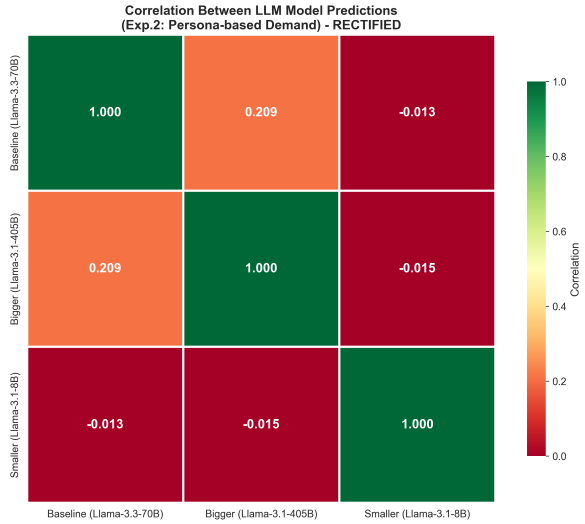

Figure 7: Scale discordance matrix showing pairwise correlations between three LLM sizes (8B, 70B, 405B). Values near zero indicate minimal agreement, demonstrating that model scale does not induce smooth distributional refinement. Different models encode fundamentally different world priors.

Figures 13–18 provide detailed iteration-by-iteration heatmaps demonstrating the modes-first, tails-later convergence pattern.

E Parsing Methodology for Smaller Models

Smaller LLMs (8B parameters) exhibit different output formatting behaviors compared to larger models, requiring specialized parsing strategies. This section documents the parsing challenges encountered and the rectification methodology applied.

E.1 Original Parsing Failures

With the original single-tier parsing approach (JSON bracket extraction only), the smaller model (Llama-3.1-8B) exhibited significant parsing failures:

Metric	Original	Rectified
Persona-based Analysis		
Success Rate	15% (3/20)	100% (20/20)
Unique Destinations	27	131
Avg Destinations/Persona	13.0	23.9
Total Records	39	477
Convergence Analysis		
Iterations Parsed	3/20	20/20
Total Destinations	63	308
Final Unique Destinations	—	100

Table 10: Comparison of smaller model (8B) results before and after parsing rectification.

E.2 Failure Case Analysis

The smaller model's outputs frequently deviated from the requested JSON format in three characteristic ways:

1. Markdown-Embedded JSON: Instead of returning a clean JSON array, the model embedded JSON objects within markdown formatting:

```
Here are my travel preferences:
- **Orlando**: `{"departure_airport": "IND",
  "arrival_airport": "MCO",
  "travel_probability": 0.15}`
- **Denver**: `{"departure_airport": "IND",
  ...
```

2. Natural Language Format: Some responses used descriptive text rather than structured JSON:

```
1. Arrival Airport: MCO (Orlando)
   Travel Probability: 0.15
2. Arrival Airport: DEN (Denver)
   Travel Probability: 0.12
...
```

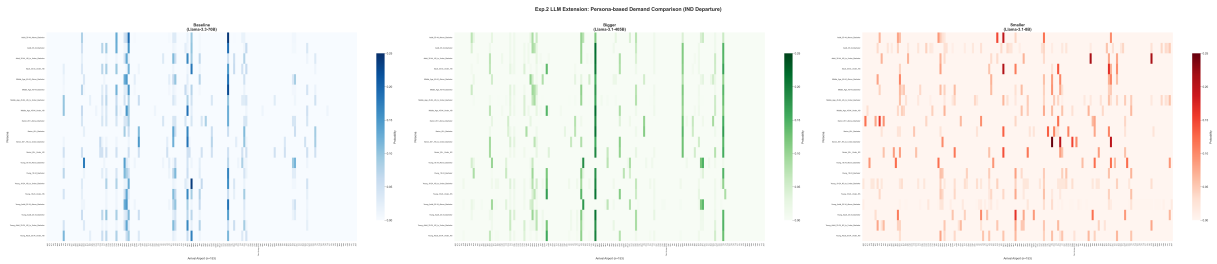

Figure 8: Side-by-side comparison of persona-based distributions across three LLM sizes with rectified parsing. Left: Baseline (70B, 46 destinations); Center: Bigger (405B, 72 destinations); Right: Smaller (8B, 131 destinations). The smaller model produces the highest diversity when properly parsed.

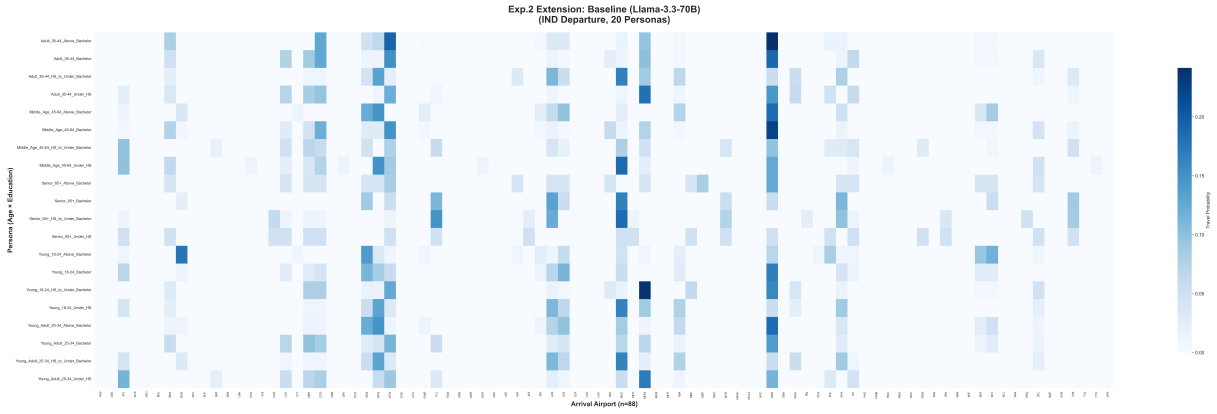

Figure 9: Baseline model (Llama-3.3-70B) predictions for 20 personas. The heatmap shows consistent preferences with moderate tail diversity.

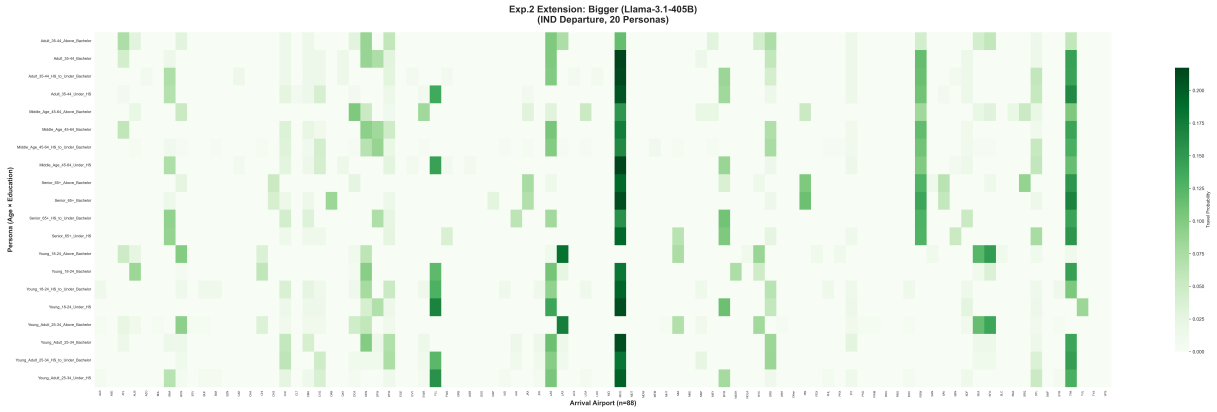

Figure 10: Bigger model (Llama-3.1-405B) predictions for 20 personas. The larger model produces more dispersed distributions with increased diversity.

3. Truncated Responses: With `max_tokens=1024`, verbose responses were cut off mid-JSON, producing malformed output.

E.3 Three-Tier Parsing Algorithm

To address these failures, we implemented a three-tier parsing approach:

Tier 1: Bracket Extraction (Original Method)

1. Search for content between `[...]` brackets
2. Attempt JSON parsing with trailing comma handling
3. If successful, normalize probabilities and return

Tier 2: Regex JSON Extraction (New)

1. If Tier 1 fails, apply regex pattern matching to find JSON-like objects containing airport fields embedded in the text

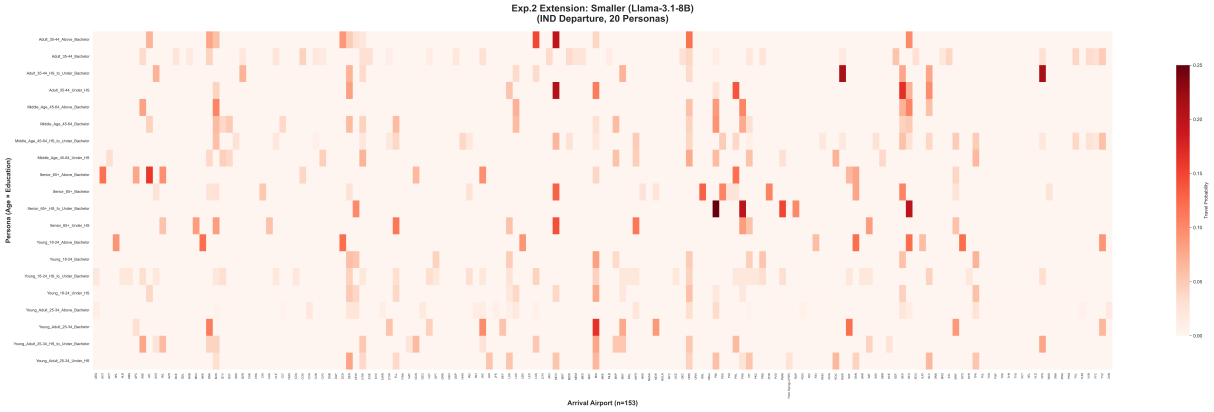

Figure 11: Smaller model (Llama-3.1-8B) predictions for 20 personas with improved three-tier parsing. Successfully parses all 20 personas with 131 unique destinations and 23.9 average destinations per persona, the highest diversity among all model sizes. See Appendix E for parsing methodology.

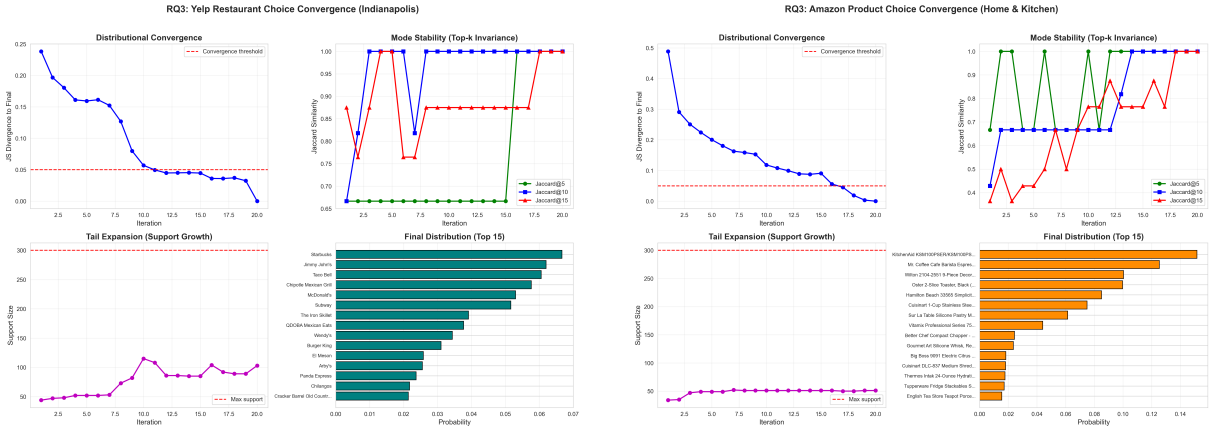

Figure 12: Cross-domain convergence analysis. Left: Yelp restaurant choice (300 restaurants). Right: Amazon product choice (300 products). Both domains exhibit rapid mode stabilization (Jaccard@5 $\rightarrow$ 1.0 by iteration 4) followed by continued tail expansion, confirming the generality of the modes-first, tails-later pattern.

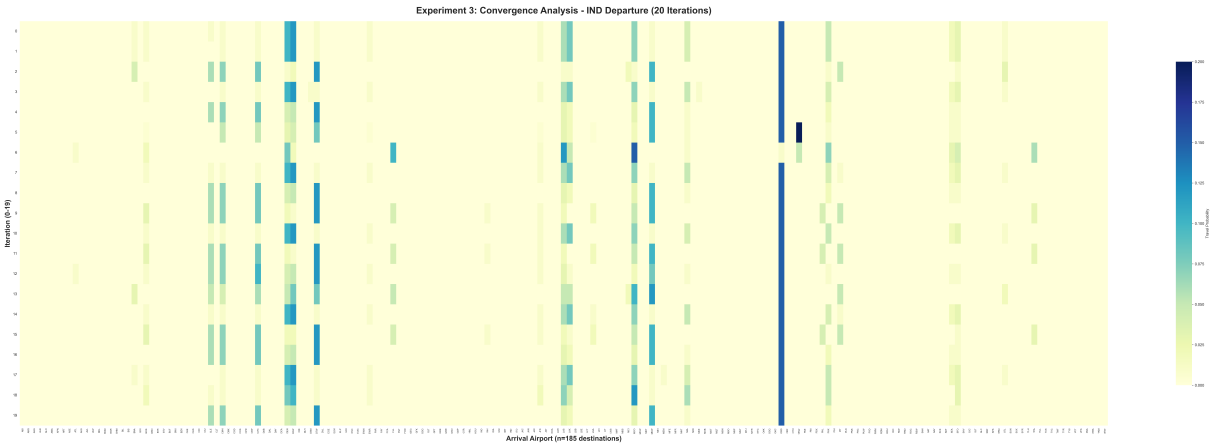

Figure 13: Convergence heatmap for Regional Context A (IND) over 20 iterations. Rows: iterations (0-19). Columns: outcomes. Bright vertical bands show stable high-probability modes; lighter peripheral colors show gradual tail expansion.

2. Extract all matching objects and validate that they contain required fields (departure airport, destination, probability)
3. Clean formatting issues: convert single quotes to double quotes, remove trailing commas

Tier 3: Natural Language Extraction (New)

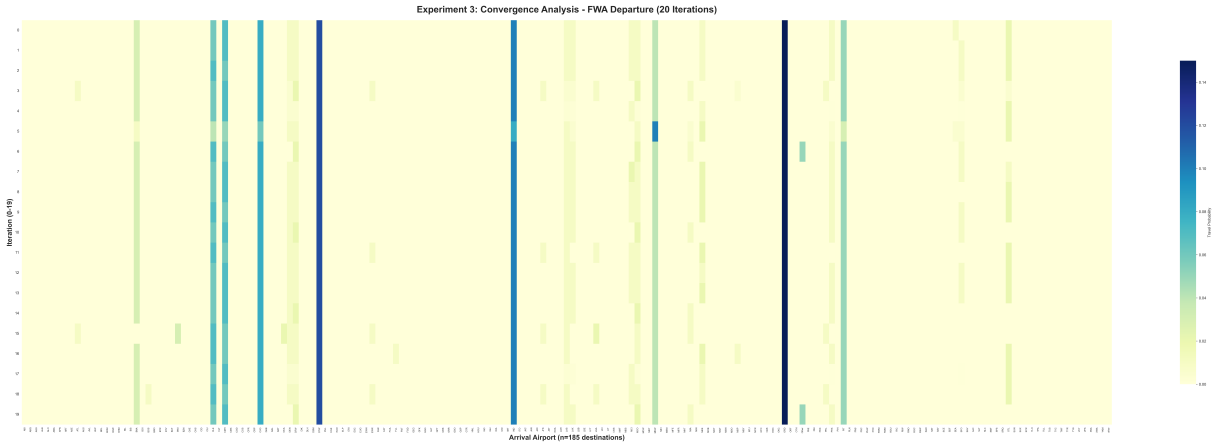

Figure 14: Convergence heatmap for Regional Context B (FWA) over 20 iterations. Similar pattern to Figure 13: rapid mode stabilization with continued tail expansion.

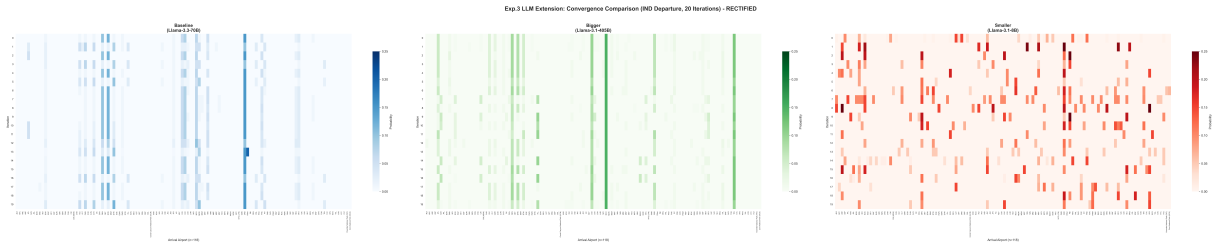

Figure 15: Side-by-side comparison of convergence patterns across three LLM sizes with rectified parsing. Left: Baseline (70B, 32 final destinations); Center: Bigger (405B, 54 final destinations); Right: Smaller (8B, 100 final destinations). The smaller model achieves highest diversity but requires more iterations for stabilization.

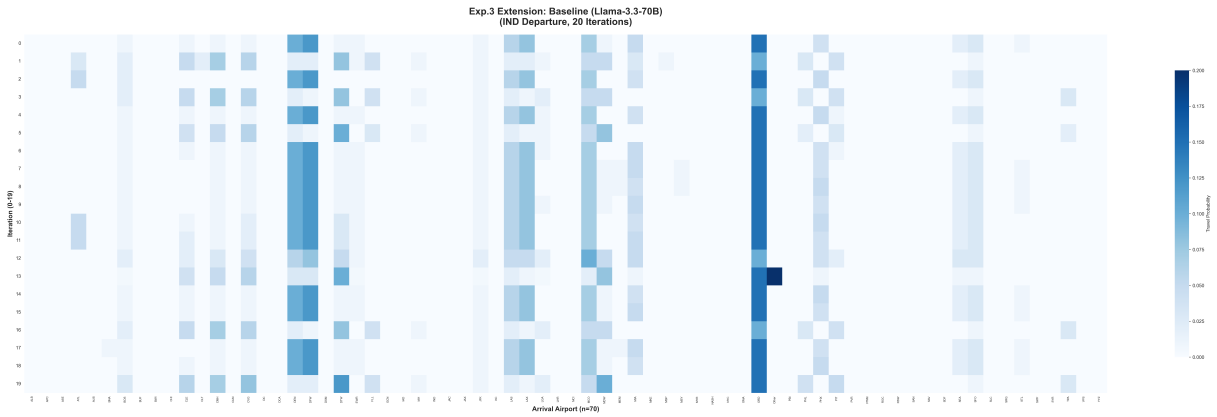

Figure 16: Baseline model (70B) convergence over 20 iterations. Demonstrates rapid mode stabilization with consistent patterns.

1. If Tier 2 fails, search for natural language patterns
2. Patterns include: "Arrival: XXX", "Destination: Name (XXX)", "Travel Probability: 0.XX"
3. Extract airport codes and probability values via regex
4. Construct structured records from extracted data

Additionally, `max_tokens` was increased from 1024 to 4096 to prevent truncation of verbose responses.

E.4 Implications for Model Comparison

The parsing rectification reveals important insights about smaller model behavior:

- **Output Diversity:** The 8B model actually

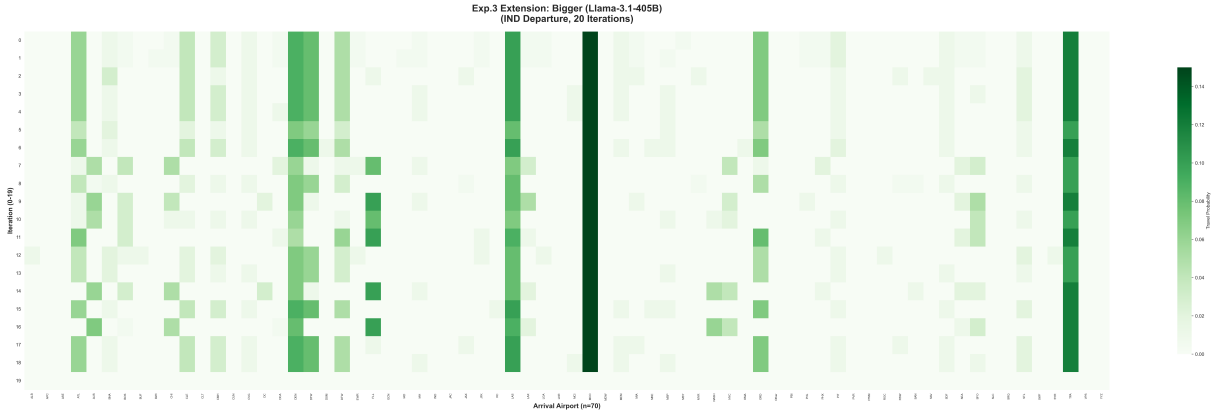

Figure 17: Bigger model (405B) convergence over 20 iterations. Shows stable convergence with broader outcome diversity.

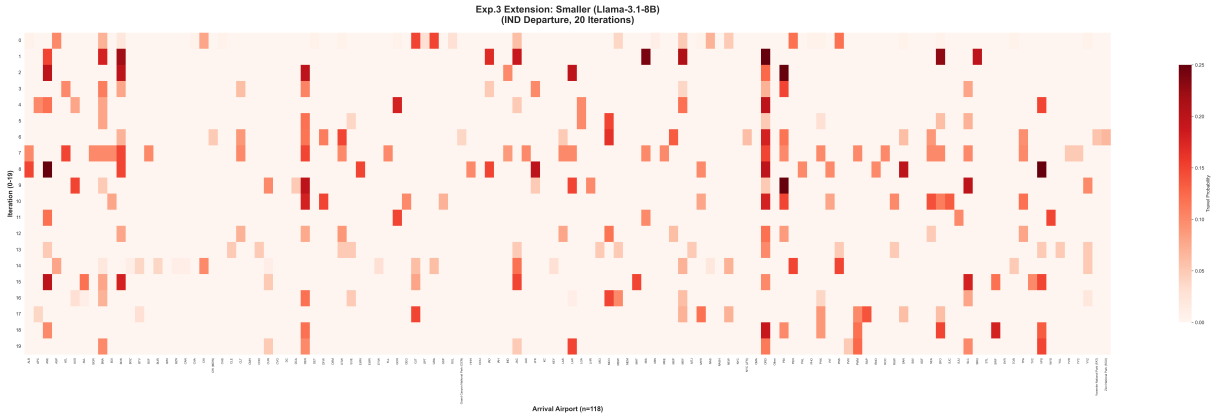

Figure 18: Smaller model (8B) convergence over 20 iterations with improved three-tier parsing. Successfully parses 20/20 iterations with 308 total destination records and 100 unique destinations at final iteration. See Appendix E for parsing methodology.

generates *more* diverse destinations (131) than larger models (72 for 405B, 46 for 70B), but this was masked by parsing failures.

- **Format Compliance:** Smaller models exhibit lower instruction-following fidelity for output formatting, requiring robust post-processing.
- **Semantic Content:** Despite format issues, the underlying semantic content is coherent and analyzable once properly extracted.
- **Convergence Behavior:** The smaller model achieves 100 unique destinations after 20 iterations (vs. 54 for 405B, 32 for 70B), demonstrating higher exploratory diversity.

E.5 Original Results (Without Rectification)

For transparency, we present the original smaller model results obtained with single-tier parsing. These results illustrate what researchers may encounter without specialized parsing for smaller models.

Original Persona Analysis Results: Only 3/20 personas (15%) were successfully parsed, yielding 27 unique destinations with 13.0 avg destinations per persona. The remaining 17 personas returned malformed JSON.

Original Convergence Analysis Results: Only 3/20 iterations (15%) were parsed, producing 63 total records. Convergence analysis was not possible due to insufficient data.

Figure 19 visualizes the parsing failure: with single-tier parsing, only 3 of 20 persona rows contain data, with the remaining 17 rows empty due to malformed outputs. This sparse heatmap illustrates the severity of parsing failures when working with smaller models without robust post-processing.

These results highlight the importance of robust parsing strategies when working with smaller LLMs, which may produce semantically valid but

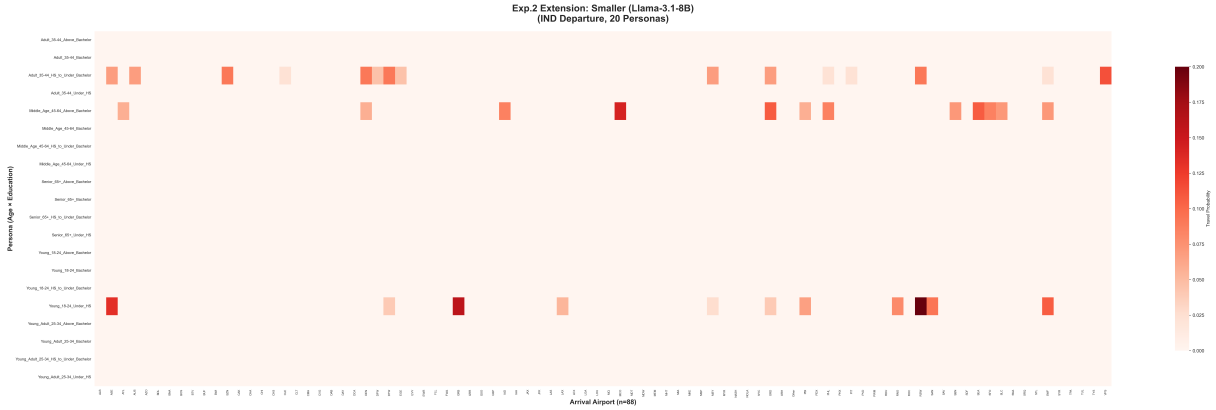

Figure 19: Original smaller model (8B) results with single-tier parsing for persona-based analysis. Only 3/20 personas (15%) were successfully parsed, yielding a sparse heatmap with 17 empty rows. Compare with Figure 11, which shows the same data after three-tier parsing rectification (20/20 personas, 131 unique destinations).

syntactically non-compliant outputs.

F Extended Prompt Sensitivity Analysis

F.1 Model Sensitivity vs Prompt Sensitivity

Comparing model choice against prompt perturbations reveals that model selection has a larger impact on output distributions than persona conditioning.

Model Sensitivity: Cross-model JS divergences are substantial: 70B vs 405B yields JS=0.365 (95% CI: [0.33, 0.41]), 70B vs 8B yields JS=0.474 (95% CI: [0.42, 0.53]), and 405B vs 8B yields JS=0.501 (95% CI: [0.46, 0.54]).

Prompt Sensitivity: In contrast, prompt perturbations produce smaller effects: persona conditioning yields JS=0.134 (95% CI: [0.12, 0.14]), age perturbations yield JS=0.330 (95% CI: [0.28, 0.38]), and education perturbations yield JS=0.261 (95% CI: [0.23, 0.29]).

These results indicate that switching between model sizes (e.g., 8B to 70B) induces distributional shifts 2–4$\times$ larger than demographic prompt conditioning, highlighting model choice as a critical factor in LLM-based simulations.

G Extended Stability and Convergence Analysis

G.1 Stability Metrics Over Iterations

Figure 20 shows the evolution of stability metrics (JS divergence, Jaccard@5, Jaccard@10) over 20 iterations across four airport contexts. The IND airport exhibits an alternating pattern in early iterations before converging, reflecting the model's exploration of alternative mode configurations.

G.2 Stability Metrics Across Model Sizes

Figure 21 compares stability metrics across three LLM sizes. The smaller model (8B) exhibits significantly higher instability (mean JS=0.483) compared to the baseline (70B, JS=0.171) and larger (405B, JS=0.152) models.

G.3 Mode-Tail Decomposition Across Models

Figure 22 presents the mode-tail decomposition across model sizes. The smaller model (8B) has 60% mode contribution to total divergence, indicating higher mode instability compared to baseline (35%) and bigger (13%) models.

G.4 Tail Dynamics Across Models

Figure 23 shows the evolution of tail dynamics (support size, entropy, tail mass) across model sizes over 20 iterations. All models show gradual convergence, but with different rates and final distributions.

H Control Experiment: Technical Details

This section provides technical details for the control experiment comparing LLM iterative sampling against i.i.d. sampling from fixed distributions.

H.1 Experimental Design

To test whether the modes-first, tails-later pattern is a trivial consequence of sampling from any probability distribution, we constructed synthetic categorical distributions and simulated i.i.d. sampling under matched conditions.

Parameters:

- Outcome space: $|\mathcal{Y}| = 300$ (matching our travel domain)

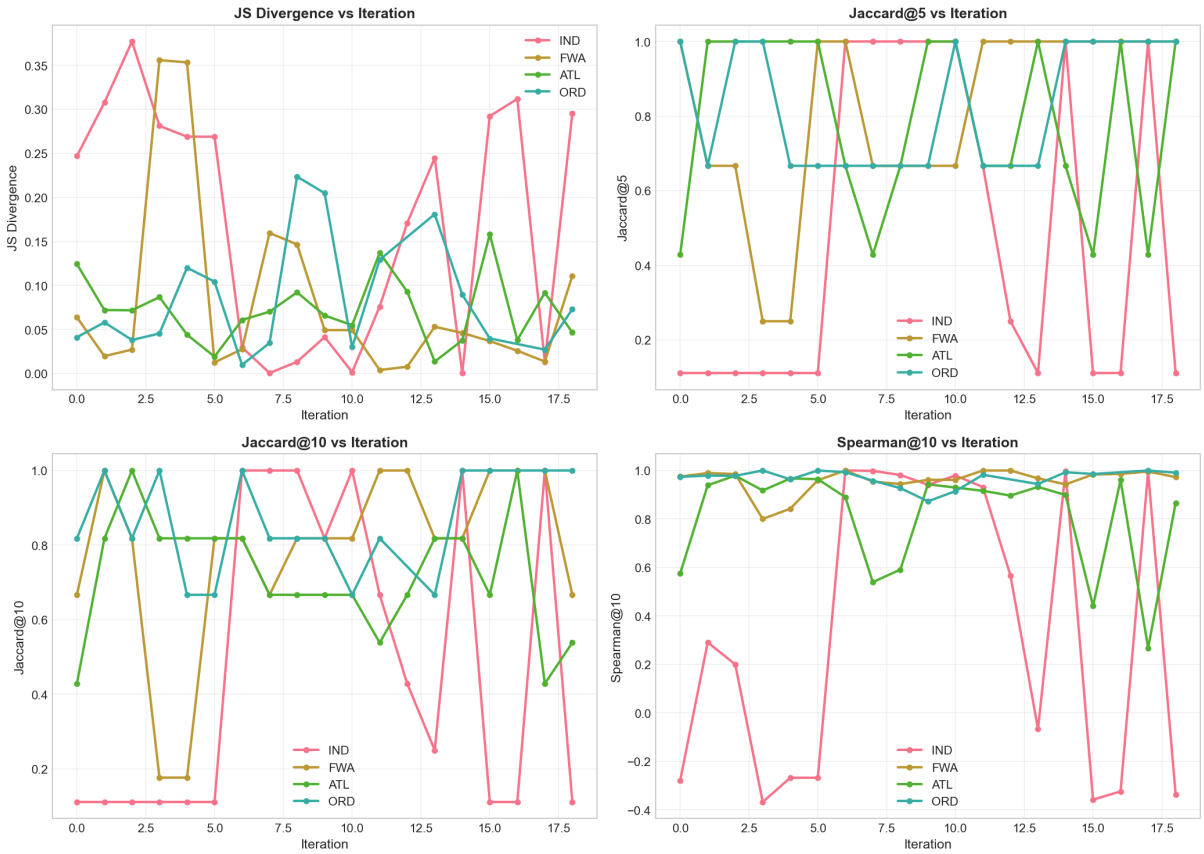

Figure 20: Stability metrics evolution over 20 iterations across four airports (IND, FWA, ATL, ORD). JS divergence decreases rapidly in early iterations, while Jaccard@k metrics increase, confirming mode stabilization. IND shows an alternating pattern in iterations 1-5 before converging.

- Iterations: 20 (matching our experimental protocol)
- Samples per iteration: 50 (for empirical distribution estimation)

H.2 Synthetic Distributions

Three distribution types were used as controls:

Zipf Distribution ($\alpha = 1.5$)**:** A heavy-tailed power-law distribution commonly observed in real-world preference data:

$$p(y_i) \propto \frac{1}{i^{\alpha}}, \quad i = 1, \ldots, 300 \tag{3}$$

Uniform Distribution: Equal probability across all outcomes:

$$p(y_i) = \frac{1}{300}, \quad \forall i \tag{4}$$

Multimodal Distribution: Concentrated probability mass on 5 randomly selected modes with remaining mass distributed uniformly:

$$p(y_i) = \begin{cases} \text{Dirichlet}(1, \ldots, 1) \times 0.6 & \text{if } i \in \text{modes} \\ \frac{0.4}{295} & \text{otherwise} \end{cases} \tag{5}$$

H.3 Metrics Computed

For each iteration t, we computed:

- **Jaccard@5**: Overlap of top-5 outcomes between consecutive iterations
- **JS Divergence**: Jensen-Shannon divergence between consecutive empirical distributions
- **Support Size**: Number of unique outcomes with probability $> 0.1\%$

H.4 Key Findings

Comparing i.i.d. sampling baselines with actual LLM generation reveals striking differences. All three synthetic distributions (Zipf, Uniform, Multimodal) exhibit near-zero JS divergence after convergence (JS < 0.01 for iterations 10–20), gradual mode stabilization, and tail expansion that plateaus. In contrast, Llama-3.3-70B shows elevated post-convergence JS divergence (> 0.08), abrupt mode stabilization (iterations 3–5), and continuous tail expansion throughout all 20 iterations.

The comparison reveals that LLM iterative generation exhibits:

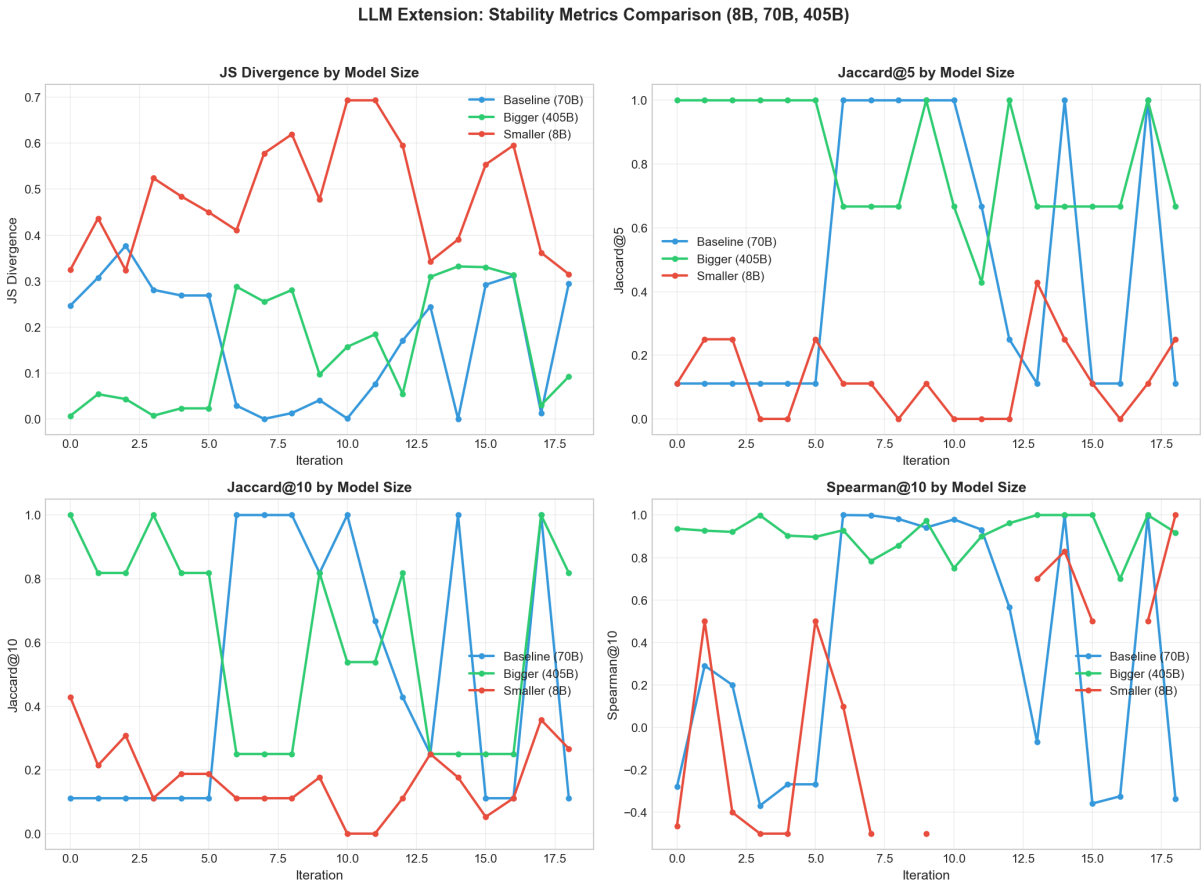

Figure 21: Stability metrics comparison across three LLM sizes (8B, 70B, 405B). Smaller model shows highest instability (JS=0.483), while larger models achieve more stable distributions. This suggests that model scale correlates with distributional consistency.

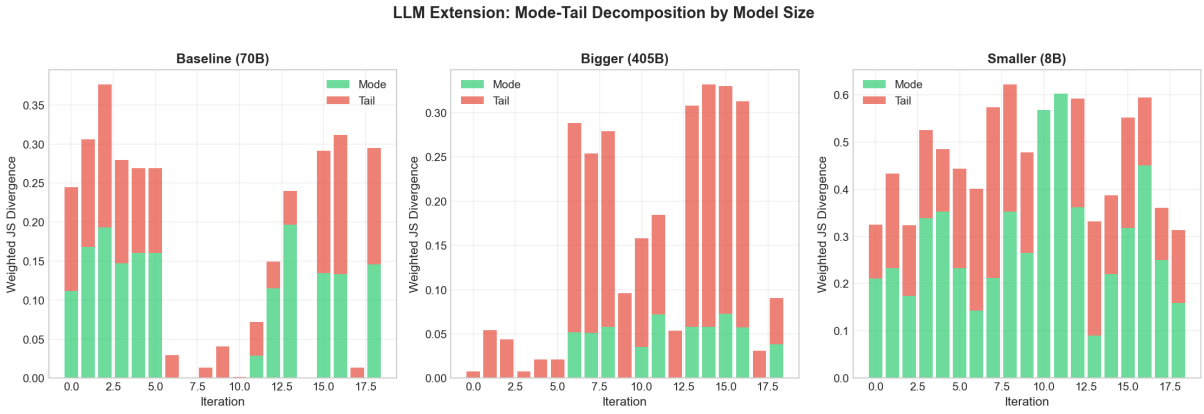

Figure 22: Mode-tail decomposition across model sizes. Smaller model (8B) has 60% mode contribution, indicating more mode instability compared to baseline 70B (35%) and bigger 405B (13%). Larger models exhibit "modes-first" behavior more strongly.

1. **Abrupt mode stabilization** (Jaccard@5 jumps to 1.0 by iteration 3-5) vs. gradual i.i.d. convergence

2. **Elevated post-convergence JS divergence** (>0.08) vs. near-zero i.i.d. divergence (<0.01)

3. **Continuous tail expansion** (support grows from 19 to 32) vs. i.i.d. plateau behavior

These differences confirm that the modes-first, tails-later phenomenon is a distinctive property of LLM generation, not a sampling artifact.

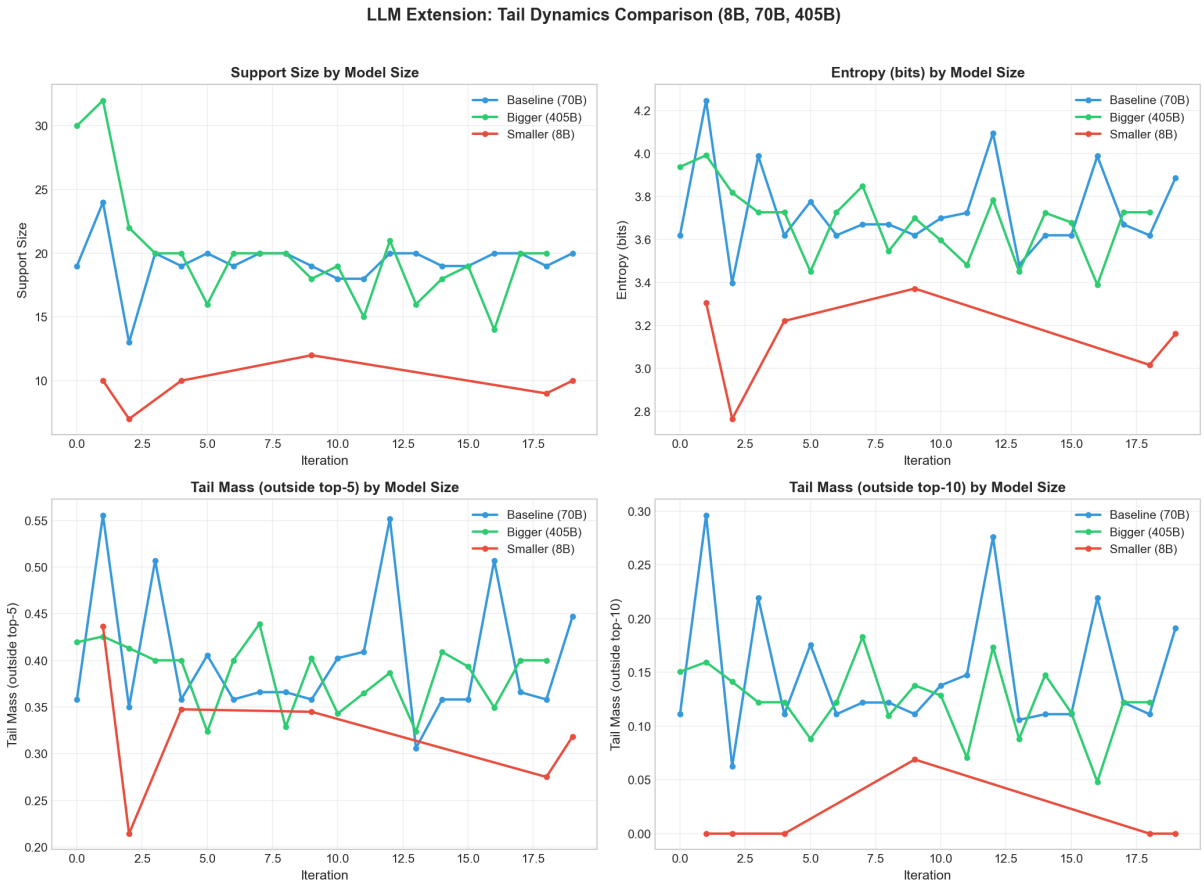

Figure 23: Tail dynamics (support, entropy, tail mass) across model sizes over 20 iterations. Smaller model maintains higher support and entropy throughout, consistent with its higher exploratory diversity.

I Cross-Domain Dataset Descriptions

This section provides detailed descriptions of the datasets and sampling procedures used for the cross-domain generalization experiments.

I.1 Yelp Restaurant Dataset

Source: Yelp Open Dataset (https://www.yelp.com/dataset)

Data Extraction: We extracted restaurant data from the Yelp Open Dataset, focusing on Indianapolis, Indiana to match the geographic context of our primary travel demand experiments.

Filtering Criteria:

- **Location:** Businesses located in Indianapolis (city = "Indianapolis")
- **Category:** Businesses containing "Restaurants" in their category list
- **Status:** Only open businesses (is_open = 1)
- **Name Quality:** Excluded entries with generic or missing names
- **Category Quality:** Excluded entries with only generic categories (e.g., "Restaurants" alone without cuisine specification)

Final Dataset:

- 300 restaurants sampled from Indianapolis
- 54 unique cuisine categories represented
- Each restaurant record includes: name, cuisine type, star rating, review count, and location attributes

Cuisine Distribution: The extracted restaurants span diverse cuisine types including American (Traditional), American (New), Mexican, Italian, Chinese, Japanese, Indian, Thai, Mediterranean, and 45 other cuisine categories, providing a semantically rich choice set for LLM evaluation.

I.2 Amazon Product Dataset

Source: Amazon Product Data (McAuley et al., 2015; He and McAuley, 2016) (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., Kitchen & Dining, Bedding, Bath, Storage & Organization, Home Décor, etc.)

2. **Embedding Generation:** Product descriptions 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)
- Each product record includes: title, description, category hierarchy, price, and rating attributes

Subcategory Distribution:

Subcategory	Count
Kitchen & Dining	20
Bedding	20
Bath	20
Storage & Organization	20
Home Décor	20
Furniture	20
Cleaning Supplies	20
Heating, Cooling & Air Quality	20
Lighting & Ceiling Fans	20
Event & Party Supplies	20
Wall Art	20
Kids' Home Store	20
Irons & Steamers	20
Vacuums & Floor Care	20
Other	20
Total	**300**

Table 11: Amazon Home & Kitchen subcategory distribution.

I.3 Domain Comparison

The three domains used in this study differ in key dimensions:

This diversity enables testing whether the observed distributional patterns (modes-first convergence, scale discordance, prompt sensitivity bounds) generalize beyond the primary travel demand setting.

J Cross-Family Methodology

This section documents the methodology for extending scale discordance analysis beyond the Llama family to include Qwen and Mistral model families.

J.1 Model Configuration

We evaluate nine models spanning three families, each with small, medium, and large variants. The **Llama family** (Meta, US) includes Llama-3.1-8B, Llama-3.3-70B, and Llama-3.1-405B, all using dense architectures. The **Qwen family** (Alibaba, China) includes Qwen2.5-7B, Qwen2.5-72B (both dense), and Qwen3-235B which uses a Mixture-of-Experts (MoE) architecture with 22B active parameters. The **Mistral family** (Mistral AI, France) includes Mistral-7B-v0.3, Mistral-Small-24B (both dense), and Mixtral-8x7B which uses MoE with 12B active parameters out of 46B total. All models were accessed via the Together AI API.

J.2 Output Normalization

Different models produce destination outputs in varying formats, requiring normalization for cross-model comparison. While most models output standard three-letter IATA codes (e.g., "ATL"), some models (notably Qwen 235B and Mistral 24B) output full airport names with embedded codes (e.g., "Hartsfield Jackson Atlanta International Airport (ATL) Georgia").

We implemented a normalization function that extracts three-letter IATA codes from any format using the following procedure:

Characteristic	Value
Travel Domain	
Choice Set Size	300 airports
Geographic Scope	National (US)
Decision Frequency	Low (occasional)
Price Range	High ($100+)
Category Type	Transportation
Yelp Domain	
Choice Set Size	300 restaurants
Geographic Scope	Local (Indianapolis)
Decision Frequency	Medium (weekly)
Price Range	Medium ($10-50)
Category Type	Food & Dining
Amazon Domain	
Choice Set Size	300 products
Geographic Scope	N/A (online)
Decision Frequency	High (frequent)
Price Range	Variable
Category Type	Consumer Goods

Table 12: Comparison of domain characteristics across experimental settings.

1. If the string is exactly three uppercase letters, return it directly

2. Search for a three-letter uppercase code within parentheses using regex pattern `\([A-Z]{3}\)`

3. If found, extract and return the matched code

This normalization restored cross-model comparability. Without normalization, models using different output formats showed zero common destinations and maximum JS divergence (0.693). After normalization, typical common destination counts ranged from 15 to 30 between model pairs.

Note on maximum JS divergence: The value 0.693 equals $\ln(2)$, which is the theoretical maximum of Jensen-Shannon divergence. This maximum occurs when two distributions have completely non-overlapping supports (i.e., zero shared outcomes with non-zero probability). In Table 6, the asterisked 0.693 values indicate model pairs where, even after normalization, the distributions placed probability mass on entirely different destination sets. This extreme discordance reflects fundamental differences in how certain model pairs represent geographic preferences, rather than output format artifacts.

J.3 Mistral 24B Behavioral Adaptation

The Mistral-Small-24B model exhibited atypical behavior during data collection. Instead of generating travel demand distributions, it frequently requested clarification (e.g., "Could you please specify your travel preferences?"). This behavior, characterized as RLHF over-alignment, resulted in an 11.2% success rate (20 valid responses from 179 attempts).

Two mitigations were applied:

1. **Research context prefix:** Adding "You are generating results for scientific research projects." to the prompt reduced clarification behavior by establishing a legitimate research context.

2. **Retry algorithm:** A retry mechanism with up to 500 attempts and incremental progress saving ensured collection of the target 20 valid samples.

This behavior was not observed in persona-conditioned experiments (Section B), suggesting that detailed demographic context naturally prevents clarification behavior by reducing prompt ambiguity.

J.4 Experimental Parameters

All cross-family experiments used consistent parameters matching the original Llama experiments:

- **Temperature:** 0.2 (low temperature for reproducibility)
- **Iterations:** 20 per model (convergence analysis)
- **Max tokens:** 4096 (increased from 1024 to accommodate verbose outputs)
- **Departure context:** Indianapolis International Airport (IND)
- **Choice set:** 300 US airports

All models were accessed via the Together AI API to ensure consistent infrastructure across families. Total data collection comprised 179 API calls for Mistral 24B (due to retry requirements) and 20 calls each for the remaining eight models, yielding 20 valid samples per model for analysis.

K Detailed Metric Definitions

This section provides complete mathematical definitions for the evaluation metrics summarized in Section 3.4.

K.1 Distributional Stability Metrics

Jensen–Shannon Divergence. To measure overall distributional similarity between runs, we use Jensen–Shannon (JS) divergence (Lin, 1991):

$$\mathrm{JS}(p^{(r)}, p^{(r')}) = \frac{1}{2} D_{\mathrm{KL}}(p^{(r)} \| m) + \frac{1}{2} D_{\mathrm{KL}}(p^{(r')} \| m) \tag{6}$$

where $m = \frac{1}{2}(p^{(r)} + p^{(r')})$ is the mixture distribution and D_{KL} denotes Kullback–Leibler divergence. JS divergence is symmetric, bounded in $[0, 1]$ when using base-2 logarithm, and well-suited for comparing discrete distributions. Lower values indicate greater similarity; $\mathrm{JS} = 0$ implies identical distributions.

Mode Stability (Top-k Invariance). We measure mode stability using the Jaccard index (Jaccard, 1912):

$$\mathrm{Jaccard}_k(r, r') = \frac{|\mathcal{M}_k(p^{(r)}) \cap \mathcal{M}_k(p^{(r')})|}{|\mathcal{M}_k(p^{(r)}) \cup \mathcal{M}_k(p^{(r')})|} \tag{7}$$

where r and r' denote two sampling runs, $\mathcal{M}_k(p^{(r)})$ is the set of top-k outcomes from run r, $\cap$ and $\cup$ denote set intersection and union, and $|\cdot|$ denotes set cardinality. Jaccard ranges from 0 (no overlap) to 1 (identical sets). Higher values indicate stable dominant outcomes across samples.

Rank Consistency. For ranked mode sets, we compute Spearman rank correlation (Spearman, 1904) over the top-k outcomes to assess ordering stability. Spearman's ρ ranges from -1 to $+1$, where $+1$ indicates perfect rank agreement and values near 0 indicate no systematic relationship.

K.2 Tail Dynamics Metrics

We employ standard information-theoretic measures (Cover and Thomas, 2006) to characterize tail behavior:

Support Growth:

$$|\{y : p^{(r)}(y) > \epsilon\}| \tag{8}$$

This counts the number of outcomes with non-negligible probability (we use $\epsilon = 0.001$). Higher values indicate broader distributional coverage.

Entropy:

$$H(p^{(r)}) = -\sum_{y} p^{(r)}(y) \log p^{(r)}(y) \tag{9}$$

Shannon entropy measures distributional uncertainty (Cover and Thomas, 2006). Higher entropy indicates more uniform (less concentrated) distributions; entropy is minimized at 0 for deterministic distributions.

Tail Mass:

$$\mathrm{TailMass}_k(p^{(r)}) = 1 - \sum_{y \in \mathcal{M}_k} p^{(r)}(y) \tag{10}$$

Tail mass is defined as the complement of the cumulative probability in the mode set $\mathcal{M}_k$, following standard probability concentration analysis (Cover and Thomas, 2006). It quantifies the probability allocated to non-mode outcomes. Values range from 0 (all mass in top-k) to near 1 (highly dispersed). Higher values indicate greater exploration of low-probability alternatives.

K.3 Mode–Tail Divergence Analysis

To localize sources of instability, we compute JS divergence (Lin, 1991) separately on the mode and tail regions of the distribution. Given two distributions $p^{(r)}$ and $p^{(r')}$ from consecutive iterations, we define three metrics:

Total JS Divergence measures overall distributional shift:

$$\mathrm{JS}_{\mathrm{total}} = \mathrm{JS}(p^{(r)}, p^{(r')}) \tag{11}$$

Mode JS Divergence measures instability in the high-probability region. We restrict both distributions to the mode set $\mathcal{M}_k(p^{(r')})$ (using the later iteration as reference), renormalize to sum to 1, and compute:

$$\mathrm{JS}_{\mathrm{mode}} = \mathrm{JS}(\tilde{p}^{(r)}_{\mathcal{M}}, \tilde{p}^{(r')}_{\mathcal{M}}) \tag{12}$$

where $\tilde{p}_{\mathcal{M}}$ denotes the renormalized distribution over $\mathcal{M}_k$.

Tail JS Divergence measures instability in the low-probability region, computed analogously on $\mathcal{T}_k = \mathcal{Y} \setminus \mathcal{M}_k$:

$$\mathrm{JS}_{\mathrm{tail}} = \mathrm{JS}(\tilde{p}^{(r)}_{\mathcal{T}}, \tilde{p}^{(r')}_{\mathcal{T}}) \tag{13}$$

Important: These three metrics are computed independently on different subsets of the outcome space. They do not form an additive decomposition (i.e., $\mathrm{JS}_{\mathrm{total}} \neq \mathrm{JS}_{\mathrm{mode}} + \mathrm{JS}_{\mathrm{tail}}$). Rather, they provide complementary views: $\mathrm{JS}_{\mathrm{total}}$ captures overall divergence, while $\mathrm{JS}_{\mathrm{mode}}$ and $\mathrm{JS}_{\mathrm{tail}}$ isolate divergence within each region after renormalization. This allows us to identify whether instability arises from semantic core beliefs (mode divergence) or from peripheral exploration (tail divergence).

Interpretation of metric magnitudes: It is expected that $\text{JS}_{\text{tail}} > \text{JS}_{\text{total}}$ in practice. The tail region contains many low-probability outcomes that are inherently noisy; after renormalization (scaling tail probabilities to sum to 1), small absolute differences become large relative differences, magnifying the measured divergence. Conversely, $\text{JS}_{\text{mode}} < \text{JS}_{\text{total}}$ because high-probability outcomes stabilize quickly and exhibit less variability. The key diagnostic insight is the *trajectory* of these metrics: when JS_{mode} approaches zero while JS_{tail} remains elevated, this indicates that core semantic beliefs have converged and only peripheral outcomes continue to fluctuate.

K.4 Prompt Sensitivity Analysis

To quantify distributional shift under controlled prompt perturbations, we use JS divergence (Lin, 1991) between distributions under different prompt conditions. We consider persona vs no-persona prompts, demographic attribute changes (age, education), and departure location changes. Higher JS values indicate greater sensitivity to the perturbation. By comparing JS divergence across perturbation types, we can quantify prompt effect sizes and distinguish meaningful conditioning effects from sampling noise.

K.5 Scale Discordance Analysis

To assess whether different model scales encode consistent world priors, we compare distributions across models m_i, m_j using JS divergence (Lin, 1991):

$$\text{JS}(p_{m_i}(\cdot \mid x), p_{m_j}(\cdot \mid x)) \tag{14}$$

We additionally compute agreement@k, defined as the proportion of contexts where two models share identical top-k outcome sets. Agreement@k ranges from 0 (no shared predictions) to 1 (perfect agreement); values near chance level (≈ 0 for large outcome spaces) indicate independent distributional priors.

Joint Interpretation. The combined interpretation of these metrics reveals the nature of cross-scale relationships:

- Low JS + high agreement@k: Similar beliefs with consistent top predictions (scale refinement)
- High JS + low agreement@k: Fundamentally different world priors (scale discordance)
- Low JS + low agreement@k: Similar overall distributions but different mode compositions
- High JS + high agreement@k: Shared top predictions but divergent probability allocations

L Experimental Configuration Details

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

L.1 Persona Configuration

We define personas along two demographic dimensions:

Age Groups (5 levels):

- 18–24 (Young)
- 25–34 (Young Adult)
- 35–44 (Adult)
- 45–64 (Middle Age)
- 65+ (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 ($5 \times 4 = 20$).

L.2 Research Question Details

Distributional Stability and Convergence: We perform 20 iterations at multiple departure contexts (IND, FWA, ATL, ORD) using a fixed persona, measuring JS divergence, Jaccard@k stability, and mode–tail decomposition. Extension experiments test convergence at hub contexts and across model sizes.

Sensitivity to Prompt Perturbations: We test perturbations including: (a) persona demographics (age, education), (b) temporal context (season), and (c) geographic anchors (regional vs hub). We compute JS divergence with bootstrap confidence intervals. Seasonality experiments compare summer (June–August) vs winter (December–February) across 4 airports with 20 iterations per condition (160 total runs).

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	235B (MoE)	Large
Mistral	Mistral-7B-Instruct-v0.3	7B	Small
	Mistral-Small-24B-Instruct	24B	Medium
	Mixtral-8x7B-Instruct-v0.1	46B (MoE)	Large

Table 13: Complete model configuration across three families. All models accessed via Together AI API with temperature $T = 0.2$.

Scale Discordance and Generalization: We compare distributions across model sizes using pairwise JS divergence and agreement@k. Cross-family analysis extends to the Qwen and Mistral families to test universality. Cross-domain replication uses restaurant choice (Yelp, 300 Indianapolis restaurants) and product shopping (Amazon, 300 Home & Kitchen products). See Appendix I for dataset details.