The Story Shapes the Agent: Narrative Priors in LLM Behavior

Yixuan Wang
University of North Carolina at Chapel Hill
yxwang@cs.unc.edu

James Lester
North Carolina State University
lester@ncsu.edu

Shashank Srivastava
University of North Carolina at Chapel Hill
ssrivastava@cs.unc.edu

Abstract

Persona prompting is widely used to steer LLM agent behavior (Tseng et al., 2024; Chen et al., 2024), yet the narrative framing of a task can matter more than the assigned persona. We isolate this effect through structural isomorphism, constructing three text-based investigation games that share the same action space, stage progression, and resource constraints while varying only task narrative: disease investigation, IT troubleshooting, and murder mystery. Across 1,890 sessions spanning 3 models and 10 personas, we identify *narrative priors*: systematic action tendencies activated by a task's story framing, independent of its decision structure. Narrative priors explain 5–31$\times$ more behavioral variance than persona, are consistent across model architectures, and in two of three domains are negatively associated with task success. Persona effects that do transfer across narratives arise from *behavioral anchors*, persona descriptions whose language maps directly onto shared actions. Causal interventions confirm this: removing anchor words from a high-transfer persona reduces cross-narrative consistency by 95%. Our framework also generalizes to a held-out fourth narrative and yields a persona-selection method that improves cross-narrative transfer. These results suggest that LLM behavior that survives narrative changes should be grounded in concrete actions rather than abstract descriptions.

1 Introduction

A practitioner building an LLM-based agent faces a natural question: how do I make it behave the way I want? The dominant lightweight answer is *persona prompting*: assign a dispositional description such as "methodical and thorough" or "prefers learning through conversation," and expect the agent to behave accordingly. This expectation rests on an untested assumption: that persona-induced behavior will transfer when the same agent is redeployed in a different task domain. Prior work has studied whether LLMs can *maintain* an assigned persona within a conversation (Frisch & Giulianelli, 2024; Wang et al., 2024b; Samuel et al., 2024), but not whether the behavior generalizes to different task domains.

We test this assumption with a sharper question: when an agent's behavior changes across task domains, is the persona adapting or is another process taking over? To answer this requires holding decision structure constant while varying only the surface-level story the agent is given. We implement this *structural isomorphism* through three text-based investigation games that share identical action spaces, stage progressions, and resource constraints but differ in narrative: disease investigation, IT troubleshooting, and murder mystery (Figure 1). This design isolates narrative as a behavioral factor rather than confounding it with task structure. Running 10 learning-style personas across all three games on 3 models produces 1,890 sessions in which persona and narrative effects can be cleanly separated.

We find that task narrative is not merely a confound; it is the primary driver of agent behavior, explaining 5–31$\times$ more variance than the assigned persona. We term these implicit

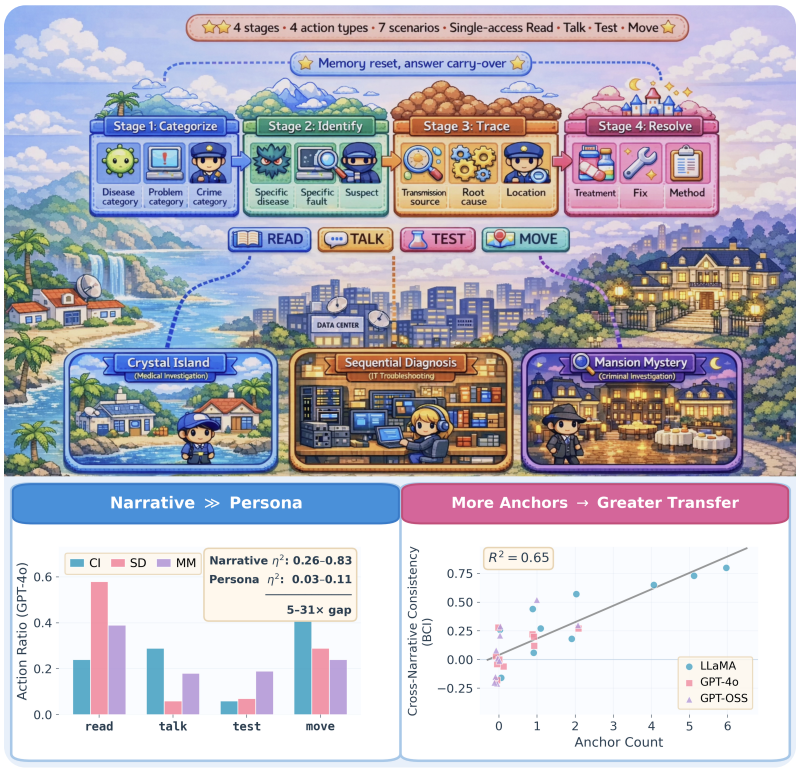

Figure 1: **Top:** Three structurally isomorphic interactive narrative games share identical decision structure (4 stages, 4 action types, 7 scenarios) but differ only in surface narrative (domain-level framing, e.g., medical vs. IT vs. crime). **Bottom left:** Despite shared structure, LLM agents exhibit very different action profiles across narratives, with narrative explaining 5–31$\times$ more variance than persona. **Bottom right:** Personas with more behavioral anchors (concrete action words) achieve higher cross-narrative consistency ($R^2 = 0.65$).

biases *narrative priors*: behavioral tendencies that LLM agents inherit from pretraining corpora, activated by the story framing of a task independently of its decision structure. Medical narratives elicit more social interaction, IT narratives more document reading, crime narratives more diagnostic testing, mirroring domain stereotypes in pretraining text. Critically, these priors are not strategic adaptations: in two of three narratives, the narrative-biased action is negatively correlated with task success (§4.1).

Personas are still powerful. Within a fixed narrative, they explain substantial behavioral variance and can produce large differences in success. But their influence is narrative-confined: a persona that shapes behavior in one game often reshapes itself in another, and cross-narrative transfer classification averages 10.4%, barely above the 10% chance baseline. The persona effects that do survive narrative change are explained by what we call *behavioral anchors*: descriptions whose concrete action language (e.g., "prefers conversation" $\rightarrow$ `talk`) binds directly to the shared action space, producing portable behavioral signatures. A causal intervention confirms this mechanism: abstracting away the anchor words while preserving the persona's psychological content reduces cross-narrative consistency by 95% (§5).

These findings reframe how to design reliable LLM agents. Persona prompting works, but only within the behaviors the narrative already specifies. For behavioral control that survives task changes, personas must be grounded in concrete actions rather than abstract traits. More broadly, our structural isomorphism framework diagnoses narrative influence, our behavioral anchor analysis reveals which personas are robust to it, and our persona-selection method generalizes to a held-out fourth narrative domain, improving cross-narrative identifiability in all evaluation settings (§4.4, §5).

2 Related Work

Persona and Role Prompting. Persona prompting is a standard interface for shaping LLM behavior in multi-agent simulation (Park et al., 2023; 2024), dialogue personalization (Zhang et al., 2018; Shao et al., 2023), and role-based task assignment (Xu et al., 2024; Wang et al., 2024a). Surveys (Tseng et al., 2024; Chen et al., 2024) distinguish role-playing (assigning personas to LLMs) from personalization (adapting LLMs to user profiles). A shared premise across this body of work is that assigned personas reliably govern behavior. We test that assumption by comparing persona influence to the task's narrative framing.

Persona Consistency and Evaluation. Recent work has shown that LLM personas are more fragile than often assumed: models fail to maintain personas under prompt rephrasing (Gupta et al., 2023), in multi-agent debate (Baltaji et al., 2024), across extended dialogue (Frisch & Giulianelli, 2024), and under adversarial probing (Wang et al., 2024b). Recent benchmarks (Tu et al., 2024; Samuel et al., 2024) measure persona consistency via linguistic consistency and knowledge adherence. We instead ask whether persona-induced behavior *transfers* across task contexts, and pinpoint the factors responsible when it does not.

Prompt Sensitivity and Framing Effects. LLMs are broadly sensitive to surface-level prompt variation: instruction phrasing affects task performance (Lu et al., 2022; Webson & Pavlick, 2022), output format alters reasoning accuracy (Zhao et al., 2021), and example ordering changes few-shot results (Lu et al., 2022). Recent work on cognitive biases in LLMs (Jones & Steinhardt, 2022) suggests that these sensitivities may reflect systematic priors rather than random noise. We extend this line of inquiry from textual sensitivity in static tasks to behavioral sensitivity in interactive agent settings, showing that narrative framing activates domain-specific behavioral priors that override persona instructions.

LLM Agents in Interactive Environments. LLM-based agents have been deployed in text-based games (Shridhar et al., 2020; Trivedi et al., 2024), social simulations (Park et al., 2023; 2024), and multi-agent workflows (Qian et al., 2024; Hong et al., 2023). Narrative-centered learning environments have also served as testbeds for studying player decision-making and engagement (Rowe et al., 2011; Wang et al., 2018). While this literature focuses primarily on agent *capabilities*, the question of how task framing shapes agent behavior independently of task structure remains largely unexplored.

3 Experimental Framework

Most behavioral studies of LLM agents conflate *what* the agent must decide with *how* the task is described. A medical investigation and an IT troubleshooting scenario differ in action spaces, information structures, and success criteria, so behavioral differences are hard to interpret: they may reflect the decision problem, the narrative wrapper, or both. To identify *narrative priors* rather than domain-specific behavior, we need tasks that vary in narrative surface but share the same decision-theoretic structure. We call this principle *structural isomorphism* and implement it through three text-based investigation games.

3.1 Isomorphic Game Design

All three games share a skeleton of structural invariants. Each consists of four investigation stages completed in fixed order, with per-stage turn limits. Between stages, the agent's context is cleared and only the answer from the preceding stage carries forward, preventing cumulative advantages from compounding across the investigation. Every information source (a document, a character, or a diagnostic test) may be accessed at most once per stage, forcing selectivity rather than exhaustive search. At each turn, the agent chooses from a discrete menu offering the same four core action types: `read` (consult documents), `talk` (interact with characters), `test` (run diagnostics), and `move` (navigate between locations). Options are pre-labeled by the game engine, so no manual or post-hoc action annotation is required. Across environments, the four stages map onto the same abstract decision sequence: *categorize* $\rightarrow$ *identify* $\rightarrow$ *trace* $\rightarrow$ *resolve*. What differs is the surface narrative wrapper around the shared decision scaffold.

Persona	Core tendency	Action
quick_intuitive	Skims, trusts patterns	–
methodical_thorough	Reads carefully, verifies	`read`
social_collaborative	Conversation-first	`talk`
coverage_focused	Covers every source once	–
confident_decisive	Commits quickly	–
verification_seeking	Seeks convergent evidence	–
hands_on_practical	Prefers doing and testing	`test`
hypothesis_driven	Guesses then tests	∼`test`
big_picture_concept.	High-level patterns	–
effortful_learner	Reads slowly, persists	`read`

Table 1: The 10 personas. *Action* = action type most directly implied by the description; "–" = no single action implied. The relation to cross-narrative consistency is examined in §4.3.

Crystal Island (CI) places the agent on a tropical research island in the role of a student investigating a disease outbreak. Across four stages the agent identifies the disease category, the specific disease, the transmission source, and the appropriate treatment. Our environment builds on the Crystal Island framework (Rowe et al., 2011), originally designed for AI in education research. We extend the codebase with multi-stage investigation, stage-level memory reset, single-access constraints, and automated behavioral logging.

Sequential Diagnosis (SD) recasts the same structure in a corporate IT environment. The agent plays a support engineer who receives an urgent ticket and must identify the problem category, the specific fault, the root cause, and the fix.

Mansion Mystery (MM) recasts the same structure as a detective scenario. The agent plays an investigator solving a diamond theft in a Victorian mansion, determining the crime category, the suspect, the location, and the method.

All three games share 4 stages, 6 locations, 4 action types, and 7 scenarios. Minor differences in character count (5–7) and document count (12–15) reflect narrative needs rather than structural asymmetry; crucially, these resources are functionally equivalent, each providing single-access information within the same action-type category (environment property details in Appendix A). Thus, any behavioral differences across environments must stem not from different action inventories or investigation scaffolds, but from how the same scaffold is interpreted under different narrative frames.

3.2 Persona Design

We define 10 personas grounded in learning-style theory from educational psychology (Kolb, 2014; Felder et al., 1988). The choice of *personality-level descriptions* is deliberate. Role-based personas (e.g., "you are a cautious analyst") and explicit behavioral directives (e.g., "always read before talking") would make transfer trivial and reveal little about implicit narrative effects. Our design instead mirrors a realistic deployment scenario in which practitioners assign personality traits and expect those traits to shape behavior across task domains.

The 10 personas vary along four dimensions: processing speed (*quick_intuitive* vs. *methodical_thorough*), source preference (*social_collaborative* vs. *coverage_focused*), decision confidence (*confident_decisive* vs. *verification_seeking*), and investigation strategy (*hands_on_practical*, *hypothesis_driven*, *big_picture_conceptual*, *effortful_learner*). Table 1 lists each persona's core tendency alongside the action type most directly implied by its description, a distinction that becomes central to the analysis of cross-narrative transferability (§4.3).

Each persona prompt has three components: (1) a personality description (∼80 words) identical across narratives; (2) a stage-specific behavioral reminder of 1–2 sentences; and (3) task instructions describing the agent's role, available resources and constraints. Only the third component changes across narratives, and describes available actions non-prescriptively. See Appendix B for persona descriptions.

3.3 Models and Configuration

We evaluate three models spanning different architectures and scales: LLaMA-3.1-70B (open-source), GPT-4o (proprietary), and GPT-4o-OSS-120B (open-source). This allows us to test whether narrative priors are architecture-specific or reflect shared properties of large-scale pretraining. For each model–narrative pair, we run all 10 personas, each across 7 scenarios with 3 independent trials, 10 personas $\times$ 7 scenarios $\times$ 3 trials $\times$ 9 model-environment pairs $= 1,890$ sessions. Each trial is an independent API session with no shared state, and a sampling temperature of $T = 0.7$.

3.4 Behavioral Features

For each session we extract 16 normalized behavioral features organized in six categories:

1. **Speed & efficiency**: average turns per stage; exploration efficiency (unique locations / total moves).
2. **Information gathering**: `read`, `talk`, and `test` ratios; information depth (information actions / unique sources).
3. **Exploration pattern**: `move` ratio; location coverage; revisit rate.
4. **Social behavior**: social breadth (unique characters interacted with / available); social dependency (`talk` / information actions).
5. **Decision patterns**: action diversity; decision consistency; hesitation index.
6. **Thoroughness**: resource coverage (unique resources / total available); experimentation rate (`test` / information actions).

Four of the 16 features are computed as proportions of total core action types: $\text{ratio}_a = \frac{n_a}{n_{\texttt{read}}+n_{\texttt{talk}}+n_{\texttt{test}}+n_{\texttt{move}}}$. These ratios sum to 1, giving a compositional snapshot of each agent's action type profile. The remaining 12 features describe how an agent navigates and decides rather than which actions it takes (e.g., exploration efficiency, decision consistency, resource coverage). This ensures that the narrative dominance we report in §4.1 is not driven solely by action-type encoding.

These features serve two roles. First, they provide a measurement language for describing how behavior differs across personas and narratives. Second, they define a behavioral space in which we measure transfer. Full feature definitions are provided in Appendix C.

4 Results

4.1 Task Narrative Dominates Behavior

The first question is straightforward: between persona and task narrative, which one actually controls what the agent does?

Variance decomposition. A two-way ANOVA (Narrative $\times$ Persona) on trial-averaged data (70 observations per model–narrative cell; $N = 630$ per model) reveals a stark asymmetry. For the three information-gathering actions (`read`, `talk`, `test`), task narrative explains 5–31$\times$ more variance than persona, with all narrative effects highly significant ($p < 10^{-15}$). The single exception is the `move` ratio for LLaMA, where persona slightly exceeds narrative ($\eta^2 = .25$ vs. .21). Persona effects are significant in every case ($p < .05$); personas do shape behavior, but narrative overwhelms them (Figure 2).

Random Forest classifiers trained on the full 16-feature profiles reinforce this picture: task narrative is predicted at 99.7–100% accuracy across all models, while persona classification reaches only 24.6–41.3% (chance: 33% and 10% respectively). This gap holds even when action-type ratios are excluded (§6). We note that ANOVA η^2 is the primary evidence for narrative dominance; this comparison is conservative for narrative, which has only 3 levels versus persona's 10 and would therefore be expected to explain less variance.

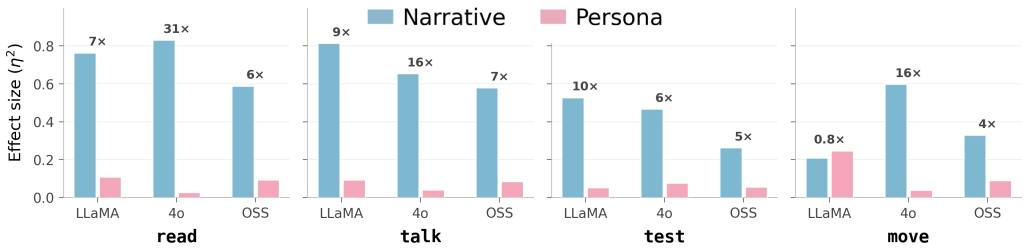

Figure 2: Two-way ANOVA effect sizes (η^2) for Narrative and Persona on the four core action ratios. Labels show the ratio of narrative to persona η^2. Task narrative explains 5–31× more variance than persona for information-gathering actions (`read`, `talk`, `test`). The sole exception is `move` for LLaMA, where persona slightly exceeds narrative (0.8×). Classification accuracy (Random Forest, 16 features): narrative 99.7–100%, persona 24.6–41.3%.

Narrative priors are not strategic adaptations. The behavioral signatures induced by narrative are consistent across all three model architectures: SD elicits the highest `read` ratio, CI the highest `talk` ratio, and MM the highest `test` ratio (Appendix G). This pattern mirrors domain stereotypes in pretraining text: medical contexts co-occur with consultation language, IT contexts with documentation, crime contexts with forensic investigation.

If these patterns reflected rational adaptation, the narrative-biased action should predict task success. It generally does not. Reading more in SD does correlate with success ($r \approx +.34$, $p < .001$ across all models), but the `talk` bias in CI is negatively correlated with success in 2 of 3 models ($r = -.29$, $p < .001$ for GPT-4o), and the `test` bias in MM is negatively correlated in 2 of 3 models ($r = -.25$, $p < .001$ for GPT-OSS). Full correlations are in Appendix G. The story framing of a task causes agents to adopt domain-stereotypical behaviors that, more often than not, hurt their performance.

Persona effects are narrative-confined. Within a fixed narrative, persona explains substantial variance (mean η^2 up to 0.64 in SD) and drives success rates from 0% to 100% across personas ($\chi^2 = 92.3$, $p < .001$ for CI in LLaMA). Within-narrative persona classification reaches 31.9% (3.2× chance), with anchor-rich personas most identifiable (*social_collaborative*: 90.5% in SD; Appendix H). But this signal collapses across narrative boundaries: cross-narrative transfer averages 10.4%, barely above the 10% baseline. Personas shape behavior powerfully within a narrative; the narrative determines which behavioral repertoire the persona operates within.

4.2 Cross-Narrative Persona Consistency

We define the Behavioral Consistency Index (BCI) to measure which personas maintain recognizable signatures across narratives. For persona p in narrative g, let $\vec{b}_p^{(g)} \in \mathbb{R}^{16}$ be the mean behavioral feature vector. We z-normalize within each narrative to remove narrative-level shifts:

$$\hat{b}_{p,f}^{(g)} = (b_{p,f}^{(g)} - \mu_f^{(g)})/\sigma_f^{(g)} \tag{1}$$

The BCI for persona p is the mean pairwise Pearson correlation of z-normalized profiles across all $\binom{3}{2} = 3$ narrative pairs:

$$\text{BCI}_p = \frac{1}{3} \sum_{(g_i, g_j)} \text{corr}\left(\hat{\vec{b}}_p^{(g_i)}, \hat{\vec{b}}_p^{(g_j)}\right) \tag{2}$$

A persona with high BCI deviates from the narrative mean in the same direction and by similar magnitudes across all three games; a persona with BCI near zero or negative reshapes its behavioral profile with each new narrative.

Figure 3(a) reports the results. Five personas maintain positive BCI across all three models, with *social_collaborative* achieving the highest cross-model mean (+.44). At the other extreme,

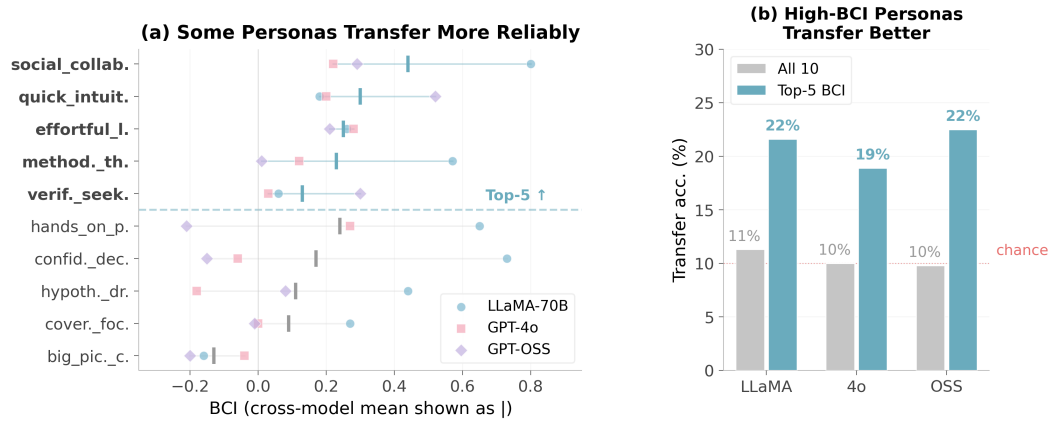

Figure 3: **(a)** Behavioral Consistency Index (BCI) by persona. Each point shows a single model; the vertical bar marks the cross-model mean. The top-5 personas (above dashed line) maintain positive BCI across all three models. **(b)** Cross-narrative classification accuracy using all 10 personas vs. the top-5 BCI subset. High-BCI filtering doubles transfer accuracy relative to the full set (chance = 10%).

big_picture_conceptual is the only persona with consistently negative BCI ($-.13$). A permutation null baseline (1,000 random persona-label shuffles) places the 95th percentile at $+.33$, exceeded by several persona–model combinations (marked with † in Appendix F). Notably, BCI varies across models for the same persona (e.g., *confident_decisive*: $+.73$ in LLaMA, $-.15$ in GPT-OSS), indicating that transferability is partially model-dependent.

We validate BCI through cross-narrative classification. Random Forest classifiers trained on one narrative and tested on another achieve 10.4% accuracy across all 10 personas. Restricting to the top-5 BCI personas doubles this to 21.0% (Figure 3(b)), and the advantage is fully preserved under a leave-one-narrative-out protocol (+11.6pp; Appendix G). BCI also shows negligible correlation with human-rated prompt specificity ($\rho = 0.12$, $p = 0.75$), ruling out prompt writing quality as a confound.

4.3 What Explains BCI? Behavioral Anchors

What separates a persona that survives narrative change from one that dissolves into it? We hypothesize that the answer lies in *behavioral anchors*: features that deviate consistently from the narrative-specific mean in the same direction across all narratives. Formally, feature f is an anchor for persona p if $|\hat{b}_{p,f}^{(g)}| > \tau$ with consistent sign across all g ($\tau = 0.3$; robust across $[0.2, 0.5]$).

Social_collaborative in LLaMA has 6 anchors: "prefers learning through conversation" maps directly to `talk`, producing a stable signature across games. *Big_picture_conceptual* has 0 anchors: "high-level thinking" lacks a stable action-level operationalization. Regression across all 30 model–persona pairs confirms this ($\text{BCI}_p = \alpha + \beta \cdot A_p + \epsilon$; $R^2 = 0.65$, $p < 10^{-7}$). But correlation does not establish causation; §5 tests this directly.

4.4 Persona Selection Algorithm

The anchor analysis yields a practical method for selecting transferable personas without target-narrative data. Given behavioral data from two or more source narratives, we compute BCI from source data only, rank personas by source-BCI, and deploy the top-K in the target narrative. Across all 9 model–target combinations with $K = 5$, source-BCI selection improves cross-narrative identifiability by +12.0pp on average, winning in all 9 conditions (Table 5 in Appendix G). The improvement is monotonic in K: at $K = 3$, +25.4pp;

at $K = 7$, +3.0pp. Source-only selection agrees with an oracle on 3.9/5 personas on average, with perfect agreement in 3 of 9 conditions.

5 Intervention Experiments

Sections 4.1–4.4 establish three empirical regularities: task narrative can shape behavior more strongly than persona, some personas transfer more reliably than others, and behavioral anchors strongly predict that transfer. But these analyses are primarily observational. We now ask three sharper questions. First, can a narrative prior be predicted in a previously unseen domain? Second, is the link between behavioral anchors and cross-narrative consistency genuinely causal? Third, does BCI capture something more informative than a simple text-level heuristic? We address each through experiments with GPT-4o-based agents.

Predicting Narrative Priors in a New Domain If narrative priors reflect domain-level associations in pretraining corpora, then the behavioral bias of an unseen narrative should be predictable. We test this by constructing a fourth isomorphic game, **Cooking Kitchen (CK)**, which preserves all structural invariants. The agent plays a kitchen inspector investigating a dish preparation problem. Because cooking contexts often co-occurs with collaboration, coordination, and social activity, we pre-register the prediction that CK will elicit a higher `talk` ratio than the mean of the original three narratives.

This prediction is confirmed: `talk` ratios form a graded hierarchy (CI .286 > CK .213 > MM .183 ≫ SD .062), with CK significantly exceeding the 3-game mean of .177 ($t = 5.64$, $p < .001$, $d = 0.48$; Appendix G). The BCI framework also generalizes cleanly: persona rankings are remarkably stable across 3 vs. 4 narratives ($\rho = 0.952$, $p < .001$), the top-5 persona set remains identical, and the anchor count regression improves slightly ($R^2 = 0.68$ vs. 0.65). These results suggest that both narrative priors and behavioral anchors capture regularities that extend beyond the initial three environments.

Testing the Causal Role of Behavioral Anchors The anchor count–BCI regression ($R^2 = 0.65$) is observational. To test whether anchors *cause* consistency, we directly manipulate anchor count while holding psychological content constant. We select the *social_collaborative* persona (BCI = +.44, 2 anchors: "conversation" and "discussion") and construct an abstract rewrite that preserves the collaborative disposition but removes concrete action words:

Original: "Strongly prefers learning through conversation and discussion. Naturally seeks others' perspectives, starts by getting oriented through conversation before consulting other resources."

Abstract: "Values collaborative epistemic exchange and dialogic sense-making. Gravitates toward intersubjective knowledge construction, preferring to establish shared understanding before engaging with static information sources."

Both describe a collaborative learner who prioritizes shared understanding before written sources; the difference is that the original maps this disposition to a concrete action (`talk`), while the abstract version does not. We run 63 sessions per version (3 games × 7 scenarios × 3 trials). The effect is sharp. Removing the anchor words reduces BCI from +.22 to +.01, a 95% drop (Table 2). The collapse is most dramatic in the SD-MM pair, where the original persona's "conversation" anchor created a stable social-action signal that vanishes without the concrete word. At the feature level (Table 3), the original persona maintains consistently elevated `talk` behavior across all three narratives ($z > +0.3$), while the abstract version's social signal is substantially attenuated.

Multiple inferential checks support the same conclusion. A trial-level bootstrap (10,000 resamples) yields a ΔBCI 95% CI of $[-.38, -.04]$, excluding zero. A Fisher z-transformation on the SD-MM pair yields $z = 1.94$, $p = .026$ (one-tailed), and a permutation test (5,000 label shuffles) yields $p = .034$. The observed ΔBCI of -0.21 also matches the prediction from the anchor count regression almost exactly: $\beta \times \Delta A = 0.11 \times (-2) = -0.22$. The convergence

Pair	Orig.	Abst.	Δ
CI-SD	−.23	−.18	+.05
CI-MM	+.13	−.02	−.15
SD-MM	+.76	+.22	−.54
BCI	**+.22**	**+.01**	**−.21**

Table 2: Pairwise correlations for *social_collaborative*: original vs. abstract rewrite. Removing anchors reduces BCI by 95%.

Narr.	Orig. $z_{\texttt{talk}}$	Abst. $z_{\texttt{talk}}$
CI	+0.31	+0.12
SD	+0.88	+0.34
MM	+0.72	+0.15

Table 3: Z-normalized `talk` scores. The abstract version's social signal largely disappears.

of manipulation, bootstrapping, and regression prediction provides strong evidence that behavioral anchors are causal.

We replicate this intervention on a second persona, *coverage_focused*, which uses concrete action vocabulary ("examine each available resource once") but has a very different behavioral profile. Removing concrete action words via an abstract rewrite reduces BCI from $+.862$ to $+.370$ ($\Delta = -0.492$, permutation $p < 0.0001$, effect size 7.1 SD vs. null), with the drop concentrating in CI-related pairs (CI-SD: $+.80 \rightarrow -.05$; CI-MM: $+.91 \rightarrow +.34$). This independently confirms that behavioral anchors are causally responsible for cross-narrative consistency.

Behavioral vs. Textual Predictors of Transfer A natural objection is that BCI may simply discover a superficial text-level heuristic: personas transfer when they contain more action-related words. We test this with an action-word (AW) baseline that counts manually specified action-linked words in each persona description (Appendix D). BCI significantly predicts cross-narrative transfer accuracy ($\rho = 0.68$–0.76, $p < .05$), while AW does not ($\rho = 0.38$–0.55, $p > .10$). BCI-based top-5 outperforms AW-based top-5 by +4.8pp on average. Two concrete cases make this constrast intuitive: *quick_intuitive* contains zero action words yet ranks 2nd in BCI (its "skim and move on" style is behaviorally distinct), while *coverage_focused* contains 4 action words but ranks 9th (its words map to every action type, producing a flat profile). More fundamentally, BCI is model-adaptive (cross-model $\rho < 0.24$); a fixed text metric cannot capture these model-specific behaviors.

Robustness to Stronger Prompts and Abstract Labels Two additional experiments test the boundaries of narrative priors. First, we replace personality-level personas with explicit behavioral directives mandating at least 60% of actions be a target type (`read_directive`, `talk_directive`, `test_directive`; 63 sessions on GPT-4o). The directives are effective: target actions increase in 8 of 9 cells (Cohen d up to $+5.05$). Yet the cross-narrative range of each directive's target ratio (0.17–0.27 across games) matches or exceeds the original cross-narrative range without any directive (0.13–0.33). Even under explicit behavioral mandates, narrative framing governs how much the agent complies. Notably, `test_directive` on MM fails entirely to raise the test ratio above its narrative baseline ($\Delta = -0.006$, $p = 0.71$), because MM's forensic-testing prior already saturates that action.

Second, we replace all action verbs with abstract codes (`read` → `ACTION_ALPHA`, `talk` → `ACTION_BETA`, etc.) via an API-layer wrapper (21 sessions on GPT-4o). None of the three narrative priors weaken; two intensify. In CI, the `talk` ratio rises from .310 to .518 (+67%, $p = .014$) while `read` collapses from .291 to .036. In SD, the dominant `read` ratio is preserved at .587 ($p = .68$). In MM, the `test` ratio rises from .178 to .319 (+79%, $p = .27$ with high variance at $N = 7$). Removing verb-level cues makes the model rely more heavily on narrative context, ruling out the surface-semantics interpretation and localizing narrative priors to the domain framing itself.

6 Robustness Analyses

We address five potential concerns. First, the four action-type ratios might trivially encode narrative identity; however, replicating all analyses with only the 12 non-ratio features

yields unchanged narrative prediction accuracy (99.7–100.0%) and strongly correlated BCI values (Spearman $\rho = 0.84$–0.99). Second, personas do affect more than action selection: 4–6 of 6 structural features show significant persona effects ($p < .05$, η^2 up to 0.29) within each narrative, including exploration patterns, decision consistency, and resource coverage. Third, personas are not simply ineffective prompts: within-narrative η^2 reaches 0.64, and per-game success rates range from 0% to 100% across personas. Fourth, anchor count is not confounded by prompt quality: it shows negligible correlation with prompt length ($\rho = 0.08$) and human-rated specificity ($\rho = 0.15$). Fifth, to verify that ANOVA results are not driven by distributional assumptions on bounded proportion data, we reran all 12 model $\times$ action-ratio comparisons under Kruskal-Wallis tests and binomial GLM with logit link. All narrative-dominance conclusions hold ($p < 10^{-9}$ in 11 of 12 cells), with effect-size ratios comparable to or larger than those under ANOVA.

7 Discussion

Narrative priors as implicit behavioral bias. Our central finding is that LLM agents exhibit *narrative priors*: implicit behavioral biases activated by story framing, independent of decision structure. The biases are consistent across architectures ($r = .96$), hurt task success in two of three narratives, modulate structural features beyond action frequencies, and are predictable in a novel domain (§5, $\rho = 0.952$). This poses a concrete reliability challenge. When a practitioner deploys the same persona across domains, the behavioral profile will be driven more by narrative than by the assigned persona, and the narrative-induced behaviors may actively hurt performance ($r = -.29$, $p < .001$ for `talk` in CI).

Behavioral anchors. Persona descriptions that create direct semantic–action bindings ("prefers conversation" $\rightarrow$ `talk`) produce portable behavioral signatures; abstract descriptions do not. The causal experiment (§5) confirms this on two independent personas: removing concrete action words reduces BCI by 95% for *social_collaborative* and by 57% for *coverage_focused*, with the observed drops matching regression predictions. This yields a design principle: for cross-narrative consistency, personas should reference concrete actions rather than abstract dispositions. More broadly, it suggests that the effectiveness of natural-language behavioral instructions may depend critically on how directly they map onto the available action space.

Practical implications. These findings yield concrete guidance for agent builders: to achieve portable behavior across task domains, persona descriptions should reference concrete actions (e.g., "prioritize reading documents") rather than abstract dispositions (e.g., "be thorough"). Even explicit behavioral directives mandating specific action frequencies do not override narrative priors (§5), and narrative priors persist when action labels are fully abstracted, localizing their source to domain framing rather than action-verb semantics.

Limitations. Our narratives belong to the sequential investigation genre; extending the isomorphism framework to structurally different task families such as planning, tool-use, or multi-agent collaboration is an important next step. The 10 personas vary along learning-style dimensions; our supplementary directive experiment suggests narrative priors persist even under explicit behavioral mandates. We analyze action-level behavior rather than linguistic output; analyzing chain-of-thought content could reveal how narrative priors form at the generation level. Finally, BCI rankings show limited cross-model consistency ($\rho < 0.24$), suggesting persona selection may need per-model calibration.

Even with these limitations, the main message is clear: persona prompting is not a context-invariant control interface. The story wrapped around a task can shape agent behavior more strongly than the persona itself. Understanding, predicting, and mitigating that effect is likely to be important for building LLM agents with reliable behavior across domains. Our structural isomorphism framework provides a diagnostic for measuring this effect, and our behavioral anchor analysis offers a practical path toward personas that are robust to it.

References

Razan Baltaji, Babak Hemmatian, and Lav Varshney. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent llm collaboration. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pp. 17–31, 2024.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*, 2024.

Richard M Felder, Linda K Silverman, et al. Learning and teaching styles in engineering education. *Engineering education*, 78(7):674–681, 1988.

Ivar Frisch and Mario Giulianelli. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pp. 102–111, 2024.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*, 2023.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.

David A Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15174–15186, 2024.

Jonathan P Rowe, Lucy R Shores, Bradford W Mott, and James C Lester. Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 21(1-2):115–133, 2011.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 8(9), 2024.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, 2024.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, 2024.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, 2024a.

Pengcheng Wang, Jonathan P Rowe, Wookhee Min, Bradford W Mott, and James C Lester. High-fidelity simulated players for interactive narrative planning. In *IJCAI*, volume 18, pp. 13–19, 2018.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1840–1873, 2024b.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 2300–2344, 2022.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. Character is destiny: Can large language models simulate personadriven decisions in role-playing. *arXiv preprint arXiv:2404.12138*, 2024.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

A Structural properties

Property	CI	SD	MM
Stages	4	4	4
Locations	6	6	6
Characters	7	5	7
Documents	12	12	15
Core action types	4	4	4
Scenarios	7	7	7

Table 4: Structural properties of the three narrative environments. Core decision architecture is identical across environments; minor differences in resource counts reflect narrative packaging rather than changes to the underlying investigation structure.

B Persona Descriptions

Each persona prompt consists of three components: (1) a personality description (∼80 words), (2) a stage-specific behavioral reminder, and (3) task instructions (game rules, single-access constraints). Below we provide the personality descriptions and reminders. The task instructions are identical across personas and enforce the structural invariants described in §3.1. Across task narratives, only the role label (student/engineer/detective) and domain-specific resource names change; the personality content is held constant.

Quick Intuitive. Learns quickly through pattern recognition. Trusts first impressions, skims for key points, comfortable with partial information. Prefers to gather information efficiently and move on rather than lingering. *Reminder: Trust your pattern recognition and move efficiently. Skim for key patterns and connections.*

Methodical Thorough. Careful, systematic, prefers solid evidence before deciding. Reads important information carefully and verifies key facts by consulting different source types. Balances thoroughness with efficiency. *Reminder: Be thorough on first pass, focus careful attention where it matters most. Verify by consulting different sources, not re-reading.*

Social Collaborative. Strongly prefers learning through conversation and discussion. Naturally seeks others' perspectives, starts by getting oriented through conversation before consulting other resources. Understands technical details eventually require written sources. *Reminder: Start with conversations to get oriented, then use other resources for specific details.*

Effortful Learner. Finds dense content genuinely challenging. Reads slowly and carefully, focuses on main ideas and key facts. Persistent in completing tasks even when content is difficult. Accepts partial understanding rather than getting stuck. *Reminder: Read carefully on first pass, focus on main points, keep moving forward even with partial clarity.*

Confident Decisive. Trusts own judgment, commits quickly once sufficient information is gathered. Does not second-guess or proactively seek contradicting information. Willing to change mind if new information contradicts conclusion, but does not dwell on uncertainty. *Reminder: Make clear decisions with confidence. Once decided, trust your judgment and move forward.*

Verification-Seeking. Wants convergent evidence from multiple different sources before deciding. Seeks corroboration across diverse source types. Cross-checks across sources, not within the same source. *Reminder: Look for convergence across different source types. Cross-check across sources, not within the same source.*

Hands-On Practical. Learns best by doing and testing. Prefers direct verification over reading. Finds practical experimentation more engaging than lengthy explanations. Reads when necessary but keeps it brief. *Reminder: Prefer hands-on investigation when possible. Keep reading brief and move to practical action.*

Coverage-Focused. Systematic and comprehensive in information gathering. Wants to examine each available resource once but thoroughly. Feels more confident when coverage is complete. Thoroughness means covering many sources once, not re-examining. *Reminder: Be systematic and comprehensive. Cover all available resources once but thoroughly.*

Big-Picture Conceptual. Focuses on overall concepts and patterns rather than specific details. Skims fine details, good at seeing connections between concepts. Sometimes misses important details but understands overall logic. *Reminder: Focus on overall concepts and main ideas. Look for the big picture rather than getting caught up in details.*

Hypothesis-Driven. Forms quick hypotheses and tests through action rather than extensive research. Comfortable with initial guesses, sees incorrect attempts as informative. Adaptable and willing to change mind based on feedback. *Reminder: Form quick hypotheses and test them. Gather enough to make a reasonable guess, then act.*

Domain Mapping. Across task narratives, personas use the following role and resource mappings: CI uses a 10th-grade student investigating disease (read books/posters, talk to NPCs, test items); SD uses a tech support engineer troubleshooting IT issues (read logs/docs, talk to users/experts, run diagnostics); MM uses a detective solving a criminal case (read documents/evidence, interview witnesses/suspects, verify alibis).

C Behavioral Feature Definitions

All 16 features are organized into six categories:

Speed & Efficiency (2). `avg_turns_per_stage`: mean turns per completed stage. `exploration_efficiency`: unique locations / total moves.

Information Gathering (4). `read_ratio`, `talk_ratio`, `test_ratio`: proportion of core actions devoted to each type. `info_depth`: information actions / unique sources.

Exploration Pattern (3). `move_ratio`: proportion of core actions that are moves. `location_coverage`: unique locations / 6. `revisit_rate`: revisits / total visits.

Social Behavior (2). `social_breadth`: unique characters / available characters. `social_dependency`: talk actions / information actions.

Decision Patterns (3). `action_diversity`: unique action types / available types. `decision_consistency`: $1 - \mathrm{CV}(\text{turns per stage})$. `hesitation_index`: location revisits / moves.

Thoroughness (2). `resource_coverage`: unique resources / total available. `experimentation_rate`: test actions / information actions.

D Action-Word List

The action-word (AW) baseline (§5) counts occurrences of the following terms in each persona description: *read*, *reads*, *conversation*, *discussion*, *test*, *testing*, *tests*, *verify*, *verifies*, *examine*, *check*, *cross-check*, *hands-on*, *practical*, *doing*, *systematic*, *thorough*, *comprehensive*.

E Abstract Persona Rewrite

Section 5 compares the original *social_collaborative* persona with an abstract rewrite. Below we provide the full prompt components for both versions.

Original – Personality Description. "Strongly prefers learning through conversation and discussion. Naturally seeks others' perspectives, starts by getting oriented through conversation before consulting other resources. Understands technical details eventually require written sources."

Original – Stage Reminder. "Start with conversations to get oriented, then use other resources for specific details."

Abstract – Personality Description. "Values collaborative epistemic exchange and dialogic sense-making. Gravitates toward intersubjective knowledge construction, preferring to establish shared understanding before engaging with static information sources. Recognizes that formalized knowledge eventually requires engagement with codified materials."

Abstract – Stage Reminder. "Initiate through collaborative sense-making to establish orientation, then engage with codified sources for specific elaboration."

The abstract version preserves the same cognitive disposition (collaborative, shared-understanding-first, written-sources-second) while removing all concrete action words (*conversation*, *discussion*, *talk*) that could directly map to the `talk` action type. The task instructions component (game rules, available actions, structural constraints) is identical across both versions.

F Per-Model BCI Tables

Full pairwise correlations for each model (format: CI-SD, CI-MM, SD-MM, BCI):

LLaMA-3.1-70B.

social_collab.: +.82, +.76, +.81, **+.80**†.
confident_dec.: +.69, +.81, +.69, +.73†.
hands_on_prac.: +.64, +.82, +.48, +.65†.
method._thor.: +.56, +.61, +.54, +.57†.
hypoth._driv.: +.57, +.26, +.48, +.44†.

GPT-4o.

effortful_learn.: +.32, −.09, +.61, +.28.
hands_on_prac.: +.29, +.34, +.19, +.27.
social_collab.: −.23, +.13, +.76, +.22.
quick_intuit.: −.00, +.37, +.23, +.20.
hypoth._driv.: −.02, −.31, −.21, −.18.

GPT-4o-OSS-120B.

quick_intuit.: +.62, +.35, +.61, +.52†.
verif._seek.: +.19, +.26, +.44, +.30.
social_collab.: +.58, −.03, +.32, +.29.
big_pict._conc.: +.02, −.02, −.59, −.20.
hands_on_prac.: −.20, +.14, −.57, −.21.

G Additional Results

Model	Target	All 10	Top 5	Δ
LLaMA	CI	11.4%	21.4%	+10.0
	SD	10.2%	19.5%	+9.3
	MM	13.6%	23.8%	+10.2
GPT-4o	CI	11.0%	21.4%	+10.5
	SD	10.5%	15.7%	+5.2
	MM	9.8%	20.5%	+10.7
GPT-OSS	CI	9.5%	19.0%	+9.5
	SD	9.3%	34.8%	+25.5
	MM	11.4%	28.1%	+16.7
Average		10.7%	**22.7%**	**+12.0**

Table 5: Persona selection using source-narrative BCI only (no target data). The selected subset improves transfer identifiability in all 9 conditions.

Narrative	talk	Domain
CI	.286	Medical
CK	.213	Cooking
MM	.183	Crime
SD	.062	IT
Mean	.177	—

Table 6: talk ratio by narrative (GPT-4o). CK exceeds the 3-game mean (d=0.48, p<.001).

Persona	BCI-3	BCI-4	Δ
social_col.	+.22	+.25	+.03
effortful_l.	+.28	+.24	−.04
hands_on_p.	+.27	+.23	−.04
quick_int.	+.20	+.22	+.02
method._th.	+.12	+.14	+.02
verif._seek.	+.03	+.07	+.04
cover._foc.	−.00	+.02	+.02
confid._dec.	−.06	−.03	+.03
big_pic._c.	−.04	−.08	−.04
hypoth._dr.	−.18	−.14	+.04

Table 7: BCI over 3 vs. 4 narratives. Rankings stable (ρ=0.952); top-5 identical.

Feature	Model	CI	SD	MM
read_ratio	LLaMA	.21	**.52**	.36
	GPT-4o	.24	**.58**	.39
	GPT-OSS	.27	**.63**	.46
talk_ratio	LLaMA	**.39**	.08	.21
	GPT-4o	**.29**	.06	.18
	GPT-OSS	**.30**	.06	.14
test_ratio	LLaMA	.03	.06	**.15**
	GPT-4o	.06	.07	**.19**
	GPT-OSS	.03	.04	**.14**

Table 8: Mean action ratios by task narrative and model. Bold marks the highest value per row. All three models exhibit the same pattern: SD elicits the most reading, CI the most social interaction (talk), and MM the most testing.

H Confusion Matrices

Tables 15–16 show per-persona classification accuracy (diagonal) from the within-narrative and cross-narrative settings described in §4.1. We report the diagonal (correct classification rate per persona) for compactness; full 10 × 10 matrices are available in our code release.

Narrative bias	Model	r	Opt.?
SD → `read`	LLaMA	+.35***	✓
	GPT-4o	+.32***	✓
	GPT-OSS	+.34***	✓
CI → `talk`	LLaMA	+.18*	∼
	GPT-4o	−.29***	✗
	GPT-OSS	−.13	✗
MM → `test`	LLaMA	−.14*	✗
	GPT-4o	+.04	✗
	GPT-OSS	−.25***	✗

Table 9: Correlation between narrative-biased action and task success. Only `read` in SD is beneficial; the biases in CI and MM hurt or do not help performance.

Model	All 10	Top 5	Δ
LLaMA-70B	11.4%	21.4%	+10.1
GPT-4o	10.0%	19.4%	+9.4
GPT-OSS-120B	9.8%	25.1%	+15.3
Average	10.4%	**22.0%**	+11.6

Table 10: Leave-one-narrative-out validation. BCI computed from two task narratives; transfer tested on the held-out third.

Feature	Model	CI	SD	MM
`read`	LLaMA	+.30***	+.35***	+.21**
	GPT-4o	+.44***	+.32***	−.13
	GPT-OSS	+.24***	+.34***	+.35***
`talk`	LLaMA	+.18*	−.18**	+.44***
	GPT-4o	−.29***	−.01	+.03
	GPT-OSS	−.13	−.02	−.15*
`move`	LLaMA	−.43***	−.34***	−.35***
	GPT-4o	+.05	−.35***	+.08
	GPT-OSS	−.06	−.56***	+.05

Table 11: Feature–success correlations by task narrative. * $p < .05$, ** $p < .01$, *** $p < .001$.

	LLaMA	GPT-4o	GPT-OSS
CI→SD	20.0%	12.4%	7.6%
CI→MM	20.0%	20.0%	17.1%
SD→CI	20.0%	21.0%	19.0%
SD→MM	28.6%	21.0%	25.7%
MM→CI	21.9%	23.8%	19.0%
MM→SD	19.0%	15.2%	46.7%
Mean	21.6%	18.9%	22.5%

Table 12: Full cross-narrative transfer accuracy for top-5 BCI personas.

I Cooking Kitchen Game Details

The Cooking Kitchen (CK) game (§5) preserves all structural invariants of the original three games. Below we summarize its domain-specific configuration.

Setting. A kitchen inspector investigates a dish preparation problem across a professional kitchen facility.

Model	ρ_{BCI}	p	ρ_{AW}	p
LLaMA-70B	.76	.011	.55	.10
GPT-4o	.68	.031	.38	.28
GPT-OSS-120B	.72	.019	.41	.24

Table 13: Correlation of BCI and AW count with per-persona transfer accuracy. BCI is significant in all models; AW is not.

Model	All 10	AW Top-5	BCI Top-5	Δ
LLaMA-70B	11.3%	17.8%	**21.6%**	+3.8
GPT-4o	10.0%	14.2%	**18.9%**	+4.7
GPT-OSS-120B	9.8%	16.5%	**22.5%**	+6.0
Avg.	10.4%	16.2%	**21.0%**	+4.8

Table 14: Transfer accuracy by selection method. BCI top-5 outperforms AW top-5 in all models. Random-5 mean: 13.1% [10.8%, 15.6%].

Persona	CI	SD	MM
social_collab.	42.9%	**90.5%**	54.0%
hands_on_prac.	36.5%	**63.5%**	39.7%
coverage_focused	36.5%	44.4%	14.3%
verification_seek.	31.7%	38.1%	25.4%
confident_dec.	33.3%	33.3%	31.7%
methodical_thor.	20.6%	20.6%	28.6%
big_picture_conc.	27.0%	23.8%	22.2%
quick_intuitive	22.2%	25.4%	15.9%
hypothesis_driv.	14.3%	27.0%	19.0%
effortful_learn.	19.0%	27.0%	**0.0%**
Mean	28.4%	39.4%	25.1%

Table 15: Within-narrative per-persona classification accuracy (diagonal of confusion matrix). Personas with concrete behavioral anchors (*social_collaborative*, *hands_on_practical*) are most identifiable. *Effortful_learner* is unidentifiable in MM.

Direction	Accuracy	Dominant Prediction
CI → SD	6.7%	social_collab. (76.3%)
CI → MM	8.6%	social_collab. (83.0%)
SD → CI	10.0%	social_collab. (96.7%)
SD → MM	9.8%	social_collab. (71.4%)
MM → CI	9.5%	hands_on_prac. (35.7%)
MM → SD	10.0%	effortful_learn. (29.4%)

Table 16: Cross-narrative transfer: overall accuracy and the single most-predicted class. In 4 of 6 directions, the classifier collapses to predicting *social_collaborative* for nearly all inputs, reflecting the dominant behavioral bias of the target task narrative.

Stages. (1) Cuisine category → (2) Specific dish → (3) Problem source → (4) Correction.

Locations (6). Main Kitchen, Prep Station, Cold Storage, Pantry, Dining Floor, Office.

Action Mapping. `read`: recipe cards, inspection logs, ingredient lists, supplier documents. `talk`: head chef, sous chef, line cooks, servers, kitchen manager. `test`: taste dishes, check temperatures, inspect ingredients, verify storage conditions. `move`: navigate between the 6 locations.

Scenarios (7). Each scenario involves a different cuisine category (e.g., Italian, Japanese, Mexican) with a unique dish, problem source, and correction. Scenario content is generated dynamically from a fixed configuration to ensure structural equivalence with the original games.

Role. The agent plays a kitchen inspector. The persona descriptions are identical to those used in the original three games (Appendix B), with only the role label and domain-specific resource names changed.