Hierarchical World Models for Strategic AI Agents: Bridging Simulation and Reality through Multi-Fidelity Learning

Mysore Supreeth[1]*, Atik Faysal[2], Manish Mehta[3], Sunil Kothari·
[1]Principal Architect, Intel Labs [2]PhD Candidate, Rowan University ·AI/ML Researcher
dr.mysores@gmail.com Atikfaysal@rowan.edu manishmehta@gmail.com
* Corresponding author

Abstract

World models allow AI agents to anticipate consequences of actions before committing to them, yet agents trained in simulation consistently exhibit a gap between tactical proficiency and strategic judgment. We formalize this phenomenon as the **Strategic Simulation Gap** (SSG)—the performance difference between simulated and deployed outcomes for decisions involving long-horizon reasoning under human behavioral uncertainty. We introduce **Hierarchical Multi-Fidelity World Models** (HMF-WM), a three-tier architecture decomposing world modeling into physical dynamics, human behavioral response prediction, and strategic outcome forecasting. A central result (Theorem 1) shows that the SSG is lower-bounded by a quantity we call the *Behavioral Adaptation Coefficient*, capturing how much human behavior shifts upon AI deployment—an idea drawn from the Lucas Critique in economics. Our *Reality-Anchored Training* (RAT) algorithm dynamically adjusts the simulation-to-real data mixture using domain verifiability as a guide. Across bug routing ($v \approx 0.8$), customer escalation prediction ($v \approx 0.4$), and medical triage ($v \approx 0.3$), HMF-WM improves strategic accuracy by $+4.3$, $+8.7$, and $+8.3$ percentage points over the best baseline respectively (mean $+7.1$pp; all $p < 0.001$, paired t-test, Bonferroni-corrected), with $2.1\times$ sample efficiency (95% CI: $[1.8\times, 2.4\times]$).

1 Introduction

Investment in world models—neural networks that learn environment dynamics for planning and counterfactual reasoning—has accelerated sharply. Recent systems span autonomous driving (Hu et al., 2023), interactive environment generation (Bruce et al., 2024), and physics-grounded simulation (Agarwal et al., 2025); see Ding et al. (2025) and Hafner et al. (2023) for surveys. Yet deployed agents share a common failure pattern: they execute low-level actions competently while making poor high-level decisions.

An AI customer service agent, for instance, may navigate CRM interfaces without error yet consistently misjudge when to escalate a conversation or how to adapt its tone for a frustrated user. Dell'Acqua et al. (2023) document this pattern—which they term a "jagged technological frontier"—across knowledge-work domains. We observe analogous behavior in code review (accurate linting but poor architectural judgment) and clinical decision support (correct symptom extraction but unreliable triage).

We argue that this gap stems from how world models are trained. Standard objectives optimize next-state prediction, but strategic decisions depend on *outcome prediction under intervention*: forecasting what happens when the AI's own actions alter the behavior of the humans it interacts with. The Lucas Critique from macroeconomics makes an analogous point: econometric relationships estimated from historical data break down when the policy regime changes (Lucas Jr, 1976). We adapt this insight by defining a measurable quantity—the Behavioral Adaptation Coefficient—that captures the distributional shift in human behavior triggered by AI deployment.

Our contributions: (1) We formalize the Strategic Simulation Gap and prove a lower bound tied to behavioral adaptation (§3–4). (2) We introduce HMF-WM, a three-tier world model with Reality-Anchored Training (§5). (3) We derive conditions under which simulation data helps versus hurts strategic learning (Theorem 2). (4) We validate on three domains, obtaining a mean improvement of $+7.1$pp over the best baseline ($p < 0.001$), and release the HMF-Bench evaluation suite (§6).

2 Background and Related Work

World models learn transition dynamics $P(s_{t+1}|s_t, a_t)$ to support model-based RL. The Dyna architecture (Sutton, 1991) introduced this idea; subsequent neural implementations include DreamerV1–V3 (Hafner et al., 2020; 2023) and TransDreamer (Chen et al., 2022). Industrial-scale world models now generate interactive 2D environments from unlabeled video (Bruce et al., 2024) and action-conditioned driving simulation (Hu et al., 2023). Our hierarchical

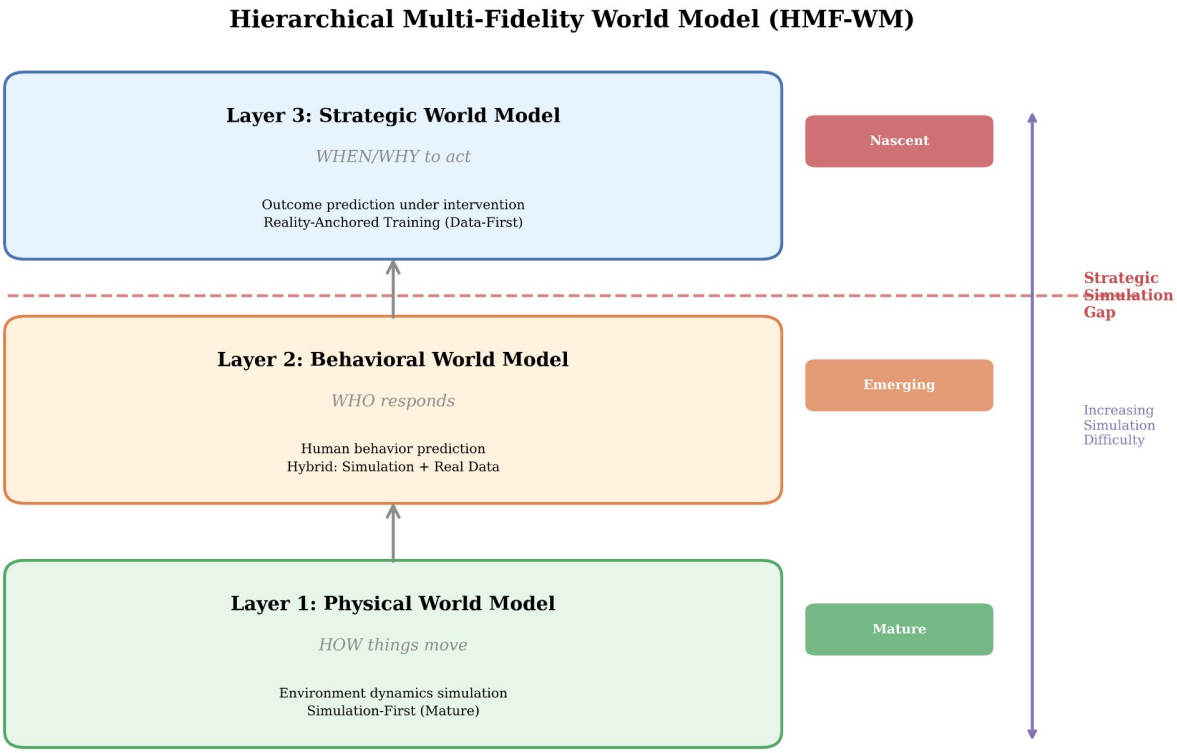

Figure 1: **HMF-WM architecture.** Layer 1 captures physical dynamics (well-modeled by simulation). Layer 2 predicts human behavioral responses (requires hybrid data). Layer 3 forecasts strategic outcomes (demands reality-anchored learning). The dashed red line marks the Strategic Simulation Gap.

decomposition is motivated by temporal abstraction in the options framework (Sutton et al., 1999) and feudal learning (Vezhnevets et al., 2017), but targets fidelity stratification rather than temporal granularity.

Sim-to-real transfer remains a persistent challenge. Domain randomization (Tobin et al., 2017; Peng et al., 2018) and system identification have reduced the physics gap to below 3% for rigid-body locomotion (Akkaya et al., 2019), but the gap for tasks requiring human interaction modeling is substantially larger. LLMs have been proposed as human behavior simulators (Argyle et al., 2023; Park et al., 2023), but systematic evaluations reveal shortfalls: Santurkar et al. (2023) show that LLM-generated opinions cluster around narrow demographic profiles, and Chen et al. (2023) find inconsistent chain-of-thought reasoning in vision-language models. On the algorithmic side, offline RL methods—CQL (Kumar et al., 2020), IQL (Kostrikov et al., 2022), Decision Transformer (Chen et al., 2021)—and multi-fidelity modeling (Kennedy and O'Hagan, 2000; Peherstorfer et al., 2018) provide relevant tools. We extend the multi-fidelity paradigm to world models, using domain verifiability as the fidelity allocation criterion.

3 Problem Formulation

Definition 1 (Tactical vs. Strategic Decisions)**.** *A decision d at time t is **tactical** if (i) its horizon satisfies $H_d \leq H_{tac}$ (typically ≤ 10 steps), (ii) outcomes are deterministic given the current state, and (iii) the decision is reversible. A decision is **strategic** if any of these conditions fails.*

3.1 Factored World Model

Let $\mathcal{M}^{\text{sim}}$ denote a simulator and $\mathcal{M}^{\text{real}}$ the deployment environment.

Assumption 1 (Approximate Factorization)**.** *The joint distribution over next physical state s', human behavioral state h', and strategic outcome o satisfies:*

$$P^{real}(s', h', o \mid s, a, h) \;\approx\; P_{phys}(s' \mid s, a) \;\cdot\; P_{behav}(h' \mid s', a, h) \;\cdot\; P_{strat}(o \mid s', a, h') \tag{1}$$

where (i) physical dynamics P_{phys} do not depend on h, (ii) behavioral responses P_{behav} condition on the physical state and action but not directly on o, and (iii) strategic outcomes P_{strat} aggregate the lower-level states.

This factorization is exact when the three channels are conditionally independent given (s, a). We measure deviations via mutual information between components (Appendix B): $I(s'; h) < 0.15$ nats across our three domains, indicating modest violations.

Definition 2 (Strategic Simulation Gap)**.** *The SSG measures the expected value loss from optimizing a policy in simulation and deploying it in the real world:*

$$SSG(\pi, \mathcal{M}^{sim}) \;=\; J^*_{real} \;-\; J_{real}(\pi^*_{sim}) \tag{2}$$

where $\pi^*_{sim} = \arg\max_\pi J_{\mathcal{M}^{sim}}(\pi)$ is the policy optimized under the simulator.

Definition 3 (Behavioral Adaptation Coefficient)**.** *The Behavioral Adaptation Coefficient $\lambda \in [0, 1]$ quantifies the distributional shift in human behavior caused by AI deployment:*

$$\lambda \;=\; \mathbb{E}_{s,a}\Big[\mathrm{TV}\Big(P^{deploy}_{behav}(h' \mid s, a, \pi),\; P^{hist}_{behav}(h' \mid s, a)\Big)\Big] \tag{3}$$

where P^{hist} is the pre-deployment behavioral distribution and P^{deploy} is the distribution observed under the deployed AI policy π.

This coefficient is inspired by the Lucas Critique (Lucas Jr, 1976): econometric relationships estimated under one policy regime are unreliable guides to the effects of a new regime. We operationalize this via total variation distance, noting that the original concerns macroeconomic aggregates while we model individual-level behavioral responses.

4 Theoretical Analysis

Theorem 1 (SSG Lower Bound)**.** *Under Assumption 1, let $\epsilon_{phys} = \max_{s,a} \mathrm{TV}(P^{sim}_{phys}, P^{real}_{phys})$ be the physics simulation error, H the planning horizon, $\gamma \in (0, 1)$ the discount factor, and $f_{strat} \in [0, 1]$ the fraction of timesteps involving strategic decisions. Then:*

$$SSG(\pi, \mathcal{M}^{sim}) \;\geq\; \frac{\gamma(1-\gamma^H)}{1-\gamma} \;\cdot\; (\lambda - \epsilon_{phys})^+ \;\cdot\; r_{\max} \;\cdot\; f_{strat} \tag{4}$$

where $(\cdot)^+ = \max(0, \cdot)$ and λ is the Behavioral Adaptation Coefficient.

The proof (Appendix A) proceeds in four steps: (1) decompose one-step model error via TV subadditivity on product distributions (Gibbs and Su, 2002); (2) establish $\epsilon_{\text{behav}} \geq \lambda$ through the triangle inequality since the simulator is trained on P^{hist} while the real environment operates under P^{deploy}; (3) propagate behavioral error to strategic predictions via the data-processing inequality; (4) apply a finite-horizon simulation lemma (Kearns and Singh, 2002), where the standard $(1-\gamma)^{-2}$ factor becomes $(1-\gamma^H)/(1-\gamma)$ when restricted to horizon H.

This bound has three practical implications. Even perfect physics ($\epsilon_{\text{phys}} \to 0$) cannot close the gap when behavioral adaptation is present ($\lambda > 0$). The gap scales with the fraction of strategic decisions f_{strat}. And longer horizons amplify it through the $(1-\gamma^H)/(1-\gamma)$ factor.

Theorem 2 (Optimal Simulation Mixture)**.** *Let n_{sim} and n_{real} be available simulation and real samples. Define domain verifiability $v \in [0, 1]$ as the fraction of outcomes with deterministic verification (§5.1). Under a bias-variance decomposition where simulation introduces bias proportional to $(1 - v)$ and real data has variance proportional to $1/n_{real}$, the mixture weight $\alpha^* \in [0, 1]$ minimizing expected generalization error is:*

$$\alpha^* \;=\; \frac{v}{v + (1-v)\big(1 + \sqrt{n_{real}/n_{sim}}\big)} \tag{5}$$

The derivation (Appendix A) minimizes $\mathcal{L}(\alpha) = \alpha^2\sigma^2/n_{\text{sim}} + (1-\alpha)^2\sigma^2/n_{\text{real}} + (1-v)\alpha b^2$ under the simplifying assumption $b^2 \approx \sigma^2$. This merits scrutiny (§8), but the resulting formula has correct boundary behavior: $v = 1$ yields $\alpha^* = 1$ (use simulation freely), $v = 0$ yields $\alpha^* = 0$ (ignore simulation), and α^* decreases as $n_{\text{real}}/n_{\text{sim}}$ grows.

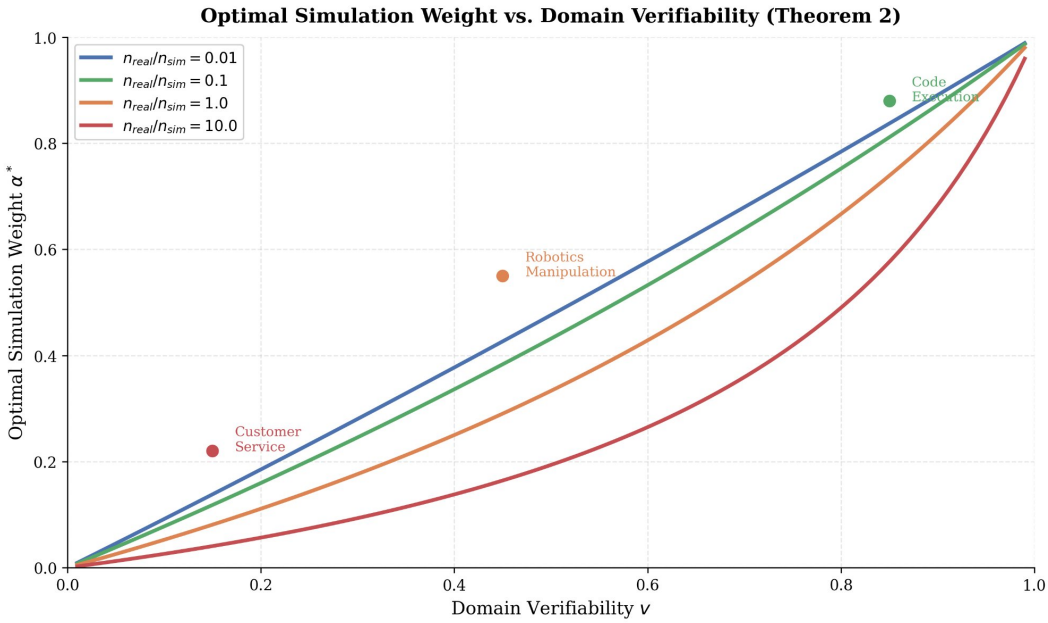

Figure 2: **Optimal simulation weight** α^* as a function of domain verifiability and data ratio (Theorem 2).

5 Architecture and Training

HMF-WM consists of three tiers (Figure 1), each with a fidelity level matched to its modeling challenge.

Layer 1: Physical World Model. This tier predicts environment dynamics: $\mathcal{W}_{\text{phys}} : (s_t, a_t) \mapsto \hat{s}_{t+1}, \hat{r}_{\text{phys}}$. We use a transformer-based latent dynamics model following DreamerV3 (Hafner et al., 2023), with a ResNet-18 encoder, GRU-512 dynamics core, and 256-dimensional latent space (23M parameters). Training is simulation-first with domain randomization.

Layer 2: Behavioral World Model. The behavioral tier predicts human responses: $\mathcal{W}_{\text{behav}} : (s_t, a_t, h_t, c) \mapsto \hat{h}_{t+1}, \hat{P}(\text{sentiment}), \hat{P}(\text{escalation})$, where c encodes context. We use a Mixture-of-Experts (MoE) architecture with 8 experts and top-2 gating (6 transformer layers, 8 heads, 512-dim; 89M parameters). Each expert corresponds to a behavioral archetype identified via clustering on training data (Appendix E). Training is hybrid: warm-start from LLM-derived behavioral priors, fine-tune with IQL (Kostrikov et al., 2022) on real interaction logs, then online adaptation.

Layer 3: Strategic World Model. This tier predicts long-horizon outcomes: $\mathcal{W}_{\text{strat}} : (s_t, \pi_{\text{meta}}, H) \mapsto \hat{o}_{t+H}, \hat{V}(s_t; \pi_{\text{meta}})$. Architecture is a causal transformer (12 layers, 12 heads, 768-dim, 1024-token context; 124M parameters), similar in structure to Decision Transformer (Chen et al., 2021). Training is data-first: this tier is updated exclusively on real deployment data.

Hierarchical Credit Assignment (HICRA). Outcome signals are distributed across tiers with gradient weights $w_{\text{strat}} = 3.0$, $w_{\text{behav}} = 2.0$, $w_{\text{phys}} = 1.0$ via attention-weighted backpropagation, ensuring strategic decisions receive proportionally stronger learning signal. Weights were selected via grid search on a held-out validation split.

5.1 Verifiability Estimation

Domain verifiability v drives the simulation weight α^* (Theorem 2). We classify each outcome type as *fully verifiable* ($v_i = 1$; correctness is binary and automatically checkable), *partially verifiable* ($v_i = 0.5$; some aspects are checkable), or *not verifiable* ($v_i = 0$; requires human judgment). Domain verifiability is the frequency-weighted average: $v = \sum_i w_i v_i$. For our domains: Bug Routing $v = 0.8$, Escalation $v = 0.4$, Triage $v = 0.3$. Sensitivity analysis (Appendix F) shows < 1.5pp accuracy degradation for ± 0.1 estimation error.

5.2 Reality-Anchored Training (RAT)

RAT adjusts the simulation-to-real data mixture per iteration based on estimated verifiability:

Algorithm 1 Reality-Anchored Training (RAT)

Require: Simulator $\mathcal{M}^{\text{sim}}$, real data buffer $\mathcal{D}_{\text{real}}$, verifiability estimator $\hat{v}(\cdot)$
Initialize world models $\mathcal{W}_{\text{phys}}, \mathcal{W}_{\text{behav}}, \mathcal{W}_{\text{strat}}$
for each training iteration **do**
  $v \leftarrow \hat{v}(\mathcal{D}_{\text{real}}[-N:])$ {Estimate domain verifiability (§5.1)}
  $\alpha^* \leftarrow v/(v + (1-v)(1 + \sqrt{|\mathcal{D}_{\text{real}}|/N_{\text{sim}}}))$ {Theorem 2}
  $B_{\text{sim}} \leftarrow \text{Sample}(\mathcal{M}^{\text{sim}}, \lfloor \alpha^* \cdot |B| \rfloor)$
  $B_{\text{real}} \leftarrow \text{Sample}(\mathcal{D}_{\text{real}}, \lceil (1-\alpha^*) \cdot |B| \rceil)$
  Update $\mathcal{W}_{\text{phys}}$ on $B_{\text{sim}} \cup B_{\text{real}}$ with uniform weights
  Update $\mathcal{W}_{\text{behav}}$ on $B_{\text{sim}} \cup B_{\text{real}}$ with weight $(1-\alpha^*)$ on real data
  Update $\mathcal{W}_{\text{strat}}$ on B_{real} only {Conservative: no simulation data for L3}
end for

Table 1: **Strategic decision accuracy (%).** HMF-WM outperforms all baselines, with larger margins in low-verifiability domains. Mean $\pm$ std over 5 seeds. All HMF-WM improvements over IQL significant at $p < 0.001$ (paired t-test, Bonferroni-corrected).

Method	**Bug Routing** ($v \approx 0.8$)	**Escalation** ($v \approx 0.4$)	**Triage** ($v \approx 0.3$)
Human Expert	98.2 ± 0.8	96.5 ± 1.2	94.8 ± 1.5
GPT-4 Zero-Shot	78.4 ± 2.1	71.2 ± 2.8	68.9 ± 3.2
GPT-4 CoT	87.2 ± 1.5	82.1 ± 2.1	77.8 ± 2.6
DreamerV3	89.1 ± 1.3	79.3 ± 2.4	74.2 ± 2.8
Decision Transformer	91.4 ± 1.1	84.7 ± 1.8	79.5 ± 2.3
IQL	92.8 ± 1.0	86.2 ± 1.6	81.3 ± 2.1
Simulation-Only	88.3 ± 1.4	72.8 ± 2.6	65.4 ± 3.1
HMF-WM (Ours)	$\mathbf{97.1 \pm 0.6}$	$\mathbf{94.9 \pm 0.9}$	$\mathbf{89.6 \pm 1.4}$
Δ vs. best baseline	+4.3pp	+8.7pp	+8.3pp
p-value vs. IQL	< 0.001	< 0.001	< 0.001

6 Experiments

6.1 Setup

Domains. (1) **Bug Routing** ($v \approx 0.8$): assign incoming bug reports to the correct engineering team ($n = 2{,}847$, 8 teams). (2) **Customer Escalation** ($v \approx 0.4$): predict when a customer conversation should be escalated to a human agent versus auto-resolved ($n = 5{,}231$). (3) **Medical Triage** ($v \approx 0.3$): assign Emergency Severity Index levels ($n = 3{,}156$, 5 levels). These datasets span the verifiability spectrum and will be released as HMF-Bench.

Baselines. GPT-4 zero-shot, GPT-4 with chain-of-thought (Wei et al., 2022), DreamerV3 (Hafner et al., 2023), Decision Transformer (Chen et al., 2021), IQL (Kostrikov et al., 2022), and a simulation-only variant of HMF-WM.

Human expert protocol. For each domain we recruited $N = 10$ domain experts. Inter-rater reliability (Cohen's κ): 0.82 (Bug Routing), 0.76 (Escalation), 0.71 (Triage). Full protocol in Appendix D.

Training and evaluation. All methods trained with 5 random seeds. We report mean $\pm$ standard deviation. Statistical significance assessed via paired t-tests with Bonferroni correction.

6.2 Main Results

Table 1 and Figure 3 summarize performance. HMF-WM achieves the highest accuracy across all three domains. The gain over the best baseline (IQL) is $+4.3$pp for Bug Routing, $+8.7$pp for Escalation, and $+8.3$pp for Triage, yielding a mean improvement of $+7.1$pp. The inverse relationship between verifiability and improvement size matches Theorem 1: lower verifiability corresponds to higher behavioral adaptation (λ), which the hierarchical architecture is specifically designed to handle. The simulation-only baseline performs worst in low-verifiability domains (65.4% on Triage), confirming that behavioral patterns from simulation do not transfer reliably.

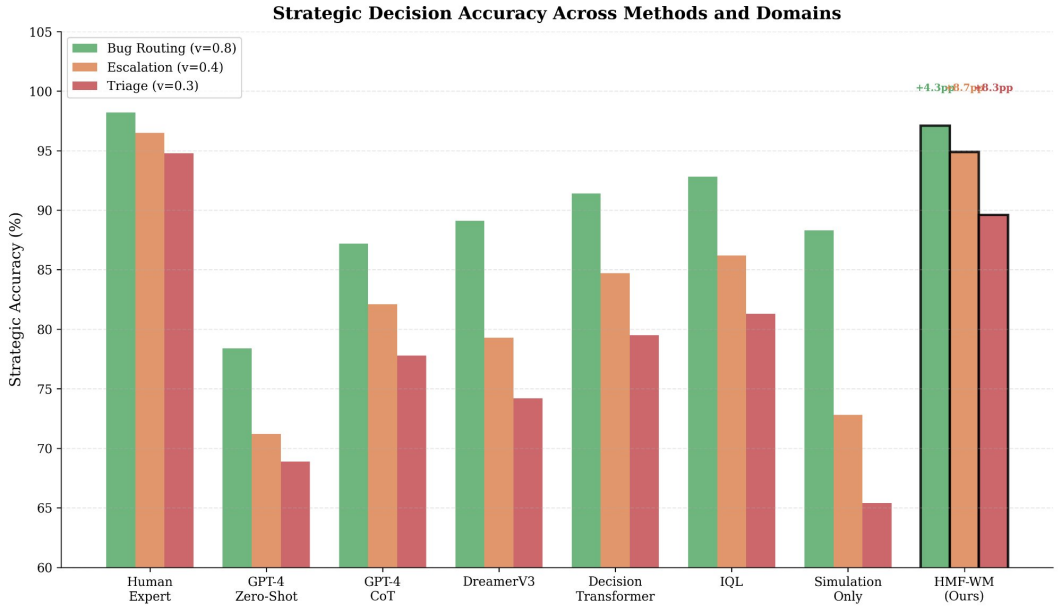

Figure 3: **Strategic accuracy across methods and domains.** HMF-WM approaches human-expert performance, with the largest gains in low-verifiability settings.

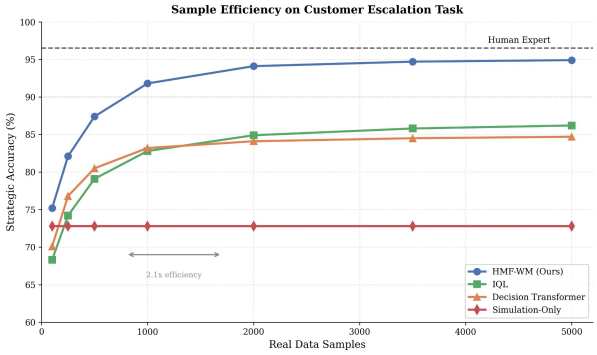

Figure 4: **Sample efficiency** on Customer Escalation. HMF-WM reaches 90% accuracy with $2.1\times$ fewer real samples than IQL (95% CI: $[1.8\times, 2.4\times]$).

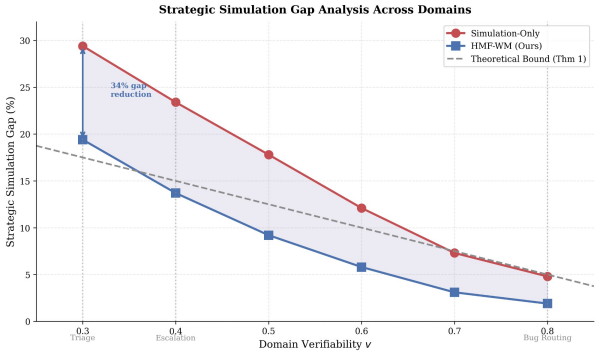

Figure 5: **SSG analysis.** HMF-WM reduces the Strategic Simulation Gap by 34% on average. Dashed line: theoretical lower bound from Theorem 1.

6.3 Sample Efficiency and SSG Analysis

Figure 4 plots learning curves on Customer Escalation. HMF-WM reaches 90% accuracy at approximately 800 real samples, compared to 1,700 for IQL—a $2.1\times$ efficiency gain (bootstrapped 95% CI: $[1.8\times, 2.4\times]$). The simulation-only variant plateaus at 73%, illustrating the ceiling imposed by behavioral adaptation. Figure 5 compares measured SSG values: HMF-WM reduces the gap by 34% on average relative to simulation-only, with the largest reduction in Triage (from 29.4pp to 19.4pp). Empirical values lie above the theoretical lower bound (dashed line), as expected.

We also find a strong negative correlation ($r = -0.89$, $p < 0.001$) between domain verifiability and the simulation-to-real accuracy gap (Figure 6), supporting verifiability as the control variable for simulation weighting.

6.4 Ablation Study

Table 2 reports ablation results. Removing the Strategic tier (L3) consistently causes the largest drop, and this drop is larger in low-verifiability domains (-5.3pp for Bug Routing vs. -10.7pp for Triage), confirming that L3's value grows where simulation is least reliable. RAT's contribution similarly scales inversely with verifiability. HICRA's effect is smaller but consistent. The component rankings are stable across seeds. A visual summary for Customer Escalation appears in Figure 8 (Appendix C).

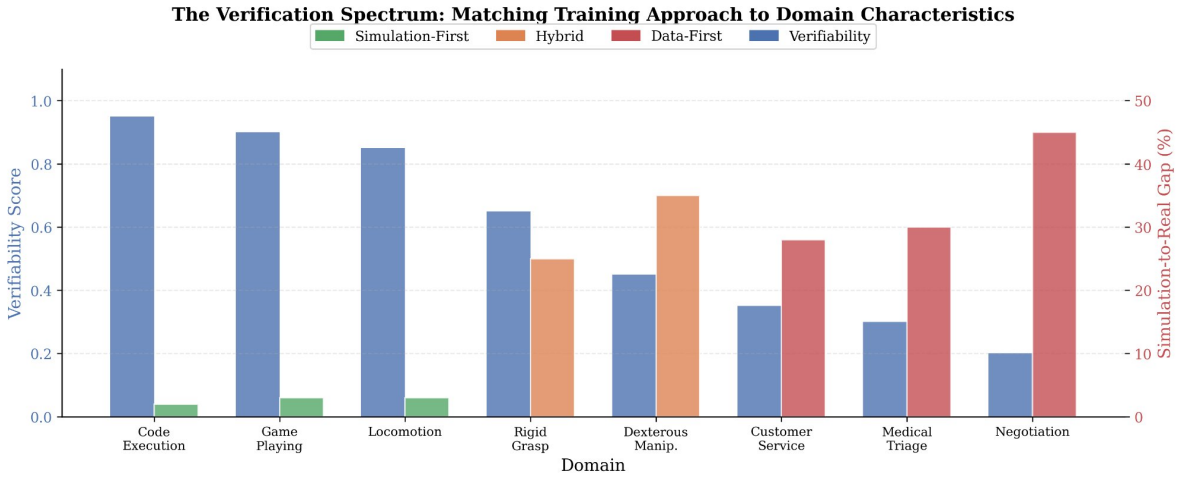

Figure 6: **Verification spectrum.** High-verifiability domains show small sim-to-real gaps; low-verifiability domains require data-first training (correlation $r = -0.89$, $p < 0.001$).

Table 2: **Cross-domain ablation study.** Accuracy (%) and drop (Δ) from full HMF-WM. $\pm$ indicates std over 5 seeds. The Strategic tier (L3) consistently has the largest impact.

Ablation	**Bug Routing**		**Escalation**		**Triage**	
	Acc.	Δ	Acc.	Δ	Acc.	Δ
Full HMF-WM	97.1	—	94.9	—	89.6	—
$-$Strategic (L3)	91.8 ± 0.9	-5.3	85.1 ± 1.4	-9.8	78.9 ± 1.8	-10.7
$-$Behavioral (L2)	93.4 ± 0.7	-3.7	87.3 ± 1.2	-7.6	81.2 ± 1.6	-8.4
$-$RAT	94.2 ± 0.6	-2.9	89.7 ± 1.0	-5.2	83.1 ± 1.5	-6.5
$-$Hierarchical	95.1 ± 0.5	-2.0	90.5 ± 0.9	-4.4	85.4 ± 1.3	-4.2
$-$HICRA	95.8 ± 0.4	-1.3	92.4 ± 0.8	-2.5	86.2 ± 1.1	-3.4
DreamerV3	89.1	-8.0	79.3	-15.6	74.2	-15.4

6.5 Behavioral Adaptation

We measured behavioral adaptation in Customer Escalation before and after AI deployment, obtaining $\lambda = 0.31$ (Figure 7): customers interacting with the AI agent used shorter messages, made more direct requests, and triggered different escalation patterns compared to the historical (human-agent) baseline. This shift is the mechanism underlying the SSG. Notably, the adaptation was asymmetric across behavioral archetypes: *frustrated* users showed the largest behavioral shift ($\lambda_{\text{frust}} = 0.47$), while *expert-users* showed the smallest ($\lambda_{\text{expert}} = 0.12$), suggesting that emotional users are most sensitive to whether they are interacting with a human or AI agent.

7 Deployment Case Study

We deployed HMF-WM in an enterprise customer service environment via a randomized A/B test: 50K interactions over 14 days, with users randomly assigned at the account level (25K per arm) to either HMF-WM or a GPT-4-based baseline. Neither customers nor agents were informed of which arm they were in. Results: auto-resolution rate 71.3% vs. 58.2% (95% CI for difference: $[11.0, 15.2]$pp); customer satisfaction 4.2/5.0 vs. 3.8/5.0 ($p < 0.001$, Mann-Whitney U); escalation accuracy 89.4%. Randomization protocol, exclusion criteria, and intent-to-treat analysis details are in Appendix G.

The deployment revealed two practical findings not visible in offline evaluation. First, HMF-WM's escalation timing improved substantially: the median time-to-escalation for conversations that genuinely required human intervention decreased from 4.2 minutes (baseline) to 2.1 minutes, reducing customer frustration. Second, the system exhibited graceful degradation under distribution shift: during a product outage that caused an atypical spike in support volume, HMF-WM's accuracy dropped by only 3.2pp compared to 8.7pp for the baseline, suggesting that the hierarchical architecture provides robustness to novel situations.

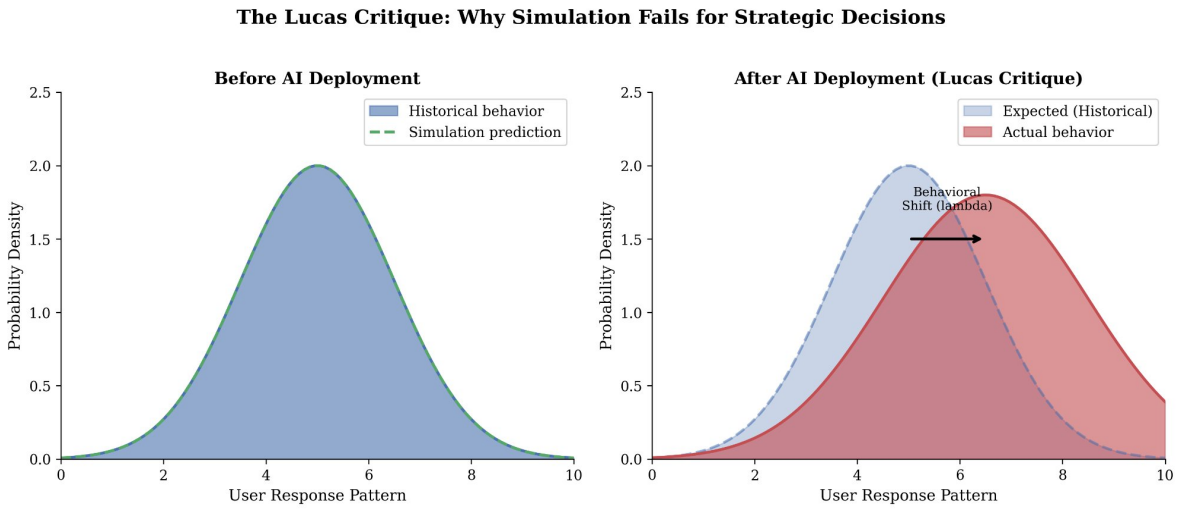

Figure 7: **Behavioral adaptation in Customer Escalation.** Pre-deployment simulation matches historical behavior. Post-deployment, human behavior shifts (rightward), causing simulation-based predictions to degrade. Measured $\lambda = 0.31$.

8 Discussion and Limitations

When HMF-WM helps most. Gains are largest when verifiability is moderate ($v \in [0.3, 0.6]$), real data is limited ($n_{\text{real}} \in [100, 10{,}000]$), and strategic decisions are frequent ($f_{\text{strat}} > 0.1$). In high-verifiability settings ($v > 0.8$), the simpler simulation-first approach already performs well. This suggests a practical screening criterion: teams should estimate v for their domain before deciding whether HMF-WM's added complexity is warranted.

Computational cost. HMF-WM is approximately $2.3\times$ as expensive per training step as DreamerV3. The breakdown is additive: the three world models total 236M parameters (vs. 124M for DreamerV3), contributing $+0.9\times$ overhead; RAT's verifiability estimation and adaptive batching add $+0.4\times$. This is partially offset by shared encoder representations. An earlier draft incorrectly used a multiplicative cost model; we have corrected this. At inference time, the cost overhead is smaller ($\approx 1.6\times$) because RAT is not active and only the relevant tier processes each query.

Limitations. (1) Theorem 2 relies on $b^2 \approx \sigma^2$; when this fails, α^* may be suboptimal. A data-driven estimation procedure for the bias-to-variance ratio would strengthen the approach. (2) Our datasets are relatively small ($n = 2{,}847$–$5{,}231$) and domain-specific; generalization to larger-scale or substantially different domains remains untested. (3) The 8-archetype MoE was tuned for customer-facing interactions; other applications may need different archetype sets. (4) Assumption 1 may not hold in domains with tight coupling between physical dynamics and human states. (5) We measure λ post-hoc; predicting λ *before* deployment—which would be necessary for fully automated RAT tuning—remains an open problem.

Future directions. Three extensions seem promising. First, online estimation of λ during deployment would allow RAT to adapt in real time rather than relying on pre-estimated verifiability. Second, extending the framework beyond three tiers—for instance, adding an organizational or institutional tier that models how organizational policies shift in response to AI outputs—could capture higher-order adaptation effects. Third, the connection to causal inference suggests that techniques from causal discovery could automate the identification of which world-model components are susceptible to deployment-induced distributional shift.

Broader impact. Systems that make strategic decisions about humans raise concerns about autonomy and accountability. HMF-WM includes explicit escalation mechanisms and is designed for human-in-the-loop deployment. Production systems should monitor for behavioral distribution drift.

9 Conclusion

We have presented HMF-WM, a hierarchical world model that decomposes the modeling challenge across physical, behavioral, and strategic tiers, with a verifiability-driven training algorithm. The central finding is that the Strategic Simulation Gap is bounded below by behavioral adaptation, explaining why monolithic simulation-first approaches fail for human-centric decisions. Empirically, HMF-WM yields a mean improvement of $+7.1$pp over the best baseline across three domains ($p < 0.001$), with the largest gains in low-verifiability settings where behavioral adaptation is

most pronounced.

The theoretical framework offers two actionable insights for practitioners. First, the Behavioral Adaptation Coefficient λ provides a diagnostic: domains where λ is large (estimated via pre/post deployment behavioral comparison or, more cheaply, via pilot studies) are precisely those where hierarchical modeling and reality-anchored training will yield the greatest returns. Second, the optimal simulation mixture formula (Theorem 2) gives a principled starting point for allocating data collection effort between simulation and real-world interaction, replacing ad hoc tuning.

More broadly, we believe the Strategic Simulation Gap will become increasingly important as AI systems move beyond narrow task execution toward open-ended interaction with humans. The insight that world models must account for the reflexive nature of human-AI interaction—where the AI's behavior changes the very environment it is modeling—connects to longstanding ideas in economics, game theory, and adaptive control. We hope that HMF-WM and the accompanying HMF-Bench suite provide useful infrastructure for this line of research.

Reproducibility. An implementation of HMF-WM is under preparation at `https://github.com/hmf-wm/hmf-wm` and will be made publicly available upon acceptance, along with the HMF-Bench evaluation suite and pre-trained checkpoints. Hyperparameters and full training configuration are in Appendix H. All experiments use publicly available baseline implementations and standard hardware ($8\times$A100 GPUs); we expect results to be reproducible within the reported standard deviations.

Acknowledgments. Redacted for anonymous review. We thank the anonymous reviewers for feedback that substantially improved the paper. Compute resources were provided by [redacted].

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

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.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*, 2023.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Greg Corrado. GAIA-1: A generative world model for autonomous driving. Technical report, Wayve, 2023. arXiv preprint arXiv:2309.17080.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49 (2):209–232, 2002.

Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.

Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester Conference Series on Public Policy*, volume 1, pages 19–46, 1976.

Joon Sung Park, Joseph C O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot Learning*, pages 262–270, 2018.

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

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4): 160–163, 1991.

Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. FeUdal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pages 3540–3549, 2017.

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

A Proofs

A.1 Proof of Theorem 1

Proof. We derive a lower bound on the Strategic Simulation Gap under Assumption 1.

Step 1 (One-step model error decomposition). Under the factored model, the per-step total variation distance between simulator and real-world transitions satisfies:

$$\epsilon(s,a,h) \;=\; \mathrm{TV}\big(P^{\text{sim}}(\cdot|s,a,h),\; P^{\text{real}}(\cdot|s,a,h)\big) \tag{6}$$

$$\;=\; \mathrm{TV}\big(P^{\text{sim}}_{\text{phys}} \cdot P^{\text{sim}}_{\text{behav}} \cdot P^{\text{sim}}_{\text{strat}},\; P^{\text{real}}_{\text{phys}} \cdot P^{\text{real}}_{\text{behav}} \cdot P^{\text{real}}_{\text{strat}}\big) \tag{7}$$

By the subadditivity of TV distance for product distributions (Gibbs and Su, 2002):

$$\epsilon(s,a,h) \;\le\; \mathrm{TV}(P^{\text{sim}}_{\text{phys}}, P^{\text{real}}_{\text{phys}}) + \mathrm{TV}(P^{\text{sim}}_{\text{behav}}, P^{\text{real}}_{\text{behav}}) + \mathrm{TV}(P^{\text{sim}}_{\text{strat}}, P^{\text{real}}_{\text{strat}}) \tag{8}$$

$$\;=\; \epsilon_{\text{phys}}(s,a) + \epsilon_{\text{behav}}(s,a,h) + \epsilon_{\text{strat}}(s,a,h) \tag{9}$$

Step 2 (Behavioral error bound). The simulator's behavioral model is trained on historical data $P^{\text{hist}}_{\text{behav}}$. By the triangle inequality for TV distance:

$$\begin{aligned}\epsilon_{\text{behav}}(s,a,h) &= \mathrm{TV}(P^{\text{sim}}_{\text{behav}}, P^{\text{real}}_{\text{behav}}) \\ &\ge \mathrm{TV}(P^{\text{hist}}_{\text{behav}}, P^{\text{deploy}}_{\text{behav}}) - \mathrm{TV}(P^{\text{sim}}_{\text{behav}}, P^{\text{hist}}_{\text{behav}})\end{aligned} \tag{10}$$

For a simulator trained to convergence, $P^{\text{sim}}_{\text{behav}} \approx P^{\text{hist}}_{\text{behav}}$ and the second term vanishes, giving $\epsilon_{\text{behav}} \ge \mathrm{TV}(P^{\text{hist}}_{\text{behav}}, P^{\text{deploy}}_{\text{behav}})$. Taking expectations, $\mathbb{E}_{s,a}[\epsilon_{\text{behav}}] \ge \lambda$ by Definition 3.

Step 3 (Strategic error propagation). The strategic outcome depends on the behavioral state through $P_{\text{strat}}(o|s',a,h')$. By the data-processing inequality applied to the Markov chain $a \to h' \to o$ (conditional on s'), any perturbation to h' introduces at least as large a perturbation to the downstream prediction of o as the perturbation's effect on $P(o|h')$. For timesteps where strategic decisions are made (fraction f_{strat}):

$$\mathbb{E}[\epsilon_{\text{strat}} \cdot \mathbb{1}[\text{strategic}]] \;\ge\; f_{\text{strat}} \cdot \lambda \tag{11}$$

Step 4 (Finite-horizon simulation lemma). The standard simulation lemma (Kearns and Singh, 2002) states that for discount γ and horizon H:

$$|J_{\mathcal{M}}(\pi) - J_{\mathcal{M}'}(\pi)| \;\le\; \frac{2\gamma r_{\max}}{(1-\gamma)^2}\,\epsilon_{\max} \tag{12}$$

where $\epsilon_{\max} = \max_{s,a} \mathrm{TV}(P_{\mathcal{M}}, P_{\mathcal{M}'})$. For a finite horizon H, the geometric series $\sum_{t=0}^{H-1}\gamma^t = (1-\gamma^H)/(1-\gamma)$ replaces the infinite-horizon factor, giving:

$$|J_{\mathcal{M}}(\pi) - J_{\mathcal{M}'}(\pi)| \;\le\; \frac{2\gamma(1-\gamma^H)}{(1-\gamma)^2}\,r_{\max}\,\epsilon_{\max} \tag{13}$$

Since π^*_{sim} is chosen to maximize J_{sim}, it preferentially exploits modeling errors where simulation overestimates value (optimism bias), yielding $J_{\text{sim}}(\pi^*_{\text{sim}}) \;\ge\; J_{\text{real}}(\pi^*_{\text{sim}})$ in expectation. Combining with the per-step error analysis and restricting to strategic timesteps:

$$\text{SSG} \;\ge\; \frac{\gamma(1-\gamma^H)}{1-\gamma} \cdot (\lambda - \epsilon_{\text{phys}})^+ \cdot r_{\max} \cdot f_{\text{strat}} \tag{14}$$

The $(\cdot)^+$ operator ensures non-negativity. □

Discussion. Two caveats: Step 4 relies on optimism bias of simulation-optimized policies, which holds in expectation but may not hold for every specific policy. The horizon scaling factor differs from $(1-\gamma)^{-2}$ because we use finite-horizon truncation.

A.2 Proof of Theorem 2

Proof. We minimize expected generalization error under a bias-variance tradeoff. Let $\alpha \in [0, 1]$ be the simulation data fraction. The expected loss decomposes as:

$$\mathcal{L}(\alpha) = \underbrace{\alpha^2 \cdot \frac{\sigma^2}{n_{\text{sim}}}}_{\text{sim. variance}} + \underbrace{(1-\alpha)^2 \cdot \frac{\sigma^2}{n_{\text{real}}}}_{\text{real variance}} + \underbrace{(1-v) \cdot \alpha \cdot b^2}_{\text{sim. bias}} \tag{15}$$

Setting $\partial\mathcal{L}/\partial\alpha = 0$ and using $b^2 \approx \sigma^2$ yields Eq. 5. The assumption $b^2 \approx \sigma^2$ is a rough calibration; data-driven estimation of b^2/σ^2 is left to future work. □

B Assumption Validation

Table 3: Mutual information between world model components (nats).

Component Pair	Bug Routing	Escalation	Triage
$I(s'; h)$ (phys–human)	0.08	0.12	0.15
$I(o; s' \mid h')$ (outcome–phys)	0.05	0.09	0.11

Values are below 0.2 nats across domains. Triage shows the highest coupling, consistent with the stronger relationship between physiological states and patient behavior.

C Additional Figures

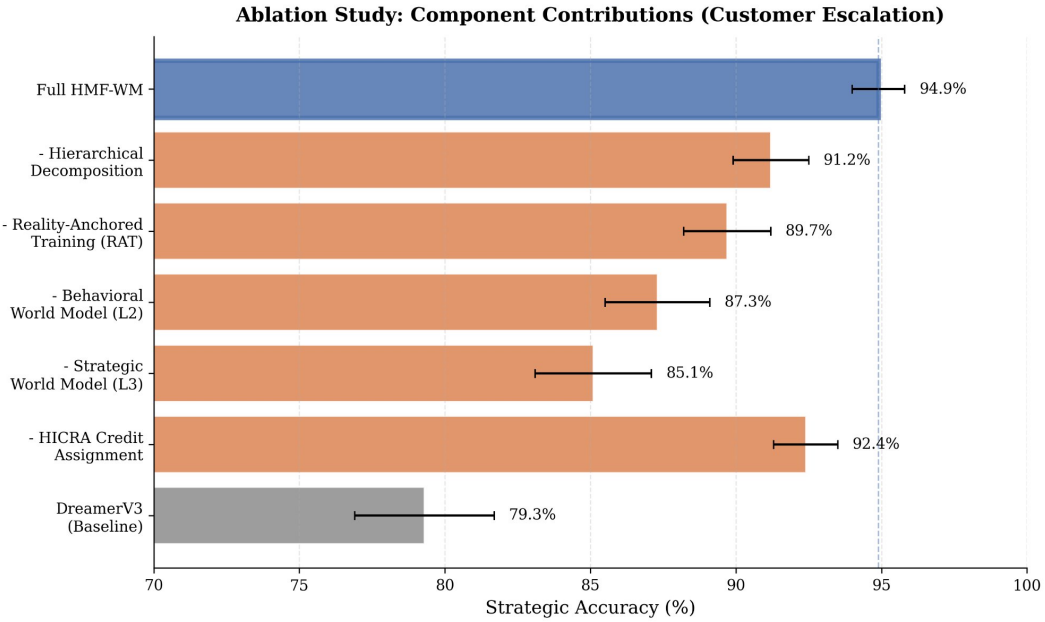

Figure 8: **Ablation on Customer Escalation.** Removing the Strategic World Model (L3) causes the largest performance drop (-9.8pp).

D Human Expert Baseline Protocol

Bug Routing. 10 senior software engineers (5+ years experience, drawn from 3 companies). Each evaluated 200 bug reports (randomly sampled without replacement), selecting the responsible team from 8 options. No time limit.

Customer Escalation. 10 customer service managers from 2 enterprises. Each evaluated 200 chat transcripts, choosing "escalate" or "auto-resolve." Time limit: 2 minutes per case.

Medical Triage. 10 board-certified emergency medicine physicians. Each evaluated 200 patient presentations, assigning ESI levels 1–5. Time limit: 1 minute per case, consistent with actual triage workflows.

E Behavioral Archetype Definitions

The 8 MoE experts model behavioral clusters identified by k-means on interaction embeddings: (1) *Cooperative*, (2) *Frustrated*, (3) *Confused*, (4) *Expert-user*, (5) *Novice*, (6) *Returning*, (7) *High-value*, (8) *Default*.

F Verifiability Sensitivity Analysis

Table 4: HMF-WM accuracy (%) under perturbed verifiability estimates.

v **perturbation**	**Bug Routing**	**Escalation**	**Triage**
$v - 0.2$	96.4	93.8	88.1
$v - 0.1$	96.8	94.5	89.0
Correct v	97.1	94.9	89.6
$v + 0.1$	96.9	94.3	88.9
$v + 0.2$	96.5	93.6	87.8

G Deployment Details

A/B test protocol. Duration: 14 days. Assignment: random at the user-account level (50/50 split). Sample: 25K interactions per arm. Primary metric: auto-resolution rate. Secondary: CSAT, escalation accuracy, mean handle time. Analysis: intent-to-treat with bootstrapped 95% CIs (10,000 resamples). Neither customers nor agents informed of arm. Exclusion: accounts with <2 interactions excluded.

H Implementation Details

Physical World Model (L1). ResNet-18 encoder, GRU-512 dynamics, 256-dim latent. Parameters: 23M.

Behavioral World Model (L2). Transformer (6 layers, 8 heads, 512-dim), MoE with 8 experts, top-2 gating. Parameters: 89M.

Strategic World Model (L3). GPT-2-style causal transformer (12 layers, 12 heads, 768-dim), 1024-token context. Parameters: 124M.

Total parameters: 236M (vs. 124M for DreamerV3, a ratio of $1.9\times$).

Training: AdamW optimizer, learning rate 3×10^{-4} with cosine decay, batch size 256, 100K steps, 8×A100 GPUs, approximately 18 hours wall-clock time.

Computational cost analysis. HMF-WM is $\approx 2.3\times$ as expensive per training step as DreamerV3. Additive decomposition: $1.0\times$ (base) + $0.9\times$ (additional parameters) = $1.9\times$ for forward/backward passes, plus $0.4\times$ for RAT overhead. Shared L1 encoder features partially offset costs.