More Context Is Not Validation: A Baseline-First Audit of Action-Grounded LLM Listener Agents

Chinmay Rawat[†]
University of Illinois Urbana-Champaign

Taaha Kazi **Vasu Sharma**
Pocket Entertainment

[†] Work conducted while interning at Pocket Entertainment (Pocket FM).

Abstract

LLM agents can generate individualized audience reactions before those reactions are behaviorally valid. We present a baseline-first audit of listener simulation for audio storytelling, separating three questions: whether a representation changes an LLM's predictions, whether the predictions beat a leakage-safe calibrated alternative, and whether the behavioral label measures the intended construct. First, on 1,199 listener–show tasks across four models, evidence-cited action memory reduces Brier error relative to summary personas for every model, but reliable ROC AUC gains appear only for Sonnet (+0.072, 95% bootstrap interval 0.044–0.101) and Opus (+0.058, 0.035–0.082). Next, on a separate 306-task cohort, a five-fold listener-grouped calibrated prior achieves Brier 0.216, compared with 0.235 for a grounded LLM hybrid. Finally, on 150 warm-user tasks, adding historical content digests changes Brier by +0.007 relative to a candidate-digest control and worsens error by +0.0129 (0.0018–0.0246) relative to a raw-candidate arm. A raw-data audit further shows that episode-position events mix listening, autoplay exposure, skipping, and missing playtime telemetry. More context can improve a representation, but does not by itself establish predictive or construct validity.

1 Introduction

An LLM can sound like an audience before it can predict one. Agents conditioned on personas or histories produce coherent individual reactions and can match some aggregate human response patterns (Park et al., 2023; Aher et al., 2023; Argyle et al., 2023). This creates a tempting shortcut for audience research: describe a listener, provide candidate content, and ask whether that listener will continue. But fluent simulation can misrepresent group opinions or depart systematically from human response behavior (Santurkar et al., 2023; Tjuatja et al., 2024; Dillion et al., 2023).

We study this gap in audio storytelling, where later platform events provide a behavioral holdout. Our original 50-case diagnostic used a short summary persona to predict continuation beyond episode position 3. It reached ROC AUC 0.558, lost to simple rules (best ROC AUC 0.638), and had worse probability error than a popularity rule. The sample was too small for a firm ranking, but it exposed a concrete failure: individualized language did not imply individualized prediction. A separate multinomial Naive Bayes study found exploratory signal in episode-opening text for *aggregate* five-minute engagement, following Reddy et al. (2021). That count-based classifier was neither an LLM nor an individual listener agent.

We now replace the proposed action-grounded roadmap with an implemented, multi-stage audit. The contribution is not a claim that action memory solves listener simulation. Instead, we test three nested propositions:

1. Does action memory improve an LLM relative to a summary persona?
2. Does the resulting signal beat a prospective, calibrated non-LLM alternative on the same cohort?

3. Does the event-log target faithfully measure continuation?

The experiments successively weaken an easy success narrative. Stronger models benefit from action memory; a calibrated prior remains the safer probabilistic predictor on a separately frozen cohort; extra historical content does not help; and the target is an episode-position proxy rather than verified consumption.

2 Baseline-first audit protocol

For cutoff t_c, a predictor receives admissible evidence $x_{\leq t_c}$ and returns $p(y_{>t_c} = 1)$ before the later label is exposed. We assess ranking with ROC AUC and probability error with Brier score and log loss. User-clustered resampling or user-grouped folds prevent repeated tasks from masquerading as independent evidence.

baseline-first validation has three gates. **Temporal validity** requires pre-cutoff inputs and a hidden holdout. **Predictive validity** requires beating an inexpensive, prospective alternative on the same target and cohort, not merely another LLM prompt. **Construct validity** requires the logged outcome to measure the claimed behavior. An evaluation-set prevalence is an oracle diagnostic; because it uses holdout labels, it is not a deployable prior. Relative improvement at one gate cannot compensate for failure at another.

3 Representations and target

Summary persona. The original representation compresses pre-cutoff history into minutes-weighted genre, language, and format affinities; listening style; completion; top and abandoned shows; optional demographics; and coverage flags. Missing attributes remain null.

Action memory. The richer representation adds a bounded, recency-ordered sequence of show actions, including episode-position, playtime, recency, and outcome buckets. Behavioral facets cite source record identifiers, and schema checks reject malformed personas. Recent action records are cutoff-bounded. Both arms see the same candidate-show packet. This design draws on history-grounded agents for recommendation and individual simulation (Wang et al., 2024; Park et al., 2024).

Candidate content. Later experiments represent episodes 1–3 either as raw opening text or an evidence-cited digest of pacing, hooks, unresolved questions, and stakes. Historical-show digests allow content-to-content comparison, while fixed prompts forbid treating generic hook quality as proof of personal taste.

Outcome. The binary label is one when maximum observed episode position is at least four during the outcome window. This is initially useful as a reproducible cliff, but Section 7 audits its behavioral meaning.

4 Experiment 1: action memory at scale

4.1 Design

We compare summary and action-memory personas on 1,199 listener–show tasks, one per listener, spanning 136 shows at natural 0.395 positive prevalence. Haiku 4.5, Gemini 3.5 Flash, Sonnet 5, and Opus 4.8 each score both arms. Table 1 uses paired complete cases and 2,000 paired bootstrap resamples. One task per user makes row and listener resampling equivalent.

4.2 Results

Action memory improves discrimination for Sonnet and Opus; Haiku and Gemini intervals include zero. It reduces Brier error relative to the corresponding summary arm for all four models: differences are $-.037$, $-.023$, $-.063$, and $-.045$, respectively, with paired

Table 1: Paired summary (S) versus action-memory (A) results. Δ is A minus S; parentheses are 95% intervals for $\Delta ROCAUC$. Lower Brier is better.

		ROC AUC			Brier	
Model	N	S	A	Δ (95% CI)	S	A
Haiku	1,199	.495	.502	+.007 (−.016, .029)	.317	.280
Gemini	1,199	.482	.497	+.014 (−.007, .037)	.404	.381
Sonnet	1,197	.500	.572	+.072 (.044, .101)	.312	.249
Opus	1,197	.500	.558	+.058 (.035, .082)	.301	.256

bootstrap intervals excluding zero. An exploratory warm-user slice, restricted to users with pre-cutoff listening actions, strengthens the representation pattern: action memory improves ROC AUC by .015 for Haiku ($N = 809$), .039 for Gemini ($N = 809$), .063 for Sonnet ($N = 808$), and .057 for Opus ($N = 808$); intervals exclude zero except for Haiku. Cold-start rows show no comparable Sonnet or Opus gain. This supports an interpretation based on usable behavioral history rather than a universal action-memory effect.

The study nevertheless answers only proposition 1. Shared account-level summary fields were not independently snapshotted at the fixed history cutoff; recent actions pass the cutoff audit and holdout labels remain hidden, so the paired design is informative about representation sensitivity but not a clean end-to-end temporal validation. No same-cohort, cross-fitted calibrated baseline was run. The 0.395 evaluation prevalence is therefore reported only as an oracle diagnostic, not as the bar the agent cleared.

5 Experiment 2: calibration stress test

We next use a separately frozen 306-task cohort (79 listeners, 181 shows, positive rate .307) with an explicit prediction date, observation endpoint, and 31-day outcome window. A horizon-matched input prior, deterministic behavioral features, candidate-content digest features, and bounded grounded LLM signals are combined using five-fold listener-grouped cross-fitting: a listener never appears in both train and test within a fold.

Table 2: Calibration stress test on 306 tasks. All "cross-fit" rows are listener-grouped out-of-fold predictions.

Predictor	ROC AUC	Brier
Cross-fit calibrated prior	.549	**.216**
Cross-fit prior + behavior	.542	.217
Cross-fit prior + behavior + digest	.512	.228
Grounded LLM hybrid	.548	.235
Legacy clean persona diagnostic	**.615**	.230
Legacy full persona diagnostic	.615	.234
Compact two-stage LLM	.564	.251
Compact direct raw-read LLM	.556	.254

Table 2 separates ranking from calibration. The grounded hybrid recovers about the same ROC AUC as the cross-fitted prior (.548 versus .549) but has materially worse Brier error (.235 versus .216). The paired listener-clustered Brier difference for calibrated prior minus the original input prior is $-.0227$, with 95% interval $[-.0426, -.0028]$. Adding deterministic behavior or digest features does not improve on that calibrated prior.

The lower half is an architecture diagnostic, not a causal ablation. The legacy and compact arms differ in persona fields, prompt contract, intermediate summarization, and raw-reading behavior. Their ROC AUC gap shows that interface choices can move results, but cannot identify which choice caused the movement. More importantly, the best-ranking

legacy arm still has worse probability error than the calibrated prior. Proposition 2 is not satisfied.

6 Experiment 3: does more history help?

On a frozen 150-task warm-user cohort, we hold the structured user evidence fixed and vary candidate and historical content. The candidate-plus-history arm has historical digests on 146 rows and cites them on every available row. Table 3 shows that context volume is not a monotonic quality lever.

Table 3: Content-history experiment on 150 warm-user tasks.

Arm	ROC AUC	Brier
Input prior (no LLM call)	.575	**.238**
Structured user + raw candidate	**.598**	.242
Structured user + candidate digest	.573	.248
Candidate + historical digests	.580	.255

Historical digests minus the candidate-digest control change Brier by $+.0070$, with listener-clustered interval $[-.0004, .0139]$. Compared with the raw-candidate arm, history worsens Brier by $+.0129$ $[.0018, .0246]$. The treatment raises probabilities by .0196 on average: this helps the 45 positives but hurts the 105 negatives by the same aggregate Brier amount, consistent with overprediction. The result does not prove historical content is intrinsically harmful; it rejects the simpler assumption that adding plausible, cited context reliably improves behavioral prediction.

7 Construct and data-quality audit

The final gate fails upstream of modeling. In the source export, 11.3% of rows record under three seconds of daily playtime; 10.5% of recent action-memory entries have zero minutes. Repeated zero-second events can be autoplay-queue exposures, while separate outages can freeze playtime as episode position advances. Maximum position can also jump through skipping. Thus position ≥ 4 mixes exposure, navigation, missing telemetry, and actual listening. It cannot by itself establish consumption or preference.

A separate 318-task baseline audit across 80 users reinforces the uncertainty. Ordinary pre-cutoff user and category predictors obtain ROC AUC values around .50–.55, with listener-clustered intervals all including .5; out-of-fold logistic baselines have Brier near .24. This does not prove impossibility, but it provides little evidence that the current episode-position cliff is strongly rankable from conventional metadata.

A defensible next target should require a qualified exposure, measurable playtime, and an inactivity gate before labeling a drop. Exposure, start, continuation, and preference should be evaluated separately. The resulting task should be replicated on a later temporal split with episode-level telemetry and a fully snapshotted feature store.

8 Discussion and limitations

The experiments produce three practical lessons. First, raw behavioral evidence can matter, but its benefit depends on model capability and history availability. Second, calibrated priors are not ceremonial baselines: they can be superior probability estimators even when an LLM ranks some cases better. Third, target validity is upstream of agent validity. A precise score against an ambiguous event is not a precise statement about a listener.

The studies use different frozen cohorts and prompt families, so their metrics should not be compared as a single leaderboard. The platform data are observational and recommendation affects exposure. The scale study has shared metadata timing limitations; the 306-task

architecture comparison is multi-factor; and the warm-history experiment is a single finite cohort. Private data limit reproducibility and external validity. We make no causal claim about why action memory helps and no claim that these agents should replace audience research or score individual users.

9 Conclusion

Action memory makes some LLM listener predictions better than summary personas, but more context is not validation. On the audits where the comparison is available, calibrated non-LLM prediction remains difficult to beat, and extra historical content can increase error. The underlying event label also overstates what telemetry can support. Behavioral simulation should earn trust only after clearing temporal, predictive, and construct-validity gates on the same frozen task.

References

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

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.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023. doi: 10.1016/j.tics.2023.04.008.

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. *arXiv preprint arXiv:2304.03442*, 2023. URL https://arxiv.org/abs/2304.03442.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024. URL https://arxiv.org/abs/2411.10109.

Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. Modeling language usage and listener engagement in podcasts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 632–643. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.52. URL https://aclanthology.org/2021.acl-long.52/.

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. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

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.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model-based agents for recommender systems. *ACM Transactions on Information Systems*, 2024. doi: 10.1145/3708985.

Ethics Statement

We report aggregate statistics and do not include raw listening histories, user or device identifiers, per-show performance, or script excerpts. Missing personal attributes remain missing. Simulated reactions can stereotype listeners, and uncertain outcome semantics can turn platform artifacts into apparently personal judgments. We treat these systems as research diagnostics, not replacements for real audience research, human judgment, or decisions about individual users.

LLM Usage

LLMs were used as writing and editing aids and to assist implementation of experiment code from author-specified designs. The authors designed the study and verified claims, numbers, citations, and references against the underlying artifacts. The listener agents evaluated in the experiments are objects of study, not authoring tools.

Reproducibility Statement

The listener data are private platform records and cannot be released. We report task definitions, cutoffs, cohort sizes, model aliases, prompt-family differences, metrics, folds, and resampling procedures. Internally, schema-validated predictions, prompt audits, frozen labels, and aggregate scoreboards are versioned. De-identified schemas, prompt templates, and evaluation code may be released where permitted, but event-level histories cannot be shared.