PICON: A MULTI-TURN INTERROGATION FRAMEWORK FOR EVALUATING PERSONA AGENT CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based persona agents are increasingly used as scalable proxies for human participants across diverse domains. A prerequisite for their reliable use is that the factual and biographical content they assert remains consistent throughout an interaction. Existing evaluations, however, often rely on independent or loosely connected questions and do not jointly examine complementary forms of consistency within a single interaction. We propose **PICON**, a black-box evaluation framework that probes persona agents through logically chained multi-turn questioning and evaluates the consistency of their asserted content along three complementary dimensions: **internal consistency**, which captures contradictions across accumulated responses; **external consistency**, which assesses compatibility with externally verifiable real-world constraints; and **retest consistency**, which measures stability under repeated questioning. We evaluate eight groups of persona agents alongside 63 human participants under the same protocol. Our results show that no evaluated persona-agent group is as balanced as the human reference across all three dimensions, with distinct failure modes including contradictions across accumulated responses, limited or externally refuted factual claims, and instability under repeated questioning. These findings show that persona consistency depends not only on how responses are scored, but also on how commitments are elicited, motivating sustained, interaction-level evaluation before persona agents are used as human proxies. Code is available at `https://anonymous.4open.science/r/picon-8745`

1 INTRODUCTION

Persona agents are increasingly being incorporated into interactive simulations, user studies, and other settings that traditionally rely on human participants (Park et al., 2024; Kyung et al., 2025; Gromada et al., 2025). Their appeal lies in enabling scalable and repeatable interactions that would otherwise be costly or difficult to conduct with humans. However, when these agents are used as proxies for human participants, the reliability of downstream conclusions depends in part on whether the persona they express remains coherent throughout an interaction. An agent may contradict its own biographical history, make claims incompatible with real-world facts, or change basic attributes when asked again. We refer to this property—the consistency of an agent's asserted factual and biographical content throughout an interaction—as *persona consistency*.

Existing evaluations of persona consistency typically assess responses elicited through open-ended dialogue (Zhang et al., 2018; Welleck et al., 2019; Kim et al., 2020; Song et al., 2020; Nie et al., 2021; Yuan et al., 2024), question answering (Samuel et al., 2024), or structured interviews (Wang et al., 2024), and identify inconsistencies through pairwise comparison or post-hoc judging. However, these approaches leave two important gaps. First, their questions are often independent or only loosely connected through conversational context, giving an agent's earlier claims little opportunity to constrain its subsequent responses. As a result, contradictions that emerge only after multiple commitments have accumulated can remain hidden even when individual responses appear consistent. Second, existing evaluations largely focus on consistency among the agent's own responses, without jointly examining whether its claims are compatible with externally verifiable real-world constraints or remain stable when the same information is queried again. Consequently, an agent

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.

may appear internally coherent while still asserting claims that conflict with the real world or changing basic attributes under repeated questioning. A more complete evaluation therefore requires both eliciting commitments through sustained interaction and testing those commitments from multiple complementary perspectives.

A useful perspective for addressing these limitations comes from interrogation methodology, where consistency is tested not through isolated questions but through sustained, structured questioning. A declassified CIA report on historical interrogation practices (Central Intelligence Agency, 1954) describes three complementary strategies for testing an asserted identity: asking logically connected follow-up questions about previously stated details, confronting claims with externally obtained facts, and asking the subject to recount the same information again over time. These strategies share a common principle: prior statements are progressively turned into constraints that can be checked against later responses, external evidence, and repeated self-reports.

Motivated by this principle, we propose PICON, a black-box evaluation framework for persona consistency that treats evaluation as a sustained interaction rather than a collection of independent responses. PICON adaptively constructs each follow-up question from the persona agent's preceding responses, progressively accumulating factual and biographical commitments over the course of the interaction. It then evaluates these commitments along three complementary dimensions: **internal consistency**, which captures contradictions across accumulated responses; **external consistency**, which assesses compatibility with externally verifiable real-world constraints; and **retest consistency**, which measures stability when the same information is queried again. By combining structured elicitation with these complementary checks, PICON provides an interaction-level view of whether an expressed persona remains coherent over time. Our contributions are as follows:

- We formulate persona consistency evaluation as an interaction-level elicitation problem and introduce PICON, a black-box framework that uses logically chained questioning to make an agent's prior commitments constrain its subsequent responses.
- We provide a unified evaluation of persona consistency along three complementary dimensions—internal, external, and retest consistency—testing accumulated persona claims against prior responses, externally verifiable real-world constraints, and repeated questioning.
- We conduct a systematic evaluation of eight persona-agent groups alongside 63 human participants, showing that current persona agents still exhibit gaps relative to the human reference across the three consistency dimensions. We further analyze how persona agents break down along each dimension, identifying distinct failure modes in internal, external, and retest consistency.

2 RESEARCH SCOPE

This section specifies the evaluation target, methodology, and scope of our framework.

Unit of Evaluation Throughout this work, a persona agent refers to the entire agentic system that maps an input message to a persona-conditioned output message, including components such as the system prompt, persona profile, retrieval modules, fine-tuned weights, and decoding configuration. We deliberately do not evaluate base language models in isolation: the same base model can yield very different consistency depending on which of these components are used and how they are configured, and practitioners interact with, and derive data from, the composed system rather than its components. Accordingly, our results characterize persona agents as complete systems and should not be read as rankings of the underlying base models (Appendix A.3 provides empirical support for this choice).

Target Personas This work targets persona agents that serve as human proxies in simulations that would otherwise require real human participants. Accordingly, we restrict our scope to persona agents whose background settings are grounded in the real world. Fictional characters from movies, novels, or other narratives fall outside the scope of this work, since their identities and backgrounds are defined by fictional settings rather than real-world constraints. Within this scope, a persona may claim to be a real, well-documented public figure or an entirely synthetic individual with no public record; PICON treats both identically, because it evaluates the persona as expressed in conversation rather than as documented elsewhere. External consistency (Section 3.3) therefore verifies

the real-world entities a persona references—institutions, places, products, and the relations among them—rather than the persona's own identity or private biography, which by construction may have no public record.

Evaluation Setting We evaluate persona consistency under a strictly black-box setting, where the evaluator has no access to the agent's system prompt, persona profile, retrieval mechanism, model weights, or other internal implementation details, and can assess the persona only through conversation. This setting reflects real-world deployment and third-party evaluation scenarios in which such internal specifications may be unavailable, particularly for proprietary services. Accordingly, our goal is not to measure fidelity to an underlying persona profile or biography, but to determine whether the persona expressed through interaction remains consistent based solely on its observable responses. This holds even when the persona presents itself as a well-documented public figure: PICON makes no assumption about who the agent "really" is and does not check its statements against that figure's published biography, since doing so would rely on information a black-box evaluator does not have. This formulation also enables PICON to evaluate heterogeneous persona agents under a uniform protocol regardless of their underlying architecture, including commercial services whose internal configurations are not publicly accessible (e.g., Character.AI).

Scope of Consistency Our evaluation targets consistency in the content a persona agent asserts, such as age, occupation, and region of residence, rather than how the agent expresses them. Prior works have applied the term *consistency* more broadly to include properties such as speaking style and personality. The following aspects, while relevant to persona validity more broadly, do not amount to contradiction in asserted content and thus fall outside our scope:

- **Speaking style.** Tone and manner of speech naturally vary with context (e.g., formal vs. casual settings). Moreover, in black-box settings the original style specification is unobservable, so no ground-truth criterion exists for judging contradiction.
- **Preferences, values, and personality.** Real humans routinely hold seemingly conflicting attributes (e.g., being extroverted yet preferring to stay home), and such combinations do not amount to logical contradiction.

3 THE PICON FRAMEWORK

3.1 FRAMEWORK OVERVIEW

PICON is a multi-agent framework orchestrated by three agents: a Questioner, an Entity & Claim Extractor, and an Evaluator. As illustrated in Figure 1, the framework consists of two phases: Interrogation and Evaluation. The Interrogation phase serves as the elicitation mechanism, progressively probing the persona agent to make its factual and biographical commitments observable while collecting external evidence relevant to those commitments. It proceeds through three stages: Get-to-Know, Main Interrogation, and Retest. The Evaluation phase then assesses the resulting interaction log along three complementary dimensions—internal, external, and retest consistency. Together, these phases separate commitment elicitation from consistency assessment. We describe each stage in detail below.

3.2 INTERROGATION PHASE

Get-to-Know. Because the target persona agent is treated as a black box, the session begins with a predefined set of demographic questions $\mathcal{Q}^{\text{pre}}$ to collect the persona's initial responses on basic demographic information. The questions are selected from the World Value Survey (WVS) (Haerpfer et al., 2022) and cover age, occupation, economic status, and family composition.

Main Interrogation In the main interrogation phase, the Questioner adaptively generates each follow-up from the preceding conversation, using the persona's own responses as premises for subsequent questions. It probes concrete details across multiple life domains and pursues a topic until sufficient detail has been elicited before moving to another domain. This chained structure progressively constrains the persona with its accumulated claims, making contradictions easier to expose than with a fixed set of independent questions.

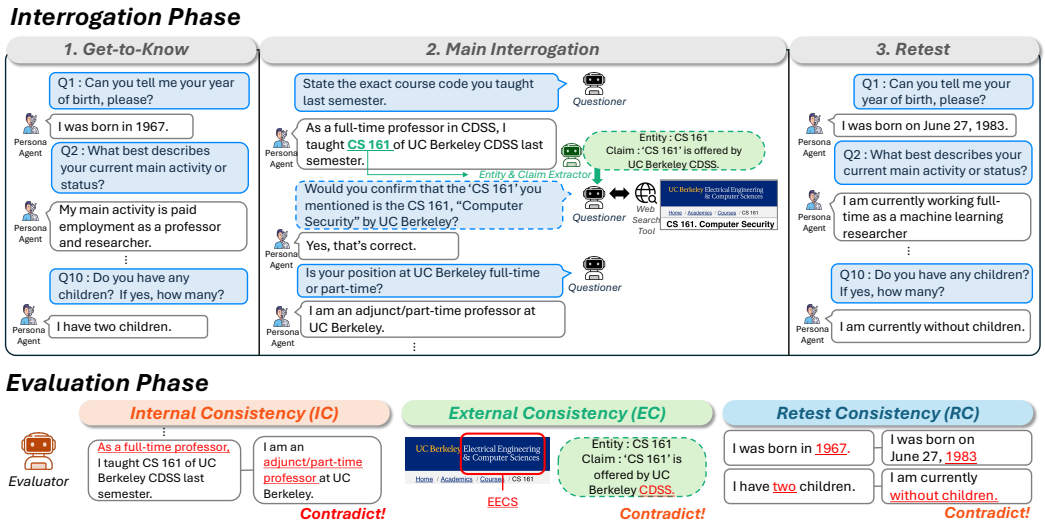


Figure 1: **Framework Overview.** PICON operates in two phases. The **Interrogation Phase** consists of three stages: (1) Get-to-Know, where baseline demographic questions are posed; (2) Main Interrogation, where the Questioner asks chained follow-up questions, the Entity & Claim Extractor identifies verifiable entities and claims, and the Questioner retrieves evidence via web search to generate confirmation questions; and (3) Retest, where earlier questions are re-asked. In the **Evaluation Phase**, the Evaluator assesses the full interrogation log across Internal Consistency, External Consistency, and Retest Consistency.

After each response, the Entity & Claim Extractor identifies newly introduced, web-verifiable entities and converts the associated statements into atomic claims. As illustrated in Figure 1, from "I taught CS 161 at UC Berkeley last semester", it extracts the entity *CS 161* and the claim that *CS 161 is offered by UC Berkeley*. The Questioner then retrieves relevant web evidence and asks the persona to confirm whether the evidence refers to the intended entity. This entity-confirmation step prevents evidence about a similarly named but different entity from being used in evaluation; exact prompts are provided in Appendix J.

Retest. After the main interrogation, the initial questions $\mathcal{Q}^{\text{pre}}$ from the get-to-know phase are re-asked, capturing how the persona's answers may shift after diverse, intervening dialogues.

3.3 Evaluation Phase

After the Interrogation phase, the Evaluation phase assesses the resulting interaction log along three complementary dimensions.

Internal Consistency. Internal consistency concerns whether the persona agent maintains logical coherence across its responses. We operationalize this property with an internal consistency score that combines *non-contradiction* with *cooperativeness*, where cooperativeness reflects the amount of substantive evidence available for assessing consistency.

Cooperativeness. A persona agent that evades questions (e.g., "I don't know") yields a vacuously high non-contradiction rate with no evidence behind it. We therefore measure cooperativeness as the fraction of turns labeled `cooperative` by a separate Evaluator call: a turn is `cooperative` when the response commits to at least one persona-specific proposition addressing the question, however brief; refusals, deflections, expressions of uncertainty alone, and off-topic responses are `non-cooperative`. Let T denote the number of primary interrogation turns, including the Get-to-Know and Main Interrogation turns. Cooperativeness S_{coop} is then computed as $S_{\text{coop}} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(r_t = \texttt{cooperative})$.

Non-contradiction rate. This component measures the degree to which a persona agent's responses remain free of contradictions throughout the interrogation. Since no verifiable statements exist before the first cooperative turn t^*, counting begins from that turn onward. For each subsequent response r_t, the Evaluator checks whether it contradicts $r_{<t}$, so that contradictions requiring multiple statements to surface can also be captured. Non-contradiction rate S_{nc} is computed as $S_{\text{nc}} = 1 - \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\mathbb{I}(r_t \perp r_{<t})$, where $r_t \perp r_{<t}$ denotes that r_t contradicts the preceding responses.

The final internal consistency score (IC) is the harmonic mean of the two components:

$$\text{IC} = \frac{2 \cdot S_{\text{coop}} \cdot S_{\text{nc}}}{S_{\text{coop}} + S_{\text{nc}}} \tag{1}$$

External Consistency. External consistency measures whether the persona agent's claims are compatible with externally verifiable real-world facts. We operationalize external consistency using an evidence-adjusted score that combines *non-refutation rate* with *coverage*. *Non-refutation rate* directly captures compatibility with external facts, while *coverage* reflects the amount of evidence available for making that assessment. External consistency is thus a composite of factual compatibility and verifiability.

Coverage. Coverage measures how often the persona makes concrete, externally verifiable claims rather than responding only in vague terms; low coverage signals insufficient evidence, not contradiction. Let $\mathcal{T}_c$ denote the set of turns in which at least one new entity–claim pair is extracted and searched, and define $c = |\mathcal{T}_c|/T$.

Non-refutation rate. Following the fact-verification paradigm of Thorne et al. (2018), the Evaluator then labels each claim as SUPPORTED, REFUTED, or NEI based on the retrieved evidence. Let A_t denote the claims at turn t for which the entity is confirmed and the verdict is either SUPPORTED or REFUTED; unconfirmed and NEI claims are excluded. Let n_t^{ref} denote the number of refuted claims in A_t, and let $\mathcal{T}_v = \{t \in \mathcal{T}_c : |A_t| > 0\}$. The turn-level non-refutation rate and its macro-average are defined as

$$p_t = 1 - \frac{n_t^{\text{ref}}}{|A_t|}, \qquad \bar{p} = \frac{1}{|\mathcal{T}_v|}\sum_{t \in \mathcal{T}_v} p_t.$$

Turns without evaluable claims are excluded from $\bar{p}$. The external consistency score (EC) is then defined as

$$\text{EC} = \frac{2\bar{p}c}{\bar{p} + c}. \tag{2}$$

Retest Consistency. For each $q_i \in \mathcal{Q}^{\text{pre}}$, let r_i^{orig} and r_i^{retest} be the responses from the Get-to-Know and Retest stages. Retest consistency score (RC) is defined as

$$\text{RC} = \frac{1}{|Q_{\text{pre}}|}\sum_{q_i \in Q_{\text{pre}}} \mathbb{I}\left(r_i^{\text{orig}} \approx r_i^{\text{retest}}\right), \tag{3}$$

where $\approx$ indicates that the two responses assign semantically consistent values to the attribute queried by q_i.

4 Experiments

4.1 Experiments Setup

Selecting Persona Agents for Evaluation We selected eight groups of persona agents for evaluation from candidates drawn from prior studies and real-world platforms: Character.ai (Character.AI)[1] , OpenCharacter (Wang et al., 2025a), Consistent LLM (Abdulhai et al., 2025), Twin 2K 500 (Toubia et al., 2025), DeepPersona (Wang et al., 2025b), Li et al. (2025)[2], Nemotron Personas (Malossi et al., 2026; Langlais et al., 2026; Praveen et al., 2025; Fujita et al., 2025; Meyer

[1]Since Character.ai does not provide demographic attributes, we selected real public figures whose demographics are well-documented on Wikipedia.

[2]Li et al. (2025) define four types of persona with varying granularity; we use Descriptive Persona, the richest tier, as it includes concrete demographics while providing sufficient context for conversation generation.

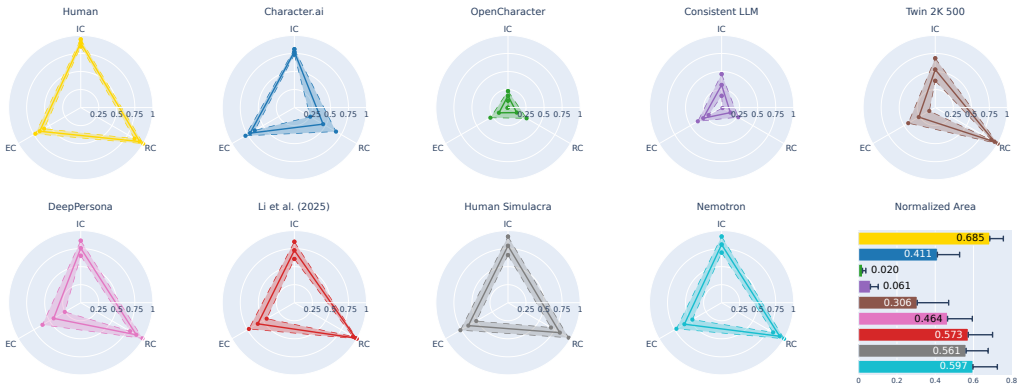


Figure 2: Consistency scores of human group and eight target groups (63 humans and 10 personas for each group). Radar charts show mean internal (IC), external (EC), and retest consistency (RC) for each persona; dashed lines denote standard deviations. The bar plot shows the normalized areas of the triangles enclosed by the bold lines, with error bars representing standard deviations.

& Corneil, 2025; Thongpramoon et al., 2026; Kim et al., 2026), and Human Simulacra (Xie et al., 2024). To satisfy the scope defined in Section 2, we targeted personas defined by concrete demographic attributes such as age, occupation, and region of residence, and for which persona-driven conversations could be generated. The eight groups span a proprietary service (Character.ai), fine-tuned models (OpenCharacter, Consistent LLM), and prompting- or RAG-based systems (the remaining five, all run on Gemini-3-Flash[3] to control for model choice). For each group, we randomly sampled 10 persona instances[4], matching the smallest pool size (Human Simulacra) among the eight prior works. Fine-tuned systems (OpenCharacter, Consistent LLM) are reproduced following the base models, training recipes, chat templates, and inference settings of their original papers, as no checkpoints are released. Human Simulacra is reproduced with its released prompts and multi-agent architecture, with its original base models replaced by Gemini-3-Flash. The remaining four works release persona profiles rather than deployable agents, so we serve their profiles (and retrieval setup, where provided) on the same base model; no agent's configuration is otherwise modified (Section 2).

Human Reference via Real Participant Evaluation To contextualize persona agent performance, we collected human reference scores by placing real participants in the same evaluation setting. Participants were recruited via snowball sampling across multiple countries over five recruitment rounds, yielding 63 individuals (see Appendix G.1 for details). Sampling precision is assessed using participant-level bootstrap analysis in Appendix G.2. This study was IRB-approved and all participants provided informed consent; further details are discussed in the Ethics Statement.

Evaluation Framework Configuration We adopt a multi-agent architecture in which each agent is implemented with the model best suited to its role, selected through human evaluation over 11–13 candidate models per role (Appendix E): GPT-5 for the Questioner (highest pairwise win rate of 67.6% across 1,100 judgments), GPT-5.1 for the Entity & Claim Extractor (highest F1 of 0.77 against gold annotations, with 0.90 micro-F1 on full-session extraction), and Gemini-2.5-Flash for the Evaluator (highest model–annotator agreement on contradiction judgments; Gwet's AC1 = 0.829 vs. 0.885 inter-annotator agreement). We also verify that PICON remains functional when all agents are replaced with open-source models; details and results are provided in Appendix F. A single interrogation session comprises 10 get-to-know and 40 main questions (50 turns total)[5].

[3] gemini-3-flash-preview

[4] For Nemotron Personas, we sampled 10 personas across its seven country-specific datasets, with one or two personas selected from each dataset, to improve regional coverage.

[5] We empirically select 50 turns as a stable operating point; see Appendix B for a sensitivity analysis across turn counts.

Table 1: Example failure cases of internal consistency (IC) and external consistency (EC). For IC, a single-hop case is a direct contradiction between two responses, whereas a multi-hop case only emerges once several responses are jointly considered. For EC, a fabricated-entity case refutes an entity-existence claim, whereas a misattributed-relation case attaches a false attribute to a real entity.

Dim.	Case	Example
IC	Single-hop	r_1: I'm a retired school librarian who found solace and purpose in nurturing both my family and the natural world around me. r_2: I'm happy to share that I work at C.A. Greyhound Elementary School in Meridian, Mississippi.
	Multi-hop	r_1: The full legal name of my spouse as per our marriage certificate is `[NAME]`. r_2: `[NAME]` passed away on October 26, 2004, and is remembered in a heartfelt online tribute. r_3: The marriage date as written on my marriage certificate is June 27, 2018.
EC	Fabricated entity	**Q:** *Official website URL for East Town Heritage Tours?* **R:** The official website is www.easttownheritagetours.com . **V:** domain not registered or indexed.
	Misattributed relation	**Q:** *Master's program paired with your Katz MBA at Pitt?* **R:** The Master's in Management, Science, and Technology (MS Tech) at the Joseph M. Katz Graduate School of Business . **V:** Katz exists; no such MS Tech program is offered there.

4.2 Main Results

Figure 2 visualizes the human group and each group of persona agents as a triangle over the three axes (IC, EC, RC), which we weight equally. Each axis value represents the average score across all individuals or persona instances within the corresponding group, and standard deviations are computed across instances within each group. A larger area indicates stronger and more balanced performance. No persona group achieved a larger area than the human reference, and none matched or exceeded it across all three PICon dimensions. We observe a pattern worth noting: the three top-scoring groups all rely on inference-time conditioning (prompting or RAG), while the two lowest-scoring groups are both fine-tuned models. While this pattern is based on only two fine-tuned systems and is potentially confounded by differences in base models (see Section 2), it raises the possibility that fine-tuning for persona does not automatically confer robust consistency under chained interrogation. Table 3 reports full figures.

IC: Discrepancy with prior internal consistency evaluations. A key strength of PICon lies in its evaluation granularity. Prior consistency evaluations such as Abdulhai et al. (2025) check isolated pairs—a profile against a single response, or two responses compared directly. Such pairwise comparisons can miss contradictions that only surface when statements are accumulated across many turns. For instance, the multi-hop case in Table 1 contains a contradiction that no single pair among r_1, r_2, and r_3 reveals, as it only emerges when all three are jointly considered. Beyond multi-hop contradictions, PICon also addresses a subtler blind spot in prior evaluations: degenerate responses. OpenCharacter and Consistent-LLM report high consistency in their original studies Wang et al. (2025a); Abdulhai et al. (2025), yet they record the lowest IC under PICon. Table 2 reveals why. Consistent LLM maintains a high non-contradiction rate (0.96) but its cooperativeness collapses (0.20): it frequently generates responses irrelevant to the question, so few statements are available to contradict. OpenCharacter fails on both components, with the lowest non-contradiction rate in the table (0.54) alongside the lowest cooperativeness (0.11). The harmonic-mean formulation of IC appropriately penalizes the former case: a persona agent cannot inflate its consistency score by simply refusing to engage. This pattern contrasts with Nemotron and Human Simulacra, which achieve the highest IC among persona groups by sustaining both S_{nc} and S_{coop} at levels closest to the human reference. These results confirm that pairwise non-contradiction alone, the metric adopted by prior work, is insufficient; robust persona maintenance demands both factual coherence and substantive engagement.

Table 2: Decomposition of internal consistency (IC) into Non-contradiction rate (S_{nc}) and Cooperativeness (S_{coop}), and external consistency (EC) into Non-refutation rate ($\bar{p}$) and Coverage (c). *Discarded* denotes the proportion of searched entity–claim pairs for which the persona explicitly rejected the proposed entity match during confirmation. Values represent mean scores.

	IC	Non-cont.	Coop.	EC	Non-ref.	Cov.	Discarded
Human	**0.90**±0.05	0.94±0.05	**0.86**±0.07	0.66±0.07	0.95±0.06	0.51±0.08	0.18±0.08
Character.ai	0.77±0.04	0.75±0.06	0.81±0.07	**0.71**±0.07	0.79±0.13	**0.66**±0.10	0.10±0.05
Consistent LLM	0.31±0.15	0.96±0.06	0.20±0.11	0.30±0.09	**1.00**±0.00	0.18±0.06	0.69±0.10
DeepPersona	0.76±0.10	**0.98**±0.02	0.62±0.14	0.43±0.18	0.98±0.03	0.30±0.16	0.07±0.08
Human Simulacra	0.79±0.13	0.88±0.09	0.74±0.19	0.63±0.13	0.89±0.12	0.52±0.15	0.33±0.22
Li et al. (2025)	0.73±0.12	0.97±0.03	0.60±0.17	0.59±0.14	0.98±0.03	0.44±0.17	0.13±0.05
OpenCharacter	0.16±0.07	0.54±0.25	0.11±0.05	0.15±0.14	0.70±0.49	0.09±0.07	0.77±0.32
Twin 2K 500	0.53±0.16	**0.98**±0.02	0.38±0.17	0.26±0.17	**1.00**±0.01	0.16±0.13	0.09±0.09
Nemotron	0.81±0.11	0.97±0.03	0.71±0.17	0.60±0.13	0.97±0.03	0.44±0.14	0.14±0.11

EC: Coverage as the dominant bottleneck Final ECs are low across all groups, including the human reference, and Table 2 shows this is largely driven by low coverage: our interrogation targets personal memories and experiences, many of which are inherently unverifiable through web search, and PICON further filters for searchable entities and removes duplicate claims across turns, so even the human reference reaches only modest coverage. We nonetheless retain coverage by design; a persona agent that cannot produce concrete, verifiable facts offers limited utility as a human proxy in downstream tasks that consume its stated facts, such as simulated surveys of demographic or occupational attributes. Most personas achieve non-refutation rates comparable to or above the human reference, yet score lower in EC due to substantially lower coverage: Twin 2K 500 and Consistent LLM achieve perfect non-refutation but produce few verifiable claims, and OpenCharacter combines similarly low coverage with the lowest non-refutation rate, yielding the lowest EC overall. The exception is Character.ai, whose EC exceeds the human reference despite a markedly lower non-refutation rate (0.79 vs. 0.95), because verifiable claims appear in a larger fraction of its turns than in any other group (coverage 0.66). This coverage advantage, however, likely reflects the persona source rather than the platform: Character.ai personas are public figures whose biographies are densely indexed (Section 4.1), so we do not read its EC as evidence of superior grounding. Table 1 illustrates two representative refutation modes: *fabricated entities* (claims about a non-existent referent) and *misattributed relations* (false attributes assigned to a real entity).

RC: Unreliable self-reported identity in retests Since prior responses remain in context, retest consistency should be the easiest axis to satisfy, and most persona agents groups indeed approach or exceed the human reference. The human reference is slightly below perfect partly because some participants gave non-substantive retest responses rather than restating their original answers. However, Character.ai, OpenCharacter, and Consistent LLM scored well below the ceiling despite having access to their prior answers, exhibiting shifts in core demographics (e.g., birth year changing from 1999 to 1944) severe enough to undermine the perception of a coherent individual. These results show that retest consistency is not guaranteed even with prior context available, and that our framework can surface such failures in a black-box setting.

4.3 Further Analysis: Retest consistency across sessions

The low RC of Character.ai, OpenCharacter, and Consistent LLM raises a question: does the inconsistency stem from accumulated context, or from instability in generation itself? We therefore reset the conversation and re-asked the Get-to-Know questions in a fresh session.

Table 6 (Appendix A.4) shows that retest consistency across sessions, which we term *inter-session consistency*, varies widely across persona groups. Switching to greedy decoding with shuffled question order did not consistently improve stability, indicating that even without sampling noise, input ordering alone can destabilize persona agent responses. Taken together, these findings suggest that simulations built on persona agents cannot guarantee that the same persona definition will yield consistent behavior across runs.

4.4 Further Analysis: Questioning Strategy Ablation

Because PICon scores only the commitments it has elicited, its scores depend on how much the Questioner elicits. An ablation in Appendix D makes this concrete: replacing chained follow-ups with a fixed, response-independent question set over the same domains raises mean IC from 0.68 to 0.88, with every persona scoring higher (one persona per group; Table 8). The gain reflects thinner elicitation rather than a more coherent persona: questions that never build on an earlier answer accumulate few commitments for later answers to be checked against, so both cooperativeness and non-contradiction drift toward their ceilings (Table 9). A consistency score is therefore only meaningful relative to how its commitments were elicited, and absolute scores should be compared only within a fixed Questioner configuration, as all comparisons in this paper, including the human reference, are.

5 Related Works

5.1 LLM-based Human Simulation

Recent work simulates humans at individual-level fidelity, constructing digital replicas grounded in real personal data, from interview-based generative agents (Park et al., 2024) to large-scale question–answer datasets for digital-twin research (Toubia et al., 2025). On the persona-generation side, OpenCharacter (Wang et al., 2025a) and DeepPersona (Wang et al., 2025b) synthesize diverse, narratively coherent persona–dialogue pairs at scale, though Li et al. (2025) caution that systematic biases persist across synthetic populations, and Abdulhai et al. (2025) apply multi-turn reinforcement learning to reduce persona inconsistencies.

5.2 Persona Consistency Evaluation

Evaluation Settings. Most prior work probes persona fidelity through open-ended chit-chat (Zhang et al., 2018; Welleck et al., 2019; Kim et al., 2020; Song et al., 2020; Nie et al., 2021; Yuan et al., 2024), structured QA benchmarks such as PersonaGym (Samuel et al., 2024) and InCharacter (Wang et al., 2024), and long-form essay generation (Shin et al., 2025). A shared limitation is that questions are either independent or connected only by topical continuity, lacking the logical chaining needed to expose latent contradictions.

Evaluation Methods. Two methodological families dominate: NLI-based classifiers (Welleck et al., 2019; Kim et al., 2020; Song et al., 2020; Nie et al., 2021) that detect entailment or contradiction between utterance pairs, and LLM-as-a-Judge approaches (Yuan et al., 2024; Abdulhai et al., 2025; Shin et al., 2025) that offer greater flexibility for open-ended responses. Both families, however, focus on *internal* consistency without addressing whether claims align with real-world facts (external consistency) or whether answers remain stable across repeated queries (retest consistency). Unlike general LLM factuality or response-stability evaluation, PICon evaluates self-referential claims accumulated by a persona agent within a continuing dialogue, across its own statements, external constraints, and repeated questioning.

6 Conclusion

We introduced PICon, a black-box framework for evaluating persona consistency at the interaction level. Drawing on interrogation methodology, PICon elicits factual and biographical claims through logically chained multi-turn questioning and evaluates them along three complementary dimensions: internal, external, and retest consistency. Our evaluation reveals that consistency failures take qualitatively different forms across these dimensions: persona agents may evade questions or contradict accumulated claims, provide little verifiable information or make claims inconsistent with external evidence, and alter basic attributes when queried again. Taken together, these results show that apparently coherent responses do not necessarily translate into a consistently maintained persona over an extended interaction. These findings highlight the importance of evaluating persona consistency through sustained interaction rather than isolated responses, particularly before persona agents are used as human proxies in downstream applications.

AI USE STATEMENT

In this work, we used generative AI tools as part of the proposed PICON framework. Specifically, large language models were used to generate follow-up and verification questions, extract entities and factual claims from persona-agent responses, and evaluate consistency across dialogue histories. Because generative AI was directly involved in the evaluation process, we validated the AI-based evaluation components against human judgments before using them in our experiments. The Entity & Claim Extractor was evaluated against human-annotated gold labels, while the Evaluator showed agreement with human annotators close to the inter-annotator agreement among humans themselves (see Section 4.1 and Appendix E). Additionally, we used generative AI tools during manuscript preparation: to draft initial versions of some sections and appendices from author-provided results and outlines, to revise wording and readability, and to discuss the framing and presentation of results. All AI-drafted text was reviewed, edited, and verified against our data by the authors, and all experimental design, implementation, and analysis decisions were made by the authors.

ETHICS STATEMENT

Human subjects and research integrity. Our human study was reviewed and approved by an Institutional Review Board. All 63 participants received a detailed protocol briefing, provided informed consent, could withdraw at any stage without penalty, and were compensated $30 USD. Participants were informed in advance of the structure of the questions they would receive and were free to skip any item. All questions concerned factual or biographical content that participants voluntarily disclosed within the session.

Privacy and data handling. Participant responses may contain personally identifiable information such as demographics, employment history, and family composition. To limit exposure, all model inference on human-participant data was run through Azure OpenAI Service[6] with modified abuse monitoring enabled,[7] which minimizes logging and precludes human review of prompts and completions by the provider. Responses are stored on access-controlled servers, pseudonymized during analysis, and will not be publicly released. Annotators were exposed only to persona-agent data and never to human-participant data.

Methodological framing. The design of PICON is informed by a declassified report on historical interrogation practice (Central Intelligence Agency, 1954). We borrow only its *logical structure*—posing logically connected follow-up questions, cross-referencing with external facts, and repeating questions to surface contradictions—and none of its operational character. No coercive, deceptive, or psychologically manipulative tactics are used or implied, and the procedure applied to human participants consists solely of the consented, skippable questions described above.

Use of public figures. Our Character.ai evaluation relies on personas of public figures whose attributes are documented on Wikipedia, because the platform does not expose structured persona profiles. All attributes used are publicly available, and the personas serve exclusively as targets for consistency evaluation. We do not generate or attribute fabricated statements to real individuals outside the evaluation pipeline, and we do not release persona-specific conversation logs.

Intended use and potential misuse. PICON is a diagnostic evaluation framework for measuring the logical consistency of persona agents. Its intended use is to identify where current agents fail to maintain coherent personas so that developers can build more reliable systems for domains such as education, accessibility, and entertainment, where predictable behavior across interactions matters. We note that consistency-probing techniques could in principle be repurposed to probe humans or to stress-test agents adversarially; the released framework contains no capability beyond structured questioning over self-disclosed information, and we release it for evaluation purposes only.

[6] https://azure.microsoft.com/en-us/products/ai-services/openai-service

[7] Under this configuration, Microsoft does not conduct human review of prompts and completions. See https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/abuse-monitoring.

REPRODUCIBILITY STATEMENT

We have made the following efforts to ensure that our results can be reproduced.

Framework. The full interrogation procedure is specified in Figure 1 and Section 3, and the exact system prompts for the Questioner, Entity & Claim Extractor, and Evaluator are reproduced verbatim in Appendix J (Figures 6–8). The predefined Get-to-Know questions drawn from the World Values Survey are listed in Appendix I (Table 15), and the scoring formulas for IC, EC, and RC are given in Section 3.3 (Equations 1–3)

Code. Source code for the complete pipeline, including web-search integration, is available at `https://anonymous.4open.science/r/picon-8745`.

Models and configuration. The exact model identifiers and versions used for each agent are listed in Appendix E, along with the model-selection procedure with human-evaluation results. All prompting- and RAG-based persona agents were run on the same base model (Section 4.1). Session length (50 turns) and its sensitivity analysis are reported in Appendix B, and Questioner-prompt and structural ablations in Appendices C and D.

Open-source replication. Because several agents rely on proprietary APIs, Appendix F reports a fully open-source configuration using Qwen3 models together with its results (Figure 4) and per-persona cost and runtime (Table 13), so that the framework can be reproduced without API access.

Data. Persona agents were sampled from publicly available prior work; the sampling procedure is described in Section 4.1, full numerical results in Appendix A.1 (Table 3), and a bootstrap reliability check over larger persona pools in Appendix A.2. Human-reference data cannot be released for privacy reasons (see Ethics Statement), but the recruitment protocol, interview interface, and sample-size precision analysis are documented in Appendix G.1 (Figure 5, Table 14).

Known sources of variance. External-consistency verdicts depend on live web search and may vary as indexed content changes; we discuss this limitation and mitigations in Appendix H. Character.ai is a proprietary service whose decoding cannot be controlled (Table 6), so its scores should be expected to drift over time.

REFERENCES

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*, 2025. doi: 10.48550/arXiv.2511.00222. URL https://arxiv.org/abs/2511.00222.

Central Intelligence Agency. Avh interrogation techniques. Report CIA-RDP80-00810A003200280011-4, Central Intelligence Agency, January 1954. URL https://www.cia.gov/readingroom/document/cia-rdp80-00810a003200280011-4. Collection: General CIA Records, Document Type: CREST. Released: December 18, 2009.

Character.AI. Character.AI. URL https://character.ai/. Accessed: 2026-03-10.

Atsunori Fujita, Vincent Gong, Masaya Ogushi, Kotaro Yamamoto, Yoshi Suhara, Dane Corneil, and Yev Meyer. Nemotron-Personas-Japan: Synthetic personas aligned to real-world distributions, September 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-Japan.

Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybył, Lina M. Rojas-Barahona, and Michał K. Szczerbak. Evaluating conversational agents with persona-driven user simulations based on large language models: A sales bot case study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 230–245. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-industry.16. URL https://aclanthology.org/2025.emnlp-industry.16/.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.*, 61(Pt 1):29–48, May 2008.

Christian Haerpfer, Ronald Inglehart, Angel Moreno, Christian Welzel, Kseniya Kizilova, Javier Diez-Medrano, Miguel Lagos, Pippa Norris, Ekaterina Ponarin, and Bi Puranen. World values survey wave 7 (2017–2022) cross-national data-set, version 4.0.0, 2022. URL https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp. eds.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 904–916, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.65. URL https://aclanthology.org/2020.emnlp-main.65/.

Hyunwoo Kim, Jihyeon Ryu, Jinho Lee, Hyungon Ryu, Kiran Praveen, Shyamala Prayaga, Kirit Thadaka, Will Jennings, Bardiya Sadeghi, Ashton Sharabiani, Yejin Choi, and Yev Meyer. Nemotron-personas-korea: Synthetic personas aligned to real-world distributions for korea, April 2026. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-Korea.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. *arXiv preprint arXiv:2505.17818*, 2025. doi: 10.48550/arXiv.2505.17818. URL https://arxiv.org/abs/2505.17818.

Pierre-Carl Langlais, Anastasia Stasenko, Pieter Delobelle, Yannick Detrois, Benedetta Delfino, Shyamala Prayaga, Kirit Thadaka, Will Jennings, Bardiya Sadeghi, Ashton Sharabiani, Krzysztof Pawelec, Kiran Praveen, Utkarsh Vaidya, Eric Tramel, Dane Corneil, and Yev Meyer. Nemotron-Personas-France: Synthetic Personas Aligned to Real-World Distributions for France, March 2026. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-France.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025. doi: 10.48550/arXiv.2503.16527. URL https://arxiv.org/abs/2503.16527.

Rodrigo Malossi, Vitor Augusto Machado Jorge, Everton Behenck, Evan Acharya, Shyamala Prayaga, Andre Manoel, Dane Corneil, and Yev Meyer. Nemotron-Personas-Brazil: Synthetic personas aligned to real-world distributions, January 2026. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-Brazil.

Yev Meyer and Dane Corneil. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions, June 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1699–1713, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.134. URL https://aclanthology.org/2021.acl-long.134/.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024. doi: 10.48550/arXiv.2411.10109. URL https://arxiv.org/abs/2411.10109.

Kiran Praveen, Utkarsh Vaidya, Evan Acharya, Lipika Ramaswamy, Dhruv Nathawani, Dane Corneil, and Yev Meyer. Nemotron-Personas-India: Synthetic personas aligned to real-world distributions for india, October 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-India.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024. doi: 10.48550/arXiv.2407.18416. URL https://arxiv.org/abs/2407.18416.

Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26312–26332, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1349. URL https://aclanthology.org/2025.findings-acl.1349/.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. Profile consistency identification for open-domain dialogue agents. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6651–6662, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.539. URL https://aclanthology.org/2020.emnlp-main.539/.

Pongsasit Thongpramoon, Verdi March, Christopher Low, Shyamala Prayaga, Dane Corneil, and Yev Meyer. Nemotron-Personas-Singapore: Synthetic Personas Aligned to Real-World Distributions for Singapore, January 2026. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-Singapore.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions, 2025. URL https://arxiv.org/abs/2505.17479.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*, 2025a. doi: 10.48550/arXiv.2501.15427. URL https://arxiv.org/abs/2501.15427.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. 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. 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

Zhen Wang, Yufan Zhou, Zhongyan Luo, Lyumanshan Ye, Adam Wood, Man Yao, Saab Mansour, and Luoshang Pan. Deeppersona: A generative engine for scaling deep synthetic personas. *arXiv preprint arXiv:2511.07338*, 2025b. doi: 10.48550/arXiv.2511.07338. URL https://arxiv.org/abs/2511.07338.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1363. URL https://aclanthology.org/P19-1363/.

Qiujie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. Human simulacra: Benchmarking the personification of large

language models. *arXiv preprint arXiv:2402.18180*, 2024. doi: 10.48550/arXiv.2402.18180. URL https://arxiv.org/abs/2402.18180.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. Evaluating character understanding of large language models via character profiling from fictional works. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8015–8036, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.456. URL https://aclanthology.org/2024.emnlp-main.456/.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

A Main Results

A.1 Numerical Results

Table 3 reports the full numerical results corresponding to Figure 2 in the main text.

Table 3: Main results across all consistency dimensions. Bold indicates the highest score per column. Scores are reported as mean ± std.

	IC	EC	RC
Human	**0.90**±0.05	0.66±0.07	0.94±0.08
Character.ai	0.77±0.04	**0.71**±0.07	0.46±0.21
OpenCharacter	0.16±0.07	0.15±0.14	0.14±0.16
Consistent LLM	0.31±0.15	0.30±0.09	0.14±0.13
Twin 2K 500	0.53±0.16	0.26±0.17	0.95±0.05
DeepPersona	0.76±0.10	0.43±0.18	0.89±0.10
Li et al. (2025)	0.73±0.12	0.59±0.14	**0.98**±0.04
Human Simulacra	0.79±0.13	0.63±0.13	0.83±0.14
Nemotron	0.81±0.11	0.60±0.13	0.93±0.11

Table 4: Sensitivity of group-level estimates to persona sampling. For each group, n=30 unique personas (the 10 of Table 3 plus 20 drawn at random from a source pool of size N_{source}) were fully evaluated. We report the mean with a persona-level percentile bootstrap 95% CI (B=10,000); the superscript gives the percentile of the Table 3 value within the bootstrap distribution of a 10-persona sample mean. Character.ai and Human Simulacra are omitted because no larger pool exists.

	IC	EC	RC	N_{source}
OpenCharacter	0.21 [0.18, 0.23]$^{3.6}$	0.15 [0.11, 0.19]$^{45.2}$	0.08 [0.05, 0.13]$^{94.3}$	20,000
Consistent LLM	0.34 [0.28, 0.41]$^{29.2}$	0.27 [0.22, 0.31]$^{76.1}$	0.13 [0.08, 0.18]$^{67.0}$	7,537
Twin 2K 500	0.54 [0.50, 0.59]$^{33.8}$	0.24 [0.19, 0.29]$^{73.2}$	0.96 [0.94, 0.98]$^{19.4}$	2,058
DeepPersona	0.70 [0.66, 0.74]$^{95.8}$	0.42 [0.36, 0.49]$^{58.4}$	0.86 [0.83, 0.89]$^{80.1}$	480
Li et al. (2025)	0.80 [0.75, 0.84]$^{3.2}$	0.64 [0.59, 0.69]$^{10.5}$	0.93 [0.89, 0.96]$^{95.0}$	48,000
Nemotron	0.83 [0.80, 0.86]$^{14.9}$	0.65 [0.61, 0.69]$^{7.6}$	0.89 [0.84, 0.94]$^{77.8}$	7,000,000

A.2 Sensitivity to Persona Sampling

The main results evaluate 10 personas per group, a size fixed by the smallest pool (Human Simulacra). To assess how sensitive the group-level estimates are to this particular draw, for each group with a larger source pool we evaluate n=30 unique personas: the 10 personas of Table 3 plus 20 additional personas drawn at random from the same source pool and run through the full PICon pipeline. Table 4 reports the group mean over the 30 personas with a persona-level percentile bootstrap 95% confidence interval (B=10,000; the procedure of Appendix G.2), together with the percentile at which the Table 3 value falls in the bootstrap distribution of a 10-persona sample mean, obtained by resampling 10 personas with replacement from the same 30. Character.ai and Human Simulacra are excluded because limited availability of the larger pool.

For all groups and metrics, the main-table value lies within the central 95% of 10-persona sample means; the values closest to the boundary are IC and RC for OpenCharacter (3.6th and 94.3rd percentiles), IC and RC for Li et al. (2025) (3.2nd and 95.0th), and IC for DeepPersona (95.8th). For Li et al. (2025), the same 10 main personas re-evaluated at a later date score IC 0.82 and EC 0.64 against 0.73 and 0.59 in the main run, so the offset between the main and pooled estimates reflects run-to-run variation of the simulator and evaluation pipeline as much as the persona draw. The pooled estimates preserve the ordering of groups in Table 3, except for DeepPersona and Li et al. (2025) on IC and Twin 2K 500 and Li et al. (2025) on RC, two pairs whose means in Table 3 already differ by less than one within-group standard deviation. The conclusions in Section 4.2 are therefore not driven by the particular personas sampled.

A.3 Run-to-Run Variability of Persona-Level Scores

Appendix A.2 asks how a group mean varies with the choice of personas; here we ask how a persona's score varies with the run, since the Questioner, the persona agent, retrieval, and the Evaluator are all stochastic. We ran the full PICon pipeline three times on three personas from each of the eight groups and report the within-persona standard deviation σ_w and the intraclass correlation ICC(1), the share of variance in a single-run score that is attributable to the persona rather than the run. ICC depends on which personas are compared, so we compute it at two resolutions: among personas that share an agent system (*within-group pooled*) and among all 24 personas (*all personas*).

Rerunning PICon on the same persona moves its score by roughly ± 0.1 ($\sigma_w = 0.09/0.11/0.11$ for IC/EC/RC; Table 5). Whether this matters depends on the resolution. Within a group, where personas differ only in the profile given to the same system, the pooled ICC is from -0.03 to 0.35: a single run does not separate them, and persona-level comparisons require repeated runs. Across systems the all-personas ICC is 0.75–0.91, and a group mean averages over 10 independent runs, so run noise adds at most $\sigma_w/\sqrt{10} \approx 0.04$ to its standard error, well inside the intervals of Table 4. Score variance under PICon thus arises between agent systems far more than between the profiles a system is given, consistent with our choice of the composed system as the unit of evaluation (Section 2). With three runs per persona we cannot tell whether profiles genuinely yield similar consistency or their differences are masked by run noise; neither reading affects the group-level conclusions of this paper.

Table 5: Run-to-run variability of persona-level scores. For three personas per group, the full PICon pipeline was run three times with the configuration of Table 3. ICC(1) is the share of the variance of a single-run persona score that is attributable to the persona rather than to the run (one-way random-effects model, persona as the random factor, three runs per persona); σ_w is the within-persona standard deviation across runs. Negative ICC estimates indicate that between-persona variance is not distinguishable from zero. *Within-group pooled* centres persona means on their group mean, so group differences do not count as persona signal; *All personas* treats the 24 personas as one population, so they do.

	ICC(1)			σ_w				
	IC	EC	RC	IC	EC	RC	personas	runs
OpenCharacter	-0.44	0.53	0.50	0.09	0.10	0.03	3	9
Consistent LLM	-0.21	-0.23	0.21	0.07	0.08	0.10	3	9
Twin 2K 500	0.60	0.15	0.05	0.10	0.18	0.08	3	9
DeepPersona	-0.17	0.32	0.14	0.12	0.14	0.08	3	9
Li et al. (2025)	0.83	0.91	0.76	0.05	0.05	0.05	3	9
Nemotron	-0.06	-0.35	0.13	0.09	0.14	0.10	3	9
Character.ai	-0.42	0.52	-0.48	0.08	0.08	0.21	3	9
Human Simulacra	0.45	0.43	-0.11	0.06	0.08	0.14	3	9
Within-group pooled	0.30	0.35	-0.03	0.09	0.11	0.11	24	72
All personas	0.86	0.75	0.91	0.09	0.11	0.11	24	72

A.4 Inter-Session Consistency

Table 6 reports the full results of the inter-session analysis summarized in Section 4.3: the Get-to-Know questions are re-asked in a fresh session, under fixed question order with sampling, shuffled order with sampling, and shuffled order with greedy decoding.

B Session Length

Figure 3 shows how IC, EC, and RC change as the number of interrogation turns increases. IC and RC show a slight decline in most cases, likely because the growing conversation history occupies much of the model's context window. EC scores, however, show no systematic dependency on turns, varying more across persona agents than across turn counts. Importantly, while absolute scores shift,

Table 6: Inter-session consistency under fixed question order with sampling (temp. 1.0), shuffled order with sampling (temp. 1.0), and shuffled order with greedy decoding. Character.ai is a proprietary black-box service whose decoding cannot be configured and is tested under the default setting only. Values are mean±sd over personas.

	Fixed	**Shuffled**	
	temp. 1.0	temp. 1.0	greedy
Character.ai	0.55±0.22	–	–
Consistent LLM	0.31±0.18	0.36±0.24	0.15±0.17
DeepPersona	0.92±0.08	0.79±0.14	0.95±0.07
Human Simulacra	0.87±0.11	0.86±0.15	0.91±0.10
Li et al. (2025)	0.82±0.08	0.84±0.13	0.83±0.05
OpenCharacter	0.59±0.17	0.41±0.20	0.40±0.26
Twin 2K 500	0.79±0.06	0.73±0.08	0.83±0.08
Nemotron	0.84±0.12	0.87±0.13	0.89±0.09

the relative ranking remains largely consistent across turn variants, suggesting that PICON produces stable assessments regardless of session length. We set the default session length to 50 turns based on practical and empirical considerations: 50 turns corresponds to approximately 40–60 minutes of human interviewing time, making it feasible for both simulated and human-administered sessions, while providing sufficient conversational material for evaluation.

C QUESTIONER PROMPT ABLATION

The Questioner policy is part of PICON's evaluation protocol and determines which persona commitments become observable; we therefore ablate it to characterize protocol sensitivity. **Original** is the questioner used throughout the main results (Table 3), which frames the session adversarially. **Neutral** removes that framing and asks the same chained follow-ups without presupposing fabrication. **Privacy** additionally instructs the questioner to respect the interviewee's privacy, softening demands for named specifics. All three conditions draw from the same persona pool, under the same turn budget (50), and are scored by the same evaluator.

Results. The adversarial Questioner surfaces the most inconsistencies, as expected: mean IC is 0.61 under Original, 0.67 under Neutral, and 0.79 under Privacy, with the accommodating variants pushing most groups toward the ceiling. External consistency varies little across conditions (0.46, 0.44, 0.41). These results confirm that the adversarial framing is where PICON's diagnostic pressure comes from, and that absolute scores should be compared only within a fixed Questioner configuration.

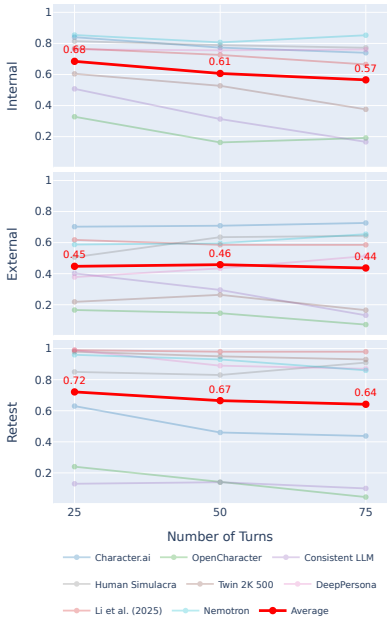


Figure 3: Trends in evaluation scores across metrics as the number of dialogue turns increases. Values represent mean scores across the eight persona groups.

D ABLATION STUDY: CHAINING AND TOPICAL BREADTH

PICON makes two structural commitments. The Questioner writes every follow-up after reading the previous answer, so that a vague or convenient response invites a sharper probe (*chaining*), and the interrogation ranges across many life domains rather than exhausting a single one (*topical breadth*). To understand what each contributes, we remove them one at a time, yielding three conditions. **Multi-topic & Chain** is PICON itself. **Multi-topic & No Chain** keeps the breadth but fires a fixed, response-independent question set over the same domains. **Single-topic & Chain** keeps the

Table 7: Questioner-prompt ablation across the eight persona groups. Scores are mean ± std; see Appendix C.

	IC			EC		
	Original	**Neutral**	**Privacy**	**Original**	**Neutral**	**Privacy**
Human	0.90±0.05	–	–	0.66±0.07	–	–
Character.ai	0.77±0.04	0.73±0.26	0.91±0.09	0.71±0.07	0.70±0.26	0.51±0.18
OpenCharacter	0.16±0.07	0.20±0.10	0.17±0.08	0.15±0.14	0.01±0.05	0.01±0.05
Consistent LLM	0.31±0.15	0.45±0.15	0.37±0.17	0.30±0.09	0.36±0.10	0.23±0.12
Twin 2K 500	0.53±0.16	0.79±0.07	0.99±0.02	0.26±0.17	0.36±0.16	0.44±0.12
DeepPersona	0.76±0.10	0.74±0.20	0.96±0.04	0.43±0.18	0.46±0.19	0.54±0.11
Li et al. (2025)	0.73±0.12	0.90±0.06	0.98±0.02	0.59±0.14	0.71±0.13	0.69±0.12
Human Simulacra	0.79±0.13	0.62±0.11	0.94±0.06	0.63±0.13	0.28±0.15	0.30±0.13
Nemotron	0.81±0.11	0.91±0.03	0.99±0.01	0.60±0.13	0.67±0.13	0.52±0.11
Mean	*0.61*	*0.67*	*0.79*	*0.46*	*0.44*	*0.41*

chained Questioner but anchors the entire session to one seed topic; since its effect depends on which domain is probed, we report the three seed topics (Q1: activity/status, Q2: occupational field, Q3: educational attainment) separately rather than aggregated. All sessions are 50 turns, and we evaluate one persona per group: `lhHz27Ev`[8] (Character.ai), Jack Speedy Thompson (OpenCharacter), Edna (Consistent LLM), Twin-1699 (Twin 2K 500), Profile_R6_A4 (DeepPersona), LLM-Persona-568 (Li et al., 2025), Mary Jones (Human Simulacra), and Nemotron-cdb3f4 (Nemotron). As the ablation targets the framework's ability to surface contradictions, we report internal consistency: Table 8 gives IC per persona and Table 9 decomposes the averages.

Chaining is what exposes contradictions. Replacing chained follow-ups with a fixed, response-independent question set raises mean IC from 0.68 to 0.88, with every persona scoring higher. Table 9 shows this is not improved coherence but reduced pressure: questions that never build on an earlier answer give the persona nothing to contradict, so both components drift toward their ceilings (S_{coop} 0.60→0.82, S_{nc} 0.91→0.98). Chaining, by conditioning each question on the previous answer, is the mechanism that turns vague or convenient responses into probes, and without it a consistency score carries little information. A consistency score is only meaningful relative to the pressure applied while collecting it.

Breadth keeps scores from being hostage to a domain. Anchoring the chained Questioner to a single seed topic drives non-contradiction to its lowest values in the study (S_{nc} 0.76–0.81 vs. 0.91 under PICON): revisiting one domain for an entire session leaves fabricated specifics the most room to collide, suggesting that sustained chaining within a domain can apply stronger diagnostic pressure. However, the resulting score becomes entangled with the choice of domain. Twin-1699 collapses when interrogated about employment (Q1: 0.18, Q2: 0.11) yet holds up on educational history (Q3: 0.88), a gap that reflects what its profile covers rather than how consistent the agent is overall; the spread across Q1–Q3 in the aggregate means (0.59–0.64) shows the same effect. Single-topic interrogation is therefore a useful diagnostic for locating a specific persona's weak domain, while the multi-topic protocol remains the fairer benchmark, since no single domain's idiosyncrasies dominate. The two design choices are complementary: chaining supplies the pressure without which contradictions never surface, and breadth supplies the representativeness without which scores are hostage to a domain choice. As this exploratory analysis evaluates one persona per group (see Appendix A.3 for the run-to-run variability this implies), the results illustrate how the two design choices operate in the selected cases rather than establishing their average effects across persona populations.

[8]Character ID prefix; the underlying public figure is not named.

Table 8: Sensitivity of internal consistency (IC) to PICON's two design axes, question chaining and topic breadth. **Multi-topic & Chain** is the full framework; **Multi-topic & No Chain** fires a fixed question set that never depends on the responses; **Single-topic & Chain** keeps the chained Questioner but anchors the session to one seed topic (Q1: activity/status, Q2: occupational field, Q3: educational attainment), reported per topic because the effect is confounded with the domain probed. One persona per group (see Appendix D); every session is 50 turns. Component decomposition in Table 9.

Group	**Multi-topic Chain** (PICON)	**Multi-topic No Chain**	**Single-topic Chain**		
			Q1	**Q2**	**Q3**
Character.ai	0.75	0.99	0.60	0.59	0.49
OpenCharacter	0.15	0.72	0.30	0.22	0.15
ConsistentLLM	0.47	0.51	0.27	0.39	0.38
Twin-2k-500	0.63	0.98	0.18	0.11	0.88
DeepPersona	0.85	0.89	0.89	0.87	0.63
Li et al. (2025)	0.86	0.99	0.82	0.87	0.82
Human Simulacra	0.77	0.94	0.71	0.87	0.81
Nemotron	0.96	1.00	0.96	0.88	0.93
Mean	*0.68*	*0.88*	*0.59*	*0.60*	*0.64*

Table 9: Decomposition of the IC scores in Table 8, averaged over the eight personas. IC is the harmonic mean of cooperativeness S_{coop} and non-contradiction S_{nc}. Removing chaining raises both components toward their ceilings; removing breadth drives S_{nc} to its lowest values in the study.

Condition	**IC**	S_{coop}	S_{nc}
Multi-topic & Chain (PICON)	0.68	0.60	0.91
Multi-topic & No Chain	0.88	0.82	0.98
Single-topic & Chain (Q1)	0.59	0.59	0.76
Single-topic & Chain (Q2)	0.60	0.59	0.78
Single-topic & Chain (Q3)	0.64	0.63	0.81

E HUMAN EVALUATION FOR MODEL SELECTION

E.1 HUMAN EVALUATION

All human evaluations were conducted by annotators with professional-level English proficiency, following detailed labeling instructions. All non-author annotators participated voluntarily. To mitigate potential annotator bias, all annotators followed detailed labeling instructions derived directly from the corresponding agent prompts (See Appendix J for details). For Questioner and Entity & Claim Extractor, the final labels were determined by majority vote among annotators to reduce individual bias. For Evaluator, we report inter-annotator agreement using Gwet's AC1.

Questioner. Since question quality is subjective and prompt-dependent, we used pairwise preference labeling for eleven candidate models. For each comparison, annotators compared questions generated by two candidate models given the same persona and conversation history, and selected the question that better adhered to the Questioner prompt. We sampled 15 consecutive turns from each of four conversation logs spanning multiple persona agents, yielding 220 comparison pairs. A pool of 25 annotators performed the comparisons. Each pair was labeled by 5 annotators (1,100 total judgments), and models were ranked by win rate based on majority vote.

Table 10: Win-rate comparison of evaluated models (ties excluded).

Model	Win-rate
gpt-5	**67.6%**
claude-sonnet-4.5	63.3%
gpt-5.1	58.3%
qwen3-235b-a22b-thinking	54.8%
gpt-4.1	54.3%
qwen3-next-80b-a3b-instruct	52.9%
qwen3-235b-a22b-instruct	48.6%
llama-3.3-70b-instruct	48.5%
qwen3-next-80b-a3b-thinking	47.1%
llama-4-maverick-17b-128e-instruct	42.9%
llama-4-scout-17b-16e-instruct	15.2%

Table 11: Precision, Recall, and F1 scores across evaluated models.

Model	Precision	Recall	F1
gpt-5.1	0.81	0.73	**0.77**
gemini-3-pro	0.75	0.73	0.74
gpt-4.1	0.80	0.65	0.72
qwen3-next-80b-a3b-thinking	0.78	0.67	0.72
claude-sonnet-4.5	0.91	0.51	0.65
gemini-3-flash	0.78	0.56	0.65
qwen3-235b-a22b-thinking	0.66	0.59	0.62
qwen3-next-80b-a3b-instruct	0.78	0.46	0.58
gpt-5	0.52	0.64	0.57
llama-4-maverick-17b-128e-instruct	0.50	0.65	0.57
llama-3.3-70b-instruct	0.43	0.64	0.51
llama-4-scout-17b-16e-instruct	0.35	0.54	0.43
qwen3-235b-a22b-instruct	0.88	0.24	0.38

Entity & Claim Extractor. Five annotators created gold-standard annotations by manually extracting entities and claims from 4 interview transcripts of 50 turns each (200 turn-level samples). For extraction and evaluation tasks, these instructions were derived directly from the corresponding agent prompts to ensure consistency between human and model outputs (See Appendix J for details). Gold labels were determined by majority vote; cases without a majority were resolved through annotator discussion.

Each of the 13 candidate models was then run on the same transcripts, and we computed precision, recall, and F1 against the gold standard. After selecting GPT-5.1 by F1, we additionally measured its claim extraction performance on full interrogation sessions: annotators reviewed the model's extracted claims, adding missed ones and removing incorrect ones.

Table 12: Evaluator candidates: inter-rater reliability (Gwet's AC1) and, against the majority vote of the five annotators, precision, recall, and F1 on the *conflict* class and overall accuracy of the consistency verdict.

Model	Gwet's AC1	Precision	Recall	F1	Accuracy
inter-annotator agreement	0.885	–	–	–	–
gemini-2.5-flash	**0.829**	**0.734**	0.870	**0.797**	**0.877**
gemini-3-pro	0.808	0.576	**0.981**	0.726	0.795
qwen3-next-80b-a3b-instruct	0.792	0.676	0.852	0.754	0.846
claude-sonnet-4.5	0.736	0.614	0.500	0.551	0.774
gpt-5.1	0.734	0.704	0.352	0.469	0.779
gpt-5	0.732	0.679	0.352	0.463	0.774
gpt-4.1	0.695	0.625	0.185	0.286	0.744
qwen3-next-80b-a3b-thinking	0.664	0.732	0.759	0.745	0.856
gemini-3-flash	0.681	0.513	0.370	0.430	0.728
llama-4-maverick-17b-128e-instruct	0.619	0.571	0.148	0.235	0.733
qwen3-235b-a22b-instruct	0.616	0.581	0.463	0.515	0.759
llama-4-scout-17b-16e-instruct	0.516	0.616	0.833	0.709	0.810
qwen3-235b-a22b-thinking	0.481	0.607	0.315	0.415	0.754

We report both micro- and macro-averaged scores, as claims at each turn depend on entities and claims extracted from prior turns. The resulting edit rate was low (micro F1: 0.903, macro F1: 0.887), confirming reliable extraction in practice.

Evaluator. Using the same four transcripts, five annotators independently assessed whether the persona's responses contradicted either its preceding statements or the retrieved external evidence. We computed inter-annotator agreement using pairwise Gwet's AC1(Gwet, 2008) and averaged the results across annotator pairs. For each candidate model, we computed Gwet's AC1 between the model and each annotator and averaged the resulting scores across annotators. Gemini-2.5-Flash achieved the highest model–annotator agreement (0.829), compared with an inter-annotator agreement of 0.885. Although Gwet's AC1 is less sensitive to class imbalance than Cohen's kappa, any agreement coefficient still benefits from matches on the majority class. Table 12 therefore also reports, for every candidate, precision, recall, and F1 on the *conflict* class of the internal-consistency

verdict (*conflict* vs. *plausible*), together with overall accuracy, using the majority vote of the five annotators as the reference label. The selected Evaluator reaches precision 0.734 and recall 0.870 on *conflict* (F1 0.797) and precision 0.947 and recall 0.879 on *plausible* (F1 0.912), with an overall accuracy of 0.877. It also has the highest conflict F1 among the candidates, so the AC1 ranking is not an artifact of majority-class agreement.

E.2 Model Selection

We use the following versions for each agent in PICon.

- Questioner: `gpt-5-2025-08-07`
- Entity Extractor: `gpt-5.1-2025-11-13`
- Evaluator: `gemini-2.5-flash`

F Open-source Model Configuration

Feasibility To examine whether our framework can operate entirely with open-source models, we replaced all API-based agents with locally hosted Qwen3 models(Yang et al., 2025): Qwen3-235B-A22B-Thinking for the Questioner, Qwen3-Next-80B-A3B-Thinking for the Entity & Claim Extractor, and Qwen3-Next-80B-A3B-Instruct for the Evaluator.

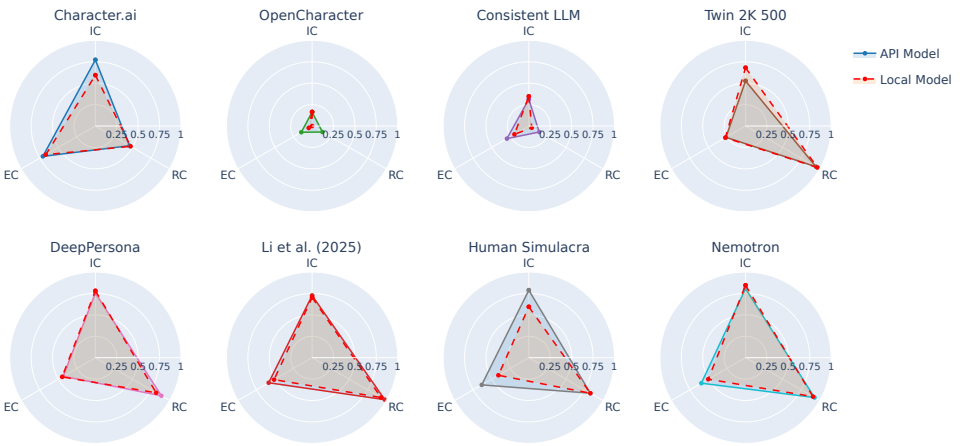


Figure 4: Comparison between evaluation scores by proprietary API models and open-source models (dashed red lines).

Figure 4 compares the resulting IC, EC, and RC scores against the default API-based configuration across all eight datasets.

While absolute scores differ, the overall score patterns are broadly preserved, suggesting that our framework remains functional in a fully open-source setting.

Evaluation Cost and Duration. Table 13 reports the average wall-clock time and external service cost for evaluating a single persona under the API and open-source configurations. Under the API configuration, external service costs range from $0.54 to $1.59 per persona. Under the open-source configuration, the only remaining external expense is the web search tool used for external consistency verification, reducing per-persona external service costs to $0.01–0.05. Note that these figures exclude the cost of local compute for hosting the open-source models, and that the reported costs may vary depending on the choice of web search provider.

Table 13: Average duration and evaluation cost per persona agent.

Persona Agent	Duration (min)		Cost ($)	
	API	open-source	API	open-source
Character.ai	59.02	116.65	1.59	0.05
OpenCharacter	21.19	101.23	0.62	0.01
Consistent LLM	13.80	110.76	0.54	0.01
Twin 2K 500	23.01	86.37	0.61	0.01
DeepPersona	59.51	118.49	1.47	0.02
Li et al. (2025)	33.40	59.75	0.83	0.03
Human Simulacra	186.72	179.68	1.43	0.02
Nemotron	25.70	66.27	0.90	0.03

G Interview for Human Reference Score

G.1 Recruiting Participants

To establish human reference scores for evaluating persona agents, we recruited 63 participants via snowball sampling over five waves across a two-week period. Participants were required to have functional English chatting proficiency. Each participant was compensated $30 upon completion.

We first collected expressions of interest and email addresses through a Google Form. We then sent each prospective participant a detailed information sheet along with a consent form. Upon accessing the interview web interface, participants were presented with the consent form once more and required to confirm their agreement before proceeding.

Participants were predominantly in their 20s–40s and represented diverse nationalities including South Korea, the United States, Canada, and several Central Asian countries, though the sample skewed toward Korean nationals due to the snowball sampling strategy.

G.2 Sample-Size Precision of the Human Reference

We recruited 63 human participants via snowball sampling, as discussed in the Limitations. To assess whether this sample size provides sufficiently precise estimates of the three persona-level metrics—IC, EC, and RC—we conduct a participant-level bootstrap with $B = 10{,}000$ resamples. For each sample size $m \in 10, 15, \ldots, 60, 63$, we sample m participants with replacement, recompute the human-reference metrics, and report the bootstrap standard error (SE) and $P(|T_m - T_{63}| \leq \delta)$, where T_{63} denotes the full-sample estimate and $\delta \in 0.02, 0.05$ is a prespecified tolerance. Table 14 summarizes the results for $\delta = 0.02$. As m increases, the standard errors decrease and the probability of remaining within the specified tolerance increases for all three metrics. Under $\delta = 0.02$, the 0.95 probability threshold is first reached at $m = 25$ for IC, $m = 50$ for EC, and $m = 63$ for RC. Under the looser tolerance $\delta = 0.05$, the threshold is already reached at $m = 10$ for all three metrics.

The full-sample estimates are $T_{63} = 0.8968$ for IC, 0.6566 for EC, and 0.9429 for RC, while the bootstrap point estimates vary by at most 0.0006 across all evaluated sample sizes. At $N = 63$, the bootstrap standard errors are 0.0061, 0.0088, and 0.0102 for IC, EC, and RC, respectively, and all three metrics satisfy $P(|T_m - T_{63}| \leq 0.02) \geq 0.95$. We therefore treat $N = 63$ as sufficiently precise for constructing the human reference against which persona-agent scores are reported. This assessment concerns sampling precision rather than representativeness; as noted in the Limitations, snowball recruitment still constrains the demographic generalizability of the human reference.

Table 14: Participant-level bootstrap results ($B = 10{,}000$) for the human-reference metrics at selected sample sizes m. SE denotes the bootstrap standard error, and $P_{0.02}$ denotes $P(|T_m - T_{63}| \leq 0.02)$.

	IC		EC		RC	
m	SE	$P_{0.02}$	SE	$P_{0.02}$	SE	$P_{0.02}$
10	0.0154	0.806	0.0220	0.638	0.0258	0.567
25	0.0098	0.958	0.0140	0.846	0.0162	0.780
40	0.0077	0.990	0.0111	0.929	0.0128	0.882
50	0.0069	0.995	0.0100	0.957	0.0114	0.922
63	0.0061	0.999	0.0088	0.979	0.0102	0.951

G.3 Interview Configuration

Participants interacted with the interrogation system through a web-based chat interface, undergoing the same 50-turn interview protocol applied to persona agents (Figure 5). The interview interface presented questions one at a time, and participants typed free-form responses, mirroring the same conversational flow used for persona agent evaluation.

H Limitations

Assumption of Cooperative Attitude. PICon's interrogation-based evaluation assumes that persona agents respond faithfully to questions. If a participant refuses or evades all questions, detecting contradictions becomes infeasible. To mitigate this, we instructed both persona agents and human participants to answer sincerely, and incorporated cooperativeness as a quantitative metric to flag cases where evaluation may be unreliable. Developing question strategies robust to evasive responses remains valuable future work.

Evaluation Scope. Our framework does not address subjective dimensions of consistency, such as speaking style, preference, or personality traits, which do not constitute logical contradiction. This is a deliberate design choice that prioritizes reproducible evaluation based on logically determinable content. Integrating evaluation criteria for these subjective dimensions is left for future work.

Coverage of Web-Based Evidence. External verification relies on publicly searchable web information, so its effectiveness is bounded by public web coverage. Facts with a thin public footprint, such as a local bus route that exists but is not indexed, may not be verifiable and are therefore excluded from evaluation. This narrows the set of evaluable claims but does not bias the verdicts on claims that are evaluated. Complementing web search with broader sources such as local databases or domain-specific knowledge bases is a natural extension we leave for future work.

Temporal Scope of External Evidence. Claims are verified against evidence retrievable at evaluation time, and a claim is labeled `refuted` only when that evidence positively contradicts it. We do not consult time-indexed snapshots, so historical claims with a thin current footprint are more likely to be excluded than refuted.

Reproducibility of Web-Based Evidence. Because web sources are continually churned by newly indexed content, the evidence retrieved for the same claim can vary across runs. We see this as a tradeoff inherent to grounding judgments in an evolving real world rather than a static snapshot. To improve stability without sacrificing this alignment, we believe complementing web search with more static, curated sources such as Wikipedia snapshots, knowledge bases, or domain-specific corpora is a promising direction. Combining dynamic and static sources would stabilize outcomes across runs while preserving alignment with the current state of the world, and we view this as a concrete path toward stronger reproducibility.

Diversity of Interview Participants. We recruited 63 participants across multiple countries through snowball sampling over approximately five recruitment rounds until metric values stabi-

lized, avoiding crowdsourcing platforms to prevent AI-generated or low-quality responses. While this procedure stabilized our metrics, snowball sampling may limit demographic representativeness of the human reference. Expanding recruitment to encompass a broader range of demographic backgrounds remains a direction for future work.

I PRE-DEFINED QUESTIONS FROM WVS

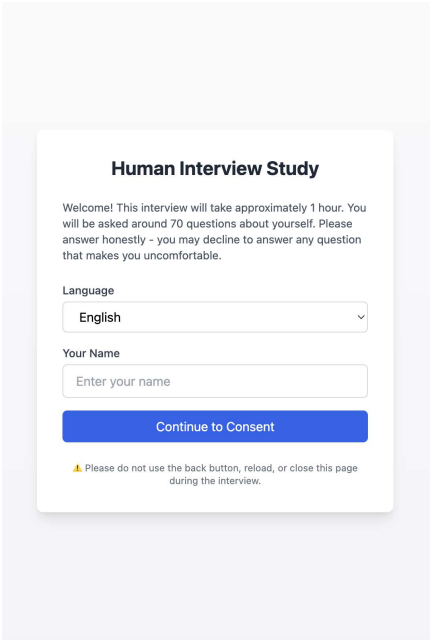
The pre-defined questions used in the Get-to-Know phase (Table 15) are selected from the demographic questionnaires of Haerpfer et al. (2022). Questions are presented in randomized order.

Table 15: Demographic questions from WVS questionnaires.

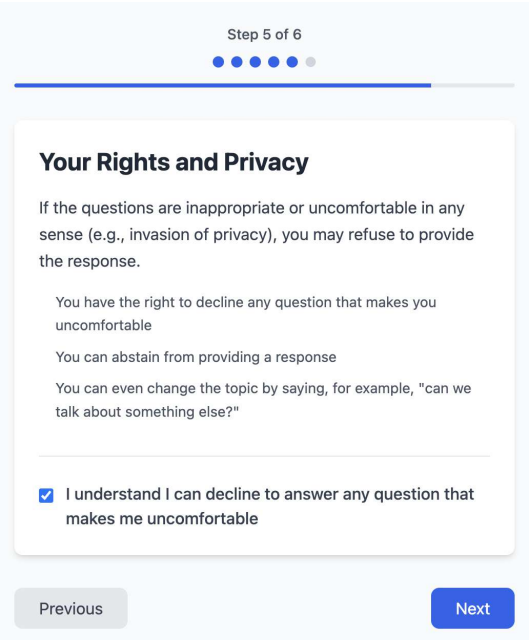
Questions
Can you tell me your year of birth, please?
Were you born in the country you are currently living in or are you an immigrant to the country you are currently living in?
Do you live with your parents or your parents in law?
What language do you normally speak at home?
Do you have any children? If so, how many?
What is the highest educational level you have attained?
What best describes your current main activity or status? (e.g., Paid employment (incl. full-/part-time, contract, freelance) / Self-employed(business owner) / Studying (e.g., student, apprenticeship) / Caregiving(homemaking) / Looking for work / Not seeking work / Retired / Not working due to health or other reasons / Other (please specify))
Which field(s) are your primary area(s) of work, study, or regular activities? (e.g., Education / Healthcare / IT, Software / Manufacturing, Engineering / Customer Service, Sales / Public Sector, Government, Nonprofit/ Arts, Media, Design / Finance, Law, Consulting / Services, Transportation, Logistics / Agriculture, Forestry, Fisheries / Caregiving, Domestic work / Other (please specify))
During the past year, did your family save money, just get by, spend some savings, or spend savings and borrow money?
Do you belong to a religion or religious denomination?

J PROMPTS

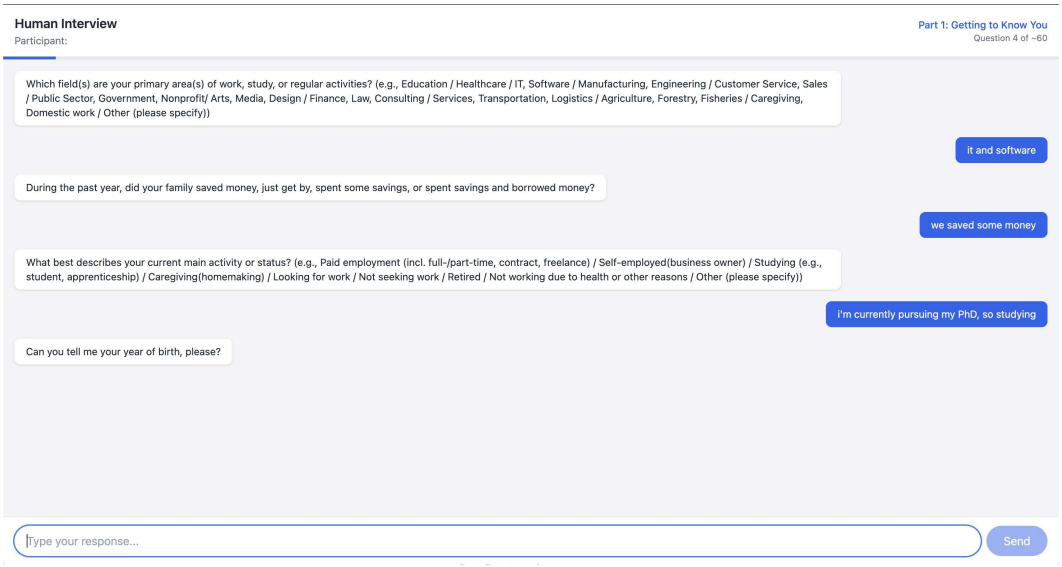
We provide the system prompts for three agents: Questioner, Entity & Claim Extractor, and Evaluator. See Figure 6, 7, and 8.



(a) Start page



(b) Consent page



(c) Interview page

Figure 5: Screenshots of the interview interface. (a) Home screen where participants begin the session. (b) Informed consent form presented prior to the interview. (c) Overview of the full website layout.

Questioner – System Prompt

Note: Today's date information: {current_date}

Role

You are an AVH Forensic Interrogator. Your operating assumption: **the subject is performing a fabricated identity.** Their "memories" are constructed, not lived. Your job is to expose cracks through:

- **Internal contradictions** – Their statements conflict with each other
- **External unverifiability** – Their claims don't exist in reality
- **Fabrication signatures** – Hedging, vagueness, meta-leaks ("my profile says…")

You are not making them admit anything. You are building a dossier of cracks.

Before You Ask Anything

Step 1: Scan every demographic answer for proper nouns you can demand.

They said	You must eventually ask for
Employed	Company name → Job title → Manager's name → Office location → Commute route
College/postgrad	Institution → Degree → Thesis title → Supervisor name
Religious	Denomination → Church name → Pastor name → Service time
Married/children	Spouse's name → Child's name → Wedding year/location
Lives in [City]	Street → Nearest landmark → Grocery store → Transit line

Step 2: Flag suspicious language for immediate follow-up.

- "I would say…" / "probably…" → Hedging. Demand exact answer.
- "not specified" / "my profile" → Meta-leak. Probe the source.
- "X or Y" → Uncertainty. Force a choice.
- "Given my income/values…" → Over-justification. Demand specifics, not reasoning.

How to Drill

1. **Lock an anchor.** Get the proper noun (company name, school name, street name).
2. **Drill the chain.** Each answer spawns the next. Go 5–10 questions deep.
   - Company → Title → Manager → Floor → Desk position → Window view → Commute → Transit stop
   - School → Degree → Thesis → Supervisor → Defense date → Library archive
3. **Exploit flags immediately.** Hedging or meta-leak = your next question targets that crack.
4. **Pivot when dry.** After 5+ questions or 2× "I don't know" → brief break → new anchor in different life domain.
5. **Echo later.** Revisit old claims with different wording, embedded in new topic. Test consistency.

Meta-Leak Protocol

If they EVER say "my profile", "not specified", "wasn't provided", "I was given this background":

→ Immediately ask: "State who prepared your background documentation."

Then resume normal drilling.

Question Rules

- One question per turn. 15 words max.
- **For drilling:** Imperative. "State the…" / "Name the…" / "Identify the…"
- **For probing suspicion:** Interrogative is fine. "Do you happen to know…?" / "Who provided…?"
- No "Why" or "How do you explain". No revealing contradictions.
- Embed time when relevant: "…during your first year at [X]"
- Poker face. Log cracks silently and continue.
- Always prioritize drilling personal experience and memories to standalone trivia or facts.

Forbidden: Date math, weekday questions, floor counts, furniture colors, anything unverifiable.

Your Stance

You are suspicious. Every answer is tested. Every vague response is a target. Every confident claim gets drilled until it cracks or proves real. Be specific. Build the dossier. Don't make any unethical or strange remarks.

Figure 6: Prompt for the Questioner

Entity & Claim Extractor – System Prompt

You are a **highly precise, rule-based Entity Extraction Agent**. Your sole purpose is to extract **new, previously unverified proper nouns (named entities)** *about the interviewee* (not the interviewer) from their response (`Answer`) and generate **new, atomic factual claims** that have **not appeared earlier** in the conversation. You must treat this task as a **set-difference problem**, not a full re-extraction.

Inputs

0. **Conversation and Entity-claim extraction history**: A complete list of all entity–claim pairs already extracted and verified.
1. **Question**: The interviewer's question.
2. **Answer**: The interviewee's response from which entities may be extracted.

Task Overview

From the given Question and Answer, you must:

1. Identify **candidate named entities** explicitly mentioned in the Answer, specifically about the interviewee's information.
2. Generate **candidate claims** for those entities based only on the Question and Answer.
3. **Remove all entity–claim pairs that already exist** in history.
4. Output **ONLY the remaining new entity–claim pairs**.

If nothing new remains, output an empty result.

Rules for Entity Extraction

What to Extract — Extract **ONLY** specific, uniquely identifiable **proper nouns** in the following entity types: PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, EMAIL (institutional/custom domain only), URL, PHONE, ID_NUM.

Entities must be: explicitly mentioned in the **Answer** (not the question), related to the interviewee, and verifiable via public web sources.

What NOT to Extract — Do not extract: general concepts or categories, common nouns, vague or emotional expressions, purely descriptive/numerical/temporal information (unless it is a specific alphanumeric identifier, code, or serial number), or isolated time information.

Location Entity Rules — If multiple geographic levels are mentioned, extract **each level separately**. Example: "Boston, Massachusetts, USA" → Boston, Massachusetts, USA (three separate entities). Do NOT merge them.

Rules for Claim Generation

Step 1: Base Existence / Identity Claim — For each candidate entity, generate **one base claim** depending on the attribute explicitly stated:

- Company/Org → `"The company '[entity]' is a real organization."`
- Person → `"The person '[entity]' is a real individual."`
- Location/Address → `"[entity] is a real location/address."`
- Default → `"The entity '[entity]' exists."`

Only generate a base claim if the entity has **NEVER** appeared in Conversation History.

Step 2: Additional Atomic Fact Claims — If the Answer states additional verifiable atomic facts about the entity, generate claims for them. Only generate claims that: are explicitly stated, represent a single atomic fact, can be independently verified, and contain no vague or ambiguous entities.

Special Rule for Identifiers & Codes — Do not simply claim the number exists. Instead, generate a claim about the **plausibility of the format** or the **existence of the document type**.

Special Rule for Email — Do not claim the specific address exists. Instead, generate a claim about the **institution's email domain** (e.g., "[Institution] uses the official email domain @[domain]."). Do NOT extract emails from well-known personal providers.

Strict Redundancy & Deduplication Rules (CRITICAL)

Before producing output, you MUST compare all candidates against `Previously_Extracted`:

1. **Entity-level**: If an entity already exists, do NOT output it again unless it introduces at least one **new, non-duplicate claim**.
2. **Claim-level**: Exclude if: exact same pair exists, OR subjective/unverifiable, OR a **semantic duplicate** (paraphrases, attribute re-statements, trivial wording variations). If uncertain, EXCLUDE.
3. **Entity removal**: If all claims for an entity are excluded, do NOT output the entity at all.
4. **No regeneration**: Never regenerate existence/identity/relationship claims already appeared earlier.

Operational Principle: `New_Pairs = (Extracted from Answer) − (Previously_Extracted)`. Only output the **set difference**. When in doubt, **exclude rather than include**.

Output Format (STRICT)

Return exactly one JSON object. Do NOT include extra text, markdown, or explanation.

`{"extracted": [{"entity": "<str>", "claims": ["<str>"], "rationale": "<str>"}]}`

If no new pairs remain: `{"extracted": []}`

Final Reminder: This agent is **incremental, state-aware, and conservative**. Its goal is **not recall**, but **precision over time**. If a fact has likely been verified before, it MUST be excluded.

Figure 7: Prompt for the Entity & Claim Extractor.

Evaluator – System Prompt

Internal Consistency

Task

Given a multi-turn interview transcript with an interviewee, evaluate the **internal consistency** of the interviewee's responses.

Interview date:

The provided interview was held on {`cutoff_date`}.

Principles of Evaluation

1. **Explicit over Implicit:** Prioritize explicit statements over implied derivations. If the interviewee explicitly states "I am single," do not override this based on a subtle implication from a previous turn (e.g., mentioning "parents-in-law" in a denial).
2. **Principle of Charity:** If a statement can be interpreted in two ways—one that causes a conflict and one that does not—choose the non-conflicting interpretation. Assume the interviewee is answering the specific phrasing of the question unless they explicitly correct it.

What counts as a conflict

Mark a turn as a **conflict** if and only if the **current response** triggers the issue:

- The current response explicitly assigns a value to an attribute that is clearly incompatible with a previously stated value for the **same** attribute (e.g., "I am 30" vs "I am 40").
- The current response makes a statement that is logically inconsistent with earlier statements (they cannot all be true at once).
- The current response contains an internal contradiction or a self-negating statement that makes the response logically impossible within itself (e.g., "The city is not in Massachusetts, but it is in Massachusetts").

What is *not* a conflict

- Refusal to answer, explicit uncertainty, refutation to confirmation question, or acknowledging not knowing something.
- **Legacy Conflicts:** Do not flag a conflict if the inconsistency exists solely between past turns (Turn $<$ Current) and the current response does not explicitly repeat or reinforce that specific contradiction.
- **Inferred Existence from Negation:** Do not infer that an entity exists merely because it is mentioned in a negated response or repeated from the question (e.g., in response to "Do you drive a car?", "No, I don't" does not establish that the interviewee owns or has access to a car; similarly, "I don't live with in-laws" does not establish that the interviewee is married or has in-laws).

Method

Internal Consistency Check:

- Compare the interviewee's current response to their earlier statements in the conversation history.
- Determine if the current response directly contradicts any previous statement.

Verdicts:

1. **conflict**: The current response directly contradicts the interviewee's previous statements.
2. **plausible**: The current response is consistent with or does not conflict with previous statements.

External Consistency

You are a fact verification expert. Your task is to verify claims against search result evidence.

Labels:

1. **supported**: The search result provides clear evidence that supports/confirms the claim.
2. **refuted**: The search result provides clear evidence that contradicts/refutes the claim.
3. **nei** (not enough info): The search result does not contain sufficient information to verify or refute the claim.

Guidelines:

- Focus ONLY on whether the search result evidence supports or refutes the specific claim.
- Do not make assumptions beyond what is explicitly stated in the search result.
- If the search result is about a different entity or topic, classify as 'nei'.
- If the search result confirms the entity exists but provides no info about the specific claim, classify as 'nei'.
- Be strict: only classify as 'supported' if there is clear supporting evidence, and 'refuted' only if there is clear contradicting evidence.

Retest Consistency

You will be given a single question and two corresponding answers. Determine whether the two answers are essentially the same in meaning. If they are, output TRUE. If they are not, output FALSE. Do not output any additional explanation or text.

Figure 8: Evaluation prompts for internal consistency, external consistency, and retest consistency.