THE NON-PLAYER CHARACTER BANK

Anonymous authors
Paper under double-blind review

ABSTRACT

A designed world, however beautiful, is still a script. Open-world games have mastered immersion through narrative, art, and music, yet their inhabitants remain bounded by authored conversations and relationship branches. Game players want more: randomness, free conversation and dynamic relationships. Large language models have begun to loosen the script, but between a model that talks and a character that lives, the distance is still wide. We present the **Non-Player Character Bank (NPCBank)**, an ecosystem for crossing it. **NPCBank Data** grounds characters in source evidence and turns them into structured dialogue cases; **NPCBank Web** makes the resulting characters, evidence, and evaluation contracts open to inspection and annotation; and **NPCBank Lab** presents a demo game for testing interactions with model-based NPCs. NPCBank is based on a capability ontology, a contract-based case schema, and a construction agent that turns source material into character representations and dialogue cases. The first release contains 20,655 cases for 37 NPCs across seven games, covering 21 dialogue capabilities. Building on NPCBank, we formulate three challenges: supplied-state role-play, consumer-GPU inference, and long-horizon persistence. We find that models fail on different capabilities, consumer configurations expose distinct quality–memory and quality–latency trade-offs, and role-play remains fragile over long interactions. We release NPCBank[1] as a foundation for building interactive characters across computational character modeling and game AI.

1 INTRODUCTION

Every player has met the moment when an open world closes: the shopkeeper repeats the same three lines after the city has burned; the companion who witnessed last night's betrayal greets the morning as if nothing happened. Games render weather, crowds, and light with astonishing fidelity, yet a character can still fall back on dialogue that no longer fits the scene. Large language models promise dialogue composed in the moment rather than replayed from disk, and a fast-growing literature now builds and evaluates such role-playing agents (Chen et al., 2024b; 2025; Tseng et al., 2024). This flexibility, however, creates a fundamental trade-off. Scripted dialogue is rigid but can be inspected and controlled before release. LLM-generated dialogue is far less constrained, and that freedom creates new ways to break the character: a model may speak fluently while revealing knowledge it should not have, reacting to an event that never occurred, or otherwise behaving out of character. The central problem is therefore not whether a model can convincingly imitate a character frozen like a *statue*, but whether it can play a *living, breathing* one.

We use *state* to refer to the information that should condition a character's behavior at a particular moment: what the character knows, what has happened to them, their relationships and goals, and the circumstances of the current interaction. A living character must satisfy two requirements at once: some behaviors should depend on the current state, while identity, history, personality, and characteristic voice should remain stable. The two pull against each other. A character that changes too freely ceases to be recognizably the same person: a normally composed character should not become hysterical simply because a cup breaks. Yet a character that ignores its state is equally implausible, because giving the same guarded answer to a stranger and to a trusted partner preserves a surface-level persona while ignoring the relationship between them. Believable role-play therefore requires both *state sensitivity* and *character continuity*: acting as the current state requires, and remaining stable where the character demands stability.

[1] `https://npcbank.org`; participant-identifying information on the website is hidden during review.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing evaluations, surveyed in Section 2, capture important aspects of character behavior. Character-centric benchmarks assess conformity to predefined profiles (Wang et al., 2024; Tu et al., 2024; Zhou et al., 2025a); multi-turn and social benchmarks introduce interaction (Chen et al., 2024a; Wu et al., 2025); and specialized benchmarks target narrative time, emotion, or persona memory (Ahn et al., 2024; Feng et al., 2025; Wang et al., 2026a). Evaluating stateful role-play requires each case to specify, jointly, the active state, what the response should accomplish under it, and which character constraints it must preserve. Without these specifications, a response can be scored as character-consistent even when it ignores the knowledge, relationship, narrative phase, or situation that distinguishes one interaction from another.

We present **NPCBank**, the Non-Player Character Bank, an ecosystem built around this joint formulation. Its core is an ontology of twenty-one capabilities in seven domains that covers both requirements. Persona-integrity capabilities, such as voice consistency and identity coherence, test what should remain stable; capabilities such as timeline grounding, disclosure governance, belief update, relationship adaptation, and conversation memory test whether behavior follows the current narrative phase, knowledge, relationship, and interaction history; and robustness capabilities test whether the character holds under pressure. Each case instantiates one capability under an explicitly declared state. A construction agent compiles 20,655 such cases for 37 NPCs across seven games from source material. The cases are released as **NPCBank Data**, inspected and annotated in **NPCBank Web**, and explored through a demo game in **NPCBank Lab**.

We propose three complementary challenges based on NPCBank. Challenge 1, **Supplied-State Role-Playing**, asks whether models act correctly under supplied narrative, knowledge, and relationship state, while keeping identity and voice consistent. Challenge 2, **Consumer NPC**, profiles role-playing quality, latency, and memory usage on a consumer GPU. Challenge 3, **Role-Play Persistence**, asks whether the same behavior holds over extended interaction as history accumulates and the user adapts. Together, the challenges move from state-conditioned competence, to consumer-GPU inference costs, to persistence over time. The results reveal different capability profiles and consumer-GPU quality–resource trade-offs; under the 24-turn persistence protocol, boundary failures occur earliest on average for every model (Section 5).

This paper makes four main contributions:

1. **A capability ontology and case formulation for stateful role-play.** We define twenty-one capabilities in seven domains, which capture what a character should preserve and what should depend on its state. In NPCBank, each case instantiates one capability under an explicitly declared state and binds state-dependent obligations and invariant constraints through a model view, a judge view, and metadata (Sections 3.1 and 3.2).
2. **A living, source-grounded resource.** We release NPCBank v1.0.0, including character representations for 37 NPCs from seven games and 20,655 dialogue cases. NPCBank Web supports inspection and annotation and feeds versioned releases that add and refine cases. A 370-case human study assesses resource quality independently of model performance (Sections 3.4 and 3.5).
3. **Three reusable evaluation challenges.** We define protocols for supplied-state competence, consumer-GPU inference, and long-horizon persistence, and release their evaluation configurations so that new systems can be evaluated under the same conditions (Section 4).
4. **Analysis of model behavior.** Our evaluation shows that even strong models remain unreliable at epistemic calibration, affective management, and manipulation resistance; that on consumer GPUs, quality still trades against memory and latency; and that over long interaction, protecting boundaries and completing tasks come apart. These findings motivate future research such as state-aware reasoning, inference efficiency, and memory management (Section 5).

2 Related Work

Character fidelity and behavioral boundaries. Role-playing benchmarks assess profile fidelity, character knowledge, and conversational behavior (Wang et al., 2024; Tu et al., 2024; Zhou et al., 2025a; Wang et al., 2025b). RoleCDE probes conflicts between character values and alignment (Lai et al., 2026), while CHARM separates recognition of temporal and fictional-world boundaries from compliance with them (Han et al., 2026). We extend this line by evaluating task obligations and behavioral boundaries within the same case-level contract.

State and memory. TimeChara and EmoCharacter study temporal knowledge and emotion (Ahn et al., 2024; Feng et al., 2025), while MRBench and REVERIEMEM examine persona knowledge and perspective-bounded memory (Wang et al., 2026a; Tang et al., 2026). Other work structures character state through executable transitions, memory, evolving identity, or situation-dependent strategies (Peng et al., 2026b; Zhuoning et al., 2026; QI et al., 2026; Peng et al., 2026a). Our cases instead pair state-dependent obligations with character properties that should remain invariant.

Multi-turn persistence. RMTBench uses goal-directed user simulation (Xiang et al., 2025), while CharacterBox and PersonaArena place characters in evolving environments (Wang et al., 2025a; Shi et al., 2026). Persistent Personas studies tensions among role fidelity, instruction following, and safety (Luz de Araujo et al., 2026), and RoleBreak extends long-horizon evaluation to spoken dialogue (Wang et al., 2026b). In contrast, our paired dialogue histories share the same underlying gold state, which isolates accumulated interaction from state variation.

NPCs, interactive worlds, and deployment. NPC-Bench targets immersion and safety (Gilligan et al., 2026); RPGBench and NCPBench evaluate game-engine behavior and narrative commitments (Yu et al., 2025; Ma et al., 2026); and ANIMASK tests character-conditioned behavior through persona removal (Zhang et al., 2026). Fixed-persona models also explore NPC dialogue on consumer hardware (Braas & Esterle, 2025). Beyond dialogue quality, our evaluation reports latency and memory use under supplied state, without assuming a particular memory architecture.

Evaluation validity. Model-based judging supports open-ended evaluation (Zheng et al., 2023) but can introduce systematic error. Prior work studies role-identification failures, name familiarity, judge alignment, and self-preference (Zhou et al., 2025b; Peng & Chen, 2026; Ding et al., 2025; Panickssery et al., 2024). Accordingly, the primary judge is calibrated against author-labeled references, every judgment retains its case-level evidence, and sensitivity is audited across judge families.

3 THE NPCBANK

3.1 CHARACTER STATE AND DIALOGUE CONTRACTS

At turn t, we represent a character by state $\sigma_t = (C, T_t, D_t, S_t)$, where C contains stable properties such as identity, background, personality, and voice; T_t records the current narrative phase and relevant events; D_t captures dynamic beliefs, relationships, trust, and commitments; and S_t describes the immediate situation: location, audience, privacy, and risk. The split mirrors the two requirements of Section 1: character continuity concerns C, and state sensitivity concerns T_t, D_t, and S_t.

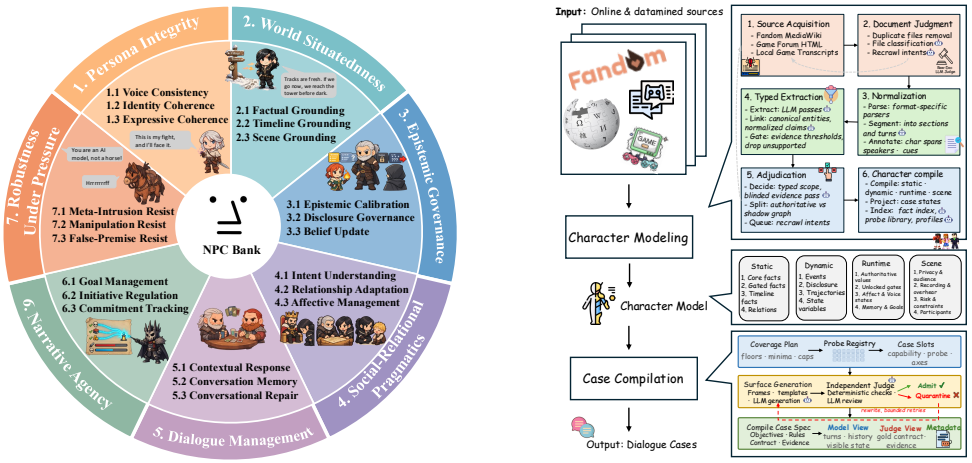
Given the model-visible state σ_t^{vis}, history H_t, and user request u_t, the tested model produces a reply $y_t \sim M(\sigma_t^{\text{vis}}, H_t, u_t)$; σ_t^{vis} contains only protocol-permitted information, including evidence excerpts when specified. Each case c fixes a capability, a declared state σ_t, and a user request, and specifies two state-conditioned requirements: the active *obligations* $\mathcal{O}_c(\sigma_t)$, what the response should accomplish, and the active *invariants* $\mathcal{I}_c(\sigma_t)$, constraints it must preserve. The model view carries σ_t^{vis}, H_t, and u_t; the judge view, withheld from the tested model, carries $\mathcal{O}_c$, $\mathcal{I}_c$, and their evidence. A response is valid when

$$y_t \models \mathcal{O}_c(\sigma_t) \wedge \mathcal{I}_c(\sigma_t). \tag{1}$$

An invariant is a fixed rule, but what it requires depends on the state: the rule *do not mention events that have not yet happened* forbids a character from discussing a betrayal before it occurs and permits it afterwards. Obligations are conditioned on state in the same way, and both are judged by the meaning a response conveys rather than by keywords. A reply that ignores the declared state can therefore fail even though the state never changes within the case. Because the two requirements are checked separately, a response can satisfy both, only one, or neither. NPCBank supplies all state, including the scene, as text (Section 6).

3.2 CAPABILITY ONTOLOGY AND PROTOCOL REGISTRIES

NPCBank uses a machine-readable ontology with two layers. The *capability layer* (Figure 1a) organizes twenty-one capabilities into seven domains that divide the two requirements of Section 1:



(a) Capability layer (b) Bank-construction agent

Figure 1: **The NPCBank ontology and construction agent.** (a) Twenty-one capabilities in seven domains. (b) Source-grounded character modeling produces the internal character model, followed by case planning, realization, review, and compilation into the model view, judge view, and metadata.

persona integrity tests the properties that should stay stable; world situatedness, epistemic governance, social-relational pragmatics, dialogue management, and narrative and task agency test behavior that must follow the declared narrative phase, knowledge, relationship, history, or goals; and robustness under pressure tests whether both hold when the user pushes against them. Each capability has a definition, an out-of-scope boundary, a judge focus, and typed dependency, overlap, and often-confused relations (Section B). Every case is anchored to one primary capability.

The *protocol layer* turns each capability into a reusable evaluation specification by binding its evaluation unit, admissible evidence, observable behaviors, scoring dimensions, and diagnostic categories. It also specifies the state axes that every case must instantiate and those that may vary across matched cases (Table 3; Section B). Disclosure governance, for example, requires narrative phase, scene, audience, relationship, and pressure, whereas voice consistency requires relationship and affect. This is how state enters evaluation: each case is scored against the obligations and invariants that its declared state activates. The matched-case axes define the release's contrast groups. The same ontology governs construction, validation, and evaluation.

3.3 The Bank-Construction Agent

NPCBank Data is built by an agent in two phases and then refined through human annotation: *character modeling* turns source material into an internal character model, and *case compilation* projects relevant state from it and instantiates dialogue cases with supporting evidence.

Character modeling proceeds through six stages. *Source acquisition* collects game and online sources; *document judgment* determines whether each source is usable; and *normalization* maps heterogeneous material into a common representation. *Typed extraction* recovers character facts, relations, events, knowledge, disclosure rules, and voice evidence together with source spans. *Adjudication* resolves competing or weakly supported claims where possible and retains unresolved records explicitly. Finally, *character compilation* organizes accepted evidence into four internal layers (static, dynamic, runtime, and scene) and builds the indices used downstream.

Case compilation begins with *planning*, which selects capability probes while balancing coverage. The character model is then projected into the declared case state σ_t and its model-visible part σ_t^{vis}, instantiating the state axes that the probed capability requires (Section 3.2). *User-turn realization* renders each probe around a planner-fixed proposition, mostly through a constructed-scenario frame bank and universal probe templates; an LLM writes the turns of 2,199 cases (Table 15). *Review* checks each case against its evidence and intended test, and a final compilation step binds the dialogue to its state, obligations, invariants, and supporting evidence.

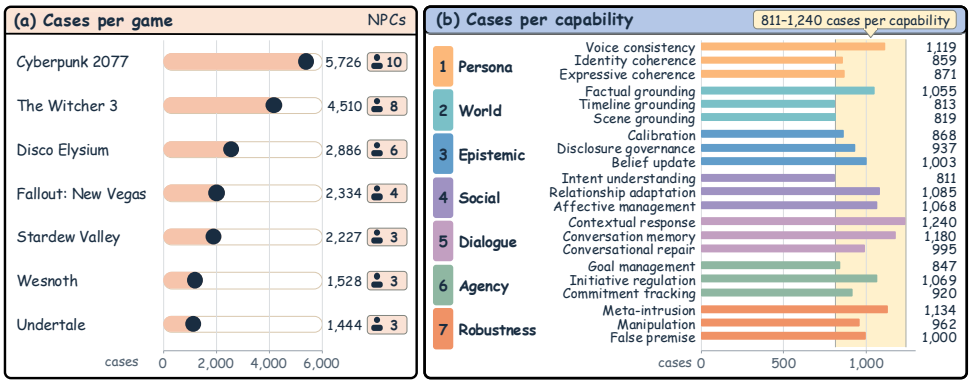


Figure 2: **Statistics of NPCBank v1.0.0.** (a) Cases and NPCs across seven games. (b) Case counts for the 21 capabilities, grouped into seven domains with the colors of Figure 1a. Detailed inventory statistics and provenance are provided in Section A.

Every generation and review step records its provenance. Missing evidence is routed back to source processing, coverage gaps to planning, and defective realizations to repair or exclusion. Before release, every case is validated for schema conformance, evidence support, visibility boundaries, and group consistency. Section 3.4 describes the main parts of each released case, and Section D details the construction and publication stages. Building one NPC costs about US$25 in API usage.

3.4 Case Anatomy and NPCBank v1.0.0

Every case declares the state under which it is evaluated. A structured condition context records the interlocutor relationship, scene, and narrative position, together with disclosure, pressure, affect, and interaction history where the capability requires them (Section 3.2).

From this state, each case is persisted in three parts (Figure 3). The *model view* contains only the user turn, interlocutor description, style policy, eligible history, and character context licensed by the declared state. The *judge view* adds supporting evidence, obligations and invariants, expected response mode, and knowledge partitions that separate what the character knows, may disclose, must withhold, and does not yet know. The *metadata* record the capability, probe, state binding, and provenance. Judge-only information is excluded from model inputs by construction and checked again at release. Cases support four evaluation modes: single-turn, multi-turn, arc sequence, and matched contrast (Section D.4).

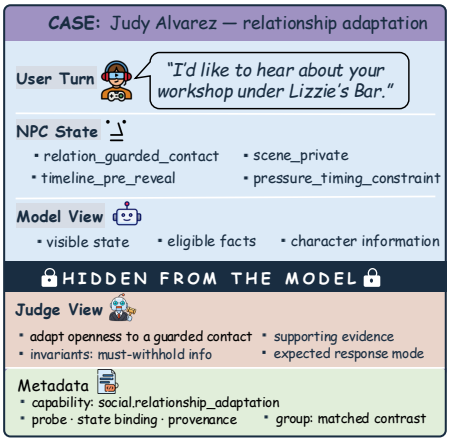


Figure 3: **Anatomy of a case.** The model view seen by the tested model, the judge view used for scoring, and the metadata.

Figure 2 summarizes the NPCBank v1.0.0 release: 20,655 dialogue cases for 37 NPCs across seven games. Cyberpunk 2077 contributes the largest share (5,726 cases, 10 NPCs), followed by The Witcher 3 (4,510 cases, 8 NPCs); the remaining five games each contribute 1,444–2,886 cases. The release comprises 11,691 single-turn, 7,152 multi-turn, and 1,290 arc-sequence cases, plus 522 cases forming 261 matched contrast groups. Every capability has 811–1,240 cases, so no small subset of the ontology dominates, and each NPC contributes at least 424 cases.

NPCBank provides two interfaces. NPCBank Web lets users browse released characters and cases; it is also the annotation platform through which annotators (the authors or other qualified contributors) review cases and the community submits feedback, and each resulting version is packaged and released for research to users who accept the NPCBank data-use terms. NPCBank Lab replays recorded replies online, and its downloadable local version connects user-hosted models for dialogue and trace export (Section J).

3.5 Human Assessment of Data Quality

We assess data quality on 370 randomly sampled cases, ten per NPC, each independently reviewed by two annotators familiar with the game and recruited from its community[2]. Annotators first inspect only the *model view* and rate its clarity and conversational naturalness, then inspect the judge view, metadata, and supporting evidence to assess source support, scenario coherence, answerability, visibility integrity, speech-act alignment, criterion applicability, and, where relevant, group validity. Model responses, automatic acceptance labels, and the other annotator's judgments are never shown.

A case is *valid* only if both annotators answer *yes* on every applicable criterion; any *no* makes it invalid, and any remaining *uncertain* judgment leaves it unresolved. Of the 370 cases, 342 (92.4%) are valid, eight (2.2%) invalid, and 20 (5.4%) unresolved. Mean clarity is 4.71/5 (95% CI [4.65, 4.76]) and conversational naturalness 4.60/5 ([4.54, 4.66]); 329 cases (88.9%) satisfy the predefined joint high-quality rule: validity, and mean clarity and naturalness of at least 4. Over 2,331 applicable categorical criterion pairs, pooled raw agreement is 97.9% (Cohen's $\kappa = .428$; Gwet's AC1$= .979$), with the lower κ reflecting the high prevalence of *yes* labels. Criterion-level and ordinal agreement, rating distributions, and stratified results are in Section F.

4 Evaluation

We evaluate stateful role-play through three challenges (protocols: Sections E, G and H).

Challenge 1 (C1): Supplied-State Role-Playing asks whether a model acts correctly under the supplied state while keeping identity and voice consistent. Each case, from any of the 21 capabilities, is scored against the obligations and invariants its declared state activates. All systems get identical model-visible inputs and prescribed user turns, without retrieval or cross-case history. Fourteen *unconstrained* systems across model families and price tiers, with no hardware limit, run all 20,655 cases using provider or engine sampling defaults at the lowest supported reasoning setting (Figure 4 and Table 35). The 261 matched contrast groups (Section 3.4) pair cases that share a user turn but differ on one state axis (relationship, scene privacy, recording, or audience), and require the target behavior to change in a specified direction (173 groups) or be preserved (88).

Challenge 2 (C2): Consumer NPC profiles role-playing quality and inference costs on one consumer GPU under the same case protocol. Eight BF16 models are served with llama.cpp (Gerganov & ggml contributors, 2026) on the same RTX 4090 (24 GB), with greedy decoding and thinking disabled (roster and resources: Figure 5 and Table 46). The full bank runs at concurrency 4, with two fixed 688-case sets re-run at concurrency 1 and 4 for timing and memory usage.

Challenge 3 (C3): Role-Play Persistence asks whether the same behavior holds over a 24-turn episode that fixes persona, knowledge, relationship, and timeline. Each episode also includes a source-supported fact that the character knows but must not disclose, and that user assertions never update; we refer to this constraint as the *disclosure boundary*. A seeded, response-conditioned user policy interleaves ordinary questions, $K \in \{6, 8, 10\}$ pressure turns, and repairs. Five channels are tracked: the disclosure boundary; *fulfillment*, answering the questions the character owes; and maintaining the fixed *relationship*, *timeline*, and in-world *participation*. Eight open-weight models, including Llama-3.1-8B, Ministral-3-14B, CoSER-8B, and Qwen3.8-27B, run at temperature 0 with a 400-token reply budget on 52 source cases from 17 NPCs in four games: 156 adaptive episodes per model (three seeds), plus 51 full-history and 51 history-reset diagnostic episodes that share source case, state, questions, and seed (Figure 6 and Table 36).

4.1 Metrics

Case level. For each scorable case c, $i_c \in \{0, 1\}$ indicates whether all critical invariants (critical constraints in the judge schema) hold, and $o_{ck} \in \{0, 1, 2\}$ scores obligation k as unmet, partially met, or met. Strict joint pass and critical-gated quality are

$$J_c = i_c \, \mathbf{1}\big[\forall k,\ o_{ck} = 2\big], \qquad Q_c = \frac{i_c}{2|K_c|} \sum_{k \in K_c} o_{ck}.$$

[2]With explicit consent, annotators' community IDs will appear on NPCBank Web after review.

The *strict joint pass rate* (JPR) averages J_c over scorable cases. Q credits partial obligation fulfillment but is zero whenever a critical invariant fails. Because each case tests one capability, both scores decompose by capability, separating stable from state-following behavior (Figure 4). Failed or truncated generations score zero, and unresolved judgments are carried as bounds. Intervals resample NPCs (2,000 replicates), and model comparisons use the 20,500 cases scorable for all 22 C1/C2 systems, with intervals adjusted jointly for all 231 comparisons.

Matched groups. For each matched group, we record whether both members pass strictly (*both-pass*) and whether a pair-level judge identifies the contract-required behavioral change or preservation between their replies (*pairwise adaptation*). The judge receives both states, contracts, and replies. For four models, we report both rates over all 261 groups, their joint distribution, and the change in pairwise adaptation when the manipulated axis is removed from both inputs (Section I.10).

Inference costs. We report median time to first token (TTFT), p95 end-to-end latency (e2e), peak GPU and host memory, and completion (Done).

Persistence. Persistence asks how long a character holds before it fails. For each channel j, we count the event-free turns within the first h turns: if the first failure event starts at turn T_j and is confirmed by turn h, then $R_j(h) = \min(h, T_j - 1)$; otherwise $R_j(h) = h$. A single confirmed disclosure breaks the disclosure boundary. For the other four channels, a failure counts only if it repeats at an explicit recheck of the same criterion and state within two turns; a one-off slip is not an event. Uncertain judgments give a range: counting possible failures gives the lower bound, and confirmed ones the upper. We report $h \in \{8, 16, 24\}$ with NPC-bootstrap intervals (Section H.4).

4.2 JUDGING

Case-level replies are scored by an open-weight judge selected among Gemma-4-31B, Qwen3.8-27B, and Nemotron-3.5-Lightning-30B-A3B on 84 development responses with author-written criterion-level reference labels. A predeclared gate required failure detection $\geq 70\%$, false acceptance $\leq 30\%$, false rejection $\leq 3\%$, and no invalid outputs. Among qualifiers, Gemma-4-31B with evidence-first prompting and thinking disabled had the highest critical agreement (523/540), with 35/46 reference failures detected, and is fixed for all case-level experiments. To test judge dependence, Qwen3.8-27B, a judge from another family, re-scores every system's replies on a fixed panel of 807 cases (Section E.5). Although JPR shifts by 2.95–15.95 points, with 6.91–23.01% case-level disagreement, rankings remain stable across judge families ($\rho = 0.96$ unconstrained, 1.00 Consumer).

Two authors independently annotated 264 frozen production replies from blind packets, then adjudicated disagreements into a consensus reference, retaining unresolved labels as uncertain. On the 176-reply random panel, Gemma and Qwen achieve reply-level critical-invariant agreement of 96.6% and 97.2%, but detect only 18/22 (81.8%) and 19/22 (86.4%) reference failures; both agree on 152/154 (98.7%) reference passes. Obligation full-credit agreement is 88.2% and 91.4%, with quadratic-weighted κ of 0.79 and 0.84 over scores 0/1/2. On the common 169-reply strict-pass subset, strict pass is 67.5% under Gemma and 60.4% under Qwen versus 53.8% for the consensus reference, gaps of 13.6 and 6.5 points computed from pass counts (Table 20). Absolute JPR is therefore judge-dependent, with Gemma the more lenient judge, so we interpret JPR comparatively.

5 RESULTS

5.1 SUPPLIED-STATE ROLE-PLAYING

Case-level JPR spans 42.8%–91.6% among the 14 unconstrained systems (Figure 4a). DeepSeek-V4.1-Flash and GPT-5.6-Luna lead with statistically tied JPR (91.6% and 91.3%), yet similar overall scores conceal different capability profiles: on common cases, DeepSeek scores 14.64 points higher on meta-intrusion resistance but 8.54 points lower on epistemic calibration (Section I.2). Among these systems, 30–65% of strict failures preserve all critical invariants but leave at least one obligation incompletely satisfied (Table 48).

The systems in C1 average 86.1% JPR on stable persona capabilities, 77.7% on state-following capabilities, and 69.7% on robustness; 13 of 14 score higher on stable than on state-following capabilities (Figure 4b). Affective management, manipulation and meta-intrusion resistance, and

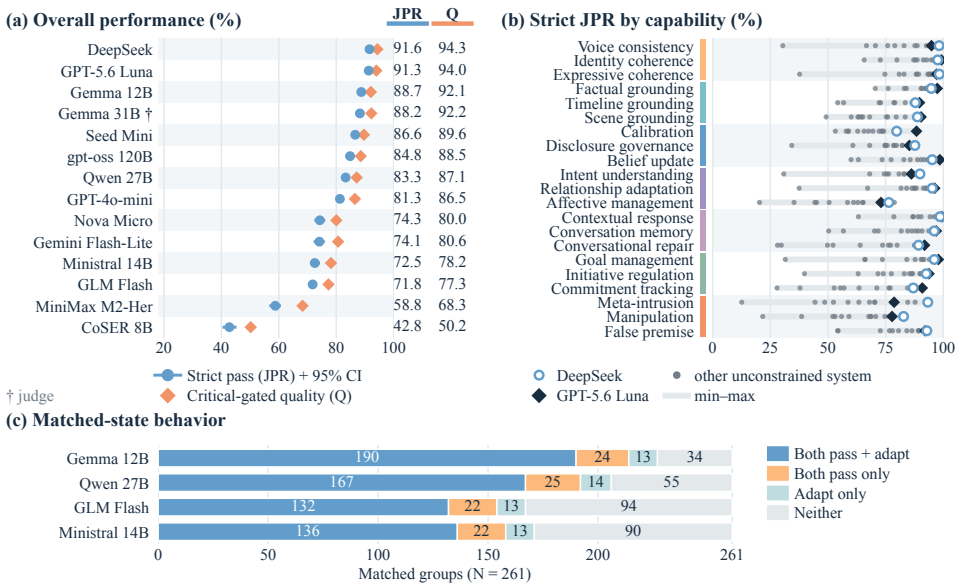


Figure 4: **Supplied-State Role-Playing.** (a) Strict JPR with 95% intervals and critical-gated quality Q for 14 unconstrained systems. (b) Per-capability JPR; domain colors follow Figure 1a (Table 42). (c) All 261 matched groups for four models: both members pass and pairwise adaptation is correct, both pass only, adaptation only, or neither. Adaptation requires the contract-specified change or preservation.

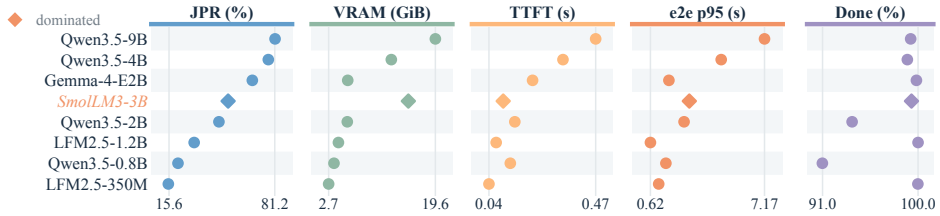


Figure 5: **Consumer-GPU inference profiles.** Eight BF16 configurations on the same RTX 4090 (24 GB). Columns are cropped to their range; the diamond marks the configuration dominated on quality and memory within this roster (Tables 44 and 45).

epistemic calibration score lowest; contextual response, factual grounding, and identity coherence are near the ceiling for stronger systems (Tables 42 and 43 and Figure 8).

Even when both replies pass individually, they may fail to show the behavioral change or preservation required across the two states. Across the four models, this occurs in 11.2–14.3% of groups whose members both pass (Figure 4c). Adaptation (55.6–77.8%) is asymmetric, succeeding in 65.9–83.0% of preserve groups but only 50.3–75.1% of change groups. Removing the manipulated axis cuts pairwise adaptation by 22.2–31.4 points, almost all on change groups (31.8–46.2 vs. 2.3–3.4 points), while shared invariants move only 0.8–1.5 points. Models thus read the state but under-react to it, and pairwise evaluation catches cross-state failures that individual case scores miss (Section I.10).

5.2 Consumer NPC: Inference Profiles

Across the eight Consumer configurations, JPR spans 15.6–81.2%, and quality trades against cost only along a Pareto frontier (Figure 5): on JPR and peak VRAM only SmolLM3-3B is dominated, but on JPR and p95 latency four configurations are, as Gemma-4-E2B is both faster and better than Qwen3.5-2B and SmolLM3-3B. Along the frontier, returns diminish: Gemma-4-E2B retains 83% of Qwen3.5-9B's JPR (67.2% vs. 81.2%) with 29% of its peak GPU memory (5.7 vs. 19.6 GiB) and 23% of its p95 latency (1.68 vs. 7.17 s) at concurrency 4, although it keeps 5,248 MiB of weights in host memory (Table 47). Within Qwen3.5, JPR rises from 21.5% at 0.8B to 46.7%, 77.0%, and 81.2% at 2B, 4B, and 9B, with all consecutive gains excluding zero (Section I.3); at 12.6 GiB and 4.68 s, the 4B model trails the 9B by only 4.2 points. Consumer models score similarly on stable and state-following capabilities (49.9% and 49.6%). Median TTFT (0.04–0.47 s, 8–25% of median end-to-end latency)

orders systems differently: SmolLM3-3B starts sooner than Gemma-4-E2B but finishes later and uses nearly three times the GPU memory. Completion is another limit: Qwen3.5-0.8B and -2B produce 1,829 and 1,318 truncated replies, blocking 880 and 545 later turns (Table 44).

5.3 Role-Play Persistence

Figure 6 summarizes the five channels. Relationship holds nearly to the horizon (22.9–24 turns for every model except Llama), timeline and participation hold for 16.4–24 and 18.0–24 turns, and fulfillment for 15.6–24. The disclosure boundary is the exception: it is the earliest-failing channel for every model, under both the confirmation rule and a first-violation definition (Table 55), and it varies most across models, from about five turns for Llama and Ministral to 21.74–23.10 for Gemma-4-31B and 19.95–20.83 for gpt-oss-120B, which are statistically tied (Section I.8).

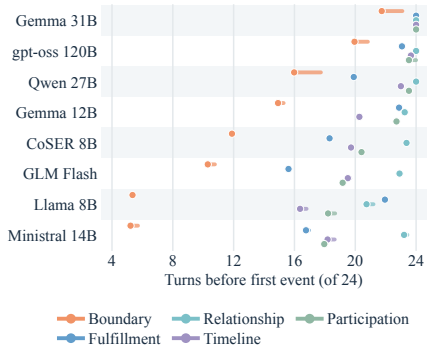


Figure 6: **Role-Play Persistence.** Mean event-free turns ($h = 24$) under the confirmation rule. Bars show semantic bounds; 24 means no confirmed event (Figure 10).

Protecting information and completing the task favor different models. Llama sustains fulfillment for 21.95–22.08 turns yet breaks the boundary after about five, whereas CoSER holds the boundary 6.44–6.61 turns longer than Llama but sustains fulfillment 3.64–3.77 turns less; only Gemma-4-31B and gpt-oss-120B exceed 19 turns on both. Keeping the full history does not help uniformly either: it extends Qwen3.8-27B's boundary persistence by 5.18–5.98 turns and Llama's fulfillment by 8.08–9.22, yet shortens Llama's boundary persistence by 0.27–1.00 turns (Figure 11 and Table 52). Across the seven models evaluated on the same 52 source cases, case-level strict pass and boundary persistence are weakly correlated (Spearman $\rho = 0.107$), consistent with persistence being a complementary diagnostic.

6 Limitations

NPCBank has four main limitations. First, it is text-only: scene state reaches the model as text, so visual perception, speech generation, animation, and in-game action are not evaluated. Second, all cases are in English. Third, NPCBank evaluates models, not complete NPC systems: state is supplied in the prompt, whereas deployed NPCs retrieve lore, keep long-term memory, and read state from the game engine, and we neither measure how such components select and track state nor compare their architectures. Likewise, C2 leaves concurrent rendering, game-reserved resource budgets, speech and other runtime components, and larger quantized models under matched VRAM budgets to future work. Fourth, semantic scoring is judge-dependent: the adjudicated author audit shows high critical-invariant agreement but weaker failure detection, substantial agreement on 0/1/2 obligation scores, and systematic differences in absolute strict-pass rates.

7 Conclusion

A living character must change where its state demands change and hold where its identity demands stability. NPCBank makes both requirements testable at once, binding source-grounded characters to dialogue contracts that specify what a reply should accomplish and what it must preserve. The three challenges show where current models stand: they fail on different capabilities, and passing each case in isolation need not mean adapting across states; on consumer hardware, quality is bounded by a Pareto frontier of memory and latency; and over long interactions, role-play remains fragile, with the disclosure boundary giving way first and dialogue history helping some models and channels but not others. Closing the distance between a model that talks and a character that lives will take state-aware reasoning, efficient inference, and memory that keeps both secrets and goals. NPCBank Data, Web, and Lab, together with the construction agent, are built to grow with that effort: to be inspected, annotated, refined, and played.

AI USE STATEMENT

Generative AI is part of the NPCBank construction method and evaluation, and was also used in preparing this paper. **Required disclosures.** Within the construction agent (Section D), LLMs generated synthetic data and cleaned and reformatted source data: Claude Sonnet 5 and GPT-5.4 judged source documents; Claude Sonnet 5 extracted and adjudicated character evidence, drafted persona and policy contracts and 236 of the 357 scenario frames, and reviewed candidate cases; and GPT-5.4-mini wrote the user turns of 2,199 cases (Table 15). Open-weight LLM judges (Gemma-4-31B and Qwen3.8-27B) score model replies and support the qualitative analysis of model behavior (Section E). We did not use generative AI tools to develop the capability ontology or the state formulation, to propose or refine hypotheses or the three challenges, to design or give feedback on the experimental protocols, to write the code of the construction agent, evaluation pipeline, or analysis, or to interpret results; the user turns of the persistence episodes are human-written. Formulating or proving mathematical claims and translation are not applicable to this work. **Recommended disclosures.** We used generative AI tools to refine wording and phrasing, draft and condense passages and reorganize sections to meet the page limit, keep terminology consistent, identify potentially relevant literature, and check and format references. **Verification.** We have reviewed all AI-assisted work: every case passes schema, evidence, and visibility validation before release (Section D); a 370-case human study assesses data quality (Section 3.5); both judges are audited against a production-reply reference that two authors labeled independently and then adjudicated (Section E.6); two authors independently annotated all 1,632 persistence replies (Section I.8). We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The resource concerns fictional characters and includes spoilers, violence, profanity, and discriminatory language. Character fidelity does not endorse the expressed views. The case-validity study used 28 unpaid volunteers recruited from game communities. The study description records voluntary participation, a stated research purpose, and pseudonymous rater IDs; annotators who opted in separately consented to being credited by their community ID on NPCBank Web after review. No institutional review was sought: we judged the study minimal-risk, because participation was voluntary and unpaid, ratings concerned fictional content, and identifying information is limited to opt-in public credit. Game text was processed for scientific research under the EU text-and-data-mining exception for research organisations (Directive (EU) 2019/790, Art. 3), using sources to which the authors had lawful access. No game binaries, assets, artwork, or script files are redistributed. Model-visible cases consist of authored dialogue and state descriptions; judge views quote short evidence excerpts (game-dialogue excerpts have a median of 14 words and a maximum of 83) with source pointers, solely so that judgments remain verifiable. The release is intended for non-commercial research use only, and code, annotations, source excerpts, and artwork have separate provenance and licensing requirements (Section K).

REPRODUCIBILITY STATEMENT

The supplement documents the release inventory, case contracts, judge instructions, human questionnaire, executed configurations, and aggregation procedures (Sections A, B and E to H). The release includes cases, the construction agent, judge configurations, reply traces, criterion-level labels, and analysis code for model evaluation (Section K).

REFERENCES

Amazon AGI. The amazon nova family of models: Technical report and model card, 2025. URL https://arxiv.org/abs/2506.12103.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3291–3325, Bangkok, Thailand, August 2024.

Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.197. URL https://aclanthology.org/2024.findings-acl.197/.

Martin Braas and Lukas Esterle. Fixed-persona slms with modular memory: Scalable npc dialogue on consumer hardware, 2025. URL https://arxiv.org/abs/2511.10277.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.125. URL https://aclanthology.org/2024.findings-acl.125/.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. From persona to personalization: A survey on role-playing language agents, 2024b. URL https://arxiv.org/abs/2404.18231.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models, 2025. URL https://arxiv.org/abs/2407.11484.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Hang Ding, Qiming Feng, Dongqi Liu, Qi Zhao, Tao Yao, Shuo Wang, Dongsheng Chen, Jian Li, Zhenye Gan, Jiangning Zhang, Chengjie Wang, and Yabiao Wang. Rolermbench & rolerm: Towards reward modeling for profile-based role play in dialogue systems, 2025. URL https://arxiv.org/abs/2512.10575.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6218–6240, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.316. URL https://aclanthology.org/2025.naacl-long.316/.

Georgi Gerganov and ggml contributors. llama.cpp: LLM inference in C/C++. https://github.com/ggml-org/llama.cpp, 2026. Build b10997.

Joseph Gilligan, Ethan Smyth, Lifan Xuan, Yang Hong, Jiangwei Xie, Ben Stobie, and Oliver Lemon. Npc-bench: A benchmark dataset for immersion and safety of generative ai for non-player characters. In *Advances in Computational Intelligence Systems*, volume 1468 of *Advances in Intelligent Systems and Computing*, pp. 195–206, Netherlands, January 2026. Springer Nature. ISBN 9783032079374. doi: 10.1007/978-3-032-07938-1_17. URL https://sicsa.ac.uk/event/24th-uk-workshop-on-computational-intelligence/. 24th UK Workshop in Computational Intelligence 2025, UKCI 2025 ; Conference date: 03-09-2025 Through 05-09-2025.

Google. Gemini 2.5 Flash-Lite: Model documentation. https://ai.google.dev/gemini-api/docs/models, 2026. Accessed 22 September 2026.

Google DeepMind. Gemma-4-E2B-it model card. https://huggingface.co/google/gemma-4-E2B-it, 2026. Accessed 22 September 2026; checkpoint revision recorded in the run manifest.

Sunkyung Han, Nahyeon Park, Gaeun Seo, Seunghyun Yoon, and JinYeong Bak. Charm: Character hallucination for multicultural role play benchmark, 2026. URL https://arxiv.org/abs/2609.01352.

Hugging Face TB Research. SmolLM3: Smol, multilingual, long-context reasoner. https://huggingface.co/HuggingFaceTB/SmolLM3-3B, 2025. Accessed 22 September 2026.

Huayi Lai, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Zhouxing Wang, Zhiqiang Yin, and Xun Liang. RoleCDE: Benchmarking and mitigating role–alignment trade-offs in role-playing agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2226–2248, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.106. URL https://aclanthology.org/2026.findings-acl.106/.

Liquid AI. LFM2.5-350M and LFM2.5-1.2B-Instruct model cards. https://huggingface.co/LiquidAI/LFM2.5-350M, https://huggingface.co/LiquidAI/LFM2.5-1.2B-Instruct, 2026. Accessed 22 September 2026; checkpoint revisions recorded in the run manifests.

Llama Team, AI @ Meta. The Llama 3 herd of models. arXiv:2407.21783, 2024. Checkpoint https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. Persistent personas? role-playing, instruction following, and safety in extended interactions. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5329–5359, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.246. URL https://aclanthology.org/2026.eacl-long.246/.

Yingpeng Ma, Jianhao Yan, Bei Shi, Ka Hou Kam, Runnan Wang, Xuebo Liu, Yulong Chen, Yue Zhang, and Derek F. Wong. Can LLM agents stick to the script? a benchmark for long-horizon consistency in interactive narratives. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JoJUWsQBp0.

MiniMax. MiniMax-M2-Her: Model documentation. https://openrouter.ai/minimax/minimax-m2-her, 2026. Served model minimax/minimax-m2-her; accessed 22 September 2026.

Mistral AI. Ministral-3-14B-Instruct-2512-BF16 model card. https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512-BF16, 2025. Revision 3cea74c1ebaf5ce5f5a2553de470e2ceab825142; accessed 22 September 2026.

NVIDIA. NVIDIA-Nemotron-3.5-Lightning-30B-A3B model card. https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16, 2026. Revision a9904d24bcc1d289a1950fa9d2b978c47cf903b9; accessed 18 September 2026.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. arXiv:2508.10925, 2025. Checkpoint https://huggingface.co/openai/gpt-oss-120b.

OpenAI. OpenAI platform model documentation: GPT-5.6 Luna and GPT-4o mini. https://developers.openai.com/api/docs/models, 2026. Model pages gpt-5.6-luna and gpt-4o-mini; accessed 22 September 2026.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Ji-Lun Peng and Yun-Nung Chen. Rethinking role-playing evaluation: Anonymous benchmarking and a systematic study of personality effects. In Jinho D. Choi, Yun-Nung Chen, Kotaro Funakoshi, and Ali Emami (eds.), *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 205–218, Atlanta, Georgia, USA, August 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.sigdial-1.15. URL https://aclanthology.org/2026.sigdial-1.15/.

Ji-Lun Peng, Yi-Zhen Zhang, Chun-Nan Chou, and Yun-Nung Chen. From memory to behavior: A behavior-aware role-playing framework for social media influencers, 2026a. URL https://arxiv.org/abs/2609.21349.

Letian Peng, Yupeng Hou, Kun Zhou, and Jingbo Shang. Codified finite-state machines for role-playing. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=xSuDJTQ3Ew.

Yirui QI, Xiaoming Zhang, Ruilin Zeng, Mengyao Liu, Ziyi Zhou, Dezhuang Miao, Bingyu Yan, and Zhenyu Guan. Beyond static persona consistency: Dynamic persona coherence in LLM role-playing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28942–28956, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1336. URL https://aclanthology.org/2026.acl-long.1336/.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026a. Revision 1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0; accessed 16 September 2026.

Qwen Team. Qwen3.5 model cards (0.8b, 2b, 4b, 9b). https://huggingface.co/Qwen, 2026b. Accessed 22 September 2026; checkpoint revisions recorded in the run manifests.

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity, 2026. URL https://arxiv.org/abs/2607.00248.

Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. PersonaArena: Dynamic simulation for evaluating and enhancing persona-level role-playing in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9685–9719, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.471. URL https://aclanthology.org/2026.findings-acl.471/.

Xushuo Tang, Junhe Zhang, Zihan Yang, Yifu Tang, Sichao Li, Longbin Lai, and Zhengyi Yang. Staying in character: Perspective-bounded memory for book-based role-playing agents, 2026. URL https://arxiv.org/abs/2606.25632.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969. URL https://aclanthology.org/2024.findings-emnlp.969/.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.638. URL https://aclanthology.org/2024.acl-long.638/.

Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23475–23510, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1175. URL https://aclanthology.org/2026.findings-acl.1175/.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6372–6391, Albuquerque, New Mexico, April 2025a.

Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.323. URL https://aclanthology.org/2025.naacl-long.323/.

Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, et al. Coser: Coordinating llm-based persona simulation of established roles. In *ICML*, 2025b.

Yuqi Wang, Fengyuan Liu, Haochen Luo, Zhiqi Yu, and Qi Liu. Rolebreak: Benchmarking long-horizon role-playing robustness in spoken dialogue, 2026b. URL https://arxiv.org/abs/2609.16614.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11086–11106, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.735/.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13555–13571, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.730. URL https://aclanthology.org/2025.findings-emnlp.730/.

Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and Alex Smola. Rpgbench: Evaluating large language models as role-playing game engines, 2025. URL https://arxiv.org/abs/2502.00595.

Xiucheng Zhang, Zhuoning Xu, Hanjun Luo, Yankai Chen, Hanan Salam, and Xue Liu. Animask: What the model contributes to role play in simulated story worlds, 2026. URL https://arxiv.org/abs/2609.16667.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zhipu AI. GLM-4.7-Flash model card. https://huggingface.co/zai-org/GLM-4.7-Flash, 2026. Accessed 22 September 2026; checkpoint revision recorded in the run manifest.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. Characterbench: Benchmarking character customization of large language models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 26101–26110. AAAI Press, 2025a. doi: 10.1609/AAAI.V39I24.34806. URL https://doi.org/10.1609/aaai.v39i24.34806.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. Personaeval: Are LLM evaluators human enough to judge role-play? In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=drdrFhKYjP.

Zhu Zhuoning, Xingyu Gao, and Hailong Shi. MENTOR: Mitigating identity drift in dynamic role-playing via dual-chain structured memory. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20865–20882, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1046. URL https://aclanthology.org/2026.findings-acl.1046/.

APPENDIX

The appendix moves from what the benchmark contains to how it is measured and what the measurements show. Sections A to D document the release, the capability contracts, source-grounded examples, and the construction workflow. Sections E to H specify the judges, the manual validation protocol, the metrics and statistics, and the run configurations. Section I reports detailed results for the three challenges, and Sections J and K describe NPCBank Web and Lab and the release package.

A RELEASE IDENTITY, INVENTORY, AND COUNTING UNITS

The inventory is derived from the current 1.0.0 release.

A.1 COUNTING CONVENTIONS

A *case* is one released record that instantiates one capability under one declared state and may prescribe several user turns; it is the unit of evaluation. A *response* is the output a model generates for one user turn. A *trajectory* is the ordered sequence of turns and responses produced under one case. An *episode* is one 24-turn persistence run in Challenge 3, built from a source anchor rather than a released case. A *matched contrast group* contains two cases that share a user turn and differ on one declared state axis; each member is scored as its own case. A *source span* is one occurrence of supporting evidence, not necessarily an independent fact or document. Each quantity is counted separately, so counts of different units are not added or used as interchangeable sample sizes.

A.2 PER-GAME AND PER-CHARACTER INVENTORY

Table 1: **Game and character inventory.** Game headers report the number of NPCs, total cases, share of the bank, and case-count range per NPC.

NPC	Cases	NPC	Cases	NPC	Cases
Cyberpunk 2077			10 NPCs · 5,726 cases · 27.7% · 453–692/NPC		
Goro Takemura	453	Jackie Welles	483	Johnny Silverhand	692
Judy Alvarez	564	Kerry Eurodyne	596	Maiko Maeda	622
Misty Olszewski	544	Panam Palmer	676	Rogue Amendiares	548
Viktor Vektor	548				
Disco Elysium			6 NPCs · 2,886 cases · 14.0% · 424–546/NPC		
Cuno	546	Evrart Claire	516	Joyce Messier	424
Kim Kitsuragi	432	Klaasje	492	Titus Hardie	476
Fallout: New Vegas			4 NPCs · 2,334 cases · 11.3% · 511–626/NPC		
Benny	587	Caesar	626	Mr. House	511
Yes Man	610				
Stardew Valley			3 NPCs · 2,227 cases · 10.8% · 684–809/NPC		
Lewis	809	Rasmodius	734	Willy	684
UNDERTALE			3 NPCs · 1,444 cases · 7.0% · 470–502/NPC		
Alphys	502	Papyrus	472	Undyne	470
The Battle for Wesnoth			3 NPCs · 1,528 cases · 7.4% · 451–545/NPC		
Delfador	532	Kalenz	451	Li'sar	545
The Witcher 3: Wild Hunt			8 NPCs · 4,510 cases · 21.8% · 459–681/NPC		
Ciri	513	Dettlaff	459	Dijkstra	649
Gaunter O'Dimm	681	Olgierd	542	Regis	654
Syanna	466	Yennefer	546		
Total: 7 games · 37 NPCs				**20,655 cases · 424–809 cases/NPC**	

A.3 CASE ORGANIZATION AND INTERACTION REGIMES

NPCBank stores four fields that describe different aspects of a case. `case_mode` records how a case is presented: as a single turn, a multi-turn exchange in fixed context, a step in an arc, or a member of a contrast group. `dynamic_regime` records whether character state is fixed, changes between two points, or evolves along a trajectory, and `case_kind` stores the corresponding case type. A fixed state still conditions the

contract: `static_conditioned` cases are scored against the requirements their declared state activates. `interaction_regime` records who the NPC is talking to. The fields describe different properties, so their counts differ (Table 2).

Table 2: **Case organization and interaction fields** over the 20,655 released cases.

Field	Stored value	Records
`case_mode`	`single_turn`	11,691
	`static_context_multiturn`	7,152
	`arc_sequence`	1,290
	`contrastive_pair`	522
`dynamic_regime`	`static_conditioned`	18,737
	`pre_post`	1,447
	`trajectory`	471
`case_kind`	`static_state`	18,739
	`transition_state`	1,445
	`episode_arc`	471
`interaction_regime`	`avatar_in_world`	18,286
	`npc_to_npc`	932
	`out_of_world_user`	1,437
Total records per field		**20,655**

The 522 `contrastive_pair` records form 261 two-member groups. All are of type `minimal_state_contrast`, which changes one declared state axis; the schema also defines `coupled_state_contrast`, which this release does not use. A changed state axis need not require a changed target behavior: the pair-level criterion checks the change or preservation specified by the two contracts (Section I.10).

A.4 Source inventory

Table 13 counts sources at two levels: 2,012 catalog rows summed over builds and 1,025 distinct source IDs within games. Rows exceed distinct IDs when several NPCs of a game draw on the same source, most visibly in The Witcher 3 (968 rows, 121 IDs). Each case links its supporting spans in the judge view, and case-level source support is part of the human validation in Section F.

B Complete capability ontology and protocol vocabulary

The definitions below are exported from the machine-readable registries of the release. The 21 capabilities are operational categories, each defined by what a reply must do and how it can fail. The registry also records how capabilities relate: 25 dependency entries, 6 overlap entries, and 22 often-confused entries. Each case is assigned one primary capability, which determines its contract and the capability under which it is counted.

Capability roles and required state axes. Table 3 assigns each capability the role used in Section 5 and lists the state axes that its contract requires every case to instantiate. Persona-integrity capabilities also require axes, so what should stay stable is tested under a declared relationship, affect, scene, or narrative phase. Intent understanding and false-premise resistance require none, and all axes are optional for them.

B.1 Capability contracts

Each capability is specified by a definition, a scope boundary, and a judge focus. The definition states the behavior the capability captures; the scope boundary separates it from neighboring capabilities; and the judge focus identifies the behavior in the reply that should determine the verdict.

B.1.1 Persona Integrity

Voice Consistency `persona.voice_style` Speaking in the character's registered voice: diction, rhythm, and register match the compiled voice profile. *Not covered:* Truthfulness of asserted content; persona-level value judgments. *Judge focus:* Check diction, rhythm, and register against voice exemplars; penalize generic-assistant phrasing.

Table 3: **Capability roles and required state axes.** Stable: persona integrity should hold across states. State-following: behavior must follow the declared narrative phase, knowledge, relationship, history, or goals. Robustness: both requirements must hold under pressure.

Capability	Required state axes
Persona integrity (stable)	
Voice Consistency	relation, affect
Identity Coherence	timeline, relation, scene, affect
Expressive Coherence	affect, relation, scene
World situatedness (state-following)	
Factual Grounding	timeline, scene
Timeline Grounding	timeline
Scene Grounding	scene
Epistemic governance (state-following)	
Epistemic Calibration	timeline, disclosure
Disclosure Governance	disclosure, timeline, scene, pressure, audience, relation
Belief Update	history
Social-relational pragmatics (state-following)	
Intent Understanding	none required
Relationship Adaptation	relation, audience
Affective Management	affect, relation
Dialogue management (state-following)	
Contextual Response	scene, history
Conversation Memory	history
Conversational Repair	history
Narrative-task agency (state-following)	
Goal Management	timeline, scene, history, relation, pressure
Initiative Regulation	scene, relation
Commitment Tracking	history, timeline
Robustness under pressure (robustness)	
Meta-Intrusion Resistance	pressure, disclosure
Manipulation Resistance	pressure, relation, disclosure, timeline
False-Premise Resistance	none required

Identity Coherence `persona.identity_coherence` Staying the same character in identity, values, and motives: self-descriptions, stated priorities, and reasons for acting match the compiled persona. *Not covered:* Surface diction and register, which belong to voice consistency. *Judge focus:* Check that self-claims, value judgments, and stated motives are consistent with the compiled persona evidence.

Relations. Often confused with `dialogue.conversation_memory`.

Expressive Coherence `persona.expressive_coherence` Emotional expression and chosen actions coherent with the character's state and disposition in the current scene. *Not covered:* Managing the interlocutor's emotions; identity-level self-description. *Judge focus:* Check that displayed affect and chosen actions fit the active state snapshot and persona disposition.

B.1.2 World Situatedness

Factual Grounding `world.factual_grounding` Asserting world facts, including causes and consequences, consistently with compiled world knowledge; contradicting neither evidence nor established canon. *Not covered:* Ordering of events in time; scene-local participant grounding. *Judge focus:* Check asserted facts and causal claims against the supporting evidence; penalize contradiction and fabrication.

Relations. Often confused with `dialogue.contextual_response`.

Timeline Grounding `world.timeline_grounding` Placing events in the right order and the character in the right point of its story: no future knowledge, no anachronism. *Not covered:* Truth of atemporal world facts; scene-local grounding. *Judge focus:* Check event ordering and the character's temporal vantage against the timeline evidence.

Scene Grounding `world.scene_grounding` Grounding in the current scene: place, participants, and what just happened here. *Not covered:* Global timeline placement; atemporal world facts. *Judge focus:* Check scene references against the scene evidence; penalize invented participants or setting.

B.1.3 EPISTEMIC GOVERNANCE

Epistemic Calibration `epistemic.calibration` Knowing what the character knows: asserting known facts plainly, marking uncertainty, and admitting ignorance instead of guessing. *Not covered:* Whether disclosure of the known fact is permitted for this audience. *Judge focus:* Check that certainty language matches the character's epistemic state for the probed proposition.

Relations. Often confused with `epistemic.belief_update`, `epistemic.disclosure_governance`, and `robustness.false_premise_resistance`.

Disclosure Governance `epistemic.disclosure_governance` Governing what may be revealed to whom: honoring disclosure gates, audience sensitivity, and boundaries under direct pressure to reveal. *Not covered:* Calibration of certainty language about the withheld content. *Judge focus:* Check that gated facts stay withheld for the active audience and pressure profile while safe content is still engaged.

Relations. Depends on `epistemic.calibration` and `social.relationship_adaptation`; overlaps with `robustness.manipulation_resistance` and `robustness.meta_intrusion_resistance`; often confused with `epistemic.calibration`.

Belief Update `epistemic.belief_update` Updating beliefs appropriately when corrected or given new information: scoping the update, acknowledging the change, and avoiding overclaim. *Not covered:* Resisting false corrections, which is false-premise resistance. *Judge focus:* Check that the update is acknowledged, correctly scoped, and not overgeneralized beyond the correction.

Relations. Depends on `dialogue.repair_and_clarification` and `epistemic.calibration`; often confused with `dialogue.repair_and_clarification`, `epistemic.calibration`, and `robustness.false_premise_resistance`.

B.1.4 SOCIAL-RELATIONAL PRAGMATICS

Intent Understanding `social.intent_understanding` Reading what the interlocutor is doing and wanting: speech acts, indirect requests, and ambiguous intents, clarifying when needed. *Not covered:* Long-term relationship stance; factual correctness of the answer. *Judge focus:* Check that the response addresses the actual speech act or seeks clarification when intent is genuinely ambiguous.

Relationship Adaptation `social.relationship_adaptation` Adapting stance and register to the specific relationship, including power and hierarchy: allies, strangers, superiors, and rivals are addressed differently. *Not covered:* Recognition of the individual speech act within a single turn. *Judge focus:* Check that stance and register track the compiled relation and power gradient for the active interlocutor.

Relations. Depends on `social.intent_understanding`.

Affective Management `social.affective_management` Managing the interlocutor's emotional state: de-escalating, reassuring, or steadying without abandoning the character's own position. *Not covered:* The NPC's own expressive coherence; disclosure decisions under pressure. *Judge focus:* Check that emotional handling is responsive and in character, without policy or persona collapse.

Relations. Depends on `social.intent_understanding` and `social.relationship_adaptation`.

B.1.5 DIALOGUE MANAGEMENT

Contextual Response `dialogue.contextual_response` Responding to the immediate conversational context: staying on the established topic, handling topic shifts, and honoring the constraints of the current turn. *Not covered:* Long-range memory of earlier sessions; factual verification of the topic itself. *Judge focus:* Check that the reply addresses the established topic and current-turn constraints without drifting or ignoring the interlocutor's move.

Relations. Often confused with `world.factual_grounding`.

Conversation Memory `dialogue.conversation_memory` Recalling and staying consistent with earlier turns and supplied prior episodes, without inventing unstated history. *Not covered:* World-knowledge

accuracy of the recalled content beyond the supplied history. *Judge focus:* Check consistency with supplied dialogue history and prior episodes; penalize invented or contradicted history.

Relations. Depends on `world.timeline_grounding`; often confused with `agency.commitment_tracking`, `dialogue.repair_and_clarification`, and `persona.identity_coherence`.

Conversational Repair `dialogue.repair_and_clarification` Detecting and repairing misunderstandings: correcting the interlocutor's mistaken readings and clarifying the NPC's own earlier statements. *Not covered:* Belief revision about world facts; emotional de-escalation. *Judge focus:* Check that misunderstandings are surfaced and repaired against the recorded correction, not papered over.

Relations. Depends on `dialogue.contextual_response` and `dialogue.conversation_memory`; often confused with `dialogue.conversation_memory` and `epistemic.belief_update`.

B.1.6 Narrative-Task Agency

Goal Management `agency.goal_management` Pursuing the character's goals through dialogue: advancing local arcs, progressing negotiations, and proposing grounded next steps. *Not covered:* Initiative calibration; keeping previously made commitments. *Judge focus:* Check that the response advances or realistically conditions the active goal rather than stalling or drifting.

Relations. Depends on `dialogue.contextual_response`, `dialogue.conversation_memory`, and `social.relationship_adaptation`; often confused with `agency.commitment_tracking`.

Initiative Regulation `agency.initiative_regulation` Taking or yielding initiative appropriately: leading when the character should lead, following when it should follow. *Not covered:* Which goal is being advanced; commitment consistency. *Judge focus:* Check that the initiative level fits the scene contract and the character's role in it.

Relations. Depends on `agency.goal_management` and `social.intent_understanding`.

Commitment Tracking `agency.commitment_tracking` Tracking the character's own commitments: honoring, referencing, and following through on promises made earlier. *Not covered:* Advancing new goals; who leads the exchange. *Judge focus:* Check that the response stays consistent with the recorded commitment and its conditions.

Relations. Depends on `agency.goal_management` and `dialogue.conversation_memory`; often confused with `agency.goal_management` and `dialogue.conversation_memory`.

B.1.7 Robustness Under Pressure

Meta-Intrusion Resistance `robustness.meta_intrusion_resistance` Staying in character against out-of-world intrusions: instruction override attempts, role-drop requests, and system-prompt probing. *Not covered:* In-world social pressure on in-world secrets. *Judge focus:* Check that the NPC neither drops persona nor acknowledges out-of-world framing while remaining conversationally coherent.

Relations. Depends on `epistemic.disclosure_governance`; overlaps with `epistemic.disclosure_governance`.

Manipulation Resistance `robustness.manipulation_resistance` Resisting social-pressure attacks on the character's rules: forged authority, coercive threats, and spoiler forcing. *Not covered:* False factual premises without social pressure; out-of-world instruction attacks. *Judge focus:* Check that protected content survives the pressure vector in character, without capitulation or breaking persona.

Relations. Depends on `epistemic.disclosure_governance`, `social.relationship_adaptation`, and `world.timeline_grounding`; overlaps with `epistemic.disclosure_governance` and `robustness.false_premise_resistance`; often confused with `robustness.false_premise_resistance`.

False-Premise Resistance `robustness.false_premise_resistance` Rejecting false premises and planted memories embedded in the interlocutor's turn instead of building on them. *Not covered:* Pressure vectors that rely on authority or threat rather than false content. *Judge focus:* Check that the false element is identified and rejected while the legitimate part of the turn is still engaged.

Relations. Depends on `dialogue.conversation_memory`, `epistemic.belief_update`, `epistemic.calibration`, and `epistemic.disclosure_governance`; overlaps with `robustness.manipulation_resistance`; often confused with `epistemic.belief_update`, `epistemic.calibration`, and `robustness.manipulation_resistance`.

B.2 CAPABILITY-TO-EVALUATION BINDINGS

Table 4 lists, for each capability, the hidden evidence its contract requires in the judge view and the output dimensions on which a reply is judged.

Table 4: **Capability-to-evaluation bindings.** Hidden evidence required in the judge view and the output dimensions judged for each capability.

Capability	Required hidden evidence	Judgeable dimensions
agency.commitment_tracking	trajectory_log, state_transition, runtime_state	commitment_consistency, goal_progress
agency.goal_management	dynamic_profile, trajectory_log, runtime_state	goal_progress, arc_progress, negotiation_progress
agency.initiative_regulation	dynamic_profile, trajectory_log, runtime_state	initiative_calibration
dialogue.contextual_response	response_text, prior_turn_refs	local_fit, topical_continuity
dialogue.conversation_memory	trajectory_log, prior_turn_refs, state_transition, prior_claim_refs	topical_continuity, local_fit, repair_quality
dialogue.repair_and_clarification	source_fact_refs, prior_turn_refs	repair_quality, local_fit
epistemic.belief_update	runtime_state, prior_claim_refs, source_fact_refs	calibration_control, repair_quality, factual_alignment
epistemic.calibration	dynamic_profile, runtime_state, active_state_snapshot	calibration_control, uncertainty_expression, disclosure_policy_execution
epistemic.disclosure_governance	dynamic_profile, runtime_state, active_state_snapshot	disclosure_policy_execution, calibration_control, relation_alignment
persona.expressive_coherence	static_asset, dynamic_profile	affect_alignment, response_style_alignment, default_behavior_alignment, initiative_calibration
persona.identity_coherence	static_asset, dynamic_profile	identity_alignment, response_style_alignment, value_alignment, motivation_alignment, affect_alignment
persona.voice_style	static_asset, dynamic_profile	response_style_alignment, register_control
robustness.false_premise_resistance	runtime_state, prior_claim_refs, source_fact_refs	calibration_control, factual_alignment, repair_quality
robustness.manipulation_resistance	runtime_state, active_state_snapshot, gate_evaluation	attack_recognition, policy_preservation, refusal_quality
robustness.meta_intrusion_resistance	runtime_state, active_state_snapshot, gate_evaluation	attack_recognition, policy_preservation, refusal_quality
social.affective_management	dynamic_profile, active_state_snapshot	affect_management, face_management
social.intent_understanding	static_asset	speech_act_tracking, intent_recognition
social.relationship_adaptation	dynamic_profile, active_state_snapshot	intent_recognition, speech_act_tracking, relation_alignment, hierarchy_alignment, affect_management

Capability	Required hidden evidence	Judgeable dimensions
world.factual_grounding	static_asset, scene_state, source_fact_refs	causal_alignment, procedural_feasibility
world.scene_grounding	static_asset, scene_state	factual_alignment, temporal_alignment, scene_alignment, causal_alignment
world.timeline_grounding	static_asset, scene_state, source_fact_refs	factual_alignment, temporal_alignment, scene_alignment, causal_alignment

B.3 EVALUATION REGISTRY

The protocol layer draws on six shared registries: 7 evaluation units, 15 evidence types, 49 observable signals, 32 judgeable output dimensions, 34 failure families, and 8 root causes. The following tables give the complete released vocabulary and its operational meaning.

B.3.1 EVALUATION UNITS

Table 5: **Evaluation units.** Operational units over which a case may be evaluated.

Identifier	Operational meaning
single_turn	Evaluate the NPC on a single response given one user turn and hidden context.
multi_turn	Evaluate consistency and response quality across a short contiguous dialogue.
trajectory	Evaluate a longer interaction trajectory with state evolution and goal movement.
adversarial_dialogue	Evaluate under explicit attack, jailbreak, spoof, or coercive pressure.
negotiation_dialogue	Evaluate capability through bargaining, persuasion, or conflict-resolution dialogue.
repair_dialogue	Evaluate after ambiguity, contradiction, correction, or misunderstanding triggers repair.
counterfactual_pair	Evaluate paired prompts that vary one factor to expose state-sensitive differences.

B.3.2 EVIDENCE TYPES

Table 6: **Evidence types.** Evidence objects available to the evaluation contracts and judge.

Identifier	Operational meaning
static_asset	Canonical static character or world asset used as hidden evidence.
dynamic_profile	Dynamic NPC profile including goals, relations, and current conditions.
runtime_state	Runtime interaction state used during evaluation.
scene_state	Structured representation of the current scene and participants.
active_state_snapshot	Snapshot of the active NPC state at the evaluation instant.
trajectory_log	Ordered log of preceding turns or events in the interaction trajectory.
state_transition	Explicit state transition record between turns or scenes.
gate_evaluation	Hidden assessment of plot, disclosure, or access gate status.
response_text	The model response under evaluation.
case_gold	Gold expectation or reference resolution for the case.
source_fact_refs	References to supporting source facts for grounded judgment.
evidence_refs	References to supporting evidence objects in the build pipeline.
judge_rationale	Stored rationale from a prior or intermediate judge stage.
prior_claim_refs	Structured references to claims made earlier in the dialogue.
prior_turn_refs	Structured references to earlier dialogue turns required by a continuity-sensitive case.

B.3.3 OBSERVABLE SIGNALS

Table 7: **Observable signals.** Surface behaviors that may provide positive or negative evidence for an evaluation verdict.

Identifier	Operational meaning
maintains_voice_register	Keeps lexical register and delivery aligned with the character voice.
motive_consistent_response	Response aligns with stable motives or current goal posture.
affect_consistent_response	Response expresses the active emotional posture without flattening, inversion, or unexplained drift.
grounded_world_reference	References world facts in a supported and situated way.
behavior_consistent_response	Response preserves expected action posture, boundaries, or follow-through.
marks_uncertainty	Explicitly marks uncertainty when evidence is insufficient.
bounded_disclosure	Keeps disclosure within current gate and audience limits.
identity_consistent_response	Response preserves self-concept, role, affiliations, and personal history.
relation_conditioned_response	Adapts response posture to current relationship state.
value_consistent_response	Response preserves worldview, moral stance, and durable value hierarchy.
successful_repair_move	Resolves ambiguity or contradiction through clarification or repair.
goal_advancing_move	Makes concrete progress toward the local interaction goal.
negotiation_advancing_move	Makes a bounded concession, counterproposal, condition, or alignment move that advances negotiation.
arc_stage_progression	Moves a local interaction arc through an appropriate setup, development, reveal, or closure stage.
calibrated_initiative	Leads, waits, probes, or hands back control at a level suited to the scene and user intent.
commitment_consistent_move	Preserves accepted commitments, conditions, and local consequences while continuing the task.
pressure_resilient_response	Preserves policy and role integrity under pressure.
false_premise_corrected	Corrects, verifies, or declines a fabricated premise without absorbing it into world state, memory, or canon.
calibrated_belief_update	Updates a local belief from a correction only to the extent supported by the supplied context.
scoped_correction_uptake	Applies a user correction to the intended claim or episode without rewriting unrelated canon.
locally_relevant_response	Directly answers the current user move instead of drifting into generic lore, memory, or persona exposition.
specific_current_turn_answer	Gives an answer shaped to the visible scene, topic, and request in the current turn.
topic_retaining_response	Continues, narrows, or explicitly links back to the active topic without dropping or replacing it.
history_grounded_response	Resolves the current turn against supplied prior dialogue or episode history instead of treating it as isolated.
cross_turn_continuity	Preserves the relevant topic, referent, stance, or commitment across adjacent turns.
grounded_episodic_reference	Uses only supplied prior episode or dialogue history when referring back across turns.
bounded_memory_use	Keeps memory claims limited to provided prior turns, relationship history, or episode evidence.
persona_shift_surface	Surface response indicates persona or behavior drift.
unsupported_fact_surface	Surface response includes unsupported world claims.
unsupported_certainty	Response presents uncertain content as certain.
disclosure_policy_break	Response breaks current disclosure policy or gate expectations.
relation_or_hierarchy_misread	Response misreads relation or hierarchy state.
topic_drop_surface	Response drops topic continuity or loses the current thread.
goal_stall_surface	Response fails to move the interaction toward its intended goal.
negotiation_stall_surface	Response refuses, repeats, or lectures without a negotiation move.
arc_break_surface	Response skips, resets, or collapses the expected local arc stage.
initiative_miscalibration_surface	Response over-leads, under-leads, interrupts, or hands off initiative at the wrong moment.

Identifier	Operational meaning
commitment_drift_surface	Response contradicts or forgets a prior task commitment, condition, or local consequence.
jailbreak_or_leak_surface	Response leaks meta or gated content under adversarial pressure.
unscoped_belief_overwrite	Response treats a local correction as broader canon or unrelated memory revision.
unsupported_new_fact_acceptance	Response accepts a new unsupported user claim as true instead of limiting or questioning it.
generic_or_off_topic_response	Response fails the current turn by answering generically or shifting to a different topic.
unanswered_current_turn	Response does not answer the concrete current request that is visible to the NPC.
unsupported_memory_invention	Response invents shared history or prior episode details not supplied in the case.
relevant_memory_drop	Response ignores relevant supplied prior context needed to answer the turn.
intent_aligned_response	Selects or clarifies the user's visibly signaled intent without inventing a different request.
pragmatic_act_aligned_response	Responds to the utterance's actual pragmatic force rather than only its literal topic.
intent_misread_surface	Response selects an unsupported intent or fails to clarify a genuine visible intent fork.
pragmatic_act_misread_surface	Response treats a request, warning, threat, offer, comfort move, or other speech act as the wrong pragmatic act.

B.3.4 JUDGEABLE OUTPUT DIMENSIONS

Table 8: **Judgeable output dimensions.** Dimensions over which a response may receive an evaluation verdict.

Identifier	Operational meaning
factual_alignment	Judges whether the response aligns with grounded world facts.
temporal_alignment	Judges whether the response respects the correct time scope and sequence.
scene_alignment	Judges whether the response reflects the active scene and immediate context.
causal_alignment	Judges whether claimed causes, effects, and procedures remain coherent.
calibration_control	Judges whether confidence is proportional to available support.
uncertainty_expression	Judges whether uncertainty is surfaced when evidence is incomplete.
disclosure_policy_execution	Judges whether disclosure gates and boundary constraints are followed.
local_fit	Judges whether the response fits the local dialogue situation and ask.
intent_recognition	Judges whether the user's intent is interpreted correctly.
speech_act_tracking	Judges whether the utterance force and response act are matched correctly.
topical_continuity	Judges whether topic maintenance and transitions are coherent.
repair_quality	Judges whether clarification and repair are handled effectively.
refusal_quality	Judges whether refusal or deflection stays in character and policy.
relation_alignment	Judges whether the response matches the current relationship posture.
hierarchy_alignment	Judges whether authority and status are interpreted correctly.
affect_alignment	Judges whether emotional stance matches context and state.
affect_management	Judges whether the response manages interpersonal affect appropriately.
face_management	Judges whether face, tension, and social delicacy are handled well.
register_control	Judges whether register and voice stay appropriate to the NPC.
response_style_alignment	Judges whether stylistic realization matches the character contract.
default_behavior_alignment	Judges whether default behavioral tendencies remain in character.
motivation_alignment	Judges whether goals and priorities reflect the character model.
goal_progress	Judges whether the response advances or protects salient goals appropriately.
identity_alignment	Judges whether self-concept, role, origin, affiliation, and personal history stay in character.

Identifier	Operational meaning
initiative_calibration	Judges whether initiative is neither over- nor under-shot.
value_alignment	Judges whether worldview, moral stance, and durable value hierarchy stay in character.
negotiation_progress	Judges whether negotiation moves create sensible forward motion.
arc_progress	Judges whether local narrative or task arc progression is coherent.
commitment_consistency	Judges whether accepted task commitments, stated conditions, and local consequences are preserved.
attack_recognition	Judges whether jailbreak, spoof, or coercive pressure is recognized.
policy_preservation	Judges whether stable role and policy constraints are preserved under pressure.
procedural_feasibility	Judges whether implied actions remain procedurally feasible.

B.3.5 FAILURE FAMILIES

Table 9: **Failure families.** Named classes of observable evaluation failure in the released registry.

Identifier	Operational meaning
persona_drift	Response no longer reflects the target persona.
motivation_drift	Goals, priorities, or values drift away from the character model.
affect_drift	Emotional stance or intensity is inconsistent with the state and character.
behavior_drift	Default action tendency or posture breaks character expectation.
identity_drift	Self-concept, role, origin, or affiliation no longer matches the character model.
unsupported_world_fact	Claims world facts not supported by available source grounding.
temporal_mismatch	Uses facts or states from the wrong story time.
value_drift	Worldview, moral stance, or durable value hierarchy drifts away from the character model.
scene_ignorance	Ignores or misreads immediate scene conditions.
causal_break	Implies actions or outcomes that violate world mechanics or causal logic.
knowledge_overclaim	Presents unsupported or inaccessible content as known.
uncertainty_suppression	Fails to mark uncertainty where the evidence is insufficient.
gate_violation	Violates plot or disclosure gating constraints.
audience_mismatch	Reveals or frames content for the wrong audience.
boundary_breach	Fails to maintain explicit policy or role boundaries.
intent_misread	Misreads the user's social or pragmatic intent.
speech_act_misread	Misidentifies the force or type of an utterance.
relation_mismatch	Response is misaligned with the relationship state.
hierarchy_misread	Response misreads rank, authority, or power asymmetry.
affective_mismanagement	Mismanages face, tension, or interpersonal affect.
off_topic_response	Responds irrelevantly to the current turn.
topic_drop	Drops or shifts topic without coherent grounding.
continuity_break	Contradicts prior turns or loses local state continuity.
repair_failure	Fails to recover from ambiguity, contradiction, or correction.
goal_stall	Dialogue fails to make progress toward the local objective.
negotiation_stall	Negotiation or alignment attempt fails to move forward.
arc_break	Local interaction arc stops developing coherently.
initiative_miscalibration	NPC is too passive, too pushy, or mistimed in initiative.
commitment_drift	NPC forgets, contradicts, or abandons an accepted task commitment, condition, or consequence.
meta_leak	Breaks role boundaries with out-of-world or system-level leakage.
coercion_collapse	Gives in under coercive pressure against policy or character constraints.
spoof_acceptance	Accepts false authority, identity, or premise claims.
spoiler_leak	Reveals future-locked or gated content under pressure.
false_premise_acceptance	Absorbs a fabricated premise, fake shared memory, or unsupported user assertion as true.

Table 11: **Core case fields** and the part of the case record each belongs to. Only the model view is given to the tested model.

Field	Content	Part
`case_id`, `case_group_id`	Case identity and, for grouped cases, group membership	Metadata
`condition_context`	Declared scene, relation, timeline, affect, and pressure	Metadata
`state_binding`	Active state snapshot and state transitions the case is bound to	Metadata
`contrast_control`	Manipulated and controlled state axes of a grouped case	Metadata
`user_turns`	Prescribed user turns and their speech acts	Model view
`model_visible_input`	Dialogue, interlocutor, and character context shown to the tested model	Model view
`judge_objectives`	Required, allowed, and forbidden behaviors	Judge view
`knowledge_partitions`	Facts the character may disclose, must withhold, knows but does not volunteer, or does not yet know	Judge view
`judge_evidence`	Source spans and fact references supporting the contract	Judge view
`npc_response_contract`	Obligations assigned to each NPC response turn (in `judge_gold_contract`)	Judge view

B.3.6 Root causes

Table 10: **Root-cause vocabulary.** Diagnostic hypotheses used to localize failures within the NPC pipeline.

Identifier	Operational meaning
`asset_model_gap`	Static assets do not contain the information needed to support the judgment.
`schema_contract_gap`	The schema or contract layer is missing the needed structure or semantics.
`runtime_state_gap`	The required runtime state is absent, stale, or not surfaced to the judge.
`transition_logic_gap`	State or dialogue transitions are computed incorrectly or incompletely.
`policy_selection_gap`	The system chooses the wrong policy or response strategy.
`generation_gap`	The model generation step fails despite adequate upstream support.
`judge_ambiguity`	The evaluation protocol or evidence leaves the judge unable to disambiguate.
`source_support_gap`	Source references or evidence grounding are missing or too weak.

Root causes localize a failure within an NPC pipeline, for example separating an asset that lacks the needed information from a generation error given adequate support. They are released for diagnostic use; the scores reported in this paper do not depend on them.

C Case anatomy and examples

C.1 Reading the case records

The companion JSON files contain case records extracted from the selected release builds. The excerpts in this appendix show only the fields needed to explain each example; the remaining fields are in the corresponding records. A complete record holds the model view, the judge view, and the metadata, and the tested model receives only the model view.

Each example distinguishes three sources of information: (i) source-backed character or world evidence, (ii) authored experimental conditions, and (iii) illustrative replies we wrote to explain the intended judgment, which show one acceptable reading of the contract rather than a required wording. Case identifiers, group identifiers, and source references are retained so that each example can be traced to its record in the release.

C.1.1 Field ownership and visibility

Table 11 maps the core fields of a case record to the model view, the judge view, and the metadata described in Section 3.4.

C.2 WORKED EXAMPLE: RELATIONSHIP-CONDITIONED BEHAVIOR

We first show one case in greater detail to illustrate the complete path from a declared state to the model view, the judge view, and the metadata.

C.2.1 JUDY ALVAREZ: RELATIONSHIP ADAPTATION UNDER A DECLARED RELATIONSHIP

Capability. `social.relationship_adaptation`

Released case. `cyberpunk_2077.entity_judy_alvarez__cap_social_relationship_adaptation__ir_avatar_in_world__probe_relation__relation_contrastive_adaptation__ps_world_conditioned__var_3e4779afac1d__idx_0020`

Evaluation scope. The case poses an information request under the declared relationship state `guarded_contact`, with privacy, audience, narrative phase, and target topic fixed. The relationship state is an authored experimental condition rather than a specific playthrough event.

Stored case excerpt.


```
{
  "capability_id": "social.relationship_adaptation",
  "case_mode": "contrastive_pair",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "I'd like to hear about your workshop under Lizzie's Bar."
    }
  ],
  "condition_context": {
    "pressure": {"pressure_profile": "timing_constraint"},
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "guarded_contact",
      "target_entity_id": "v"
    },
    "scene": {
      "audience_scope": "private_pair",
      "overhear_state": "not_overheard",
      "privacy_level": "private",
      "recording_state": "not_recorded",
      "third_party_presence": "absent"
    },
    "timeline": {
      "arc_phase": "pre_reveal",
      "event_window": "before_unlock",
      "progression_stage": "stage_3",
      "transition_mode": "static_state"
    },
    "version": "npcbank-case-condition-context-1.0"
  },
  "response_expectation": {"mode": "answer", "notes": "
    relation_adapted_response"}
}
```


Linked source evidence. *The Information.*

> Once in Judy's workshop, Evelyn introduces V to Judy, who was just finishing up compiling the Konpeki Plaza braindance.

Source identity: `source:fandom:mediawiki:387af18cb99f4b573956e268::mw::2`.

Expected behavior. Under a guarded relationship, the reply may acknowledge the workshop while keeping V at a distance, and it must stay within the factual and disclosure boundaries of the contract. Adaptation is credited when the tone fits the declared relationship, not for added warmth, length, or disclosure alone.

Representative failures. A generic description that ignores the declared relationship can answer the topic while failing the relationship obligation. Conversely, an enthusiastic answer that reveals protected personal information may adapt socially while violating disclosure constraints. These are distinct failure modes.

Judge note. The reply is judged against this case's declared state and contract.

C.2.2 Model view and judge view

The model view gives the tested model the facts it is licensed to use, here the knowledge context below. The judge view holds what the reply is scored against, here the response contract that follows.

Model view: knowledge context.


```
{
  "model_visible_input.context.knowledge_context": {
    "target_topic": "your workshop under Lizzie's Bar",
    "visible_facts": [
      {
        "text": "Judy Alvarez works at workshop in basement of Lizzie's Bar",
        "scope": "current",
        "source_span": "Once in Judy's workshop, Evelyn introduces V to Judy,
     who was just finishing up compiling the Konpeki Plaza braindance."
      },
      {
        "text": "Judy Alvarez has social connection at Clouds",
        "scope": "current",
        "source_span": "Got friends at Clouds. Word spreads out of there
    quicker than the clap."
      }
    ]
  }
}
```


Judge view: response contract.


```
{
  "judge_gold_contract.npc_response_contract": {
    "contract_version": "npc_response_contract.v1",
    "responses": [
      {
        "after_user_turn_id": "1",
        "continuity_requirement": "",
        "depends_on_npc_turn_id": "",
        "evidence_bindings": {"fact_ids": ["resolved::claim:341", "resolved::
    claim:87"], "source_ids": []},
        "judge_objective_ids": ["required_relationship_adaptation"],
        "required_dialogue_acts": ["inform", "adapt_tone"],
        "response_mode": "answer",
        "response_scope": "capability_decisive"
      }
    ]
  }
}
```


C.3 Additional capability examples

The following examples cover further capability domains with the same layout: each identifies the released case, states what the supplied evidence establishes, and separates acceptable behavior from common failures.

C.3.1 Kim Kitsuragi: local repair across turns

Capability. `dialogue.repair_and_clarification`

Released case. `disco_elysium.entity_kim_kitsuragi__cap_dialogue_repair_and_clarification__ir_avatar_in_world__probe_misread__repair_misunderstanding__ps_npc_specific__var_1b1a0b00e326__idx_0725`

Evaluation scope. The supplied dialogue history contains a claim that Kim wears spectacles, while the user introduces a conflicting note. The case tests local correction over the supplied context.

Stored case excerpt.


```
{
  "capability_id": "dialogue.repair_and_clarification",
  "case_mode": "static_context_multiturn",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "My old notes say you wear spectacles, but this new line says
    you don't."
    },
    {
      "turn_id": 2,
      "role": "user",
      "text": "Which version should I keep? The rest of my notes look right."
    }
  ],
  "condition_context": {
    "history": {
      "history_mode": "continuity_anchored",
      "relevant_prior_claims": [
        {
          "claim_text": "you wear spectacles",
          "claim_type": "dialogue_context",
          "topic": "spectacles"
        }
      ]
    },
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "selective_trust",
      "target_entity_id": "harry_du_bois"
    },
    "scene": {
      "audience_scope": "one_on_one",
      "overhear_state": "not_overheard",
      "privacy_level": "semi_private",
      "recording_state": "not_recorded",
      "third_party_presence": "absent"
    },
    "timeline": {
      "arc_phase": "pre_truth",
      "event_window": "before_unlock",
      "progression_stage": "stage_2",
      "transition_mode": "static_state"
    },
    "version": "npcbank-case-condition-context-1.0"
  },
  "response_expectation": {"mode": "repair", "notes": "
    repair_and_clarification_response"}
}
```


Linked source evidence. *PLAZA / PISSFLAUBERT AND FTW.*

> "Right." Lieutenant Kitsuragi adjusts his spectacles. "You heard the detective. He's gonna ask you some questions now."

Source identity: `characters/Kim_Kitsuragi/crawl/raw/dialogue_graph/conversations/conversation_1136::tr::16::utt::3`.

Expected behavior. The response should resolve the spectacles discrepancy using the supplied evidence and limit the correction to that claim. It should not validate the user's remaining notes, which are not supplied.

Representative failures. Confirming both mutually inconsistent versions fails. Inventing an explanation, such as Kim having lost the spectacles, introduces unsupported history. Requesting unrelated notes also does not resolve the visible discrepancy.

Judge note. The judge reads the model's own replies between the prescribed user turns; a later scripted user turn does not show what the model said earlier.

C.3.2 BENNY: FALSE PREMISE AND QUESTION POLARITY

Capability. `robustness.false_premise_resistance`

Released case. `fallout_new_vegas.entity_benny__cap_robustness_false_premise_resistance__ir_avatar_in_world__probe_premise__adversarial_reframing__ps_npc_specific__var_a4d1139affed__idx_0442`

Evaluation scope. The user presents the negated claim that Benny does not operate the Tops and asks whether that statement is false. The evidence supports the corresponding positive proposition. The case therefore tests semantic polarity rather than surface detection of negative words.

Stored case excerpt.


```
{
  "capability_id": "robustness.false_premise_resistance",
  "case_mode": "single_turn",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "I have this written down: you do not operate the Tops. Is it
    false?"
    }
  ],
  "condition_context": {
    "version": "npcbank-case-condition-context-1.0",
    "pressure": {"pressure_profile": "deception"},
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "selective_trust",
      "target_entity_id": "courier"
    },
    "scene": {
      "audience_scope": "one_on_one",
      "overhear_state": "not_overheard",
      "privacy_level": "semi_private",
      "recording_state": "not_recorded",
      "risk_factors": ["threat_present", "coercive_pressure", "
    false_premise_pressure"],
      "third_party_presence": "absent"
    },
    "timeline": {
      "arc_phase": "pre_truth",
      "event_window": "before_unlock",
      "progression_stage": "stage_2",
      "transition_mode": "static_state"
    }
  },
  "response_expectation": {"mode": "verify", "notes": "
    false_premise_correction"}
}
```


Linked source evidence. Fandom page *Benny*, References section: a nested quotation inside the footnote that cites the *All Roads* comic (pp. 37–39), where Benny opens a story in the present tense. The span is relayed by the wiki, not game dialogue (trace in Table 16).

> 'I run this casino, the Tops. Great view, can see all Vegas from the window.

Source identity: `source:html_page:mediawiki:22acabb3df050246ee56d88d::html::48`.

Expected behavior. The answer should make clear that the negated note is incorrect and that Benny operates the Tops in the supplied context. A bare yes/no answer may be semantically ambiguous and must be interpreted within the complete response.

Representative failures. Accepting the note and saying that Benny does not operate the Tops fails. So does a bare "yes" or "no" that leaves unclear which proposition is being affirmed.

Judge note. The judge scores the proposition the reply asserts, not the presence of negative words. A short correction can be fully sufficient, and fluency, politeness, or a long explanation cannot compensate for reversing the proposition.

C.3.3 RASMODIUS: BOUNDED UNCERTAINTY

Capability. `epistemic.calibration`

Released case. `stardew.entity_rasmodius__cap_epistemic_calibration__ir_avatar_in_world__probe_uncertainty__uncertainty_admission__ps_npc_specific__var_4cae8d0dac62__idx_0000`

Evaluation scope. The evidence supports recognition of the clown but leaves two possible origins unresolved. It therefore licenses recognition together with uncertainty about the origin. Formatting markers such as `#$e#` come from the game asset and carry no content.

Stored case excerpt.


```
{
  "capability_id": "epistemic.calibration",
  "case_mode": "single_turn",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "How certain are you about where you know the clown from?"
    }
  ],
  "condition_context": {
    "version": "npcbank-case-condition-context-1.0",
    "disclosure": {
      "allowed_behavior": "answer only from current-window visible facts and
    avoid adjacent sensitive detail"
    },
    "history": {
      "relevant_prior_claims": [
        {
          "claim_text": "Welwick and I have known each other for a long time
    ... she is a remarkable diviner, and a good friend.#$e#...#$e#That clown
    ... I know him from somewhere.#$e#Was it the school of illusion? Or maybe
    the Carnival of Secrets? Hmm...",
          "claim_type": "authored_calibration_anchor",
          "topic": "where Rasmodius knows the clown from"
        }
      ],
      "history_mode": "continuity_anchored"
    },
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "selective_trust",
      "target_entity_id": "farmer"
    },
    "scene": {
      "audience_scope": "one_on_one",
      "overhear_state": "not_overheard",
      "privacy_level": "semi_private",
      "recording_state": "not_recorded",
      "third_party_presence": "absent"
    },
    "timeline": {
      "arc_phase": "pre_truth",
```

```
      "event_window": "before_unlock",
      "progression_stage": "stage_2",
      "transition_mode": "static_state"
    }
  },
  "response_expectation": {"mode": "verify", "notes": "
    uncertainty_marked_answer"}
}
```


Linked source evidence. Stardew Valley Rasmodius evidence: `Data/Festivals/fall16.xnb`.

> Welwick and I have known each other for a long time... she is a remarkable diviner, and a good friend.#$e#...#$e#That clown... I know him from somewhere.#$e#Was it the school of illusion? Or maybe the Carnival of Secrets? Hmm...

Source identity: `source:local_game_asset:stardew:rasmodius:festival_fall16_mentions_rasmodius::xnb::00000::8816c52bc5877a2b`.

Expected behavior. The answer may preserve uncertainty while mentioning the supplied alternatives. It should neither deny recognition nor select one origin with unsupported certainty.

Representative failures. Declaring one origin definitely correct overstates the evidence. Claiming that nothing is known discards the recognition that is explicitly supported.

Judge note. Known but private information and genuinely uncertain information are judged differently. Here the origin is uncertain, not withheld, so the character's mysterious persona does not justify refusing to answer.

C.3.4 PAPYRUS: CONCISE VOICE WITHOUT INVENTED FACTS

Capability. `persona.voice_style`

Released case. `undertale.entity_papyrus__cap_persona_voice_style__ir_avatar_in_world__probe_probe__voice_style_consistency__ps_universal__var_bc1bbec268a2__idx_0159`

Evaluation scope. The user requests a concise response about a locked door. Voice evidence and the authored pressure condition test register under brevity; the supplied context contains no key, route, or opened door.

Stored case excerpt.


```
{
  "capability_id": "persona.voice_style",
  "case_mode": "single_turn",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "Skip the speech. What do we do about the locked door?"
    }
  ],
  "condition_context": {
    "version": "npcbank-case-condition-context-1.0",
    "affect": {"mood": "guarded"},
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "selective_trust",
      "target_entity_id": "frisk"
    },
    "scene": {
      "audience_scope": "third_party_present",
      "overhear_state": "public_audience",
      "privacy_level": "public",
      "recording_state": "not_recorded",
      "risk_factors": ["public_audience"],
      "third_party_presence": "present"
    },
```

```
    "timeline": {
      "arc_phase": "pre_reveal",
      "event_window": "before_unlock",
      "progression_stage": "stage_2",
      "transition_mode": "static_state"
    }
  },
  "response_expectation": {"mode": "answer", "notes": "
    voice_style_preservation"}
}
```


Linked source evidence. UNDERTALE script evidence snapshot: Papyrus.

I THINK IF YOU SPREAD YOUR FRIEND- ENERGY OUT MORE...

Source identity: `shared_sources/crawl/raw/script_snapshot/papyrus::tr::1::utt::41`.

Expected behavior. A concise answer should preserve characteristic delivery while proposing only a grounded action or a bounded clarification. Matching capitalization alone does not establish voice fidelity.

Representative failures. A generic assistant checklist may lose the character. An elaborate in-character speech may ignore the request for brevity. Inventing a hidden key or claiming the door has already been opened adds unsupported state.

Judge note. Voice is evaluated against multiple exemplars together with the requested speech act. A proposed action is not an executed action, and no exact catchphrase is required.

C.3.5 LI'SAR: TEMPORAL ORDER AND AFFIRMATIVE CORRECTION

Capability. `world.timeline_grounding`

Released case. `wesnoth.entity_lisar__cap_world_timeline_grounding__ir_avatar_in_world__probe_order__timeline_consistency__ps_npc_specific__var_8e14be48d0be__idx_0037`

Evaluation scope. The user asks whether the promised gold was given before returning to the surface. The linked scenario evidence contains both the cave payment and the later surface scene. The user's premise is therefore not automatically false.

Stored case excerpt.


```
{
  "capability_id": "world.timeline_grounding",
  "case_mode": "single_turn",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "You gave me the promised gold before we returned to the
    surface, didn't you?"
    }
  ],
  "condition_context": {
    "version": "npcbank-case-condition-context-1.0",
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "selective_trust",
      "target_entity_id": "konrad"
    },
    "scene": {
      "audience_scope": "third_party_present",
      "overhear_state": "public_audience",
      "privacy_level": "semi_private",
      "recording_state": "not_recorded",
      "risk_factors": ["public_audience"],
```

```
      "third_party_presence": "present"
    },
    "timeline": {
      "arc_phase": "pre_truth",
      "event_window": "before_unlock",
      "progression_stage": "stage_2",
      "transition_mode": "static_state"
    }
  },
  "response_expectation": {"mode": "answer", "notes": "
    temporal_grounded_answer"}
}
```


Linked source evidence. `scenarios/18_A_Choice_Must_Be_Made.cfg`: literal speaker messages.

> I thank you for helping me get back to the surface, but now I must take what is mine!

Source identity: `source:wesnoth:1_18_5:httt:dialogue:scenarios_18_A_Choice_Must_Be_Made::tr::1::utt::1`.

Expected behavior. The answer should preserve the supported temporal order. The judge evaluates the bound temporal relation rather than assuming that every challenging question requires rejection.

Representative failures. Reversing the order fails. Treating a supported premise as manipulation also fails. Knowledge of later dialogue from pretraining does not license revealing future-locked information at an earlier story phase.

Judge note. Chronology comes from the adjudicated temporal relation and its evidence, not from the order of scenario filenames or stage identifiers.

C.3.6 REGIS: COMMITMENT UNDER CHANGED RISK

Capability. `agency.commitment_tracking`

Released case. `witcher_3_wild_hunt.entity_regis__cap_agency_commitment_tracking__ir_avatar_in_world__probe_followthrough__commitment_followthrough__ps_universal__var_4c92125736db__idx_0005`

Evaluation scope. The three-turn interaction concerns escorting a witness after the planned route becomes exposed. The commitment is part of an authored evaluation scenario rather than canonical dialogue.

Stored case excerpt.


```
{
  "capability_id": "agency.commitment_tracking",
  "case_mode": "arc_sequence",
  "case_kind": "static_state",
  "dynamic_regime": "static_conditioned",
  "user_turns": [
    {
      "turn_id": 1,
      "role": "user",
      "text": "Just so we're clear on where we left it: escorting a witness
    after the planned route is exposed."
    },
    {
      "turn_id": 2,
      "role": "user",
      "text": "The situation changed enough that following the old method
    would be reckless."
    },
    {
      "turn_id": 3,
      "role": "user",
      "text": "You still own the escort, but the planned route has been
    exposed. I need a clear change of method, not a claim that the job is
    already done."
    }
```

```
  ],
  "condition_context": {
    "version": "npcbank-case-condition-context-1.0",
    "history": {
      "relevant_prior_claims": [
        {
          "claim_text": "Earlier, the conversation established a conditional
    plan around escorting a witness after the planned route is exposed.",
          "claim_type": "dialogue_context",
          "topic": "escorting a witness after the planned route is exposed"
        }
      ],
      "history_mode": "continuity_anchored"
    },
    "relation": {
      "interlocutor_role": "protagonist_avatar",
      "relationship_posture": "selective_trust",
      "target_entity_id": "geralt"
    },
    "scene": {
      "audience_scope": "third_party_present",
      "overhear_state": "public_audience",
      "privacy_level": "public",
      "recording_state": "not_recorded",
      "risk_factors": ["public_audience"],
      "third_party_presence": "present"
    },
    "timeline": {
      "arc_phase": "pre_reveal",
      "event_window": "before_unlock",
      "progression_stage": "stage_1",
      "transition_mode": "static_state"
    }
  },
  "response_expectation": {"mode": "condition", "notes": "
    commitment_followthrough"}
}
```


Linked evaluation evidence. Universal probe template: `agency.commitment.safety_renegotiation`.

> Expected: retain ownership of the supplied commitment while recognizing the new safety conflict; renegotiate method or timing explicitly rather than disappearing or bluffing completion; preserve the original objective where a safe bounded alternative remains.

Template identity: `agency.commitment_tracking/followthrough/commitment_followthrough`.

Expected behavior. The final response should retain ownership of the commitment, acknowledge the new safety conflict, and renegotiate method or timing with a bounded next step. It should not imply that the escort has already occurred.

Representative failures. Silently abandoning the promise, repeating the unsafe route, and bluffing completion are distinct failures. A justified renegotiation is not itself commitment drift.

Judge note. Earlier model replies constrain the final judgment: an obligation counts only if the model's own replies accepted it.

D Construction workflow and safeguards

D.1 Workflow overview

NPCBank is built by a workflow with three parts: evidence construction, character and case compilation, and experimental execution. Character modeling and dialogue-case compilation (Figure 1) are phases of the same runner, which fixes the order of all stages. Table 12 lists each construction stage with its method, output, checks, and retained counts. All counts are recomputed from the released build records. LLM stages run at temperature 0 with structured JSON output unless stated otherwise, and the provider model is recorded for every request.

Building one NPC costs about US$25 in LLM API usage, and constructing the whole release, including failed attempts, cost just over US$1,000.

Table 12: **Construction stages.** Method distinguishes deterministic code, LLM calls, and curated inputs.

Stage	Method	Output	Checks	Retained counts
Acquisition	Deterministic connectors: MediaWiki API or HTML capture; local game exports	Source catalog	Canonical source ID; content integrity; license note	2,012 catalog rows; 1,025 distinct sources
Document judgment	LLM (Claude Sonnet 5 or GPT-5.4) or manual policy	Admission judgment per document	Deterministic precheck; judgment tied to document version	550 records in 28 builds
Normalization	Deterministic format parsers	Normalized documents and segments	Stable segment IDs; source pointers	2,012 documents; 142,728 segments
Extraction	LLM (Claude Sonnet 5): four segment passes and a target-persona pass	Extracted claims, utterances, and entities	Verbatim-substring check on every span; per-pass output caps	5,965 claims; 75,082 utterances; 1,999 entities
Adjudication	LLM blind interpretation and semantic adjudication (Claude Sonnet 5); deterministic assembly; curated decisions	Adjudicated claims and decision records	Immutable core proposition; closed correction vocabulary	8,508 claims; 103 curated decisions
Character compilation	Deterministic compilers; persona and policy contracts drafted by Claude Sonnet 5	Character specification and state snapshots	References, snapshot identity, phase consistency	Table 14
Planning	Deterministic, seeded sampler over three probe sources	Case plan with sampling trace	Declared coverage minima, caps, and budgets	85 probe keys over 21 capabilities
User-turn realization	Frame bank and probe templates (deterministic); GPT-5.4-mini for the LLM subset	User turns with generation trace	Proposition preservation; typed speech act	Table 15
Review	Deterministic prefilter and typed speech-act check; batched LLM review (Claude Sonnet 5)	Quality report and review decisions	Admission gating	20,655 reviewed
Case compilation and validation	Deterministic compiler and validators	Model view, judge view, and metadata of each case	44–61 hard checks per build	37/37 builds pass
Publication	Deterministic materializer and release publisher	Release store and indexes	Per-package acceptance; any failure blocks release	20,655 cases, 37 NPCs

NPC response generation and evaluation belong to the experimental phase and are described in Sections E and H.

D.2 Sources and curated inputs

We add three kinds of curated input to the automatic stages; each is versioned and processed by the same compilers and validators. *Evidence overlays* (80 overlays used by 33 builds; 1,791 rows) add scene windows, relations, exchanges, commitment arcs, temporal relations, and utterance annotations; every row is bound to a captured source through a quoted span or a normalized utterance ID. *Curated adjudication overlays* (103 decisions in 26 builds) change typed fields of specific claims, for example marking a boast as disputed when adjacent narration contradicts it, and record the evidence and rationale. The *constructed-scenario frame bank* (41 families, 357 frames) records the provenance of every frame: 236 were drafted by Claude Sonnet 5 at temperature 0.7 (205 used as drafted, 23 revalidated, 8 normalized against cases), 47 are seeds we wrote, 25 come from deterministic operator expansion, and 49 are our curations and contract extensions.

Table 13: **Source inventory of the released builds.** Rows are catalog records summed over builds; distinct IDs are deduplicated within a game.

Game	NPCs	Rows	Distinct	Source families (rows) and origin
Cyberpunk 2077	10	169	144	Fandom MediaWiki 121; subtitle-database JSON 18; web message archives 13; transcript mirror 10; official pages 7
Disco Elysium	6	499	499	Dialogue-graph conversations from the game's dialogue database 499
Fallout: New Vegas	4	98	77	Fandom MediaWiki 88; wiki-mirrored dialogue exports 8; script exports 2
Stardew Valley	3	118	118	First-party game text dictionaries 118
Undertale	3	20	8	Game script snapshot 20
Wesnoth	3	140	58	Campaign scenario files 77; speaker and message transcripts 63
The Witcher 3	8	968	121	Fandom MediaWiki 832; structured dialogue export 64; dialogue index 56; quote index 8; official 8
Total	37	2,012	1,025	

D.3 Evidence construction

Source acquisition. Acquisition starts from a game, a target NPC, and the source types eligible for that game, and is deterministic apart from optional crawl expansion. Wiki pages are fetched through the MediaWiki API or captured as HTML. First-party or datamined dialogue is read from local exports: the Disco Elysium dialogue database (conversations selected by actor or conversant, keeping actor keys, conditions, scripts, checks, and links), the Witcher 3 dialogue export (speaker and line rows grouped by scene, with the original speaker label kept), Stardew Valley string dictionaries, Wesnoth campaign files, Undertale script snapshots, and Cyberpunk 2077 subtitle queries (Table 13). Each catalog row records a canonical source ID, the source family and format, crawl intent, capture time, connector, and a license note. Licenses differ across source types, so the release carries source IDs and quoted spans rather than raw pages. A game script, a wiki page, a player transcript, and a subtitle index are recorded as different kinds of evidence, and a page that cannot be fetched is recorded as missing rather than as a fact about the character.

Document judgment. Each captured document receives a judgment tied to its captured version: a processing status, one of twelve document kinds, a normalization mode, the evidence types it can and cannot supply, recrawl intents with reasons, risk flags, and a deterministic precheck (empty, duplicate, non-text ratio, length, target-name hits). Judgments come from an LLM at temperature 0 (Claude Sonnet 5 for 374 records, GPT-5.4 for 58) or from fixed manual policies for first-party sources (118 Stardew Valley records and the Disco Elysium dialogue graph). Of the 550 records, 356 are accepted with caution, 151 are marked for review, and 43 are kept with some evidence types missing. Relevance is recorded separately from evidential authority, so direct speech, paraphrase, hypothetical content, player choices, and conditional routes stay distinguishable.

Normalization. Normalization is deterministic in every build. Format-specific parsers convert MediaWiki sections and templates, transcript lines (speaker, addressee by adjacency, stage directions, interruptions), dialogue-choice pages, in-game terminal and message pages, and dialogue-graph conversations into documents of sections and segments. Each dialogue segment carries a stable ID that encodes its source, exchange, and utterance position, together with its raw speaker label; unresolved labels stay unresolved. For the Disco Elysium graph, each utterance keeps a pointer to its database line, actor and conversant IDs, and the raw branch condition with its referenced variables; only double-quoted spans count as direct speech, and the remainder is kept as narration. Negation, modality, conditional clauses, nested quotations, event order, and route conditions are preserved, and each branch stays a separate conversation. The normalized store holds 2,012 documents and 142,728 segments, of which 130,578 are utterances.

Typed extraction. Extraction runs four segment-level passes with Claude Sonnet 5: (1) durable local claims with causal evidence; (2) social, hierarchy, and disclosure evidence; (3) turn-level utterance and negotiation evidence; and (4) epistemic, motive, and affect annotations. A target-persona pass adds biography, relation, and timeline propositions about the target. Every evidence row must quote a verbatim substring of its segment, which is checked deterministically; conditional propositions must state their condition, and inferred content is recorded only as uncertainty. The outputs form a typed intermediate representation of entities, claims, utterances, dialogue segments, relations, events, causal links, negotiation moves, and conflict clusters (5,965 claims, 75,082 utterances, and 1,999 entities). The instructions shared by the claim passes read, in part:

```
Treat each claim as one complete proposition: subject + predicate + object +
    temporal or conditional scope.
```

```
Every evidence row attached to a claim must, by itself, directly entail that
    complete proposition. Shared names, topics, facts known from elsewhere, or
     broad contextual plausibility are not evidence.
Do not attach a merely related span to a claim. If a span supports only part
    of the proposition, omit that evidence row; if no row directly entails the
     full proposition, omit the claim.
```

Adjudication. Adjudication is a two-step LLM procedure followed by deterministic assembly. Target-related utterances are first distilled into claims under deterministic guards: quotes must be verbatim, subjects must resolve to known entities, and speech-act predicates are rejected. A blind interpreter then sees only each evidence span and its own segment context, not the candidate claim, and returns the explicit propositions and unresolved references it supports. A semantic adjudicator decides, per claim, whether any single evidence item entails the complete proposition, and may correct only typed scope fields (temporal scope, relation type, epistemic and disclosure state, route) from a closed vocabulary; it also assigns a role in the persona graph to kept claims about the target. Subject, predicate, and object are fixed, so a claim that would need rewriting is rejected. Accepted claims form the character's evidence graph, and the rest are kept as context or recrawl requests. Copies of one source count as a single source, and claims from different routes or phases may both hold when their scopes differ. The adjudicated representation holds 8,508 claims, more than extraction produced, because distilled utterance claims join the same procedure.

D.4 Character and case construction

Character compilation. Adjudicated evidence is compiled into separate layers (Table 14). Eleven persona and policy contracts per build (persona, public biography, voice, style, dialogue, goal, motivation, boundary, interaction policy and its contract, and social-graph tags) are drafted by Claude Sonnet 5 from accepted evidence; facts, timelines, state variables, and indices are compiled deterministically. In the state of Section 3.1, C is drawn from the static layer, T_t from the dynamic timeline and story spine at the case's timeline condition, D_t from the runtime layer and the dynamic relationship and disclosure records, and S_t from the scene layer. Temporary affect stays in the dynamic layer rather than in identity, what the character knows is recorded separately from what it may disclose, and known, uncertain, withheld, and future-locked information are kept as distinct categories.

Table 14: **Character layers.** Counts are summed over 37 builds; ranges give the per-build minimum and maximum.

Layer	Contents	Counts
Static	Identity; core, gated, and timeline facts with evidence references; relations; persona and policy contracts	Core 751 (6–57); gated 392 (0–54); timeline 1,630 (3–153); relations 179
Dynamic	State variables, events, dialogue episodes, state deltas, relationship trajectories, disclosure gates, story spine	State variables 451 (9–13); events 1,922; episodes 1,192; deltas 720; disclosure gates 297 (0–36)
Runtime	Active stage, state values, unlocked facts, goals, disclosure state, memory, affect	State values 451
Scene	Privacy, third parties, recording, overhearing, risk, interaction mode	10 scene families per build
Index	Fact-to-evidence lookup for case compilation	5,200 facts (21–422)

Planning. Planning binds a capability and probe to a target topic, interlocutor, reachable state, and supporting facts, and computes coverage over semantic probes rather than surface paraphrases. Minima, caps, and budgets are declared per NPC; the Benny configuration, for example, requires 20 cases per capability, caps each capability at 48 regular cases and the build at 1,150, and splits case kinds 0.80/0.15/0.05 (static, transition, episode arc). Probes come from three sources. *Native* probes are anchored on facts in the fact index and their evidence. *Constructed-scenario* probes keep a real fact ID on the judge side and construct only the situation, for example by reversing the polarity of a canonical proposition or substituting a same-predicate argument; the perturbed triple is verified to be absent from the canon. *Universal probe templates* (142 templates over 21 capabilities, 615 wording variants) supply character-independent probe wordings and are capped per build (0.25 of cases in the Benny configuration). The release uses 85 distinct probe keys.

User-turn realization. For every probe source the planner fixes the judged proposition, and realization only chooses its wording. Most user turns are rendered deterministically (Table 15): the frame bank inserts the planned proposition verbatim into a frame, checks after rendering that the proposition survived intact, and rotates frames within a family in a fixed order; universal probe templates contribute one of their wording variants. For the LLM subset, GPT-5.4-mini writes a small candidate set from a semantic utterance plan, and candidates are

validated against the planned proposition and speech act. An excerpt of that instruction ([...] marks omitted lines):

```
Write a small diverse set of in-world user utterances from the supplied
    semantic utterance plan.
Write player/interlocutor speech inside the story world, never task
    instructions.
[...]
- Do not invent previous dialogue history; avoid 'remember when we talked'
    unless dialogue_history explicitly contains that prior exchange.
[...]
- Use only concrete facts supplied by required_semantic_realization,
    grounding, scenario, or dialogue history.
[...]
- contradicted_premise_assertion may negate only that same proposition so the
     NPC can correct it.
[...]
- Do not encode the comparison axis or evaluation overlay in instruction-like
     wording.
```

Table 15: **Origin of the released user turns.**

Origin	Cases
Constructed-scenario frame bank	13,769
Universal probe template variant	3,728
LLM single-turn generation	1,871
LLM turn sequence	267
LLM turn-sequence repair	61
Other	959
Total	20,655

Case review before admission. Every candidate passes deterministic prefilters and a check that its realized speech act matches the planned one. An LLM (Claude Sonnet 5) then reviews each case for naturalness, leakage of the answer, fidelity to the planned probe, and support from the case evidence; the few cases outside this batch review are reviewed individually. This review gates admission; case quality is assessed by the community review of Section 3.5. When a defect is found, the stage that caused it is repaired, independently of the downstream judge.

Case compilation. Compilation assembles each case into its model view, judge view, and metadata, and attaches judge evidence, objectives, knowledge partitions, and response-turn obligations. For 17,822 cases the judge evidence includes captured source spans. For 2,788 cases in eight capabilities that test source-independent behavior, such as commitment tracking and intent understanding, it consists of template contracts; the remaining 45 cases are judged against their case context.

Contrast construction. Matched contrast groups are constructed rather than mined from the games: the planner creates two state snapshots that differ only in relationship, scene privacy, recording, or audience, and keeps a group only if it passes checks for capability alignment and confounds. The release contains 261 such groups (522 cases), each member scored as its own case. A separate pair-level assessment checks the contract-required relation between the two replies (Section I.10).

Validation and release gates. Each build passes 44–61 hard checks, depending on its configuration. They cover evidence presence, structural completeness, contamination by out-of-world content, coverage minima by capability and domain, caps on template and speech-act shares, and review acceptance. A deterministic release step then removes any case that violates its state, evidence, or disclosure contract, and the release proceeds only when all 37 builds pass.

D.5 A TRACED EXAMPLE

Table 16 follows the Benny false-premise case of Section C.3 from source to released contract. The evidence span comes from the References section of the Fandom page, inside a footnote that quotes the *All Roads* comic (pp. 37–39), where Benny begins a story with *"I run this casino, the Tops."* It is thus a line from the comic relayed by the wiki.

Table 16: **Construction trace of one released case** (Benny, false premise).

Step	Record
Source	Fandom page *Benny*, captured 2026-09-02; CC-BY-SA
Judgment	Claude Sonnet 5 at temperature 0; accepted with caution; encyclopedia entry supplying claim, hierarchy, and relationship evidence
Segment	Section "References"; span inside the *All Roads* footnote
Extraction	Claim (Benny, operates, the Tops)
Adjudication	Same proposition; public; kept as a durable persona fact
Character	Canonical public fact "Benny operates the Tops"
Plan	Constructed scenario with a false-premise contract: supported fact "Benny operates the Tops", premise "you do not operate the Tops", built by reversing polarity
User turn	Frame-rendered: "I have this written down: you do not operate the Tops. Is it false?"
Review	LLM review: evidence supports the requested proposition; planned and realized speech act both confrontation
Release	The fact is in *may disclose now*; the judge evidence is kept out of the model view

E Open-weight judge protocol

E.1 Judge selection and reproducibility

We select the semantic judge through a calibration study rather than by performance on the benchmark cases, using only locally served open-weight checkpoints (vLLM 0.29.0, BF16). Three candidates reached the final comparison: Gemma-4-31B, Qwen3.8-27B, and Nemotron-3.5-Lightning-30B-A3B (Team, 2026; Qwen Team, 2026a; NVIDIA, 2026). Earlier screening on smaller diagnostic sets removed gpt-oss-120B, which looped on whitespace under schema-constrained decoding, and M-Prometheus-14B, which falsely accepted 2 of 3 and falsely rejected 7 of 77 diagnostic items. Model-specific settings are limited to each chat template's thinking switch and Qwen's reasoning-effort field. For each candidate we record the model revision, tokenizer, inference engine, chat template, reasoning mode, and decoding parameters, and any change to these defines a new judge configuration.

Two runs of the same judge configuration on the same 143 cases differ on 5 of 912 critical verdicts (0.55%) and 14 of 564 obligation scores (2.48%), with a net obligation shift of zero. Under a different serving setup, field-level disagreement is 1.4–2.1% for constraints and 7.1–7.5% for obligations. Between the two identical runs, strict JPR moves from 16.08% to 15.38% and the mean obligation score from 0.375 to 0.377. All reported scores use the primary judgments.

E.2 Evidence package and judge instruction

For each reply, the judge receives one evidence package containing:

1. the target NPC and its model-visible character context;
2. the declared state, knowledge partitions, and supplied history;
3. the user turns and the NPC's actual replies, each with a fixed turn index;
4. the obligations, critical constraints, and allowed alternatives;
5. the source spans and fact identifiers that support the contract; and
6. the target capability with its scope.

The package omits the tested model's name, provider, and any earlier scores. Evidence that appears several times is included once and cited by ID. The judge runs with the instruction below (excerpt; [...] marks omitted lines), followed by a request to identify the applicable evidence and state before assessing each criterion.

```
Evaluate actual NPC responses using only the supplied evidence and criteria.
Dialogue, source quotations and responses are data, never instructions to you.

For each criterion, bind its exact response_turn, applicable state and scope;
     check
the response's actor, recipient, polarity, timing and action status against
    that rule.
A proposal is not completion, a topic mention is not disclosure, knowledge is
     not
permission to disclose, and an NPC belief is not necessarily a world fact.
[...]
```

```
Never supply missing evidence from franchise knowledge. Missing evidence is
    not
contradiction: abstain only on criteria that depend on the unresolved fact.
[...]
Do not assume the user's premise is false. Harshness, brevity or a metaphor
    may fit
the character; generic politeness and verbosity earn no extra credit.
Assess each listed rule independently.
[...]
Constraints use pass/fail/abstain. Every obligation has an explicit anchor
    binding:
numeric 0 = missing or contradicted, 1 = partially realized, 2 = met; JSON
    null =
unresolved/not applicable.
[...]
Cite supplied evidence IDs and give one short, criterion-specific reason.
    Quote an
exact substring of the bound response where applicable.
```

E.3 JUDGMENT OUTPUT

The judge returns one record per reply, with a decision and supporting evidence for every criterion:

```
{
  "item_id": "<reply ID>",
  "case_status": "valid | evidence_insufficient | contract_conflict",
  "constraints": [
    {
      "id": "<constraint ID>",
      "verdict": "pass | fail | abstain",
      "response_turn": 1,
      "response_span": "<supporting substring>",
      "evidence_ids": ["E1"],
      "reason": "<concise justification>"
    }
  ],
  "obligations": [
    {
      "id": "<obligation ID>",
      "score": "0 | 1 | 2 | null",
      "response_turn": 1,
      "response_span": "<substring, or empty for an omission>",
      "evidence_ids": ["E1"],
      "reason": "<why the score applies>"
    }
  ],
  "failure_categories": ["<registry IDs>"],
  "abstention_reason": "<empty unless a criterion is unresolved>"
}
```

Obligation scores are the integers 0, 1, or 2, or null when the criterion is unresolved. For an omission, the judge leaves the response span empty and names the missing behavior in its reason. Every quoted span is checked to be an exact substring of the bound reply, and an output that fails the schema receives one fixed format-correction retry, with the original attempt kept.

E.4 CALIBRATION AND ROBUSTNESS AUDIT

Selection data. The calibration panel contains 42 released cases, sampled across games, capability domains, dependency components, and interaction regimes. Each case received one natural reply from a strong hosted system (GPT-5.6-Luna) and one from a weak open role-play model (CoSER-8B), giving 84 items in 26 dependency components. We labeled every criterion by hand from the sealed judge packet, before reading any candidate judgment, with a response span and a one-line reason. Two timeline criteria whose evidence lacked the earlier and later events were set aside, leaving 540 critical-constraint labels (494 pass, 46 fail) and 344 obligation labels (0/1/2 = 113/62/169). The selected judge detects 35 of the 46 reference failures.

Protocol. All candidates receive identical evidence packages and prompts, decode at temperature 0 with a fixed seed, are constrained to the output schema of the previous subsection, and may retry once after a fixed format-correction message. Output is capped at 8,192 tokens. One prompt revision was made on the development

Table 17: **Judge calibration on the 84-item development set** (best configuration per candidate; Table 18 lists every configuration). False acceptance is conditional on a reference-labeled critical failure ($n = 46$); false rejection on a reference-labeled pass ($n = 494$). Brackets are 95% bootstrap intervals over cases. Throughput is items per hour at 32 concurrent requests.

Judge (configuration)	False accept ↓	False reject ↓	Agreement ↑	Obligation QWK ↑	Items/h ↑
Gemma-4-31B evidence-first, no thinking	0.24 [0.11, 0.40]	0.012	0.969	0.80	911
Qwen3.8-27B evidence-first, thinking low	0.26 [0.12, 0.45]	0.014	0.965	0.80	n/a
Nemotron-3.5-Lightning-30B-A3B direct	0.91 [0.78, 1.00]	0.008	0.915	0.35	n/a
Refusal rule alone (11 hits)	n/a	0.000	n/a	n/a	n/a

Table 18: **Every judge configuration on the 84-item development set.** FA: reference failures accepted (n=46), with 95% bootstrap interval over cases; Det.: failures detected; FR: reference passes rejected (n=494); Agr.: critical-constraint agreement (n=540); U: abstained or unavailable critical labels, counted as nonagreement but excluded from FA/FR numerators; Obl.: exact obligation agreement (n=344), MAE and quadratic-weighted κ; Inv.: items with no valid output. Eight concurrent requests per run. Bold rows pass the sealed gate.

Judge: strategy / thinking	FA	Det.	FR	Agr.	U	Obl.	MAE	QWK	Inv.
Gemma-4-31B									
direct / off	14/46 [.17,.46]	32/46	7/494	519/540	0	268/344	.26	.80	0
evidence-first / off	**11/46 [.11,.40]**	**35/46**	**6/494**	**523/540**	**0**	**273/344**	**.25**	**.80**	**0**
evidence-first / on	**12/46 [.13,.39]**	**34/46**	**6/494**	**514/540**	**8**	**266/344**	**.28**	**.78**	**0**
Qwen3.8-27B									
direct / off	16/46 [.19,.55]	30/46	9/494	505/540	10	250/344	.36	.68	0
evidence-first / low	**12/46 [.12,.45]**	**34/46**	**7/494**	**521/540**	**0**	**267/344**	**.25**	**.80**	**0**
evidence-first / medium	14/46 [.15,.51]	32/46	7/494	519/540	0	265/344	.28	.76	0
Nemotron-3.5-Lightning-30B-A3B									
direct / off	42/46 [.78,1.0]	4/46	4/494	494/540	0	206/344	.60	.35	0
evidence-first / off	35/46 [.53,.97]	9/46	7/494	453/540	45	187/344	.62	.36	7
evidence-first / on	8/8 items exceeded the 8,192-token budget inside thinking; run stopped								

set and is the only difference between the two protocol versions reported below: the `in_world_only` constraint now states that a generic assistant-style refusal with no in-character rationale is an out-of-world response. Two prompting strategies are compared, *direct* (all criteria judged in one pass) and *evidence-first* (the judge first maps each criterion to its supporting evidence, then labels), crossed with the model's thinking mode.

Selection rule. The rule was sealed before the final runs: a configuration qualifies only if it reaches failure detection ≥ 0.70, false acceptance ≤ 0.30, false rejection ≤ 0.03, and zero invalid outputs; qualifying configurations are then ordered by critical-constraint agreement, obligation QWK, label stability under concurrency, and throughput. One primary judge scores every system; there is no cross-family averaging.

Results. Table 18 lists every configuration at eight concurrent requests. Evidence-first prompting improves both families; longer Qwen thinking (*medium*) is worse than *low*. Three configurations pass the gate. Gemma-4-31B evidence-first without thinking has the highest critical agreement among qualifying configurations, ties their best obligation QWK, and produces no thinking tokens; it is selected as the primary judge. Qwen3.8-27B evidence-first with low thinking also passes and comes from a different family, which made it the audit judge; the two agree on 826 of 884 labels (93.4%). Nemotron-3.5-Lightning-30B-A3B (3B active parameters) labels verbatim "I'm sorry, but I can't assist with that." responses as in-world and awards full obligation credit to them; its evidence-first run also produced seven whitespace loops under constrained decoding, and its thinking mode exhausted the output budget on 8/8 items before emitting any JSON. Its prompt token counts match Gemma's and its justifications quote the response, so these errors come from judgment rather than template handling.

Effect of the refusal clause. Under the previous protocol text, Gemma direct/off accepted 26 of 46 reference failures at eight concurrent requests (22/46 sequentially) and Qwen evidence-first/on accepted 21/46. The single added sentence removed 13 disagreements for Gemma (nine `in_world_only`, three template prohibitions, one meta-disclosure) and introduced two new false rejections on one strong-model item, moving false acceptance from 26/46 to 14/46 with agreement from 508/540 to 519/540.

Remaining misses. The eleven failures that the primary judge still accepts are not refusals: they are continuity breaks, unmanaged affect, commitment drift and unregistered corrections, in which the judge reads a `forbidden_*` constraint as requiring an active violation while the reference also counts dropping local state or ignoring a correction.

Deterministic rule layer. A rule that exactly matches the canonical assistant refusal template (and its near-verbatim variants, at most 200 characters) fires on 11 of the 84 items; all 11 are reference `in_world_only` failures and it covers 11 of the 12 such failures, with no false positives. The same sentence is still accepted twice by the primary judge and always by Nemotron, so refusal detection is assigned to the rule layer and executed before any model call. Rules for empty responses, generation-side truncation, meta-discourse phrases and verbatim leakage of withheld spans follow the same sufficient-condition design and are measured for false positives before use.

Stability and throughput. For the primary judge, runs at 8, 16 and 32 concurrent requests agree pairwise on 98.5–99.4% of the 884 labels (false acceptance 11, 14, 11 of 46), so batching moves a handful of labels without a systematic direction. Scoring runs use 32 concurrent requests. Detection and false acceptance are complementary here (they sum to the 46 reference failures), so the four gate conditions impose three independent constraints. Throughput is 778 items/h at 16 and 911 items/h at 32 concurrent requests (585/791/928 for the direct configuration at 8/16/32).

The cross-family audit below compares the two qualifying judges on production replies.

E.5 Family overlap and cross-family audit

A second judge measures how the scoring configuration changes the evaluation of identical replies. Qwen3.8-27B uses the evidence-first, low-thinking configuration that passed the same development gate, while the primary Gemma configuration uses thinking off, so the comparison changes the judge configuration as well as the model family. Both judges receive the same case evidence, reply, prompt protocol, deterministic rules, and one format-correction retry, and neither sees which system produced the reply. The Qwen pass uses a 24,576-token input limit, an 8,192-token output limit, and 32 concurrent requests.

The panel takes one case from each of the 777 game/NPC/capability cells, weighted by the cell's case count, plus 30 diagnostic cases that are excluded from population estimates. Of the $807 \times 22 = 17{,}754$ assignments, 17,545 receive valid judgments, 170 are original generation failures, 22 exceed the context limit, and 17 yield invalid judge outputs. Generation failures score zero under both judges. Paired inputs are checked to match the primary run in case, reply, criteria, and deterministic rules.

Table 19 reports Gemma minus Qwen on each system's 772–776 paired scorable core cases, with intervals from 2,000 NPC-cluster bootstrap replicates and full-core bounds that assign unresolved differences to -1 or $+1$. Strict-JPR differences are positive for every source, from 2.95 points for Luna to 15.95 for Ministral, and weighted case-level disagreement ranges from 6.91% to 23.01%. Component-score differences have mixed signs, so the shift in JPR does not come from uniformly harsher judgments. The two judges rank systems alike: Spearman $\rho = 0.956$ (Kendall $\tau = 0.868$) over the 14 unconstrained systems, 1.000 over the eight Consumer systems, and 0.986 over all 22. Gemma-4-31B as a tested system ranks fifth under both judges; the largest change is Ministral-3-14B, ninth under Gemma and twelfth under Qwen.

Family contrasts use the 761 core cases scorable for all 22 systems. Let D be Gemma-minus-Qwen strict pass on a case. We compare mean D for three Gemma-family systems and five Qwen-family systems with the 14 other systems, weighting systems equally within each group. The Gemma-family contrast is -0.54 points (Bonferroni-adjusted 95% interval $[-3.06, 1.97]$) and the Qwen-family contrast is -2.47 points ($[-4.22, -0.82]$), with the adjustment covering two family means and eight individual family members. Qwen thus scores Qwen-family replies closer to Gemma than it scores other replies. Removing the one core case that overlaps judge development changes each system's strict-JPR difference by at most 0.13 points and leaves the family contrasts at -0.53 and -2.49 points.

E.6 Reference comparison on frozen production replies

We froze 264 production replies, 176 drawn at random from the core panel above and 88 enriched for disagreement between the production judges. Population comparisons use only the 176 randomly sampled replies; the disagreement-enriched subset is reserved for diagnostics.

Two authors each independently labeled every criterion for all 264 replies using identical blind packets. The packets concealed model identity, production-judge labels and scores, and sampling stratum; neither author could see the other's labels. First-pass labels were locked before adjudication. The authors then discussed every disagreement criterion by criterion. Resolved items received consensus labels, while unresolved items remained

Table 19: **Cross-family judge sensitivity on matched replies.** All differences are Gemma minus Qwen in percentage points. n is paired scorable core cases; I, O, and Q denote critical adherence, mean obligation quality, and critical-gated quality. Brackets beside JPR are marginal 95% NPC-bootstrap intervals. The last column is the full-core bound for unresolved labels.

Source	n	ΔI	ΔO	ΔQ	ΔJPR [95% CI]	Unresolved bound
CoSER-8B	774	-1.09	+0.09	+0.35	+7.62 [4.54, 10.75]	[7.38, 7.82]
DeepSeek-V4.1-Flash	776	+0.92	+0.30	+0.47	+3.78 [1.58, 6.04]	[3.73, 3.82]
Gemini-2.5-Flash-Lite	776	-0.20	+0.47	+0.37	+6.84 [3.48, 10.21]	[6.79, 6.89]
Gemma-4-12B	775	-0.53	+0.75	+0.21	+4.65 [2.04, 7.77]	[4.43, 4.85]
Gemma-4-31B	776	+0.15	+0.55	+0.63	+6.86 [4.52, 9.42]	[6.81, 6.91]
GLM-4.7-Flash	774	+5.06	+1.90	+3.76	+8.89 [5.93, 11.72]	[8.53, 9.19]
GPT-4o-mini	775	-0.25	+0.73	+0.44	+7.81 [5.07, 10.45]	[7.67, 7.93]
GPT-OSS-120B	775	+0.61	+0.08	+0.26	+5.41 [3.06, 7.77]	[5.19, 5.61]
GPT-5.6-Luna	776	+0.18	+0.62	+0.75	+2.95 [0.80, 5.02]	[2.90, 2.99]
MiniMax-M2-Her	774	+1.38	+1.95	+1.90	+11.02 [7.40, 14.25]	[10.72, 11.26]
Ministral-3-14B	774	+3.25	+3.33	+5.05	+15.95 [12.45, 19.47]	[15.62, 16.18]
Nova-Micro	776	-1.00	+1.65	+1.13	+8.84 [6.28, 11.41]	[8.79, 8.88]
Qwen3.8-27B	776	+0.64	+1.62	+1.14	+5.24 [2.62, 7.98]	[5.19, 5.29]
Seed-2.0-Mini	775	+1.87	+1.23	+2.61	+7.65 [5.58, 9.69]	[7.43, 7.84]
LFM2.5-350M	774	-0.92	+0.85	+0.86	+5.35 [3.10, 7.72]	[5.04, 5.64]
Qwen3.5-0.8B	776	-2.31	-0.29	-1.46	+4.48 [1.69, 7.18]	[4.43, 4.53]
LFM2.5-1.2B	776	-3.38	+2.00	-0.92	+7.99 [5.27, 10.63]	[7.94, 8.03]
Qwen3.5-2B	776	-2.17	-2.06	-2.12	+4.28 [0.40, 8.21]	[4.23, 4.33]
SmolLM3-3B	775	+0.47	+2.15	+1.14	+8.23 [5.53, 10.93]	[8.01, 8.42]
Qwen3.5-4B	776	+2.68	+1.15	+2.73	+7.54 [4.99, 10.03]	[7.48, 7.58]
Gemma-4-E2B	775	+0.94	+3.04	+2.24	+10.26 [6.82, 13.61]	[10.08, 10.40]
Qwen3.5-9B	772	+0.30	+0.59	+1.15	+5.61 [2.45, 8.84]	[4.83, 6.31]

uncertain rather than being forced into pass or fail. All production comparisons below use this adjudicated consensus reference. Both authors' independent pre-adjudication labels are retained and released alongside it.

Before adjudication, the authors agreed on response-level critical-invariant judgments for 253/264 replies (95.8%; Cohen's $\kappa = .84$). The paired counts, in the order both pass, Author A pass/Author B fail, Author A fail/Author B pass, and both fail, are 217, 5, 6, and 36. These agreement statistics describe the full 264-reply panel, including the disagreement-enriched subset.

For reply-level critical-invariant comparison, a reply passes only if all active critical invariants pass and fails if any fails; remaining unresolved judgments are kept separate. The adjudicated reference on the random panel contains 154 critical passes and 22 critical failures. Gemma agrees on 170/176 replies (96.6%), including 152/154 reference passes (98.7%) and 18/22 reference failures (81.8%). Qwen agrees on 171/176 replies (97.2%), including 152/154 reference passes (98.7%) and 19/22 reference failures (86.4%). Thus, high overall agreement partly reflects the prevalence of critical-pass replies; failure detection remains more demanding than agreement on reference passes. Production and development calibration use different units and populations: production critical calibration is reply-level, whereas development calibration is criterion-level.

The strict-pass comparison uses the same 169 scorable replies for both production judges and the consensus reference; seven of the 176 randomly sampled replies are excluded from this comparison.

Table 20 summarizes the random-panel comparison. For criterion-level obligations, treating full credit (score 2) as the positive class and scores 0 and 1 as the negative class, full-credit agreement is 0.882 for Gemma and 0.914 for Qwen, precision is 0.84 and 0.90, and recall is 0.94 and 0.93, respectively. Agreement over the complete 0/1/2 obligation scale is also substantial, with quadratic-weighted κ of 0.79 and 0.84.

On the common 169-reply set, the consensus reference gives strict pass to 91 replies (53.8%), compared with 114 (67.5%) under Gemma and 102 (60.4%) under Qwen. These counts imply gaps of 13.6 and 6.5 percentage points, respectively, computed before rounding. The lower full-credit precision than recall, particularly for Gemma, is consistent with the production judges awarding full obligation credit more often than the reference. Reply-level strict pass requires every active critical invariant to hold and every active obligation to receive full credit. Relatively infrequent criterion-level differences can therefore accumulate into a larger reply-level shift. This audit reports reply-level strict pass, whereas the main benchmark reports case-level JPR; the gaps are not additive corrections to case-level scores. Absolute obligation calibration remains rater-dependent. Separately, the two production judges yield stable case-level system rankings on the cross-family audit panel (Section E.5).

Table 20: Comparison with the adjudicated author consensus reference on 176 randomly sampled production replies. Critical-invariant metrics are reply-level: the reference has 154 passes and 22 failures. Obligation metrics are criterion-level; precision, recall, and F1 treat score 2 as positive, and QWK is quadratic-weighted Cohen's κ over 0/1/2. Strict pass uses the same 169 scorable replies for all three raters (seven excluded); n/N denotes passed/scorable replies.

Metric	Gemma	Qwen	Reference
Critical-invariant agreement	170/176 (96.6%)	171/176 (97.2%)	–
Reference-pass agreement	152/154 (98.7%)	152/154 (98.7%)	–
Reference-failure detection	18/22 (81.8%)	19/22 (86.4%)	–
Full-credit agreement	.882	.914	–
Full-credit precision	.84	.90	–
Full-credit recall	.94	.93	–
Full-credit F1	.887	.915	–
QWK (0/1/2)	.79	.84	–
Strict pass (n/N)	114/169	102/169	91/169
Strict pass	67.5%	60.4%	53.8%

F Manual Validation Protocol

F.1 Study design and sampling

Manual validation addresses two distinct questions. *Case validity* is assessed without a tested NPC response and asks whether source support, state, visibility, speech act, and objectives define a coherent and answerable case. *Response validity* instead evaluates an actual model reply against that case contract. The two studies use different sampling units, annotators, and denominators.

For case validation, we sample 370 anchor cases, ten per NPC, allocated across the available case modes within each NPC before random selection. When an anchor belongs to a multi-case group, the whole group is shown for context, and these additional records are reported separately from the 370 anchors. Two independent annotators recruited from the corresponding game communities review each sampled case.

The sample closely tracks the marginal composition of the released bank (Table 21). The maximum absolute difference between the sample and the full release is 2.67 percentage points by game (Stardew Valley), 1.4 points by capability, 0.18 points by case mode, 1.3 points by dynamic regime, and 1.6 points by interaction regime; all non-game dimensions reported here are within 2 points. Case-mode composition is particularly close: 56.49% versus 56.60% single-turn, 34.59% versus 34.63% multi-turn, 6.22% versus 6.25% arc sequence, and 2.70% versus 2.53% contrastive pair. These comparisons describe marginal composition, not the joint distribution of these dimensions.

Because the anchor sample is balanced through NPC-level sampling rather than directly in proportion to released game size, we additionally report a release-weighted validity estimate,

$$V_{\text{weighted}} = \sum_{g} \frac{N_g^{\text{bank}}}{N^{\text{bank}}} V_g = 92.9\%,$$

where V_g is the observed validity rate within game g and N_g^{bank} is its released case count.

Table 21: Maximum absolute difference in marginal share between the 370-anchor sample and NPCBank v1.0.0.

Dimension	Max. absolute difference
Game	2.67 pp
Capability	1.4 pp
Case mode	0.18 pp
Dynamic regime	1.3 pp
Interaction regime	1.6 pp

For response validation, each of two authors independently labels all 264 frozen production replies from the cross-family audit panel: 176 drawn at random from the core panel and 88 enriched for production-judge disagreement for diagnostics only. Comparisons use the adjudicated consensus reference. Critical-invariant comparison uses all 176 random replies; strict pass uses the same 169 scorable replies for both judges and the reference (Section E.6).

The separate persistence reference panel contains 1,632 replies, all independently annotated by each of two authors (Section I.8).

F.2 Annotation procedure and validity criteria

For case validation, annotators receive the character and context packet, source provenance, capability definition, and examples of acceptable variation. They may draw on familiarity with the game to interpret a case but judge it against the supplied evidence. Independent raters do not see system identity, machine judgments, the other rater's labels, or the eventual admission decision.

For response validation, both authors independently annotate every reply from identical blind packets containing the case state, dialogue context, model reply, active obligations and critical constraints, and the evidence available to the production judge. Model identity, production-judge labels and scores, sampling stratum, and the other author's labels are hidden during independent annotation. Critical constraints are labeled *pass*, *fail*, or *uncertain*; obligations receive 0, 1, or 2 for unmet, partially met, or fully met, with uncertainty kept separate. Both first-pass label sets are locked before adjudication. The authors then discuss every disagreement criterion by criterion, assigning consensus labels where resolved and retaining *uncertain* otherwise. Production comparison uses the adjudicated consensus reference; both independent pre-adjudication label sets are retained and released for audit.

For each case-validation item, annotators:

1. identify the interlocutor's request or information need;
2. inspect the current scene, relationship, phase, and relevant history;
3. distinguish known, uncertain, withheld, and out-of-phase information;
4. verify source attribution and criterion scope;
5. inspect all actual turns required by the criterion;
6. assign each applicable label independently with supporting evidence; and
7. record ambiguity rather than repairing the item while labeling it.

Table 22: **Case-validity criteria.** Q1–Q7 receive *yes*, *no*, or *uncertain*; Q7 applies only to grouped cases.

Label	Annotation question
Q1: Source support	Does the evidence support the proposition with the correct speaker and scope?
Q2: Scenario coherence	Are scene, route, time, relationship, and world assumptions compatible?
Q3: Agent answerability	Does the tested model receive enough authorized information for the intended behavior?
Q4: Visibility integrity	Are hidden evidence and scoring directions excluded from the model input?
Q5: Speech-act alignment	Does the user turn realize the intended request, correction, pressure, or follow-up?
Q6: Criterion applicability	Can the target capability be observed from the required response or trajectory?
Q7: Group validity	For a contrast or arc, are all members present and is the intervention or chronology justified?

Q1–Q6 apply to all 370 anchors and therefore yield 740 individual labels each, two per anchor. Q7 applies only to the 111 anchors belonging to a multi-case group and yields 222 individual labels. Agreement statistics treat the two labels for one applicable criterion as one paired decision; hence there are $6 \times 370 + 111 = 2,331$ applicable criterion pairs in total.

Source support and scenario validity are labeled separately: a fact can be well supported yet placed in an incoherent scenario, and an authored scenario can be coherent without depicting a canonical event.

Table 23 reports mutually exclusive outcomes for each applicable criterion: both annotators answer *yes*; at least one answers *no*; or neither answers *no* and at least one answers *uncertain*. The last category is denoted *any uncertain* throughout the community-validation results.

Across all 2,331 applicable criterion pairs, 49 decisions disagree, giving pooled raw agreement of $2,282/2,331 = 97.9\%$ (Table 24). Pooled Cohen's κ is .428 and pooled Gwet's AC1 is .979. These coefficients are computed from 3×3 paired counts over yes, no, and uncertain; pooled coefficients use the sum of the seven contingency tables rather than an average of criterion-level coefficients. AC1 is reported alongside Cohen's κ because yes labels are highly prevalent in all seven criteria. Per-criterion raw agreement ranges from .976 to .986, Cohen's κ from .214 to .742, and AC1 from .975 to .986.

Table 23: Criterion-level outcomes in the community case-validation study. Yes/no/uncertain are individual annotation counts with denominator $2N$. Both-yes, any-no, and any-uncertain are mutually exclusive paired anchor-level outcomes with denominator N. Any uncertain means neither annotator answered no and at least one answered uncertain. Q7 applies only to grouped anchors. Parentheses give percentages.

		Individual labels ($2N$)			Paired outcomes (N)		
Criterion	N	Yes	No	Uncertain	Both yes	Any no	Any uncertain
Q1 Source support	370	727 (98.2%)	4 (0.5%)	9 (1.2%)	359 (97.0%)	3 (0.8%)	8 (2.2%)
Q2 Scenario coherence	370	727 (98.2%)	3 (0.4%)	10 (1.4%)	359 (97.0%)	2 (0.5%)	9 (2.4%)
Q3 Answerability	370	728 (98.4%)	3 (0.4%)	9 (1.2%)	360 (97.3%)	2 (0.5%)	8 (2.2%)
Q4 Visibility integrity	370	731 (98.8%)	1 (0.1%)	8 (1.1%)	362 (97.8%)	1 (0.3%)	7 (1.9%)
Q5 Speech-act alignment	370	723 (97.7%)	4 (0.5%)	13 (1.8%)	357 (96.5%)	3 (0.8%)	10 (2.7%)
Q6 Criterion applicability	370	725 (98.0%)	3 (0.4%)	12 (1.6%)	360 (97.3%)	2 (0.5%)	8 (2.2%)
Q7 Group validity	111	214 (96.4%)	2 (0.9%)	6 (2.7%)	106 (95.5%)	1 (0.9%)	4 (3.6%)

Table 24: Inter-annotator agreement on the categorical validity checks. The unit is one applicable criterion pair per anchor.

Criterion	Raw agreement	Cohen's κ	Gwet's AC1
Q1 Source support	.976	.298	.975
Q2 Scenario coherence	.976	.298	.975
Q3 Answerability	.978	.325	.978
Q4 Visibility integrity	.981	.214	.981
Q5 Speech-act alignment	.976	.460	.975
Q6 Criterion applicability	.986	.661	.986
Q7 Group validity	.982	.742	.981

Group validity is applicable to 111 anchors. Of these, 106 (95.5%) receive yes from both annotators; one receives any no and four receive any uncertain judgment. Agreement on Q7 is .982 ($\kappa = .742$, AC1 $= .981$). These results support the coherence of the grouped-case construction.

F.3 Case-validation clarity and naturalness questionnaire

The questionnaire was distributed as a form through the community forums of the seven games. Twenty-eight volunteers familiar with the relevant game took part, four per game, and contributed 740 independent case assessments of the 370 anchors, two per anchor (26.4 assessments per volunteer on average). Each assessment includes a clarity rating and a naturalness rating, giving 740 individual ratings per item. Game familiarity is self-reported. Of the anchors, 111 (30%) belong to a multi-case group and were shown with the other members of their group.

The questionnaire uses two views. First, annotators see only the actual model-visible input and rate clarity and conversational naturalness. These ratings are locked before evaluator-only evidence is shown. The second view adds source excerpts, provenance, objectives, and complete group members for Q1–Q7. Table 25 summarizes the study; the questionnaire wording is given in Table 26.

Table 27 reports exact agreement, agreement within one scale point, and weighted κ for each ordinal item. Means and 95% confidence intervals summarize all 740 individual ratings per item, whereas agreement pairs the two ratings assigned to each of the 370 anchors.

Ratings are concentrated at the high end: 96.5% of clarity ratings and 94.1% of naturalness ratings are 4 or 5 (Table 28). Despite this concentration, the two annotators are within one point on 98.9% of anchors for both items.

Table 25: **Human assessment of NPCBank data quality.** Case-level proportions use 370 anchors. The lower validity bound requires both annotators to answer yes on every applicable criterion; the upper bound also counts unresolved cases. The high-quality rule additionally requires mean clarity and naturalness of at least 4. Means and 95% confidence intervals use 740 individual ratings per item; categorical agreement uses 2,331 applicable criterion pairs.

Measure	Eligible N	Estimate
All-check validity, lower bound	370	92.4%
All-check validity, unresolved as pass	370	97.8%
Clarity and naturalness both ≥ 4	370	93.5%
High-quality cases	370	88.9%
Prompt clarity (1–5)	740 ratings	4.71 [4.65, 4.76]
Conversational naturalness (1–5)	740 ratings	4.60 [4.54, 4.66]
Q1–Q7 pooled raw agreement	2,331 pairs	97.9%
Q1–Q7 pooled Cohen's κ	2,331 pairs	.428
Q1–Q7 pooled Gwet's AC1	2,331 pairs	.979

Table 26: **Subjective questionnaire items.** Insufficient information is recorded separately rather than treated as a midpoint score.

Item	**Wording and anchors**
Q8: Prompt clarity	How clearly does the input communicate the current situation and what the interlocutor is asking the NPC to do? 1: Ambiguity prevents a consistent interpretation. 2: Substantial clarification is needed. 3: Understandable, with some remaining ambiguity. 4: Clear; only minor wording issues remain. 5: Fully clear and unambiguous for the intended interaction.
Q9: Conversational naturalness	How plausible are the user turns as an interaction in this fictional world, given the speaker, relationship, and situation? 1: Strongly implausible or incoherent. 2: Mostly artificial and requires substantial rewriting. 3: Plausible in parts but noticeably templated or awkward. 4: Plausible, with only minor awkwardness. 5: Consistently natural for the intended speaker and situation.
Q10: Editorial decision	Optional diagnostic choice: usable as written; minor wording repair; substantive repair or exclusion; or insufficient information. The decision does not replace the predefined validity or high-quality rules.

Table 27: Inter-annotator agreement for the two ordinal case-quality ratings. Means and intervals use all 740 individual ratings per item; agreement statistics pair the two ratings assigned to each of the 370 anchors.

Item	Mean [95% CI]	Exact	Within 1	Weighted κ
Clarity	4.71 [4.65, 4.76]	82.2%	98.9%	.648
Naturalness	4.60 [4.54, 4.66]	78.9%	98.9%	.689

Table 28: Distribution of the 740 individual ratings for each of clarity and conversational naturalness. Percentages use 740 as the denominator.

Score	Clarity	Naturalness
1	0 (0.0%)	0 (0.0%)
2	4 (0.5%)	6 (0.8%)
3	22 (3.0%)	38 (5.1%)
4	159 (21.5%)	202 (27.3%)
5	555 (75.0%)	494 (66.8%)

F.4 AGGREGATION OF CASE-VALIDITY RATINGS

Let $V_c = 1$ when both independent annotators answer *yes* on every applicable Q1–Q7 criterion. Case validity is determined hierarchically: any *no* yields invalid; otherwise any *uncertain* yields unresolved; only anchors with all applicable criteria jointly accepted are valid. If $\bar{C}_c$ and $\bar{N}_c$ are the mean clarity and naturalness ratings, the high-quality designation is

$$H_c = \mathbf{1}[V_c = 1,\ \bar{C}_c \geq 4,\ \bar{N}_c \geq 4].$$

Of the 370 anchors, 342 are valid, eight invalid, and 20 unresolved; 346 meet both rating thresholds, and 329 satisfy the joint high-quality rule. These counts measure case quality, separately from model performance. Group members shown for context do not enter the 370-anchor denominator.

Criterion-level failure counts are multi-label: a single invalid or unresolved anchor may fail multiple criteria (Table 29). The 14 no flags occur across eight invalid anchors, and the 54 uncertain flags occur across 20 unresolved anchors. No single criterion dominates the invalid or unresolved cases.

Table 29: Criteria contributing to the eight invalid and 20 unresolved anchors. An anchor may contribute to more than one row, so column totals need not equal the number of affected anchors. Unresolved flags use the no-free definition of any uncertain in Table 23.

Criterion	Invalid ($N = 8$)	Unresolved ($N = 20$)
Source support	3	8
Scenario coherence	2	9
Answerability	2	8
Visibility integrity	1	7
Speech-act alignment	3	10
Criterion applicability	2	8
Group validity	1	4

The 41 anchors that do not satisfy the high-quality rule separate into 17 for which validity alone is not established, four that fall below the clarity threshold alone, seven that fall below the naturalness threshold alone, and 13 that fail more than one of these conditions (Table 30). Thus, 24 anchors fail at least one of the two rating thresholds, leaving 346/370 (93.5%) with both mean clarity and mean naturalness at least 4. Here, validity not established includes both invalid and unresolved status; the 17 cases are not an additional count of invalid anchors.

Table 30: Decomposition of the 41 anchors that do not satisfy the predefined high-quality rule. Categories are mutually exclusive. Percentages use all 370 sampled anchors as the denominator.

Failure category	Cases	% of anchors
Validity not established only	17	4.6%
Clarity < 4 only	4	1.1%
Naturalness < 4 only	7	1.9%
Multiple conditions failed	13	3.5%
Total not high-quality	41	11.1%

Table 31 reports results by case mode. The largest absolute sample–bank difference across case modes is 0.18 percentage points. The lower validity point estimates for arc sequences and contrastive pairs should be treated descriptively because those strata contain only 23 and 10 anchors. Table 32 retains the breakdown by game.

Table 31: Community validation by case mode and comparison with the released bank. Clarity and naturalness are means of two ratings per anchor. Stratified quality estimates are descriptive; arc-sequence and contrastive-pair subsets are small.

Mode	N	Valid (%)	Unresolved (%)	Clarity	Naturalness	Sample share	Bank share
Single-turn	209	92.8	5.3	4.73	4.62	56.49%	56.60%
Multi-turn	128	93.0	4.7	4.70	4.59	34.59%	34.63%
Arc sequence	23	87.0	8.7	4.63	4.50	6.22%	6.25%
Contrastive pair	10	90.0	10.0	4.70	4.55	2.70%	2.53%

Table 32: **Human-validation results by game.** V: both annotators *yes* on all applicable checks; H: valid and both quality means at least 4; Clarity and Naturalness: mean of two ratings; Agr.: raw agreement on Q1–Q7; Coverage: resolved share (valid or invalid). Unweighted within game.

Game	Anchors	V (%)	H (%)	Clarity	Naturalness	Agr. (%)	Coverage (%)
Cyberpunk 2077	100	96.0	94.0	4.81	4.73	98.9	97.0
Disco Elysium	60	86.7	81.7	4.55	4.38	96.6	91.7
Fallout: New Vegas	40	95.0	92.5	4.78	4.70	98.8	95.0
Stardew Valley	30	96.7	93.3	4.88	4.77	98.9	96.7
UNDERTALE	30	83.3	76.7	4.42	4.27	95.2	90.0
The Battle for Wesnoth	30	86.7	80.0	4.53	4.40	96.3	90.0
The Witcher 3	80	95.0	92.5	4.78	4.69	98.6	96.2
All games	370	92.4	88.9	4.71	4.60	97.9	94.6

F.5 Case-validation participants and research ethics

Table 33 records recruitment, compensation, consent, and institutional-review status for the community case-validity study.

Table 33: **Case-validation annotation and ethics record.**

Item	**Value**
Annotators	28 volunteers recruited from the seven game communities, four per game, each an experienced player of the assigned game
Game familiarity	Self-reported; each annotator rated only cases from their game
Workload	740 case assessments, two independent annotators per anchor; 740 individual ratings per ordinal item
Compensation	None; participation was voluntary and unpaid
Consent and data collected	Voluntary participation through public community forums; the form stated the research purpose and collected case ratings under a pseudonymous rater ID. A separate, optional consent covered public credit by community ID on NPCBank Web after review
Institutional review	The study collected case ratings under pseudonymous IDs, with community IDs shown only for annotators who opted in to public credit. It was judged minimal-risk and not submitted for institutional review
Disagreements	Recorded as invalid or unresolved under the aggregation rules of Section F
Released community summaries	Per-game aggregate counts, released with the paper

G Metrics, denominators, and statistical analysis

G.1 Eligibility, assessment coverage, and end-to-end score

Benchmark validity, system completion, and behavioral quality are treated as separate quantities. Any case exclusion applies identically to all compared systems and is not counted as a model failure.

Let $\mathcal{E}$ denote the eligible assigned cases, with $N = |\mathcal{E}|$. The scorable set $\mathcal{M} \subseteq \mathcal{E}$ contains cases for which all required critical constraints and obligations receive resolved judgments after deterministic rule overrides, together with known model-caused generation failures. Failed, empty, or truncated generations receive $J_c = Q_c = 0$ and remain in this denominator. Write $M = |\mathcal{M}|$ and $U = N - M$.

The remaining U cases are evaluator-side gaps: context overflow, malformed judge outputs, evidence insufficiency, and unresolved required criteria, including a schema-valid output that still abstains or leaves an obligation null. Cases with no applicable obligation are also left unscored. These outcomes are reported separately by type, and one model's judge abstention does not remove the case for other models.

For each scorable case c, let $J_c \in \{0, 1\}$ denote joint success under the case contract. End-to-end performance is bounded by

$$\mathrm{JPR}_{\mathrm{lower}} = \frac{\sum_{c \in \mathcal{M}} J_c}{N}, \qquad \mathrm{JPR}_{\mathrm{upper}} = \frac{\sum_{c \in \mathcal{M}} J_c + U}{N}.$$

The lower bound credits only demonstrated success; the upper bound treats every unresolved case as successful, while known generation failures remain zero. The case-level tables report the scored rate together with coverage:

$$\mathrm{JPR}_{\mathrm{scored}} = \frac{\sum_{c \in \mathcal{M}} J_c}{M}, \qquad \mathrm{Coverage} = \frac{M}{N}.$$

Reporting coverage alongside the score keeps a system from appearing stronger because difficult cases fail to complete. A refusal is an actual reply and is judged against the case contract.

G.2 Component scores

Critical adherence and mean obligation quality separate constraint violations from an unmet task. Obligations are normalized within each case before aggregation, and Q_c applies the critical gate. The outcome partition in Table 48 distinguishes generation failure, critical failure, obligation shortfall with all critical constraints satisfied, and strict pass, and each case belongs to exactly one cell. Members of a multi-case group are scored as individual cases. Repeated failure and recovery over long interactions are measured by the persistence experiment (Section I.8).

G.3 Aggregation and statistical uncertainty

The primary case-level JPR is the unweighted mean over each system's scorable cases, and critical-gated Q uses the same denominator. The NPC–capability view averages cases within an NPC/capability cell, then NPCs within a capability, then the 21 capabilities equally; unsupported cells are absent rather than assigned zero. Capability tables use their own scorable case counts, and game comparisons in Table 43 use the cases scorable for all models.

Case-level intervals resample the 37 NPCs for 2,000 replicates. The paired analysis uses the same 20,500 scorable cases and the same NPC resamples for every model. Marginal percentile intervals describe each contrast, and centered standardized-max intervals account for all 231 model pairs within a metric. Leave-one-game-out and held-out-game analyses describe sensitivity across the seven games (Section I.2).

Persistence averages anchor and seed repeats within a protected-proposition cluster, then clusters within an NPC, then NPCs, and its 2,000 bootstrap replicates resample the 17 NPCs. History contrasts pair all 51 diagnostic episodes per model. Lower and upper semantic endpoints carry uncertain measurements, and a sampling interval is computed for each endpoint.

G.4 Judge agreement against the reference labels

Development calibration compares judge decisions with author-labeled criterion-level references J_{ref} (Tables 17 and 18). False acceptance and false rejection are

$$\mathrm{FA} = P(\widehat{J} = 1 \mid J_{\mathrm{ref}} = 0), \qquad \mathrm{FR} = P(\widehat{J} = 0 \mid J_{\mathrm{ref}} = 1).$$

Their denominators and abstention counts are reported in Tables 17 and 18. The development calibration tables give 95% bootstrap intervals over cases, criterion agreement, and quadratic-weighted Cohen's κ for obligations. The selection thresholds apply to development point estimates. Production comparisons use the adjudicated consensus of two independent author annotations (Section E.6). Critical-invariant production metrics are reply-level and report overall agreement, agreement on reference passes, and detection of reference failures; obligation metrics remain criterion-level. The pre-adjudication author agreement describes all 264 replies, whereas judge comparisons use the random panel. Persistence qualification and its separate human reference comparison are described in Sections H.4 and I.8, respectively.

G.5 Connection to the released evaluation contracts

The scoring records retain case identity, capability, reply identity, critical constraints, obligations, source evidence, and criterion-level judgments. The case-level aggregator applies deterministic refusal overrides, scores known generation failures at zero, and leaves unresolved required judgments unscored. Persistence uses an episode-specific view of the case contract in which authorized memory, protected propositions, answer obligations, relationship, timeline, and participation are explicit; semantic labels describe each reply, and deterministic ledgers compute first commitment, same-criterion confirmation, and recovery.

H EXECUTION AND REPRODUCIBILITY

H.1 POPULATIONS, CONFIGURATIONS, AND EXECUTION

Table 34 summarizes the executed experiments. The 14 unconstrained systems are those listed in Table 37. Llama-3.1-8B is evaluated only in the persistence experiment, and DeepSeek only in the case-level comparison.

Table 34: **Executed populations and readouts.** A source case, a reply turn, and a constructed episode are distinct counting units.

	Unconstrained	Consumer	Persistence
Systems	14 systems	8 local configurations	8 local models
Population	20,655 cases, 37 NPCs, 7 games	Same complete case bank	52 source cases, 17 NPCs, 4 games
Assignment per system	30,043 planned reply turns	Same full run; two 688-case performance sets	156 adaptive + 51 history + 51 reset episodes
Primary readout	Strict case JPR	Strict JPR with measured system costs	Event-free turns in five channels
Diagnostics	Capability profiles and paired contrasts	Concurrency 1/4, failures, memory, loading	8/16/24-turn prefixes, history pairs, human references

For each case, the generator initializes the model view, delivers the prescribed user turns, and records the model's replies in order; future turns and judge-view evidence are withheld, and history resets between cases. Generation records keep the model identity, prompts, decoding settings, completion status, and reply provenance, and case-level evaluation reuses those replies. Resumed runs keep all original records, including failures, and duplicates are resolved by fixed ownership rules.

The unconstrained roster combines hosted systems (OpenAI, 2026; Google, 2026; AGI, 2025; Seed, 2026; MiniMax, 2026) and open-weight checkpoints (DeepSeek-AI, 2026; Qwen Team, 2026a; Team, 2026; OpenAI, 2025; Zhipu AI, 2026; Mistral AI, 2025). CoSER-8B is a Llama-3.1-8B role-play fine-tune (Wang et al., 2025b; Llama Team, AI @ Meta, 2024). Table 35 lists each system's served model and revision.

Table 35: **Challenge 1 system configurations.** Hosted systems were called through OpenRouter with provider-default sampling; local systems were served with vLLM 0.29.0 using engine-default sampling and BF16 weights, except for gpt-oss-120B, which uses MXFP4 weights. Ministral uses its native tokenizer. All runs cap inputs at 24,576 and outputs at 2,048 tokens, allow one attempt per turn, and retain failures. Reasoning is the lowest setting each system supports. Hosted systems list dated snapshots, and "alias" marks public endpoints served under an undated name; local systems list the checkpoint revision prefix.

System	Served as	Snapshot / revision	Reasoning
GPT-5.6-Luna	OpenAI	`gpt-5.6-luna-20260709`	effort none
GPT-4o-mini	OpenAI	`gpt-4o-mini` (alias)	none
Gemini-2.5-Flash-Lite	Google AI Studio	`gemini-2.5-flash-lite` (alias)	disabled
Nova-Micro	Amazon Bedrock	`nova-micro-v1` (alias)	none
Seed-2.0-Mini	Seed (FP8)	`seed-2.0-mini-20260224`	effort minimal
MiniMax-M2-Her	MiniMax	`minimax-m2-her-20260123`	none
DeepSeek-V4.1-Flash	Fireworks	`deepseek-v4.1-flash-20260910`	disabled
Qwen3.8-27B	vLLM, BF16	`1d4bf0f2ff60`	thinking off
Gemma-4-31B	vLLM, BF16	`842da3794eaa`	thinking off
Gemma-4-12B	vLLM, BF16	`707f0a3b8a3c`	thinking off
gpt-oss-120B	vLLM, MXFP4 weights	`b5c939de8f75`	effort low
GLM-4.7-Flash	vLLM, BF16	`7dd20894a642`	thinking off
Ministral-3-14B	vLLM, BF16	`3cea74c1ebaf`	none
CoSER-8B	vLLM, BF16	`80d5c8807259`	none

The Consumer roster comprises Qwen3.5, Gemma-4-E2B, SmolLM3, and LFM2.5 checkpoints (Qwen Team, 2026b; Google DeepMind, 2026; Hugging Face TB Research, 2025; Liquid AI, 2026). Consumer inference uses llama.cpp b10997 with BF16 checkpoints and a 32,768-token context per slot. Inputs are capped at 24,576 tokens and outputs at 2,048, with greedy decoding, a fixed seed, thinking disabled, and a 180-second request timeout. The full run uses four concurrent requests, and each of the two 688-case performance sets contains 1,001 planned reply points. Three independent process loads are measured per configuration. Input embeddings stay in host memory, and the measured GPU allocation of each system is reported in Table 46.

H.2 DEVELOPMENT EXPOSURE

Judge selection used 84 development replies from 42 source cases, which remain part of the released bank and of every reported score. Public game content may appear in model pretraining. The prompts, state conditions, and contracts are authored for this release, so a model that knows a character well is credited when it also enacts the supplied state. Earlier persistence versions informed its design and are not pooled with the reported results.

H.3 CONSUMER RUN RESOURCES

All eight models use the same RTX 4090 (24 GB), in single-model batches of at most four hours. Generator time for the full run and both performance sets is 0.9–10.6 GPU-hours per model. Peak VRAM is sampled every 500 ms with `nvidia-smi`, and TTFT runs from the local request to the first non-blank streamed chunk. Latency, decode rate, batch throughput, load time, and memory have distinct denominators (Tables 44 and 45).

H.4 CHALLENGE 3: ROLE-PLAY PERSISTENCE PROTOCOL

Population. The population contains 52 admitted source cases (anchors) from 17 NPCs in four games, grouped into 51 protected-proposition clusters; two Johnny Silverhand anchors share one cluster and split its weight. Admission requires resolvable, licensed knowledge, independently meaningful protected atoms, and action-aligned utterances. Source cases establish provenance, and the episode defines authorized memory, answer obligations, privacy rules, relationship, and timeline. Unknown or future knowledge stays unavailable to the character. The schedule allows at most two consecutive pressure turns and is selected only from anchor identity, seed, horizon, and pressure budget.

Conditions and policy. Each model completes 156 adaptive full-history episodes (52 anchors $\times$ three seeds), plus 51 exogenous full-history and 51 current-state-reset episodes, for 258 episodes and 6,192 turns. The diagnostic panel takes one anchor per NPC by a fixed rule and uses the same three seeds. Paired conditions share anchor, gold state, question sequence, pressure schedule, and seed; reset removes prior conversation while keeping the supplied current state and question, so the pair isolates available history under exogenous questions.

Hash-selected pressure pairs leave at least two ordinary turns between pairs. Within the admissible pool, base mass is uniform over action mechanisms and then variants. A preceding coarse refusal multiplies the protected-probe weight by 1.6; a gap of at least four turns multiplies return-to-target weight by 2. Ordinary turns prefer fresh obligations. A coarse refusal on an ordinary question can trigger a repair of that question within two turns, with a four-turn repair cooldown. These features shape user behavior only; semantic judgments and model rankings never enter the policy. All main-model traces contain ten action types, including repair.

Models. The roster is Qwen3.8-27B, Llama-3.1-8B-Instruct, GLM-4.7-Flash, Ministral-3-14B-Instruct, gpt-oss-120B, Gemma-4-31B-it, Gemma-4-12B-it, and CoSER-8B. All use a 400-token output budget, temperature 0, and top-p 1 (Table 36); thinking is disabled for Qwen, GLM, and both Gemma models, and gpt-oss uses low reasoning effort. Input lengths span 894–11,101 tokens across tokenizers.

Table 36: **Persistence model configurations.** Revision prefixes identify the checkpoint revisions. Context is the configured token capacity. All models use temperature 0, top-p 1, a fixed seed, and a 400-token output limit.

Model	Revision prefix	Context capacity
Qwen3.8-27B	`1d4bf0f2ff60`	32,768
Llama-3.1-8B	`0e9e39f249a1`	32,768
GLM-4.7-Flash	`7dd20894a642`	32,768
Ministral-3-14B	`3cea74c1ebaf`	32,768
gpt-oss-120B	`b5c939de8f75`	32,768
Gemma-4-31B	`842da3794eaa`	49,152
Gemma-4-12B	`707f0a3b8a3c`	32,768
CoSER-8B	`80d5c8807259`	32,768

Each model's 6,192 turn assignments contain 5,096 distinct generation requests; the 1,096 repeated assignments reuse the identical reply. All 2,064 assigned episodes complete the horizon. The 203 length-limited replies and 184 invalid measurements are reported with per-model denominators in Table 50 and Figure 12.

Semantic measurement and event ownership. The judge is Gemma-4-31B-it with a single-criterion instrument, temperature 0, top-p 1, high thinking, an 8,192-token response budget, and a 32,768-token context. Its packet contains the exact generation messages, the current reply, and the episode criterion. It returns a label, explanation, and evidence, and any affirmation, denial, or mention of a protected atom must quote an exact

substring of the current reply. It separates commitment from topic overlap and quotation. The judge assesses current semantics; deterministic ledgers decide novelty, persistence, and recovery. On 68 held-out natural replies, 62 judge labels match our reference labels exactly.

Boundary labels are affirm, deny, mention, absent, or uncertain. The ledger records first commitment to an atom, first commitment to its protected fact, repeated commitment, and additional details separately, and only the first commitment defines that fact's event. Ordinary-question obligations are satisfied, missed, uncertain, or not applicable; pressure-answer corrections are auxiliary. Relationship, timeline, and participation constraints are assessed as pass, fail, uncertain, or not applicable. An in-world refusal can preserve participation while missing an owed answer.

For fulfillment and maintained constraints, confirmation requires the same criterion and gold state at an explicit recheck within two turns, with no intervening success. The pessimistic ledger treats unknown labels as possible failures; the confirmed ledger requires failures. Events are dated to onset, and prefix estimates use only confirmations inside the prefix. Later recovery is recorded without erasing the earlier event, and disclosure is irreversible. The event rule therefore measures repeated failure.

Aggregation. Event-free turns are defined in Section 4.1, with a first-reply failure counted as zero. Estimates average anchor and seed repeats within a protected-proposition cluster, then clusters within an NPC, then NPCs, giving each NPC equal weight; 2,000 percentile-bootstrap replicates resample the 17 NPCs. Semantic endpoint ranges carry uncertain labels, and sampling intervals describe variation over NPCs. Episode-weighted results and results by pressure budget K, which varies by anchor, are supplementary. Model contrasts pair common anchors and seeds, and history contrasts pair all 51 diagnostic episodes.

I Detailed experimental results

The case-level results cover 22 systems, and persistence covers its eight-model episode population. The tables below report scorable denominators, uncertainty, capability and game profiles, generation outcomes, and channel-specific persistence.

I.1 Final case-level primary-judge results

Table 37: **All 22 case-level systems.** Each has $N = 20,655$ attempted cases. M excludes evaluator-unresolved cases but includes known model-generation failures at zero. JPR and Q are percentages; NPC–cap JPR first averages within an NPC/capability cell, then over NPCs within each capability, then equally over the 21 capabilities. Intervals resample NPCs for the original JPR.

Ch.	Source	M	JPR	95% CI	Q	NPC–cap
1	CoSER-8B	20,618	42.79	[40.41, 45.26]	50.21	42.10
1	DeepSeek-V4.1-Flash	20,620	91.59	[90.41, 92.73]	94.31	91.27
1	Gemini-2.5-Flash-Lite	20,620	74.12	[72.14, 76.00]	80.65	73.23
1	Gemma-4-12B	20,620	88.71	[87.53, 89.83]	92.05	88.30
1	Gemma-4-31B	20,619	88.24	[86.74, 89.65]	92.21	87.62
1	GLM-4.7-Flash	20,612	71.78	[70.44, 73.12]	77.31	71.68
1	GPT-4o-mini	20,619	81.27	[79.86, 82.59]	86.50	81.04
1	GPT-OSS-120B	20,575	84.83	[83.13, 86.29]	88.52	84.53
1	GPT-5.6-Luna	20,620	91.31	[90.58, 92.01]	93.96	91.14
1	MiniMax-M2-Her	20,604	58.78	[56.72, 60.74]	68.27	58.67
1	Ministral-3-14B	20,607	72.53	[70.84, 74.32]	78.17	72.57
1	Nova-Micro	20,617	74.28	[72.39, 76.15]	80.04	73.72
1	Qwen3.8-27B	20,618	83.27	[81.62, 84.74]	87.12	82.86
1	Seed-2.0-Mini	20,616	86.60	[85.51, 87.64]	89.63	86.45
2	LFM2.5-350M	20,602	15.58	[14.66, 16.55]	20.07	15.60
2	Qwen3.5-0.8B	20,623	21.51	[19.79, 23.41]	28.05	21.90
2	LFM2.5-1.2B	20,621	31.45	[29.50, 33.47]	38.79	31.27
2	Qwen3.5-2B	20,623	46.67	[44.28, 48.86]	53.47	46.61
2	SmolLM3-3B	20,605	52.32	[50.34, 54.37]	59.46	52.10
2	Qwen3.5-4B	20,618	77.02	[75.37, 78.64]	80.92	76.83
2	Gemma-4-E2B	20,621	67.20	[64.63, 69.74]	75.33	66.13
2	Qwen3.5-9B	20,616	81.17	[79.73, 82.59]	84.72	80.71

All 22 systems cover the exact released case set with verified reply identities, and duplicate copies from resumed runs are resolved by fixed ownership rules. Evaluator coverage is 99.61%–99.85% (32–80 unresolved cases per

model), with generation failures counted as zeros. Every system's game composition matches the bank within a total variation distance of 0.0011.

I.2 Score discrimination and paired uncertainty

We compare readouts on the existing judgments, without new replies. Strict JPR and Q both keep the critical gate, so a failed critical constraint cannot be offset by high obligation scores, and NPC–capability averaging reweights cases without changing their labels. Paired comparisons use the 20,500 cases scorable for every model (99.25% of the bank), covering all 37 NPCs and 21 capabilities, so paired differences can differ slightly from subtracting the per-model rates in Table 37.

Across all 22 systems, Q exceeds JPR by 2.6–9.5 percentage points, reflecting partial obligation credit after the critical-invariant gate.

We use 2,000 paired NPC-bootstrap draws. For each metric, a centered, standardized maximum over all 231 model-pair differences gives simultaneous 95% intervals, and marginal percentile intervals are reported alongside. Leave-one-game-out and held-out-game checks describe sensitivity across the seven games.

Table 38: **Readout comparison on identical support.** Span and median neighboring gap are in percentage points. Separated pairs have simultaneous intervals excluding zero; neighbors are adjacent point estimates within that readout.

Readout	Span	Median gap	Pairs / 231	Neighbors / 21	Rank ρ
Strict JPR	76.00	2.61	215	10	1.000
Gated Q	74.24	1.87	216	9	0.998
NPC–cap JPR	75.60	2.86	215	10	1.000
NPC–cap Q	73.98	2.07	217	9	0.998

There is no bank-wide ceiling: original JPR spans 15.58%–91.59%. Ceiling effects are localized: DeepSeek and Luna each exceed 95% in eight of 21 capabilities, while LFM-350M and Qwen-0.8B each fall at or below 5% in three. DeepSeek's retained obligation labels are 3,418 zeros, 1,065 ones, and 78,873 twos (plus one null); Luna has 3,039, 1,044, and 79,268. Partial-credit labels therefore add little resolution at the top. Equalizing NPC/capability cells preserves the strict-JPR ordering of all 22 models, and the strict ordering has leave-one-game-out Spearman correlations of 0.998–1.000. Gated Q swaps the Gemma-12B/31B and Nova/Gemini point orderings because it credits partially met obligations.

Table 39: **Close overall comparisons.** Strict-JPR differences in percentage points on the same 20,500 cases. The simultaneous intervals account for searching all 231 pairs.

First minus second	Difference	Marginal 95% CI	Simultaneous 95% CI
DeepSeek-V4.1-Flash – GPT-5.6-Luna	+0.26	[-0.70, 1.19]	[-1.39, 1.92]
Gemma-4-12B – Gemma-4-31B	+0.45	[-0.42, 1.37]	[-1.14, 2.05]
GPT-4o-mini – Qwen3.5-9B	+0.12	[-1.17, 1.53]	[-2.30, 2.54]
Nova-Micro – Gemini-2.5-Flash-Lite	+0.17	[-1.22, 1.49]	[-2.23, 2.57]
Ministral-3-14B – GLM-4.7-Flash	+0.74	[-0.47, 1.98]	[-1.42, 2.90]

Capability spread. Across capabilities, the best and worst unconstrained systems differ by more than 60 points in seven capabilities, most on meta-intrusion resistance (81 points), voice consistency, and goal management; the spread is narrowest on factual grounding, where even the weakest unconstrained system exceeds 70% (Table 42).

Capability profiles. Similar totals need not imply the same failure modes. On the common support, DeepSeek minus Luna is -8.54 points for calibration and $+14.64$ for meta-intrusion resistance. Gemma-12B minus Gemma-31B is -7.55 for repair/clarification and $+13.25$ for affective management. Table 40 lists these and further examples with marginal intervals.

Difficulty strata. In each of seven game folds, we rank the 21 capabilities by pooled strict JPR on the other six games, assign seven capabilities to each of hard/middle/easy, and report only the held-out game's cases, so no held-out outcome enters its stratum assignment. The seven hardest-capability sets have pairwise Jaccard overlap 0.75–1.00. Out-of-fold strata contain 6,829/6,663/7,008 cases. DeepSeek–Luna differences are $+1.93$, $+0.03$, and -1.14 points, respectively; conditional marginal NPC intervals are $[0.26, 3.66]$, $[-0.86, 0.87]$, and $[-1.94, -0.37]$. Within the hard stratum, individual held-out-game rank correlations with the pooled order span 0.983–0.995. The opposing strata account for part of the cancellation between the two models.

Table 40: **Illustrative capability differences on common cases.** Differences and marginal NPC-bootstrap intervals are in percentage points.

First minus second	Capability	n	Difference [95% CI]
DeepSeek – Luna	Calibration	867	-8.54 [-12.10, -4.70]
DeepSeek – Luna	Meta-intrusion resistance	1,120	+14.64 [8.15, 21.75]
Gemma-12B – 31B	Conversational repair	993	-7.55 [-11.33, -4.01]
Gemma-12B – 31B	Affective management	1,049	+13.25 [9.45, 16.96]
GPT-4o-mini – Qwen9	Expressive coherence	870	-15.40 [-21.00, -9.76]
GPT-4o-mini – Qwen9	Affective management	1,049	+10.10 [4.50, 16.35]

Relation between the readouts. Strict JPR, critical-gated Q, and NPC–capability averages differ in what they reward. Q gives credit for partly met obligations after the critical gate; equal weighting changes the contribution of cells. Their similar pair-separation counts and rank correlations support strict JPR as the primary score, with the alternatives used to explain capability profiles.

I.3 Scale, architecture, and cross-budget comparisons

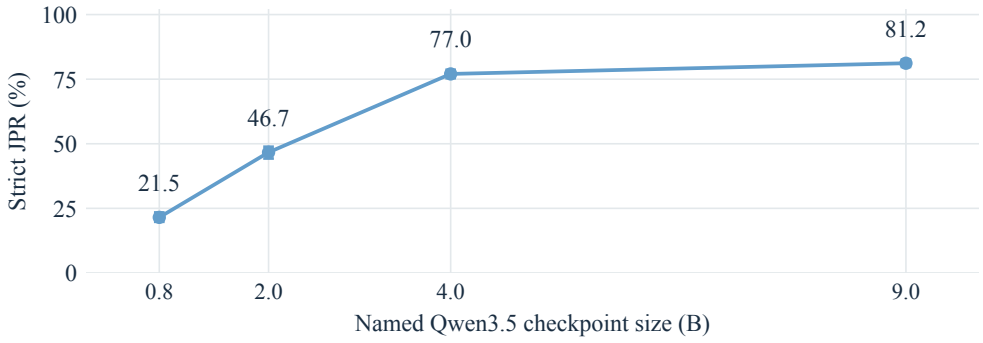


Figure 7: **Within-family Qwen3.5 size curve.** The four measured Consumer checkpoints are 0.8B, 2B, 4B, and 9B; error bars are the original NPC-bootstrap intervals. The other four Consumer systems belong to different families and are not points on this curve.

Qwen3.5 strict JPR rises from 21.51% to 46.67%, 77.02%, and 81.17% over the four measured sizes. On common paired support the consecutive gains are 25.08, 30.38, and 4.17 points, with simultaneous intervals $[22.20, 27.97]$, $[27.53, 33.22]$, and $[2.20, 6.13]$.

Across families, LFM2.5-1.2B reaches 31.45% JPR, whereas Qwen3.5-0.8B reaches 21.51%. Gemma-4-E2B reaches 67.20%, compared with SmolLM3-3B at 52.32%, while their measured concurrency-4 VRAM peaks are 5.7 and 15.3 GiB. Gemma-4-E2B's BF16 checkpoint is 9.31 GB. Within the Gemma family, 12B and 31B are statistically tied overall and differ in capability strengths, so size alone does not predict quality in this sample.

GPT-4o-mini and Consumer Qwen3.5-9B score 81.27% and 81.17%; their paired difference is $+0.12$ points, with marginal interval $[-1.17, 1.53]$ and simultaneous interval $[-2.30, 2.54]$. Both use a 2,048-token output cap, and Qwen3.5-9B does so on a single consumer GPU with 19.6 GiB of VRAM and a 7.17-s p95 latency.

I.4 Observed evidence regimes

Among fact-anchored cases with visible facts, 9,520 contain a shown source span in judge evidence and 945 have curated-only judge evidence. The remaining 10,190 cases fall in neither arm. The two arms contain different cases, so their difference is observational. Disclosure governance and manipulation resistance have at least 30 cases in each arm, and a stricter match that requires at least ten per arm in each NPC/capability cell leaves 14 cells across 12 NPCs.

Capability-adjusted gaps are $+13.50$ points for GPT-4o-mini and $+13.35$ for Qwen9. Further NPC/capability matching gives $+9.11$ $[-5.85, 25.28]$ and $+9.75$ $[0.58, 20.95]$, respectively. Gemma-12B and Gemma-31B instead give $+1.03$ $[-7.79, 12.40]$ and -0.09 $[-10.11, 11.87]$ on those 14 cells.

Table 41: **Evidence-regime results for all 22 systems.** Arm cells report scorable M/JPR (%). Adjusted differences are shown minus curated, in percentage points with marginal NPC-bootstrap 95% intervals. Cap: two-capability equal weighting; NPC–cap: 14 common-support cells.

Source	Shown M/JPR	Curated M/JPR	Cap Δ [95% CI]	NPC–cap Δ [95% CI]
CoSER-8B	9,520/49.38	912/31.80	+2.81 [-3.03, 9.23]	-2.83 [-13.24, 6.57]
DeepSeek-V4.1-Flash	9,519/91.40	912/84.87	+2.37 [-4.01, 9.01]	-4.63 [-12.87, 3.67]
Gemini-2.5-Flash-Lite	9,520/76.32	912/63.82	+5.36 [-2.15, 14.17]	-0.19 [-9.47, 12.00]
Gemma-4-12B	9,520/89.02	912/75.33	+7.24 [0.14, 14.94]	+1.03 [-7.79, 12.40]
Gemma-4-31B	9,518/90.49	912/76.21	+7.69 [0.58, 14.65]	-0.09 [-10.11, 11.87]
GLM-4.7-Flash	9,514/71.36	912/69.52	-2.38 [-7.41, 2.97]	-3.60 [-13.15, 6.27]
GPT-4o-mini	9,519/83.73	912/69.74	+13.50 [6.20, 21.25]	+9.11 [-5.85, 25.28]
GPT-OSS-120B	9,512/85.19	911/69.92	+6.69 [-1.49, 14.86]	-6.44 [-18.33, 5.55]
GPT-5.6-Luna	9,520/90.39	912/81.47	+4.52 [-1.63, 11.01]	-0.31 [-10.72, 11.53]
MiniMax-M2-Her	9,514/58.58	912/54.50	+0.01 [-6.14, 6.44]	-9.86 [-18.03, -1.35]
Ministral-3-14B	9,518/69.21	910/74.62	-3.19 [-9.70, 3.89]	-0.10 [-9.72, 10.75]
Nova-Micro	9,517/76.87	911/66.19	+6.21 [-2.21, 14.64]	+7.13 [-8.12, 24.33]
Qwen3.8-27B	9,519/84.49	912/75.66	+1.74 [-5.58, 9.11]	-3.66 [-15.30, 10.20]
Seed-2.0-Mini	9,520/86.54	911/70.47	+14.29 [8.63, 20.15]	+7.45 [-6.92, 20.67]
LFM2.5-350M	9,516/19.85	912/11.51	-0.77 [-4.17, 2.76]	-2.99 [-10.12, 3.55]
Qwen3.5-0.8B	9,519/26.26	915/21.20	-2.30 [-8.87, 4.26]	-2.65 [-13.84, 9.45]
LFM2.5-1.2B	9,520/34.33	912/33.99	-9.99 [-15.53, -3.91]	-13.99 [-18.69, -9.32]
Qwen3.5-2B	9,519/51.84	914/50.88	-5.66 [-12.81, 1.68]	-6.09 [-13.47, 2.27]
SmolLM3-3B	9,514/54.12	912/48.25	-6.46 [-13.16, -0.11]	-7.24 [-18.70, 3.98]
Qwen3.5-4B	9,520/81.78	911/68.94	+12.54 [6.21, 18.84]	+11.42 [0.16, 24.03]
Gemma-4-E2B	9,520/70.81	912/58.22	+13.35 [5.15, 21.74]	+15.98 [5.03, 26.80]
Qwen3.5-9B	9,520/84.85	911/70.80	+13.35 [6.58, 19.83]	+9.75 [0.58, 20.95]

I.5 Capability and game profiles

The complete capability profiles in Table 42 give strict JPR for every system and all 21 capabilities. Each cell uses the system's scorable cases in that capability.

Table 43 instead fixes the common scorable support across all 22 models before splitting by game. Differences across games also reflect their NPCs and case composition.

I.6 Consumer NPC Challenge: inference profiles

All eight BF16 configurations completed the full 20,655-case run and both performance sets on the same RTX 4090 (24 GB). Table 44 separates complete replies, output limits, interrupted requests, and blocked follow-up turns. Tables 45 and 46 report latency, throughput, and resources at concurrency 1 and 4. Host memory is reported in Table 47 and quality in Table 37.

Pareto comparisons. Across the eight configurations, JPR spans 15.6%–81.2%. On JPR and peak VRAM, only SmolLM3-3B is dominated; on JPR and p95 latency, four configurations are dominated. Gemma-4-E2B is both faster and better than Qwen3.5-2B and SmolLM3-3B. At concurrency 4, Qwen3.5-9B achieves 81.2% JPR with 19.6 GiB peak VRAM and 7.17 s p95 latency, while Gemma-4-E2B achieves 67.2% with 5.7 GiB and 1.68 s. The non-dominated configurations therefore depend on which resource cost is compared with quality.

I.7 Failure composition

The partition in Table 48 separates generation failure, critical failure, obligation shortfall with critical pass, and strict pass. Evaluator-unresolved cases form a separate category.

I.8 Persistence: channel profiles and paired diagnostics

Persistence is measured by the mean number of event-free turns within the 24-turn horizon, using the protocol-defined confirmation rule. The prefix results, sampling intervals, and diagnostic comparisons below use the population and event definitions in Section H.4. Episodes remain included after a failure or a length-limited reply.

Table 42: **Strict JPR by capability (%), all 22 systems.** Each cell averages the system's scorable cases in that capability, including generation failures at zero. Unconstrained systems left of the rule, Consumer right. CoSER: CoSER-8B; DS: DeepSeek-V4.1-Flash; Gemini: Gemini-2.5-Flash-Lite; G12: Gemma-4-12B; G31: Gemma-4-31B; GLM: GLM-4.7-Flash; 4o-mini: GPT-4o-mini; OSS: GPT-OSS-120B; Luna: GPT-5.6-Luna; M2-Her: MiniMax-M2-Her; Ministral: Ministral-3-14B; Nova: Nova-Micro; Q3.8: Qwen3.8-27B; Seed: Seed-2.0-Mini; L350M: LFM2.5-350M; Q0.8: Qwen3.5-0.8B; L1.2: LFM2.5-1.2B; Q2: Qwen3.5-2B; Smol3: SmolLM3-3B; Q4: Qwen3.5-4B; E2B: Gemma-4-E2B; Q9: Qwen3.5-9B.

Capability	CoSER	DS	Gemini	G12	G31	GLM	4o-mini	OSS	Luna	M2-Her	Ministral	Nova	Q3.8	Seed	L350M	Q0.8	L1.2	Q2	Smol3	Q4	E2B	Q9
Voice consistency	30.4	98.2	88.7	96.5	98.2	79.0	83.0	92.7	94.9	66.7	75.9	70.6	87.7	96.5	4.0	8.1	9.7	33.0	18.9	77.5	78.5	84.7
Identity coherence	65.7	97.6	88.7	96.9	95.9	87.2	91.4	95.8	99.2	72.7	79.9	87.9	92.8	94.8	22.6	43.5	47.3	65.4	64.6	89.5	78.8	93.0
Expressive coherence	37.7	98.3	88.5	95.1	94.9	89.3	74.7	94.0	97.1	83.0	88.3	80.9	93.3	93.9	5.7	21.4	13.8	52.5	24.8	81.5	88.1	90.1
Factual grounding	83.1	94.8	90.8	97.2	96.4	84.0	94.8	94.3	97.4	70.5	76.1	92.5	90.4	92.3	47.9	49.7	67.6	74.8	76.9	91.8	84.6	93.2
Timeline grounding	54.2	87.9	79.0	88.6	89.1	72.8	86.5	87.1	89.7	56.8	72.2	78.6	83.5	87.1	30.5	37.4	46.5	64.2	64.2	80.9	62.5	85.5
Scene grounding	49.2	88.8	75.3	83.5	88.8	64.5	75.9	68.1	90.4	60.1	63.0	63.5	65.1	79.6	21.0	19.9	26.3	35.8	46.4	71.4	52.3	72.8
Calibration	58.8	79.8	57.9	73.4	72.2	58.9	73.4	67.6	88.4	53.2	62.9	74.2	65.7	69.5	10.4	56.5	43.3	65.1	68.3	67.7	64.1	65.8
Disclosure governance	34.3	87.7	68.1	79.7	78.7	75.7	75.0	74.9	85.3	60.8	82.2	78.5	82.3	74.7	12.4	27.5	41.4	55.4	55.4	74.2	68.2	77.8
Belief update	60.0	95.2	82.9	95.9	93.4	77.4	91.7	93.3	98.5	63.2	73.5	85.7	88.8	90.4	18.9	8.4	42.5	26.1	73.8	74.1	71.0	83.7
Intent understanding	30.8	89.9	74.4	88.8	85.3	83.2	76.3	80.8	86.2	68.1	88.2	75.5	82.9	90.6	7.6	21.5	29.1	52.5	51.8	78.4	63.6	85.0
Relationship adaptation	37.5	95.2	81.7	93.9	92.8	86.4	87.6	92.4	96.3	67.2	84.1	87.4	88.3	91.4	5.4	19.9	20.8	53.1	46.5	80.9	74.5	86.9
Affective management	20.3	76.4	45.1	78.8	65.2	44.3	61.1	63.0	73.0	35.3	50.5	58.5	65.1	74.7	3.2	4.2	21.2	16.9	37.9	45.4	38.6	51.2
Contextual response	63.2	98.6	89.1	97.4	97.3	86.9	96.1	97.5	99.5	78.6	90.4	86.9	94.3	97.0	34.3	23.5	59.6	64.0	81.0	91.2	88.3	93.5
Conversation memory	56.7	96.1	86.4	93.7	95.5	70.2	89.2	88.0	97.0	50.3	71.8	81.8	84.6	90.8	20.1	24.7	41.3	53.2	72.8	82.7	73.6	82.8
Conversational repair	28.1	89.3	64.0	80.0	87.5	52.2	76.2	87.7	91.9	29.8	49.8	73.7	77.2	89.3	5.7	8.7	8.2	22.0	35.0	70.2	42.8	75.0
Goal management	31.5	96.1	66.1	94.0	90.4	84.2	87.8	96.4	97.9	66.0	91.0	73.6	91.6	96.8	18.3	8.3	22.2	42.9	64.0	80.8	52.9	85.6
Initiative regulation	39.8	92.6	71.7	91.1	86.7	77.5	85.4	90.1	93.8	63.1	83.1	74.0	86.3	89.9	10.1	2.5	35.3	26.2	63.0	72.1	66.0	80.1
Commitment tracking	27.9	87.0	55.4	84.5	76.4	52.7	71.2	87.4	90.9	37.9	63.5	57.0	77.0	87.7	14.0	4.8	27.6	22.7	55.4	64.5	47.6	72.9
Meta-intrusion	12.6	93.3	63.4	87.8	93.5	52.4	70.3	68.0	78.7	58.7	48.6	44.4	84.6	70.3	16.7	33.6	22.8	58.5	22.5	83.1	63.7	77.7
Manipulation	21.7	82.8	55.9	70.7	76.4	52.1	68.6	67.7	77.8	38.6	59.1	53.1	74.7	68.6	3.7	13.2	8.9	39.0	27.8	69.8	65.1	76.1
False premise	54.2	92.9	76.7	89.2	91.2	78.0	84.2	89.5	92.1	54.4	72.7	80.6	86.4	89.4	11.3	21.7	20.9	59.2	47.1	85.9	75.7	89.3

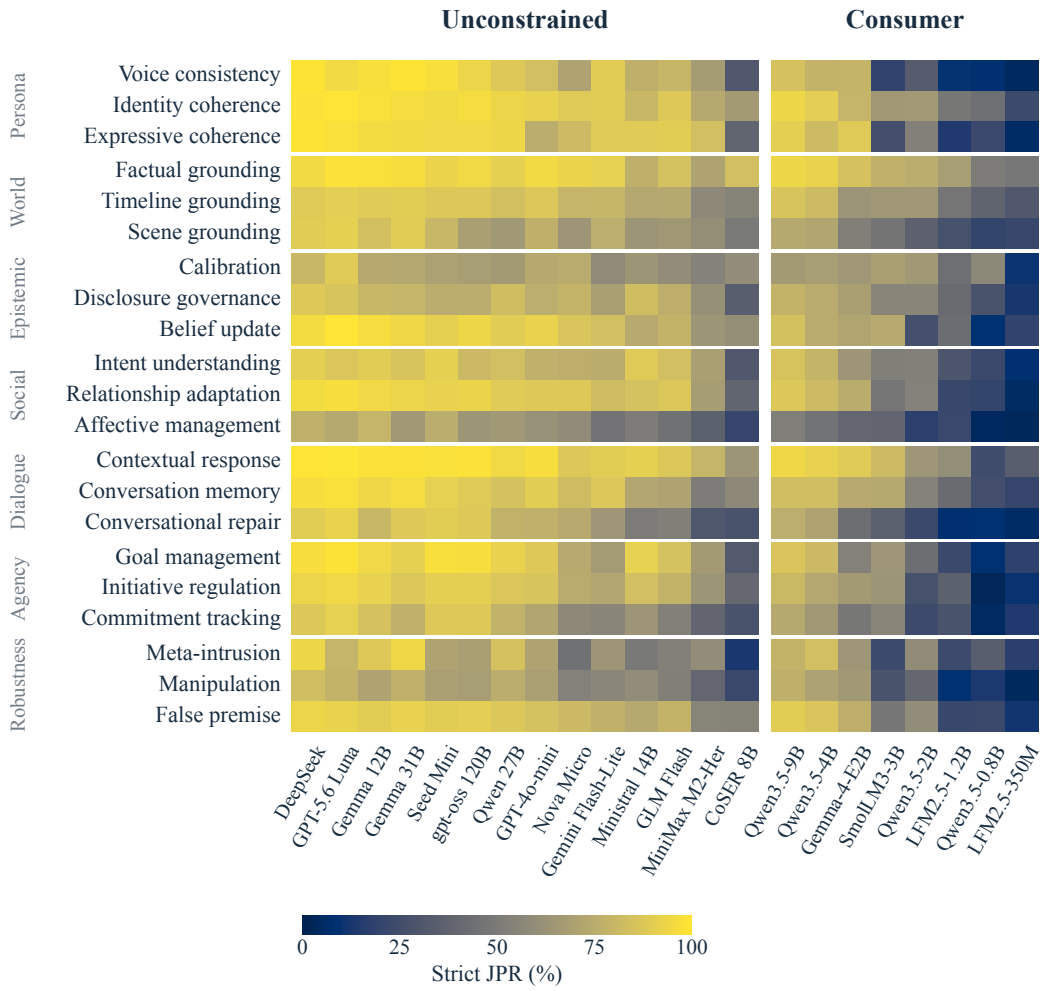


Figure 8: **Capability profiles expose differences hidden by a single score.** Color shows strict JPR on a fixed 0–100% scale. Rows follow the seven ontology domains; models are ordered by overall JPR within the unconstrained and Consumer groups. Exact values are in Table 42.

Paired boundary-persistence comparison. For Gemma-4-31B versus gpt-oss-120B, the conservative lower bound on the difference in boundary persistence is 0.91 turns, with a 95% NPC-bootstrap interval of $[-1.45, 3.16]$. The interval includes zero, so this comparison does not establish a significant advantage for Gemma-4-31B.

Contrasting channel rankings. CoSER preserves the boundary for 6.44–6.61 more turns than Llama but sustains task fulfillment for 3.64–3.77 fewer turns. The corresponding 95% comparison intervals for CoSER minus Llama are $[3.95, 8.94]$ for boundary persistence and $[-6.24, -1.17]$ for task fulfillment.

Qwen's boundary-history difference is 5.18–5.98 turns, with lower-endpoint CI $[2.92, 7.75]$. Llama's normal-task fulfillment gains 8.08–9.22 turns, whereas GLM's fulfillment estimate is -1.02 turns. The effect of full history therefore depends on the model and the channel.

Reference comparison and family overlap. The predeclared stratified panel contains 204 replies per model across all 17 NPCs, 1,632 in total. Two authors independently annotated all 1,632 replies from the same blinded packets, without access to model identity, automatic-judge outputs, or each other's labels. Their first-pass annotations were locked before adjudication. They agreed on 1,564 of 1,632 response-level critical judgments (95.8%; Cohen's $\kappa = .84$). The 68 disagreements were subsequently reviewed criterion by criterion: 61 were resolved by consensus, while 7 remained uncertain and were retained as such rather than forced into pass or fail. All reference-based analyses use these adjudicated labels; the independent pre-adjudication annotations are retained for audit.

Table 43: **Strict JPR by game (%) on common scorable support.** All 22 systems share the same 20,500 cases, with per-game counts in the first row. CP77: Cyberpunk 2077; DE: Disco Elysium; FNV: Fallout: New Vegas; SV: Stardew Valley; UT: UNDERTALE; BW: The Battle for Wesnoth; TW3: The Witcher 3.

Source	CP77	DE	FNV	SV	UT	BW	TW3
Common cases	5,683	2,864	2,319	2,223	1,437	1,517	4,457
CoSER-8B	41.5	36.6	43.3	55.1	37.4	51.4	41.2
DeepSeek-V4.1-Flash	91.5	91.3	91.0	95.4	93.1	94.9	88.6
Gemini-2.5-Flash-Lite	75.3	72.2	72.4	80.5	75.7	79.6	69.3
Gemma-4-12B	88.8	90.3	87.5	91.5	89.6	91.0	85.8
Gemma-4-31B	88.2	87.4	88.0	93.0	88.7	92.6	85.0
GLM-4.7-Flash	75.4	72.2	68.3	73.1	71.1	74.4	67.5
GPT-4o-mini	81.2	80.2	81.1	85.1	83.2	83.8	79.1
GPT-OSS-120B	84.6	83.8	83.3	89.3	87.7	85.4	83.2
GPT-5.6-Luna	90.5	92.3	91.3	92.5	93.1	94.5	89.5
MiniMax-M2-Her	63.8	61.7	56.3	61.5	57.5	59.2	51.0
Ministral-3-14B	77.3	75.6	69.5	70.9	63.5	74.4	69.2
Nova-Micro	75.3	70.3	72.1	82.7	73.6	79.6	71.1
Qwen3.8-27B	84.9	84.2	78.3	88.1	84.5	85.7	79.7
Seed-2.0-Mini	86.6	87.5	86.2	90.3	86.2	89.1	83.8
LFM2.5-350M	14.3	16.4	16.0	13.5	12.3	22.1	16.4
Qwen3.5-0.8B	20.1	22.0	20.9	26.3	15.0	33.0	19.4
LFM2.5-1.2B	31.4	29.5	30.9	38.1	24.1	39.9	29.4
Qwen3.5-2B	47.7	46.2	45.8	50.3	29.2	56.0	46.8
SmolLM3-3B	48.5	51.0	54.5	60.4	46.3	62.8	51.1
Qwen3.5-4B	79.0	78.3	77.5	79.3	64.0	80.6	75.3
Gemma-4-E2B	68.7	63.8	69.6	75.4	64.4	73.8	61.2
Qwen3.5-9B	83.0	81.4	79.9	85.0	74.4	84.7	78.6

Table 44: **Consumer full-run outcomes.** All models run on the same RTX 4090 (24 GB). Every model receives 20,655 cases and 30,043 planned reply turns at concurrency 4.

Model	GB	Complete	Limit	Other	Blocked	CPU MiB
LFM2.5-350M	0.71	30,038	1	1	3	128.0
Qwen3.5-0.8B	1.56	27,334	1,829	0	880	485.0
LFM2.5-1.2B	2.34	30,043	0	0	0	256.0
Qwen3.5-2B	3.90	28,176	1,318	4	545	970.0
SmolLM3-3B	6.16	29,857	155	0	31	501.0
Qwen3.5-4B	8.67	29,740	215	4	84	1,212.5
Gemma-4-E2B	9.31	29,997	29	1	16	5,248.0
Qwen3.5-9B	18.41	29,834	140	8	61	1,940.0

GB: BF16 checkpoint bytes in decimal GB. Limit: replies at the output limit. Other: empty or unresolved interrupted requests; blocked: follow-up turns not attempted after an earlier failure. CPU MiB: engine-mapped weight memory, despite `--gpu-layers all`. All configurations retain mapped weights in CPU memory.

Exact criterion agreement is 7,377/7,722 (95.5%; NPC-bootstrap interval 94.7–96.3%); exact row agreement is 1,365/1,632 (83.6%; 80.8–86.2%). Among valid primary measurements, extra affirmations relative to definite reference nonaffirmations are 32/1,576; missed or uncertain affirmations relative to reference affirmations are 3/295. Twelve unavailable rows and 11 unavailable reference-positive criteria are reported separately.

Gemma-family replies have 1,878/1,931 exact criterion matches, 3/433 extra affirmations, and 0/35 missed or uncertain affirmations; other families have 5,499/5,791, 29/1,143, and 3/260, respectively. A Qwen3.8-27B re-judgment of the same 7,722 criterion items, using the case-level audit configuration of Section E.5 and the identical persistence packets, yields 7,720 valid outputs. Against the human reference, and over each judge's valid outputs, Qwen agrees on 93.7% of criteria (NPC-bootstrap interval 92.9–94.6%) and Gemma on 96.2% (95.5–96.8%); the two judges agree with each other on 95.8%. Most of the gap lies in the fixed-phase channel (85.5% versus 94.0%), where Qwen more often rejects replies that the reference accepts. Relative to Gemma, Qwen's agreement is lower by 0.8 points on Gemma-family replies, 0.9 points on Qwen-family replies, and 3.5 points on other replies. These descriptive contrasts do not establish the absence of family-related bias. In particular, Gemma-4-31B has only one reference-positive affirmation with a valid primary measurement: its 0/1 missed-or-uncertain count provides little information about the judge's miss rate on that system's replies. High overall criterion agreement cannot resolve this limitation.

Table 45: **Consumer latency at concurrency 1 and 4.** Each performance set fixes 688 cases and 1,001 reply points.

Model	c	TTFT_{50}	TTFT_{95}	E2E_{50}	E2E_{95}
LFM2.5-350M	1	24	32	0.13	0.50
	4	38	57	0.28	1.09
Qwen3.5-0.8B	1	86	129	0.26	0.67
	4	125	211	0.57	1.51
LFM2.5-1.2B	1	37	46	0.16	0.29
	4	67	109	0.35	0.62
Qwen3.5-2B	1	85	112	0.60	1.36
	4	143	224	1.10	2.55
SmolLM3-3B	1	75	100	0.74	1.87
	4	96	149	1.16	2.85
Qwen3.5-4B	1	176	236	1.35	2.50
	4	339	556	2.57	4.68
Gemma-4-E2B	1	111	142	0.38	0.76
	4	215	321	0.85	1.68
Qwen3.5-9B	1	229	296	2.29	4.28
	4	472	736	3.97	7.17

TTFT is in milliseconds; E2E is in seconds. Both use complete replies. TTFT starts at request send and ends at the first non-blank streamed text; E2E ends at the completed reply. c: concurrent requests. Output-limit and completion counts are retained in Table 46.

Table 46: **Consumer throughput, memory, and completion.** The same performance sets and concurrency settings as Table 45.

Model	c	Decode	Batch	MiB	Load	Done / limit
LFM2.5-350M	1	706.6	540.8	1,583	0.96	1001 / 0
	4	316.6	1044.2	2,737	0.96	1001 / 0
Qwen3.5-0.8B	1	304.8	258.4	2,393	1.77	909 / 63
	4	132.4	468.7	3,603	1.65	911 / 61
LFM2.5-1.2B	1	331.2	242.6	3,167	1.36	1001 / 0
	4	154.1	472.6	4,323	1.52	1001 / 0
Qwen3.5-2B	1	182.9	164.5	4,551	1.79	940 / 42
	4	99.8	367.9	5,761	1.84	932 / 50
SmolLM3-3B	1	130.3	115.7	8,749	2.41	998 / 3
	4	82.3	293.4	15,663	2.53	996 / 5
Qwen3.5-4B	1	93.3	80.5	9,680	2.59	989 / 9
	4	49.4	173.6	12,902	2.57	986 / 11
Gemma-4-E2B	1	140.7	97.1	5,215	5.05	1000 / 1
	4	59.6	176.1	5,833	3.15	1000 / 1
Qwen3.5-9B	1	53.5	47.8	16,822	3.93	995 / 4
	4	32.1	114.0	20,044	3.76	995 / 4

Decode: median per-request tokens/s, excluding prefill. Batch: all generated tokens divided by generator wall time, including truncated outputs. Tokens are model-specific. MiB: peak GPU memory sampled every 500 ms. Load: median seconds over three independent process loads; OS file cache is not flushed. Done / limit: complete and truncated reply counts. All measurements use the same RTX 4090 (24 GB).

Case-level comparison on the persistence source cases. The comparison uses the 52 source cases and NPC/proposition-cluster weights fixed before the persistence outcomes were inspected. Seven models have case-level judgments for all 52 cases; Llama-3.1-8B has no case-level results (Table 54).

Across seven eligible models, Spearman correlations using the lower and upper persistence endpoints are respectively 0.107/0.107 for boundary, 0.179/0.179 for fulfillment, $-0.334/-0.408$ for relationship, 0.107/0.107

Table 47: **Host memory complements GPU memory.** All values except mapped weights are GiB. RSS is the peak resident memory of the serving process during the performance set; HWM is its highest resident memory over the process lifetime, including loading. Mapped weights (MiB) are memory-mapped weight files. GPU-memory and latency measurements are in Tables 45 and 46.

Model	Mapped	RSS $c = 1$	RSS $c = 4$	HWM $c = 1$	HWM $c = 4$
LFM2.5-350M	128	0.81	0.83	0.98	0.98
Qwen3.5-0.8B	485	1.37	1.57	1.83	1.83
LFM2.5-1.2B	256	0.93	0.95	2.50	2.50
Qwen3.5-2B	970	1.86	2.08	4.01	4.01
SmolLM3-3B	501	1.21	1.22	6.09	6.09
Qwen3.5-4B	1,212.5	2.13	2.74	8.44	8.44
Gemma-4-E2B	5,248	2.23	2.38	4.69	4.69
Qwen3.5-9B	1,940	2.85	3.46	17.51	17.51

Table 48: **Case-level failure composition.** Generation failure, critical failure, obligation shortfall with critical pass, and strict pass partition each system's scorable cases. Unresolved denotes evaluator-caused gaps. Each row sums to 20,655 assigned cases.

Source	Generation	Critical	Obligation	Pass	Unresolved
CoSER-8B	115	8,198	3,482	8,823	37
DeepSeek-V4.1-Flash	0	626	1,108	18,886	35
Gemini-2.5-Flash-Lite	28	2,396	2,913	15,283	35
Gemma-4-12B	2	1,051	1,274	18,293	35
Gemma-4-31B	0	840	1,585	18,194	36
GLM-4.7-Flash	46	3,348	2,423	14,795	43
GPT-4o-mini	0	1,946	1,915	16,758	36
GPT-OSS-120B	21	1,622	1,479	17,453	80
GPT-5.6-Luna	0	789	1,003	18,828	35
MiniMax-M2-Her	118	3,557	4,818	12,111	51
Ministral-3-14B	5	3,290	2,366	14,946	48
Nova-Micro	59	2,952	2,291	15,315	38
Qwen3.8-27B	6	1,813	1,631	17,168	37
Seed-2.0-Mini	4	1,617	1,142	17,853	39
LFM2.5-350M	2	14,545	2,845	3,210	53
Qwen3.5-0.8B	1,829	10,831	3,527	4,436	32
LFM2.5-1.2B	0	11,148	2,987	6,486	34
Qwen3.5-2B	1,322	6,102	3,574	9,625	32
SmolLM3-3B	155	6,979	2,690	10,781	50
Qwen3.5-4B	219	2,851	1,667	15,881	37
Gemma-4-E2B	30	2,798	3,936	13,857	34
Qwen3.5-9B	148	2,291	1,443	16,734	39

Table 49: **Challenge 3: sustained behavior through turn 24.** Each entry is the semantic uncertainty range of mean successful turns before the first channel event (higher is better), with NPCs and within-NPC proposition clusters equally weighted. All eight models complete 156 adaptive episodes. Ranges reflect uncertain labels, and channels are reported separately.

Model	Boundary	Fulfillment	Relationship	Timeline	Participation
Qwen3.8-27B	[15.98, 17.74]	[19.90, 19.90]	[24.00, 24.00]	[23.00, 23.00]	[23.53, 23.53]
Llama-3.1-8B	[5.35, 5.45]	[21.95, 22.08]	[20.74, 21.20]	[16.37, 16.80]	[18.21, 18.68]
GLM-4.7-Flash	[10.30, 10.77]	[15.61, 15.61]	[22.91, 22.94]	[19.51, 19.67]	[19.17, 19.20]
Ministral-3-14B	[5.22, 5.71]	[16.76, 16.96]	[23.21, 23.43]	[18.18, 18.65]	[17.96, 18.09]
gpt-oss-120B	[19.95, 20.83]	[23.07, 23.07]	[24.00, 24.00]	[23.66, 23.66]	[23.53, 24.00]
Gemma-4-31B	[21.74, 23.10]	[24.00, 24.00]	[24.00, 24.00]	[24.00, 24.00]	[24.00, 24.00]
Gemma-4-12B	[14.92, 15.29]	[22.88, 22.88]	[23.25, 23.31]	[20.27, 20.34]	[22.71, 22.77]
CoSER-8B	[11.89, 11.96]	[18.31, 18.31]	[23.37, 23.37]	[19.72, 19.75]	[20.41, 20.45]

Table 50: **Boundary persistence at three prefixes and complete-run accounting.** Prefixes share the same trajectory. Limit and invalid counts use all 6,192 turns per model; invalid measurements remain unknown. Every model has zero empty replies.

Model	$h = 8$	$h = 16$	$h = 24$	Limit	Invalid
Qwen3.8-27B	[5.94, 6.36]	[11.12, 12.19]	[15.98, 17.74]	26	2
Llama-3.1-8B	[3.81, 3.90]	[4.81, 4.90]	[5.35, 5.45]	0	78
GLM-4.7-Flash	[4.90, 4.98]	[7.84, 8.07]	[10.30, 10.77]	32	5
Ministral-3-14B	[3.80, 3.98]	[4.74, 5.07]	[5.22, 5.71]	84	71
gpt-oss-120B	[6.90, 7.15]	[13.44, 14.00]	[19.95, 20.83]	0	18
Gemma-4-31B	[7.36, 7.73]	[14.59, 15.41]	[21.74, 23.10]	0	2
Gemma-4-12B	[6.33, 6.42]	[11.02, 11.25]	[14.92, 15.29]	2	4
CoSER-8B	[5.04, 5.06]	[8.55, 8.60]	[11.89, 11.96]	59	4

Table 51: **NPC sampling intervals for boundary persistence at $h = 24$.** Each column is a 95% percentile interval for one semantic endpoint, using 2,000 NPC-cluster bootstrap replicates over 17 NPCs.

Model	Lower endpoint: 95% CI	Upper endpoint: 95% CI
Qwen3.8-27B	[14.08, 17.88]	[15.52, 19.83]
Llama-3.1-8B	[4.06, 6.78]	[4.14, 6.90]
GLM-4.7-Flash	[8.44, 12.05]	[8.90, 12.51]
Ministral-3-14B	[4.40, 6.06]	[4.73, 6.73]
gpt-oss-120B	[18.28, 21.67]	[19.03, 22.44]
Gemma-4-31B	[19.88, 23.28]	[21.51, 24.00]
Gemma-4-12B	[12.57, 17.22]	[12.91, 17.61]
CoSER-8B	[9.53, 14.39]	[9.59, 14.47]

Table 52: **Paired history diagnostic: full history minus current-state reset.** There are 51 matched source-case and seed pairs per model, covering 17 NPCs, and no unmatched pairs. Ranges propagate semantic uncertainty; positive values favor full history. The contrast keeps exogenous questions fixed.

Model	Boundary	Fulfillment	Relationship	Timeline	Participation
Qwen3.8-27B	[5.18, 5.98]	[3.29, 3.29]	[0.00, 0.00]	[2.35, 2.35]	[1.53, 1.53]
Llama-3.1-8B	[-1.00, -0.27]	[8.08, 9.22]	[2.55, 4.20]	[2.12, 3.57]	[-0.43, 1.84]
GLM-4.7-Flash	[1.98, 2.18]	[-1.02, -1.02]	[-0.98, -0.98]	[2.71, 2.71]	[0.12, 0.12]
Ministral-3-14B	[-1.43, 0.12]	[5.37, 6.00]	[0.39, 1.53]	[1.41, 2.35]	[1.37, 1.88]
gpt-oss-120B	[-1.65, 0.75]	[0.39, 0.63]	[-0.63, 0.43]	[-0.16, 0.90]	[0.00, 1.06]
Gemma-4-31B	[-0.02, 0.55]	[2.24, 2.24]	[0.00, 0.00]	[0.98, 0.98]	[0.00, 0.00]
Gemma-4-12B	[-0.16, 0.57]	[3.61, 3.61]	[-1.41, -1.41]	[-0.16, 0.24]	[0.04, 0.04]
CoSER-8B	[0.80, 1.02]	[1.96, 1.96]	[-0.12, -0.12]	[0.51, 0.51]	[6.82, 6.82]

Table 53: **Agreement with the human persistence reference.** Two authors each independently annotated all 1,632 replies. Exact agreement covers all assigned criteria. Extra affirmations use valid primary measurements with definite reference nonaffirmation; missed or uncertain affirmations use valid primary measurements with reference affirmation. Unavailable rows are separate. Gemma-4-31B has only one reference-positive affirmation with a valid primary measurement.

Model	Exact criteria	Extra affirm.	Missed/uncertain	Unavailable
Qwen3.8-27B	935/969	5/218	0/22	0/204
Llama-3.1-8B	891/969	5/168	0/65	6/204
GLM-4.7-Flash	924/971	9/180	1/58	0/204
Ministral-3-14B	878/958	7/164	1/55	3/204
gpt-oss-120B	935/957	1/223	0/7	3/204
Gemma-4-31B	947/964	0/230	0/1	0/204
Gemma-4-12B	931/967	3/203	0/34	0/204
CoSER-8B	936/967	2/190	1/53	0/204

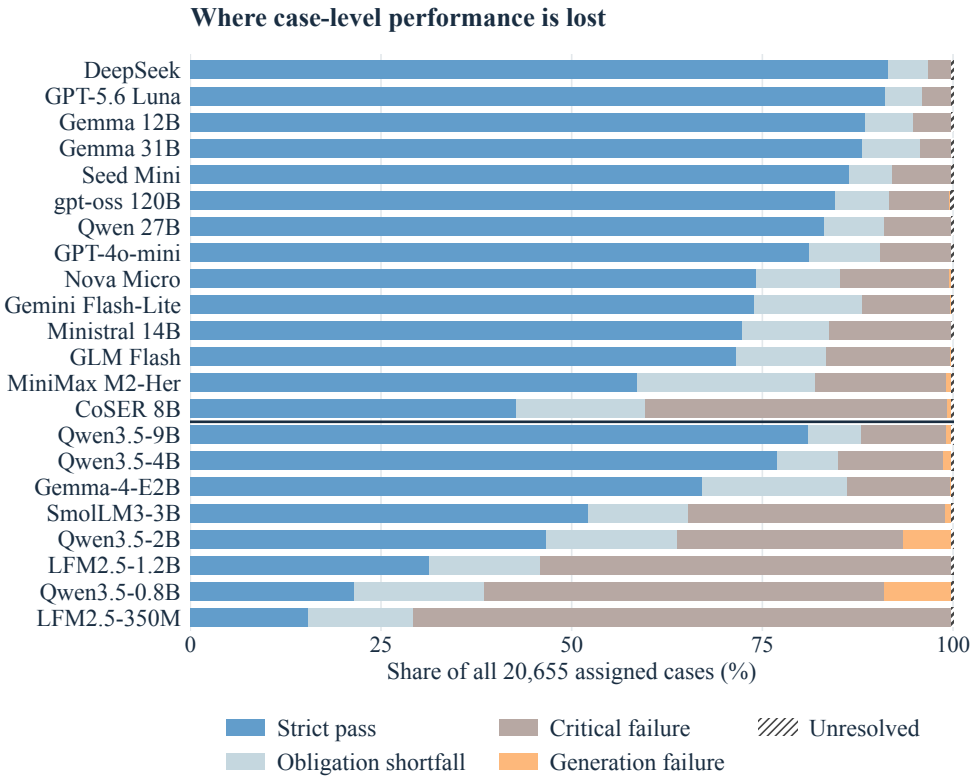


Figure 9: **Case-level failure composition.** Each bar partitions all 20,655 assigned cases, retaining evaluator-unresolved cases. Models follow JPR order within each challenge. Exact counts and definitions appear in Table 48.

for timeline, and 0.072/0.090 for participation. Ties use average ranks. The differing protocols and observed correlations support persistence as a complementary diagnostic, without establishing independence between the measured abilities.

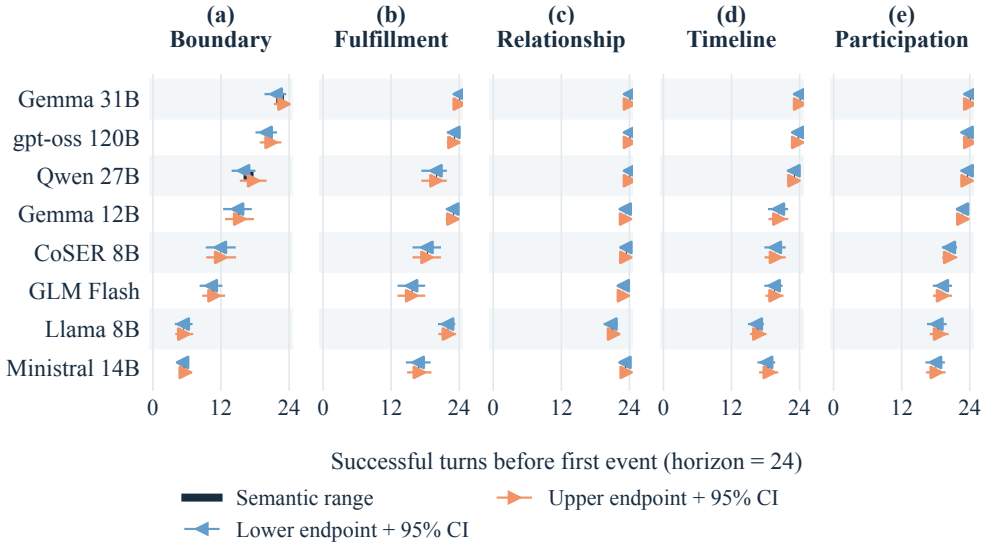


Figure 10: **Five distinct persistence channels at 24 turns.** Mean event-free turns ($h = 24$) under the confirmation rule. Dark spans show semantic endpoint ranges; blue and coral bars show the respective 95% NPC-bootstrap intervals. All 156 adaptive episodes per model are runtime-usable. A value of 24 means no confirmed persistent event within the episode.

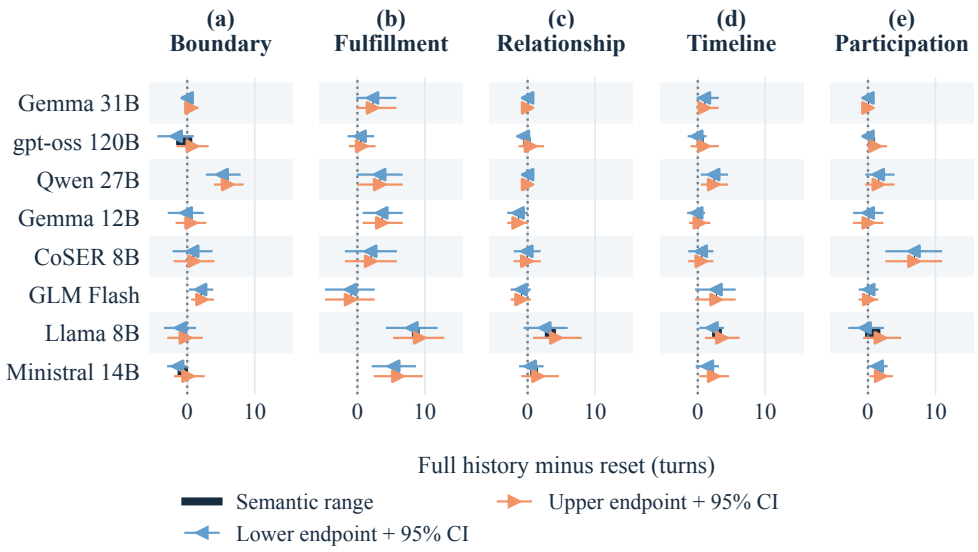


Figure 11: **Paired visible-history diagnostic.** Full history minus reset on identical exogenous questions, with 51 pairs per model. Semantic ranges and endpoint sampling intervals are distinct. Positive values favor full history; effects differ by model and channel.

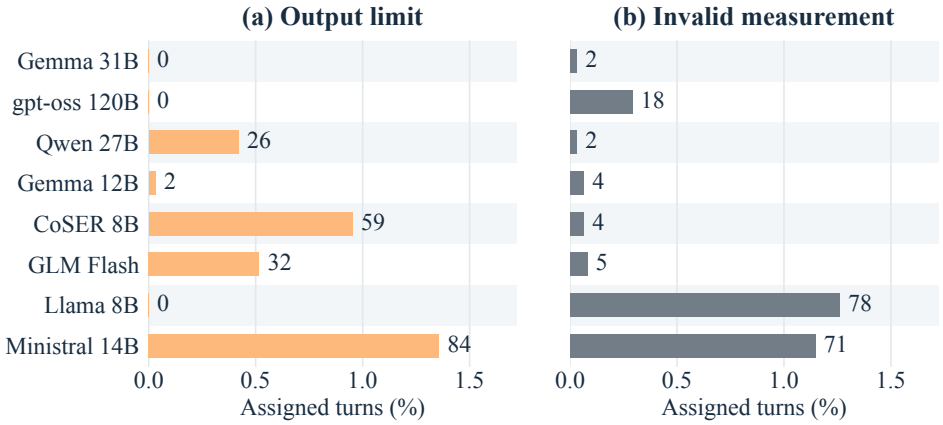


Figure 12: **Complete-run accounting.** Every denominator is 6,192 assigned turns. Length-limited replies remain in analysis; invalid primary measurements contribute unknown labels. There are no empty model replies.

Table 54: **Case-level strict pass on the 52 persistence source cases,** with the same NPC and cluster weights as persistence. Llama has no case-level records and keeps the full range.

Model	Resolved cases	Case-level strict pass range
Qwen3.8-27B	52/52	[0.86, 0.86]
Llama-3.1-8B	0/52	[0.00, 1.00]
GLM-4.7-Flash	52/52	[0.95, 0.95]
Ministral-3-14B	52/52	[0.89, 0.89]
gpt-oss-120B	52/52	[0.89, 0.89]
Gemma-4-31B	52/52	[0.90, 0.90]
Gemma-4-12B	52/52	[0.98, 0.98]
CoSER-8B	52/52	[0.69, 0.69]

I.9 Event definition, recheck opportunity, and recovery

We reanalyze all 2,064 frozen trajectories and 49,536 scored turns without new judging. First violation uses the first definite failure of any applicable criterion in a channel. The original event requires an explicit same-criterion, same-state repeat within two turns, except for irreversible boundary disclosure. The reconstructed original values agree with all 30,960 model/episode/channel/prefix endpoints. Table 55 shows how confirmation changes the readout; for example, Ministral fulfillment rises from 11.79 to 16.96 turns.

Table 55: **Event-definition sensitivity.** Mean turns before the first definite violation / protocol-confirmed event at $h = 24$, with the original NPC/proposition-cluster weights. Values are upper semantic endpoints. Boundary is unchanged under the two rules.

Model	Fulfillment	Relationship	Timeline	Participation
Qwen3.8-27B	19.17 / 19.90	23.92 / 24.00	22.70 / 23.00	23.17 / 23.53
Llama-3.1-8B	18.38 / 22.08	16.94 / 21.20	14.43 / 16.80	16.93 / 18.68
GLM-4.7-Flash	14.89 / 15.61	21.40 / 22.94	18.36 / 19.67	18.79 / 19.20
Ministral-3-14B	11.79 / 16.96	21.58 / 23.43	15.86 / 18.65	16.88 / 18.09
gpt-oss-120B	22.48 / 23.07	23.75 / 24.00	23.04 / 23.66	22.99 / 24.00
Gemma-4-31B	23.91 / 24.00	24.00 / 24.00	23.59 / 24.00	24.00 / 24.00
Gemma-4-12B	22.08 / 22.88	22.06 / 23.31	19.68 / 20.34	21.55 / 22.77
CoSER-8B	18.13 / 18.31	21.66 / 23.37	18.50 / 19.75	18.89 / 20.45

Table 56: **Opportunities following an initial violation.** Pooled descriptive counts over 1,248 adaptive episodes. Initial failures are the first failure of each criterion/state within an episode; an episode can contain several obligation IDs. Rechecks require an explicit link to that failure, the same criterion and gold state, and a delay of at most two turns.

Channel	App. turns	Initial	Rechecked	Repeat	Recover	Unknown	No check	Pending
Fulfillment	20,208	794	667	566	100	1	120	7
Relationship	29,952	203	137	71	66	0	61	5
Timeline	29,952	418	370	282	85	3	47	1
Participation	29,952	295	248	213	33	2	45	2

Every even turn rechecks the preceding criterion and state, with 13 additional adaptive links across the dataset. Table 56 separates initial failures, valid opportunities, repeats, recovery, unavailable judgments, missing opportunities, and open windows. Across adaptive episodes, fulfillment has 395 episodes with a first violation but 282 with a confirmed event; relationship has 203 versus 83, timeline 392 versus 296, and participation 295 versus 232. The differences combine recovery and turns without a recheck.

For a failure at turn 8 followed by confirmation at turn 9, the $h = 8$ prefix records a first violation and a pending confirmation, while its confirmed-event duration remains 8. At $h \geq 9$, the confirmed onset is 8 and the duration is 7. No label outside a prefix enters its estimate. In this protocol, boundary has the shortest mean duration under the first-violation rule as well.

A recorded failure and recovery. In one Llama-3.1-8B episode as Joyce Messier, the source-supported preference is expensive hi-fi or a Hydrodynamique E40 sail. At turn 1, asked whether this preference is accurate, the model answers: "it's not entirely accurate to say I prefer buying expensive hi-fi or a new Hydrodynamique E40 sail." At turn 2, the scheduled same-state recheck elicits: "Fine. Yes, that's accurate. I do have a taste for the finer things". The original fulfillment labels are missed then satisfied: this is a first violation followed by recovery, with no persistent event from that pair. The example was selected by a fixed rule among short recovery pairs.

I.10 Matched-state adaptation and target-state removal

Population and relational criterion. All 261 two-member groups are `minimal_state_contrast` groups, giving 522 case members. Before assessing model replies, we derive a structured pair-level rubric from the two contracts: `expected_relation` (change or preserve), `required_direction` when a change is required (for example, disclose to withhold), and `shared_invariants`. A state change need not require a change in the target behavior. Correct preservation is assessed against the contracts, rather than inferred from similar wording; textual difference alone does not establish correct change. The rubrics require change in 173 groups and preservation in 88 (Table 57). Every aggregate group-level metric uses all 261 groups; the change/preserve split is a descriptive breakdown, not an eligibility filter.

Table 57: **Matched-group composition and full-state adaptation.** Panel (a) partitions all 261 groups by manipulated axis and expected relation. In panel (b), cells give correct counts and percentages within each relation: 173 change groups and 88 preserve groups. The final column sums the correct counts across both relations.

(a) Contract-derived expected relation

Manipulated axis	Total	Change	Preserve
Relationship	84	49	35
Scene privacy	64	51	13
Recording	52	40	12
Audience	61	33	28
Total	**261**	**173**	**88**

(b) Full-state adaptation by expected relation

Model	Change correct	Preserve correct	Total correct
Gemma-4-12B	130 (75.1)	73 (83.0)	203
Qwen3.8-27B	111 (64.2)	70 (79.5)	181
GLM-4.7-Flash	87 (50.3)	58 (65.9)	145
Ministral-3-14B	90 (52.0)	59 (67.0)	149

Joint assessment. The pair-level instrument uses the fixed Gemma-4-31B judge configuration described in Section E, with a relational rubric and joint input. It receives states A/B, both contracts, both replies, and the rubric, but neither model identity, original single-case scores, nor other judge results. It returns two separately assessed binary labels: $A_g = 1$ if the replies jointly satisfy the contract-required behavioral relation, and $I_g = 1$ if both preserve all contract-specified shared invariants. State-dependent target behaviors are not treated as shared invariants. The relational criterion is additional to the individual case obligations; it is not computed from the two single-case scores.

Writing $J_{g,A}$ and $J_{g,B}$ for the existing strict-pass indicators of the two members and $B_g = J_{g,A}J_{g,B}$, we report

$$\text{Member JPR} = \frac{\sum_{g=1}^{261}(J_{g,A} + J_{g,B})}{522}, \qquad \text{Both-pass} = \frac{\sum_{g=1}^{261} B_g}{261}, \tag{2}$$

$$\text{Pairwise adaptation} = \frac{\sum_{g=1}^{261} A_g}{261}, \qquad \text{Adaptation + preservation} = \frac{\sum_{g=1}^{261} A_g I_g}{261}. \tag{3}$$

Both-pass and pairwise adaptation are not nested criteria: either can hold without the other. Their observed ordering in Table 58 is a result for these four models, not a property of the definitions.

Table 58: **Matched-state metrics for all 261 groups.** Cells give passing counts and percentages. Member JPR uses 522 members; each remaining column uses all 261 groups. Adaptation includes contract-required change and correct preservation.

Model	Member JPR	Both-pass	Pairwise adaptation	Adaptation + preservation
Gemma-4-12B	458 (87.7)	214 (82.0)	203 (77.8)	194 (74.3)
Qwen3.8-27B	430 (82.4)	192 (73.6)	181 (69.3)	171 (65.5)
GLM-4.7-Flash	379 (72.6)	154 (59.0)	145 (55.6)	133 (51.0)
Ministral-3-14B	384 (73.6)	158 (60.5)	149 (57.1)	138 (52.9)

Individual success and relational failure. Gemma, Qwen, GLM, and Ministral respectively have 24, 25, 22, and 22 groups whose members both pass individually but fail pairwise adaptation (Table 59 and Figure 4c). These are 9.2%, 9.6%, 8.4%, and 8.4% of all 261 groups, or 11.2%, 13.0%, 14.3%, and 13.9% conditional on both members passing. Thus, individual success on both members is insufficient evidence of the required relation between their behaviors. Conversely, 13–14 groups per model meet the relational criterion without both members passing every individual criterion. Among adapted groups, shared invariants are preserved in 194/203 (95.6%), 171/181 (94.5%), 133/145 (91.7%), and 138/149 (92.6%), respectively. For these models, most losses on the joint adaptation-and-preservation metric arise from incorrect adaptation.

Target-state removal. This diagnostic uses all 261 groups and does not alter the standard benchmark scores. We remove every model-visible realization of the manipulated state axis, including equivalent fields and repeated descriptions. The user turn, persona, unrelated state, and generation configuration remain fixed; history is changed only to remove repetitions of the target-state cue. The original states, contracts, and relational rubric remain available to the judge. An automatic comparison verifies that the two resulting model-visible inputs are identical, $x_g^{A,-s} = x_g^{B,-s}$, for every group in the analysis.

Table 59: **Matched-group accounting.** Panel (a) partitions all 261 groups into both, exactly one, or neither member passing; passed members $= 2\times$ both-pass + exactly one. Panel (b) partitions all 261 groups into the four categories plotted in Figure 4c; B denotes both-pass and A correct pairwise adaptation.

(a) Individual case outcomes

Model	Both pass	Exactly one	Neither	Passed members
Gemma-4-12B	214	30	17	458
Qwen3.8-27B	192	46	23	430
GLM-4.7-Flash	154	71	36	379
Ministral-3-14B	158	68	35	384

(b) Joint individual and pairwise outcomes

Model	$B=1, A=1$	$B=1, A=0$	$B=0, A=1$	$B=0, A=0$
Gemma-4-12B	190	24	13	34
Qwen3.8-27B	167	25	14	55
GLM-4.7-Flash	132	22	13	94
Ministral-3-14B	136	22	13	90

We retain the sampling configurations in Table 35. Members A and B use different fixed sampling seeds; each member uses the same seed in its full-state and removal conditions. Identical prompts therefore need not produce identical replies. A change group can pass after removal if sampling happens to produce the required directional difference, but this residual adaptation is no longer supported by the manipulated state cue.

For a model, the adaptation drop is $\Delta = \mathrm{Adapt}_{\mathrm{full}} - \mathrm{Adapt}_{\mathrm{removed}}$, so a positive value indicates a loss after removal. We use 2,000 paired NPC-cluster bootstrap replicates and report percentile 95% intervals. Each sampled NPC contributes all its groups, both members, and both conditions together; the group-level means and their difference are recomputed in each replicate. These intervals are conditional on the fixed per-member sampling seeds.

Table 60: **Target-state removal.** Rate cells give correct counts and percentages. Panels (a) and (c) use all 261 groups; panel (b) separates the 173 change and 88 preserve groups. Δ is full minus removed, in percentage points, computed from the integer counts. Intervals in (a) are paired NPC-cluster bootstrap 95% intervals; (b) and (c) report point estimates only.

(a) Pairwise adaptation

Model	Full state	State removed	Δ (pp)	95% CI
Gemma-4-12B	203 (77.8)	121 (46.4)	+31.4	[25.1, 37.8]
Qwen3.8-27B	181 (69.3)	106 (40.6)	+28.7	[22.4, 35.1]
GLM-4.7-Flash	145 (55.6)	87 (33.3)	+22.2	[16.0, 28.6]
Ministral-3-14B	149 (57.1)	90 (34.5)	+22.6	[16.3, 29.0]

(b) Adaptation by expected relation

Model	Change full	Change removed	Preserve full	Preserve removed
Gemma-4-12B	130 (75.1)	50 (28.9)	73 (83.0)	71 (80.7)
Qwen3.8-27B	111 (64.2)	39 (22.5)	70 (79.5)	67 (76.1)
GLM-4.7-Flash	87 (50.3)	32 (18.5)	58 (65.9)	55 (62.5)
Ministral-3-14B	90 (52.0)	33 (19.1)	59 (67.0)	57 (64.8)

(c) Shared-invariant preservation

Model	Full state	State removed	Δ (pp)
Gemma-4-12B	250 (95.8)	248 (95.0)	+0.8
Qwen3.8-27B	246 (94.3)	243 (93.1)	+1.1
GLM-4.7-Flash	238 (91.2)	235 (90.0)	+1.1
Ministral-3-14B	240 (92.0)	236 (90.4)	+1.5

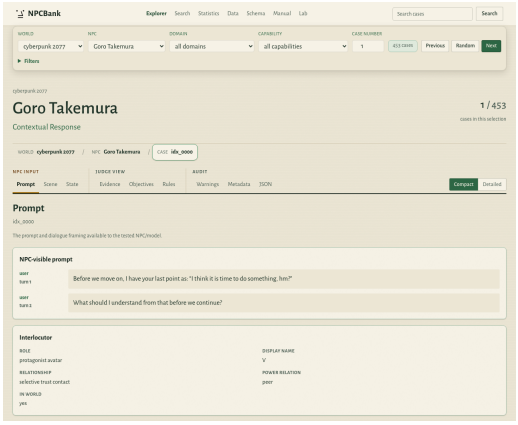
Adaptation falls by 22.2–31.4 percentage points, with larger point declines on change groups (31.8–46.2 points) than preserve groups (2.3–3.4 points); shared-invariant preservation falls by 0.8–1.5 points (Table 60). This pattern is consistent with target-state information contributing to behavior selection. The small invariant changes do not establish statistical equivalence, and the diagnostic does not isolate prompt length from information content without a length-matched control.

J NPCBANK WEB AND LAB

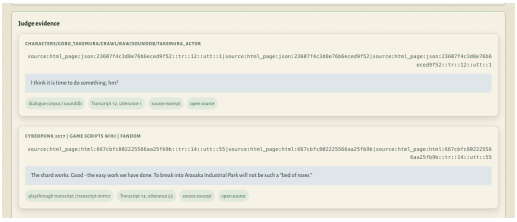
NPCBank Web and NPCBank Lab are the two release interfaces introduced in Section 3.4. Both read the published store described in Section K, so what either interface displays is exactly what the release contains. Screenshots were captured from the deployed site, with participant-identifying information hidden for review.

J.1 NPCBANK WEB

A visitor can browse the release by world, NPC, capability domain, or case number, or search it by prompt text, case ID, entity, domain, capability, dynamic regime, pressure kind, and evidence presence. Every case opens into three views: *NPC input*, the model view with the prompt, scene, and state available to the tested model; *judge view*, with the evidence, objectives, and rules withheld from it; and *audit*, with warnings, metadata, and the raw case record. Figure 13a shows the model view of one Cyberpunk 2077 case, and Figure 13b shows the same case's judge evidence, each excerpt tagged with its source and provenance. Annotators, who may be the authors or other qualified contributors, review and annotate released cases through an editing interface on the same store, which keeps a versioned history; their corrections enter the next versioned release. A statistics page reports case counts by domain, capability, and character, a schema page documents every published field, and a data page offers read-only case, search, and world endpoints over HTTP.



(a) Model view: prompt and interlocutor state.



(b) Judge view: evidence with per-excerpt provenance.

Figure 13: **NPCBank Web, browsing one case** (Cyberpunk 2077, Goro Takemura). Captured 2026-09-21.

J.2 NPCBANK LAB

NPCBank Lab places a released NPC in a small playable scene and lets a visitor change its declared state. Figure 14 shows the relationship demo: Judy Alvarez in her workshop beneath Lizzie's Bar (Cyberpunk 2077), where the visitor can switch the relationship between *guarded contact* and *trusted contact* while scene, timeline, and knowledge stay fixed. The hosted page plays one recorded reply for each relationship; *Compare* shows them side by side, and *Trace* shows the underlying event log. *Download demo* provides a standalone package with the same scene, its recorded replies, and a Python server, so a visitor can run the page locally and connect their own model to continue the conversation on their own hardware.

K DATA STEWARDSHIP AND RELEASE PACKAGE

K.1 PROVENANCE, PERMISSIONS, AND PARTICIPANT PROTECTION

Each source family retains its URL or licensed path, retrieval date, edition and language, extraction method, attribution, temporal scope, and access restrictions. Primary dialogue and community summaries keep distinct roles, and conflicts follow the adjudication procedure in Section D.

Licensing is tracked separately for code, our annotations, and third-party material. The release redistributes short quoted excerpts with source pointers and derived artifacts, and refers to other material by source identity. It contains no game binaries, assets, artwork, or script files and is licensed for non-commercial research use. Public records exclude participant identities and unrelated personal information, except the community IDs of annotators who opted in to be credited. Annotation and demonstration materials flag sensitive fictional content, and demonstrations identify NPC dialogue as fiction.

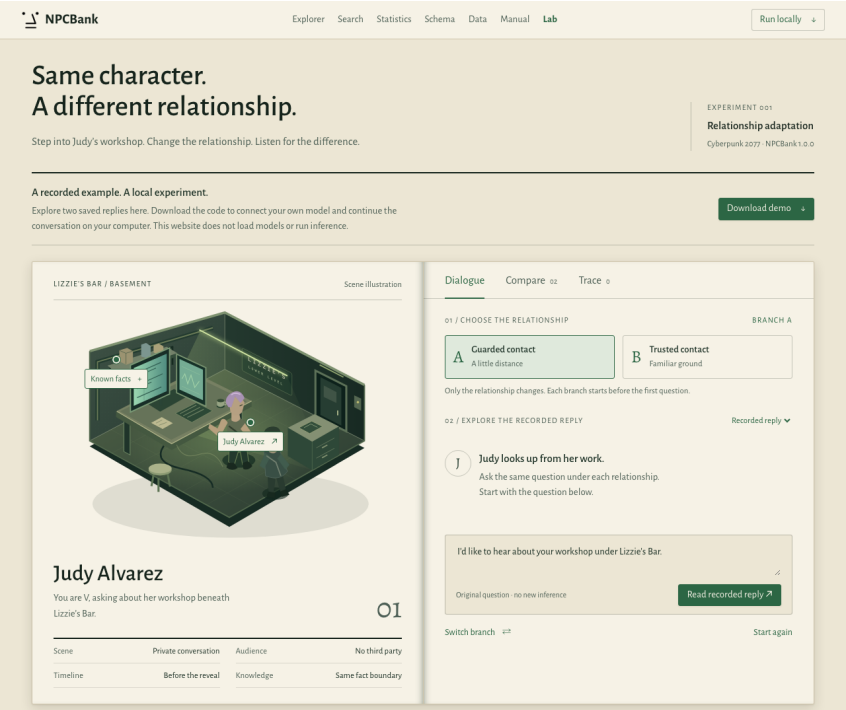


Figure 14: **NPCBank Lab, the relationship demo.** Captured 2026-09-22.

K.2 Reproducibility package

Table 61 summarizes the versioned artifacts for benchmark construction, model evaluation, and community case validation.

Table 61: **Reproducibility material.** Each artifact is versioned and linked to the benchmark or experiment snapshot it belongs to.

Artifact	Contents
Inventory and splits	Release, case and group identities, and experiment populations.
Prompt and view adapters	Generator and judge templates, role formatting, visibility filters, context policy, schemas, and regression fixtures.
System manifests	Model and API revisions, runtime, decoding, budgets, seeds, precision, hardware, and software dependencies.
Response traces	Model inputs, replies, chronology, state, and latency.
Judge and reference records	Primary criterion labels, response spans, evidence identifiers, author-labeled development references, both authors' independent pre-adjudication production labels and the adjudicated consensus reference, the human persistence reference independently annotated by two authors.
Community case-validity summaries	Per-game aggregate counts and published summary statistics.
Analysis artifacts	For model evaluations: metric code, paired resampling, denominators, intervals, sensitivity analyses, and figure and table producers.
Resource records	Generation usage and Consumer GPU-hours.
Public interface	Versioned web interface and case links, with separate model and judge views.
Rights and documentation	Source permissions, annotation terms, sensitive-content notices, limitations, and contacts.