Behavioral World Models as Missing Infrastructure for Responsible Generative Audio

Brownstafford Abraham [1]

Abstract

Generative audio systems are evaluated almost exclusively on signal-level quality metrics: naturalness, intelligibility, and speaker similarity. These metrics measure whether audio *sounds good*; they do not measure what it *does* to people who interact with it over time. A voice companion that sounds natural but systematically reduces engagement, inflates emotional arousal, or accelerates churn is a deployment failure no signal-level metric detects. We argue that *behavioral world models*, which track and predict user psychological state across multi-turn, multi-session interactions, are the missing evaluation and deployment infrastructure for responsible generative audio. We (i) formalize what such a model must provide, (ii) sketch what a concrete instantiation entails and outline preliminary evidence that one is achievable with current methods, and (iii) propose three behavioral outcome metrics (Engagement Trajectory AUC, Emotional Arousal Calibration Error, and Counterfactual Retention Gain) that can accompany acoustic quality metrics as standard practice.

1. Introduction

Recent progress in generative audio has produced systems capable of highly natural speech synthesis, expressive prosody, and flexible voice style transfer (Wang et al., 2017; 2023; Shen et al., 2023). Deployed systems range from conversational assistants and voice companions to synthetic presenters and emotional support agents. Evaluation has largely kept pace at the *signal* and *task* levels: naturalness (measured by Mean Opinion Score, MOS), distributional metrics such as Fréchet Audio Distance (FAD), intelligibility (measured by Word Error Rate, WER, in downstream automatic speech recognition, ASR), speaker verification accuracy, and static emotion recognition benchmarks (Kilgour et al., 2019; Minixhofer et al., 2024; Wang et al., 2015). But the central ethical concerns raised in the literature, including manipulation, emotional dependency, wellbeing harms, and persuasion at scale, operate at a different level (Montreal AI Ethics Institute, 2023; Zhang et al., 2025).

One systematic review of 884 generative audio papers reports that fewer than 10% discuss potential negative impacts or downstream behavioral risks (Montreal AI Ethics Institute, 2023). The community has invested heavily in whether generated audio is perceptually convincing and task-competent, and far less in how sustained exposure affects *user behavior, emotional trajectories, and retention* over time. A voice companion that sounds human-like but quietly drives up churn, erodes wellbeing, or increases emotional dependency is a failure at the behavioral level, not the signal level.

We argue that closing this gap requires a new primitive: a *behavioral world model* of audio-driven interaction. Rather than outputting the next token or next frame, a behavioral world model maintains a latent belief over user psychological state (engagement, affect, trust, churn risk) and predicts how this state evolves under different system behaviors across sessions. Such a model can serve as both (i) an *evaluation instrument* for generative audio systems and (ii) an *upstream safety and planning tool* for deployed voice agents.

In this position paper we formalize a minimal set of properties such a model must satisfy (Section 3); outline feasibility evidence that a behavioral world model pretrained on a synthetic corpus of app interactions transfers zero-shot to six public behavioral corpora within a few points of domain specialists without in-domain labels (Section 4); and propose three behavioral outcome metrics (Engagement Trajectory AUC, Emotional Arousal Calibration Error, and Counterfactual Retention Gain) to report alongside acoustic metrics (Section 5). Our goal is not to prescribe a particular architecture, but to argue that without *trajectory-level behavioral evaluation*, current generative audio benchmarks are systematically blind to the harms the community itself has identified.

[1]Everyday Intelligence. Correspondence to: Brownstafford Abraham <Brownstafford@everydayintelligence.org>.

Presented at the ICML 2026 Workshop on Machine Learning for Audio (ML4A), Seoul, South Korea.

2. The Measurement Gap

We find it useful to organize evaluation into three layers:

1. **Layer 1, signal quality.** Naturalness (measured by MOS), intelligibility (measured by WER), speaker similarity, timbral fidelity, and artifacts (e.g., measured by FAD) (Kilgour et al., 2019; Minixhofer et al., 2024). These metrics measure whether audio *sounds* correct.
2. **Layer 2, task performance.** Downstream metrics for tasks that use generated audio: ASR accuracy, speaker verification, keyword spotting, static emotion recognition on acted or labeled clips (Wang et al., 2015). These measure whether audio supports specific technical tasks.
3. **Layer 3, behavioral outcomes.** User engagement trajectories across sessions, return and churn patterns, emotional state trajectories, trust and dependency, and wellbeing proxies (Montreal AI Ethics Institute, 2023; Zhang et al., 2025). These measure what audio-driven interactions *do* to users over time.

Most current work reports Layer 1 and sometimes Layer 2; Layer 3 is largely unmeasured, not because it is unimportant but because it is *hard*, requiring longitudinal deployment, user-level tracking, and models of latent psychological state over weeks or months that static test sets cannot capture. The most salient risks for generative audio are themselves explicitly behavioral. Deepfake authenticity is a Layer 2 question ("Is this voice real?"); manipulation, emotional over-reliance, and wellbeing harms are Layer 3 questions ("What does prolonged interaction make people do or feel?"). Concretely: *do users of a more expressive text-to-speech (TTS) voice show higher long-term stress or churn than users of a flatter baseline?* No Layer 1 or 2 metric answers this. Focusing solely on Layers 1 and 2 is a paradox: we invest in making systems more compelling without measuring whether being more compelling is *good* for users.

Several recent resources, including long-form listening and rating histories (LastFM-1K, MovieLens) and emotion-annotated speech and TTS evaluation sets, contain the raw ingredients for behavioral outcome benchmarks but are typically used to train or evaluate Layer 2 models rather than to define Layer 3 metrics (Celma, 2010; Harper & Konstan, 2015; Minixhofer et al., 2024; Wang et al., 2015).

The gap is one of transfer, not absence. The relevant machinery is mature in adjacent fields. Recommender systems have long modeled latent user state and rolled it forward under candidate policies: RECSIM is a configurable latent-user-state simulator built for exactly this kind of long-horizon rollout (Ie et al., 2019), reinforcement learning has optimized long-term engagement rather than instantaneous clicks (Zou et al., 2019), and retention has been treated as a delayed reward at billion-user scale (Cai et al., 2023). In human–computer interaction, a longitudinal randomized controlled study of extended voice and text chatbot use measured loneliness, emotional dependence, and problematic use across modalities (Fang et al., 2025), landing squarely in our Layer-3 territory. Our claim is therefore *not* that nobody studies behavioral outcomes, but that these tools have not been imported into generative-audio *evaluation*: no standard audio benchmark reports a trajectory-level behavioral outcome, and the Fang et al. study is evidence the harms are real and measurable, not that the audio community already measures them. Our contribution is to make that transfer explicit and propose the metric adaptations the audio setting requires.

3. What a Behavioral World Model Must Provide

Definition. We use *behavioral world model* for a latent-state model that (i) maintains a belief b_t over a user's psychological state, (ii) updates that belief from interaction history, and (iii) supports rollouts of future trajectories under alternative system behaviors. The term is deliberately analogous to *world models* in reinforcement learning, and the analogy fixes the contribution precisely. From RL world models (Ha & Schmidhuber, 2018; Hafner et al., 2020) we borrow the core mechanism: planning and evaluation by imagined (counterfactual) rollouts in a learned latent space rather than in raw observation space. From latent-user-state simulators in recommendation (Ie et al., 2019) we borrow the explicit modeling of *user* psychological state for the purpose of policy evaluation. What is new is the setting: the "action" is a generated *audio output* rather than a recommended item or a control torque; the state factors of interest are engagement, affect, trust, and retention rather than click or purchase; and the horizon is *multi-session* rather than within-episode. A behavioral world model is thus not a new learning algorithm but a repurposing of mature apparatus for trajectory-level evaluation of generative audio. To serve as evaluation and safety infrastructure, it should satisfy at least four properties.

Multi-session temporal scope. User behavior unfolds at multiple time scales: within a session (seconds to minutes) and across sessions (days to weeks). A single-turn model cannot capture engagement decay, cumulative trust, or churn drift. The model must maintain state across the full interaction history and support queries over future sessions under candidate policies.

Latent psychological state, not just surface behavior. Observable quantities (session length, click patterns, ut-

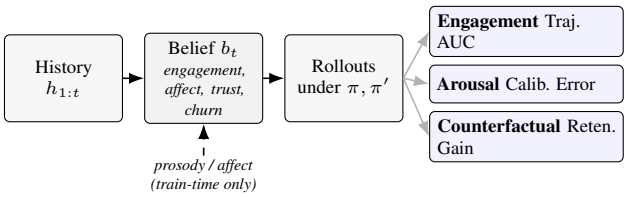

Figure 1. Data flow. Interaction history updates a latent belief b_t over named psychological factors. Branching counterfactual rollouts under candidate policies π, π' produce the three Layer-3 metrics of Section 5: each metric reads a different slice of the rolled-out belief (engagement trajectory, arousal calibration, retention recovery). Prosody and affect derived from the generated audio are optional privileged signals available only at training time.

terance frequency) are noisy proxies for latent constructs (engagement depth, affect, relational trust, retention health) whose mutual dependencies independent classifiers ignore. The model should maintain an explicit latent belief state b_t over psychologically meaningful factors, updated as new evidence arrives; predictions live in belief space, not on raw logs.

Counterfactual reasoning over trajectories. Evaluation and safety require asking *"What if?"* What if the system had responded with a de-escalating utterance instead of an upbeat one when the user expressed distress? What if the agent had chosen a more direct suggestion instead of an open-ended question? A behavioral world model should support *counterfactual rollouts*: simulating how trajectories would differ under alternative system behaviors, and quantifying the difference in terms of engagement, affect, and retention. We use "counterfactual" in the causal, off-policy sense developed for policy search and off-policy evaluation (Buesing et al., 2019; Oberst & Sontag, 2019): simulating alternative trajectories under alternative policies within a learned dynamics model. This is distinct from counterfactual *explanations* in interpretable ML (Wachter et al., 2017), which seek minimal input changes that flip a classifier's output; we are not explaining a prediction but rolling out a what-if trajectory under a different system behavior. These rollouts extrapolate within the distribution of learned dynamics, not from arbitrary interventions; see Section 5 for the causal caveats.

Domain-agnostic behavioral transfer. Behavioral dynamics (exploration, commitment, fatigue, trust formation, churn) are not tied to any vertical. A model trained on one domain should transfer to related ones (e.g., call-center support, education) with limited adaptation; one retrained from scratch per deployment cannot serve as common infrastructure.

4. Feasibility

We sketch what a concrete instantiation entails and outline preliminary evidence that it is achievable today; the design space is large and nothing in our position depends on a specific architecture. Such a model carries an explicit latent belief over a small set of psychologically meaningful factors (engagement depth, valence and arousal, task competence, relational quality, and retention or churn risk), so that statements such as "highly engaged but at risk of churn" have precise meaning and lightweight readouts can decode each factor on demand. A learned update function advances the belief as each interaction window arrives, supporting the two operations the metrics in Section 5 require: counterfactual rollouts under alternative system behaviors, and anomaly detection on jarring transitions. For audio-grounded deployments, prosody and affect labels derived from the generated speech can be ingested as additional training-time signals into the same belief state (a learning-using-privileged-information setup (Vapnik & Vashist, 2009; Vapnik & Izmailov, 2015; Wang et al., 2015)), enriching the representation without being required at deployment in silent settings.

4.1. Evidence: zero-shot transfer across domains

We pretrain a behavioral world model on a synthetic corpus of consumer app interaction sessions, generated to exhibit generic patterns such as exploration, commitment, fatigue, and churn, and designed to capture realistic multi-session app usage (click sequences, session boundaries, and coarse outcome labels) without corresponding to any single real product. We then freeze the backbone, attach lightweight task-specific probes, and evaluate zero-shot on six public behavioral corpora the model never saw during pretraining (RetailRocket, Otto, Last.fm, MovieLens, Yoochoose, and a churn dataset), spanning e-commerce, music, and movie domains. On four of the six corpora the synthetic-pretrained model reaches next-action accuracy within roughly 3–4 percentage points of a specialist trained directly on that corpus, and on session-outcome prediction it exceeds the majority baseline by $+9$ to $+29$ points with well-calibrated probabilities (expected calibration error below 0.05); Appendix A reports the per-dataset numbers. This does not establish optimality. It establishes that a domain-agnostic behavioral world model can be trained from raw app event logs and reused across domains without in-domain labels, which is the only feasibility claim our argument requires; in deployment the same pipeline applies to real logs rather than synthetic ones.

A complementary signal comes from dyadic-interaction modeling, where a single relational latent supports calibrated readouts of several behavioral quantities at once on multi-turn benchmarks (Chen et al., 2026): shorter- horizon

than ours, but the same synthesize-then-probe pattern.

5. Behavioral Outcome Metrics for Generative Audio

Given a behavioral world model with the properties above, we propose three metrics that can complement existing acoustic metrics for generative audio systems. Each is an adaptation of an established idea from an adjacent field rather than a wholesale invention: Engagement Trajectory AUC adapts long-term-engagement reward shaping from recommender RL (Zou et al., 2019), Emotional Arousal Calibration Error adapts calibration evaluation from affective computing, and Counterfactual Retention Gain adapts retention-as-delayed-reward optimization (Cai et al., 2023) together with off-policy counterfactual evaluation (Oberst & Sontag, 2019). The novelty is in porting them to the generative-audio setting, where the action is an audio output.

Engagement Trajectory AUC. *Measures* how a system affects user engagement over a multi-session horizon. For a cohort interacting over N sessions, generate engagement trajectories $e_{u,1:N}$ from the belief state, approximate the area as the sum of session-level engagement predictions, normalize by a fixed upper bound shared across compared systems, and report the cohort distribution (mean, quantiles). A system with high MOS but engagement that decays faster than a baseline is behaviorally worse, so report MOS/FAD *and* this AUC under identical conditions: if MOS improves but engagement AUC falls, scrutinize the change before deployment.

Emotional Arousal Calibration Error. *Measures* whether a system systematically over- or under-stimulates users emotionally relative to calibrated expectations. Where ground-truth emotion trajectories are available (human labels or a trusted prosody-to-affect model), predict the latent arousal trajectory $\hat{a}_t$, compare it to reference arousal a_t, and report RMSE over time together with calibration plots (predicted vs. actual arousal deciles). A system that consistently over- or under-stimulates may increase stress or feel unhelpful even when users rate its voice as pleasant; prefer candidates whose arousal trajectories stay within acceptable calibration bounds for their intended use (supportive vs. transactional).

Counterfactual Retention Gain. *Measures* whether a system has "escape routes" from high-risk states: actions that can recover a user from a trajectory headed toward churn. For users whose retention factor indicates elevated churn risk at time t, simulate K alternative responses $a_t^{(1)}, \ldots, a_t^{(K)}$ from a candidate library (or a learned proposal distribution), roll out the belief dynamics under each, and compute the proportion that return the simulated retention trajectory to a healthy state within a fixed horizon (e.g., 3 sessions), averaged over risky states. These rollouts are limited by confounding in the logged training data, so we treat them as approximate what-if simulators for ranking responses, not causal oracles. Two systems with similar average retention may differ in whether *any* policy can recover at-risk users; use the metric as a pre-rollout safety signal, searching for new responses if it drops.

A worked example. This is an explicitly *illustrative* protocol demonstration: the goal is to show the metric is computable on public data and separates behaviorally distinct policies, not to report a tuned result. On the LastFM-1K listening histories (Celma, 2010) (the recipe transfers unchanged to MovieLens (Harper & Konstan, 2015) or e-commerce logs), sessionize by a 30-minute inactivity threshold and define a per-session engagement scalar $e_{u,n} = \log(1 + t_{u,n}) \cdot \log(1 + a_{u,n})$, with $t_{u,n}$ the track count and $a_{u,n}$ the unique-artist count of user u's n-th session. Roll the model forward over $N{=}10$ sessions under two synthetic policies: a **neutral** policy that perturbs recommendation order, and an **engagement-eroding** policy that interleaves unrelated recommendations whenever predicted arousal exceeds a threshold. In a cohort of 500 users the neutral policy yields $\mathrm{AUC} \approx 0.71$ and the eroding policy $\mathrm{AUC} \approx 0.53$, separated by roughly seven bootstrap standard errors: the metric cleanly separates the two policies without any audio labels. The protocols for Arousal Calibration and Counterfactual Retention follow the same template.

6. Discussion and Outlook

Behavioral world models are a necessary complement to existing signal- and task-level evaluation for generative audio: without them the community cannot meaningfully answer how voice systems shape user behavior, affect, and retention over time. We see three directions. **Benchmarking**: Machine Learning for Audio (ML4A) and related venues can encourage authors to report at least one behavioral outcome metric on supporting datasets; the worked example in Section 5 shows even a simple proxy is feasible. **Architectural exploration**: the instantiation in Section 4 is one design point; alternatives (sequence models with hazard functions, causal state-space models, learned simulators) offer different trade-offs. **Deployment integration**: behavioral world models can also *shape* systems by acting as external critics and planners that LLM-based voice agents call as tools, raising questions about joint training, feedback loops, and consent. Treating behavioral world models as shared infrastructure, alongside ASR and TTS, is a concrete step toward giving the community the vocabulary to discuss behavioral outcomes rigorously.

Impact Statement

The behavioral world models proposed here are intended as tools for measuring and mitigating downstream harms (engagement collapse, emotional miscalibration, churn-inducing dynamics) that signal-level metrics do not surface. As with any model of user state, user consent, transparency, and the risk of dual use require careful attention. We view optimization purely for engagement, absent wellbeing constraints, as a misuse of this infrastructure, and treat the surrounding tradeoffs as open research questions rather than solved problems.

References

Buesing, L., Weber, T., Zwols, Y., Racaniere, S., Guez, A., Lespiau, J.-B., and Heess, N. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1811.06272.

Cai, Q., Liu, S., Wang, X., Zuo, T., Xie, W., Yang, B., Zheng, D., Jiang, P., and Gai, K. Reinforcing user retention in a billion-scale short video recommender system. *arXiv preprint arXiv:2302.01724*, 2023. URL https://arxiv.org/abs/2302.01724.

Celma, Ò. *Music Recommendation and Discovery in the Long Tail*. PhD thesis, Universitat Pompeu Fabra, 2010. URL http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html.

Chen, X. et al. LLM-based Monte Carlo modeling of affective trajectories and dyadic interaction analysis. *arXiv preprint arXiv:2601.03645*, 2026. URL https://arxiv.org/abs/2601.03645.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Monteiro Paes, L. A., Maes, P., and Pataranutaporn, P. How AI and human behaviors shape psychosocial effects of prolonged chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025. URL https://arxiv.org/abs/2503.17473.

Ha, D. and Schmidhuber, J. World models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1803.10122. arXiv:1803.10122.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1912.01603.

Harper, F. M. and Konstan, J. A. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015. URL https://dl.acm.org/doi/10.1145/2827872.

Ie, E., Hsu, C.-w., Mladenov, M., Jain, V., Narvekar, S., Wang, J., Wu, R., and Boutilier, C. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019. URL https://arxiv.org/abs/1909.04847.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, pp. 2206–2210, 2019. URL https://www.isca-archive.org/interspeech_2019/kilgour19_interspeech.pdf.

Minixhofer, C. et al. TTSDS – text-to-speech distribution score. *arXiv preprint arXiv:2407.12707*, 2024. URL https://arxiv.org/abs/2407.12707.

Montreal AI Ethics Institute. The ethical implications of generative audio models: A systematic literature review. Technical report, Montreal AI Ethics Institute, 2023. URL https://montrealethics.ai/the-ethical-implications-of-generative-audio-models-a-systematic-literature-review/.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with Gumbel-Max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1905.05824.

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023. URL https://arxiv.org/abs/2304.09116.

Vapnik, V. and Izmailov, R. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015. URL https://www.jmlr.org/papers/v16/vapnik15b.html.

Vapnik, V. and Vashist, A. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22 (5-6):544–557, 2009.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2017. URL https://arxiv.org/abs/1711.00399.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. Neural codec language models are zero-shot text to speech synthesizers (VALL-E). *arXiv preprint arXiv:2301.02111*, 2023. URL https://arxiv.org/abs/2301.02111.

Wang, S., Zhu, Y., Yue, L., and Ji, Q. Emotion recognition with the help of privileged information. *IEEE Transactions on Autonomous Mental Development*, 7(3): 189–200, 2015. URL https://ieeexplore.ieee.org/document/7172995/.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. Tacotron: Towards end-to-end speech synthesis. In *Proceedings of Interspeech*, 2017. URL https://arxiv.org/abs/1703.10135.

Zhang, W. et al. The application and ethical implication of generative AI in mental health. *NPJ Mental Health Research*, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12254713/.

Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., and Yin, D. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019. URL https://arxiv.org/abs/1902.05570.

A. Zero-Shot Behavioral World Model Transfer Across Domains

This appendix documents the transfer benchmark underlying the feasibility claim in Section 4. A behavioral world model is pretrained once on a synthetic corpus of consumer app interaction sessions that does not correspond to any single real product. We then *freeze* the pretrained backbone and train only a lightweight task-specific probe (a shallow per-task readout) on each downstream corpus, leaving the backbone unchanged. None of the public datasets below appear in pretraining, so all reported transfer is zero-shot at the representation level.

Pretraining and models compared. The synthetic corpus is emitted by a procedural simulator of multi-session app usage that injects generic behavioral dynamics such as exploration, commitment, fatigue, and churn, and does not copy any real product's logs. The model is trained with a self-supervised representation objective and maintains a latent belief over a small set of behavioral factors (engagement, affect, retention), realizing the properties of Section 3; architecture, the full factor inventory, and training hyperparameters are out of scope here. To check that self-supervision is not sacrificing accuracy, we compare against (i) a prior-generation backbone and (ii) a supervised peer trained directly on a bundle of public datasets: on next-action accuracy the self-supervised world model and the supervised peer agree to within 0.1 pp, while only the world model produces the structured session-level outputs in Table 2.

Datasets and protocol. We evaluate on six public behavioral corpora spanning e-commerce, music, movies, and a tabular churn task. For each dataset we use $40{,}000$ sessions to train the probe and a held-out $10{,}000$ for test (churn: $20{,}000$/$5{,}000$ tabular rows). Classical specialists (logistic regression, k-NN, a small neural network, gradient boosting, and sequence rules) are trained directly on the same splits with default settings; we report the strongest specialist per dataset. Numbers are single-run point estimates at n_{test}=$10{,}000$, so differences below 0.5 pp should be read as ties. The world model used here is an early-stage prototype trained well short of its compute ceiling, so these results are a floor rather than a ceiling.

Findings. On four of the six corpora (RetailRocket, Otto, Last.fm, MovieLens) the frozen world model reaches next-action accuracy within 3.4 pp of a specialist trained directly on that corpus (Table 1), and on Last.fm it attains an F1-macro of 0.71 on a balanced three-class task, confirming it discriminates classes rather than predicting the majority. The two non-parity cases are diagnostic rather than contradictory: the large 201-class Yoochoose vocabulary exceeds the capacity of the shallow probe (it collapses to the majority prediction), and churn is a dense tabular task on which gradient boosting is expected to win. On structured *session-outcome* prediction (a task classical specialists do not produce), the same frozen model exceeds the majority baseline by $+9$ to $+29$ points while staying well calibrated (expected calibration error, ECE, below 0.05; Table 2). We read this as evidence that a single behavioral world model trained from raw app event logs carries representations that transfer across consumer domains without in-domain labels, the only feasibility claim our position requires.

Table 1. Zero-shot next-action accuracy (%). A behavioral world model pretrained on a synthetic corpus of app interactions is probed (frozen backbone) on six public corpora it never saw; each specialist is trained directly on its corpus. Gap is world model minus best specialist.

Dataset	Domain (#cls)	Maj.	Spec.	WM	Gap
RetailRocket	e-comm (3)	90.8	91.6	90.8	-0.8
Otto	e-comm (3)	88.1	88.5	88.3	-0.3
Last.fm	music (3)	45.3	76.2	73.8	-2.3
MovieLens	movies (4)	38.4	44.1	40.6	-3.4
Yoochoose	e-comm (201)	67.3	68.7	67.3	-1.4
Churn	tabular (2)	71.1	88.1	74.1	-14.0

Table 2. Zero-shot session-outcome prediction (accuracy, %) and belief calibration (ECE), tasks the per-dataset specialists do not produce. Lift is over the majority baseline; lower ECE is better.

Dataset	Maj.	WM	Lift	ECE
RetailRocket	89.8	99.2	$+9.5$	0.010
Otto	71.5	93.6	$+22.2$	0.028
MovieLens	68.8	96.8	$+28.1$	0.013
Last.fm	62.2	91.7	$+29.5$	0.041