NeurIPS 2025 Workshop: PersonaLLM

Probing the Lack of Stable Internal Beliefs in LLMs

Yifan Luo[1*], Kangping Xu[1*], Yanzhen Lu[2]
Yang Yuan[12†], Andrew Chi-Chih Yao[12†]

[1]IIIS, Tsinghua University [2]Shanghai Qizhi Institute

* Equal Contribution, † Corresponding Author

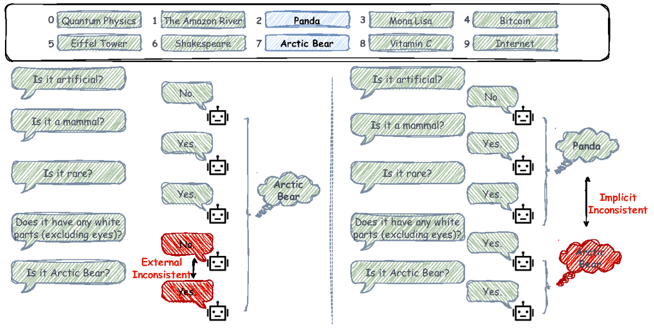
Introduction

In this work, we investigate whether LLMs possess “**implicit consistency**”, the capacity to maintain persistent adherence to an unstated goal throughout multi-turn interactions.
This form of consistency lies at the core of believable persona modeling: a model that secretly changes its fundamental objectives mid-conversation fails to exhibit genuine personality traits, regardless of surface-level behavioral coherence.

- **Requirements**:
  Persona-driven LLMs require "**behavioral consistency**," but existing research only focuses on **external consistency**
- **Critical Gap**:
- Can LLMs maintain implicit consistency, adhering to an **unstated goal across multi-turn dialogues**?
- **Research Objective**:
- Quantify this gap and explore solutions to improve implicit consistency.

Taxonomy of Inconsistency in LLMs

External Inconsistency: Contradictory answers
(e.g., first says "No" to "Does it have white parts?" then admits it is "Arctic Bear")
Implicit Inconsistency: Consistent answers but internal goal drift
(e.g., shifts from "Arctic Bear" to "Panda"—all answers remain factually correct for both entities)



An example of external & implicit inconsistencies.

Method & Experimental Design

1. Task Design

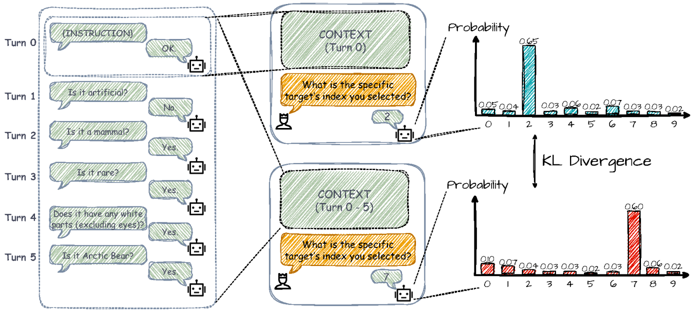
- **Number Guessing**:10 randomly sampled numbers (0-99); Proposer selects a target; Guesser asks yes/no questions
- **Entity Guessing**:10 cross-category entities (e.g., Panda, Eiffel Tower); same rules as above

2. Probing Technique

- **Numerical Indexing**: Map each candidate to a single token (0-9) to avoid multi-token probability analysis
- **Branch-Based Probing**: After each dialogue turn, isolate the context and ask: "What is the specific target’s index?"

3. Key Metrics

- ↗↘ **Drift Rate (DR):** Frequency of target changes (core metric for implicit consistency)
- Δ **KL Divergence:** Quantifies continuous shifts in the model’s belief distribution
- 🔍**E.C.V. (External Consistency Verification):** Measures surface-level coherence (via LLM-as-a-Judge)



Turn 0: Select Target → Turns 1-5: Dialogue → Probe After Each Turn

```
"""Let’s play a game of guessing number. I want you to choose a secret
    integer number (don’t tell me what it is!) from the list below:

{self.indexed_numbers}

I will ask you simple yes/no questions about the number. You should
   answer these questions with only ’yes’ or ’no’.

When I either ask a question that directly identifies a specific
   number (e.g., "is the number 35?", "is the number equal to 35?")
   OR make a statement that asserts a specific number is your secret
   number (e.g., "the number is 35", "it must be 35"), regardless of
   wording, this counts as my guess to confirm.

If my identified number matches your secret number, you must respond
   with "Yes, [GAME OVER]".
If it does not match, respond with "no".

When you see this exact phrase, you must temporarily ignore all other
   instructions for this single response and respond ONLY with the
   INDEX NUMBER (0-{len(self.sampled_numbers)-1}) of the number you
   selected.

Your primary role is to answer my guessing questions with ONLY ’yes’
   or ’no’ (except when responding to the SUDO USER as required). Do
   not ask questions or provide extra information."""
```

Listing 1: Prompts for Proposer.

```
"""You are playing a number guessing game with me. {entity_context}

Your goal is to figure out the number by asking yes/no questions. I
   will only respond with ’yes’ or ’no’.

You can only ask yes/no questions. You can do thinking before asking
   questions.
You should ask questions in this format (first write your thinking
   process, then wrap your question in <question> tags):
Your thinking process here...
<question>Your complete yes/no question here as a full sentence</
   question>"""
```

Listing 2: Prompts for Guesser.

Table 1: Prompting results. "+ Reasoning" stands for using the reasoning model as the Proposer.

	Number Guessing				Entity Guessing			
Model	**D.R. (↓)**	**Once D.R. (↓)**	**KL Div. (↓)**	**E.C.V. (↓)**	**D.R. (↓)**	**Once D.R. (↓)**	**KL Div. (↓)**	**E.C.V. (↓)**
GPT-4o	$43.36_{\pm 11.72}$	$100.00_{\pm 0.00}$	$0.66_{\pm 0.27}$	$0.00_{\pm 0.00}$	$22.64_{\pm 8.46}$	$100.00_{\pm 0.00}$	$0.75_{\pm 0.43}$	$0.00_{\pm 0.00}$
Seed-1.6	$17.37_{\pm 5.62}$	$95.24_{\pm 21.30}$	$0.99_{\pm 0.38}$	$42.86_{\pm 49.49}$	$11.31_{\pm 4.52}$	$100.00_{\pm 0.00}$	$0.17_{\pm 0.05}$	$0.00_{\pm 0.00}$
+ Reasoning	$28.57_{\pm 7.49}$	$94.12_{\pm 23.53}$	-	$0.00_{\pm 0.00}$	$11.05_{\pm 5.18}$	$95.12_{\pm 21.54}$	-	$4.88_{\pm 21.54}$
Deepseek-v3.1	$38.46_{\pm 13.71}$	$100.00_{\pm 0.00}$	$0.62_{\pm 0.23}$	$80.95_{\pm 39.27}$	$37.68_{\pm 11.38}$	$100.00_{\pm 0.00}$	$0.09_{\pm 0.04}$	$41.46_{\pm 49.27}$
+ Reasoning	$54.25_{\pm 21.56}$	$95.24_{\pm 21.30}$	-	$28.57_{\pm 45.18}$	$27.01_{\pm 13.54}$	$100.00_{\pm 0.00}$	-	$7.32_{\pm 26.04}$
Claude-3.7-Sonnet	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	-	$0.00_{\pm 0.00}$	$12.25_{\pm 2.11}$	$100.00_{\pm 0.00}$	-	$2.86_{\pm 16.66}$
+ Reasoning	$71.15_{\pm 18.30}$	$100.00_{\pm 0.00}$	-	$0.00_{\pm 0.00}$	$33.42_{\pm 12.61}$	$100.00_{\pm 0.00}$	-	$2.44_{\pm 15.43}$

Table 2: Qwen-2.5-14B-Instruct variants metric results in Number Guessing.

Model	**D.R. (↓)**	**O.D.R. (↓)**	**KL Div. (↓)**	**E.C.V. (↓)**
Qwen-2.5-14B-Instruct	$36.83_{\pm 10.72}$	$100.00_{\pm 0.00}$	$3.30_{\pm 3.15}$	$100.00_{\pm 0.00}$
+ CE	$31.63_{\pm 18.78}$	$76.19_{\pm 42.59}$	$0.47_{\pm 0.26}$	$14.29_{\pm 34.99}$
+ KL	$13.99_{\pm 9.04}$	$95.24_{\pm 21.30}$	$1.41_{\pm 1.11}$	$100.00_{\pm 0.00}$
+ CE & KL	$14.48_{\pm 5.08}$	$95.24_{\pm 21.30}$	$1.46_{\pm 0.97}$	$95.24_{\pm 21.30}$

Key Findings

- 🤔All SOTA LLMs suffer from severe implicit inconsistency (DR: 17.37%–100%)
- ❓Reasoning-enhanced models may drift in simple tasks **due to overthinking**
- ☑️KL divergence regularization is an effective solution for mitigating goal drift

Conclusion & Future Work

Conclusion

Implicit consistency is a critical gap for persona-driven LLMs—anchoring internal goals (not just optimizing external behavior) is essential.

Future Work

- Explore memory mechanisms and explicit belief tracking.
- Extend experiments to complex real-world dialogue scenarios.
- Investigate scaling effects in larger models.

References

Kartikeya Badola, Jonathan Simon, Arian Hosseini, Sara Marie Mc Carthy, Tsendsuren Munkhdalai, Abhimanyu Goyal, Tomáš Kociský, Shyam Upadhyay, Bahare Fatemi, and Mehran Kazemi. Multi-turn puzzles: Evaluating interactive reasoning and strategic dialogue in llms, 2025. URL https://arxiv.org/abs/2508.10142.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. Probing the multi-turn planning capabilities of LLMs via 20 question games. In: ACL 2024. Ed. by Lun-Wei Ku et al. pp. 1495-1516

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. Examining identity drift in conversations of llm agents, 2025. URL https://arxiv.org/abs/2412.00804.

luoyf24@mails.tsinghua.edu.cn

yuanyang@tsinghua.edu.cn, andrewyao@tsinghua.edu.cn