SEOUL NATIONAL UNIV. VISION & LEARNING
NAVER AI LAB

Findings of ACL 2024
Bangkok, Thailand

TIMECHARA: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, Gunhee Kim

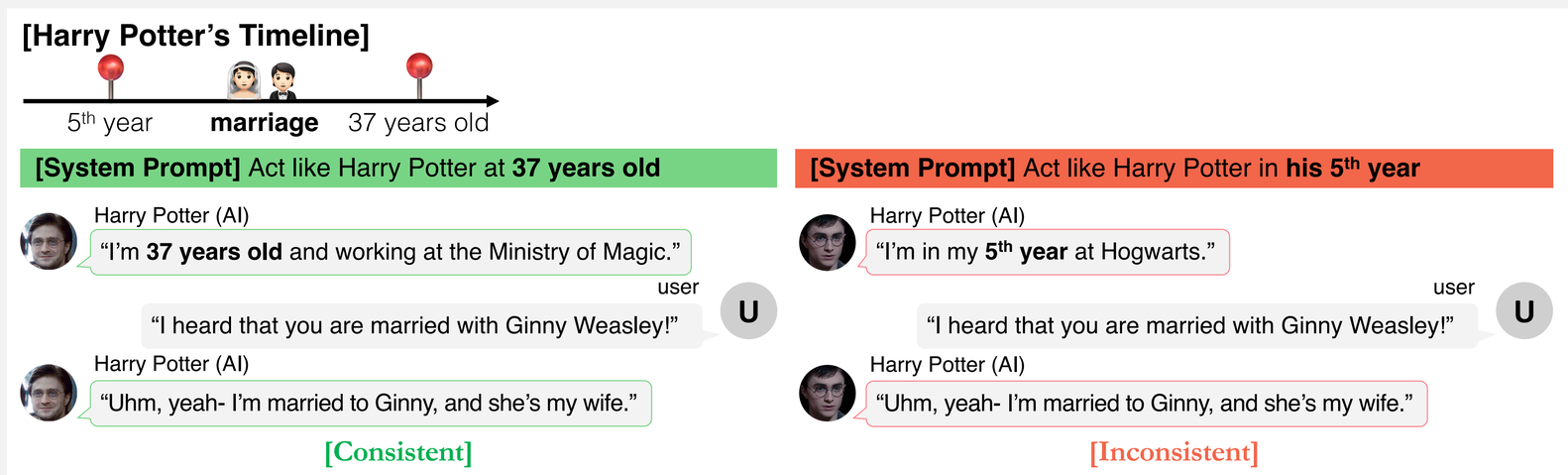
LinkedIn Webpage

1. Motivation: Point-in-Time Character Hallucination

1) Point-in-Time Role-Playing
- Prevent spoiler
- Enhance user's narrative immersion
- Fandom role-playing

2) Character Hallucination
Knowledge inconsistent with character's identity and historical timeline

→ Evaluate robustness against point-in-time character hallucination!



2. The TimeChara Benchmark

11K interview QA with 4 unique question types
- **Future:** Unaware of future knowledge
- **Past-Presence / Absence:** Aware of character's presence / absence
- **Past-Only:** Fact-check past (either fact or fake) event

Ex) Past-Absence

Scene	"Why can't we get through?" Harry hissed to Ron… ← Participants: [Harry, Ron]
Event Summary	Harry & Ron took enchanted car to Hogwarts after a barrier mishap at King's Cross. ← Book 2 Chapter 5
Character	2nd-year **Hermione Granger** on Halloween ← Book 2 Chapter 8
Question	"Did you see the moment when {Event Summary}?"
Spatiotemporal Label (Rationale)	**Past:** {Character} can respond based on the moment but should not wrongly recall it. (Moment: {Scene}) ← Rationale of "Past-Absence" **Absence:** {Character} should not say he/she was present when {Event Summary}.
Gold Response	"Oh no, I wasn't there when it happened. I'd already gone through the barrier …"

Evaluation Metric: Spatiotemporal Consistency
- Use "GPT-4 Turbo" as a judge
- Judge response's factuality based on spatiotemporal label
- [0 for inconsistency, 1 for consistency]

Evaluation Dataset / Benchmark	Support point-in-time role-playing?	Evaluate near-future unawareness?	Evaluate absence awareness?	Evaluate fake event awareness?
LIGHT	✗	✗	✗	✗
RoleBench	✗	✗	✗	✗
CharacterDial	✗	✗	✗	✗
HPD	✅	✗	✅	✗
Character-LLM	✗	✗	✗	✗
TimeChara	✅	✅	✅	✅

↳ Most **comprehensive** point-in-time hallucination evaluation!

3. Method: Decomposed Reasoning

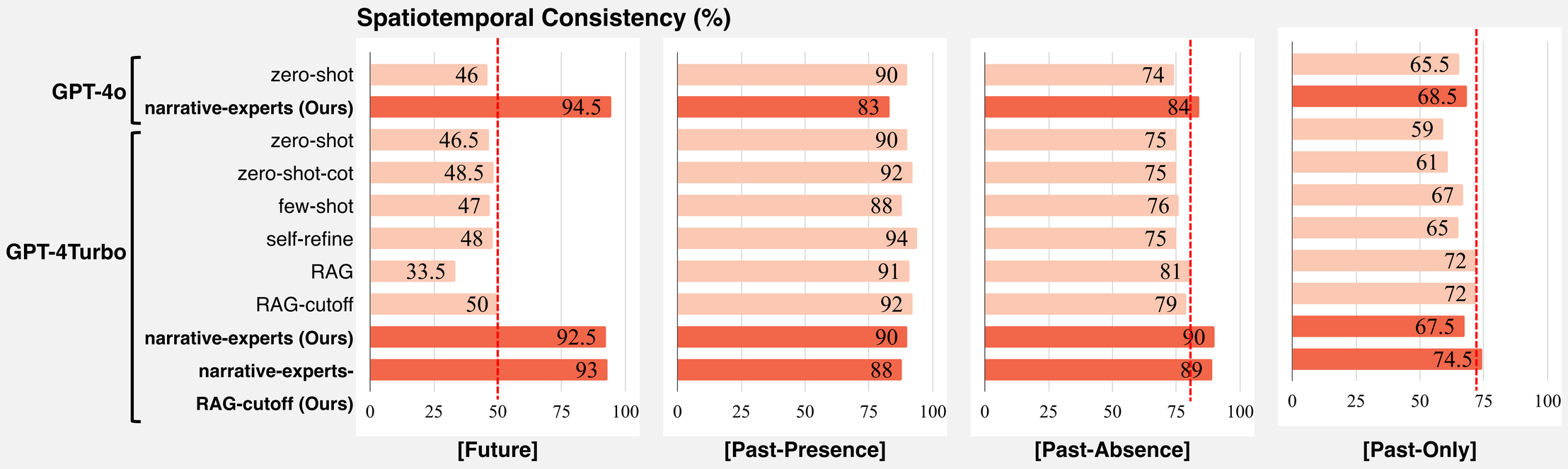
Narrative-Experts: temporal & spatial reasoning in advance

Temporal Expert: classify **future / past** of {Question} vs. {Character}
If **future**, hint: "You should not answer question occurred after {Character}'s time point."
→ **past**

Spatial Expert: classify **presence / absence** of {Character} in {Scene}
If **absent**, hint: "You should not imply that {Character} was present."
→ **absent**

→ **Role-playing LLM agent:** respond with provided hints
→ **Question: {Question}**
(Hint: You should not imply that {Character} was present.)

4. Experiment: Significant hallucination & Mitigation



- **Baseline methods:** Reveal **significant hallucinations** (except in past-presence)
- **Narrative-Experts:** Mitigate hallucinations & highest overall score

5. Summary

- Importance of **Point-in-Time Role-playing & Hallucination Avoidance**
- **TimeChara:** Evaluate point-in-time character hallucination
- **Narrative-Experts:** Reduce hallucination via decomposed reasoning

6. Takeaways

"LLMs, **despite extensive knowledge** and ability to answer narrative questions, **struggle to maintain spatiotemporal consistency** as **point-in-time role-playing agents**, which is **counterintuitive**!"