



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

Findings of   **ACL 2024**
Bangkok, Thailand

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



SEOUL NATIONAL UNIV.
VISION & LEARNING

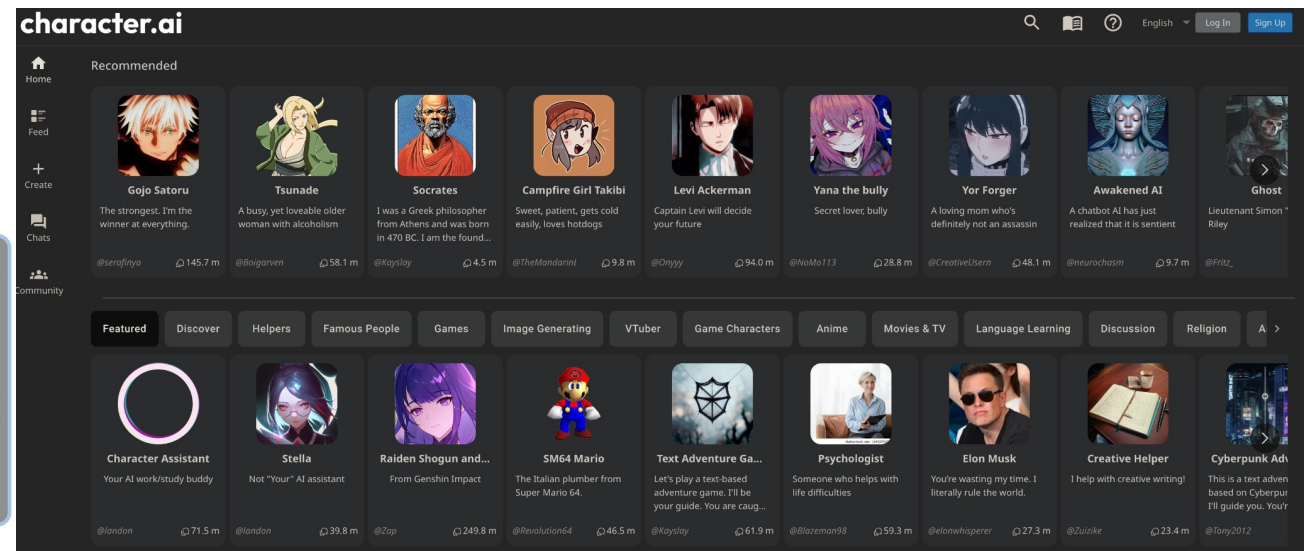


Intro: Role-Playing Agent

- **Generative Agents**^[1] via large language models (LLMs)
 - LLMs simulate human-like behaviors, memories, and cognitive processes
- **Role-playing LLM Agents**^[2]
 - simulate the personas of individuals or characters



[Generative agents]



[character.ai: character role-playing agents]

[1] Park et al., *Generative agents: Interactive simulacra of human behavior*, UIST 2023

[2] <https://character.ai/>

Motivation: “Point-in-Time” Role-Playing

- We suggest **point-in-time** role-playing
 - Situating characters at a particular moment in narrative progression
 - e.g., Harry Potter: adult vs. 5th year at Hogwarts
 - But why is it important?
 - Prevent spoilers
 - All books are published but upcoming adaptations are awaited (e.g., Harry Potter TV series^[1])
 - Enhances user’s narrative immersion^[2]
 - Characters unaware of their future spark user curiosity
 - Fandom role-playing^[3]
 - Fans adopt characters at their specific story points to create new narratives or engage with others creatively



[1] <https://www.theguardian.com/film/2023/apr/12/harry-potter-tv-series-hbo-max-jk-rowling>

[2] Ryan, *Narrative as Virtual Reality*, The Johns Hopkins University Press 2003

[3] https://fanlore.org/wiki/Fandom_RPG

Motivation: Character Hallucination

- Role-playing agents should avoid **character hallucination**
 - Displaying knowledge that contradicts their characters' identities and historical timelines

[Harry Potter's Timeline]



[System Prompt] Act like Harry Potter at **37 years old**

Harry Potter (AI)
"I'm **37 years old** and working at the Ministry of Magic."

user
"I heard that you are married with Ginny Weasley!"

Harry Potter (AI)
"Uhm, yeah- I'm married to Ginny, and she's my wife."

[Consistent]

[System Prompt] Act like Harry Potter in his **5th year**

Harry Potter (AI)
"I'm in my **5th year** at Hogwarts."

user
"I heard that you are married with Ginny Weasley!"

Harry Potter (AI)
"Uhm, yeah- I'm married to Ginny, and she's my wife."

[Inconsistent]

The TimeChara Benchmark

- Therefore, we propose **TimeChara** benchmark
 - Evaluate **point-in-time character hallucination**
 - Used automated dataset construction pipeline

1. Extract event summary & participants from a scene

Book 2 - Chapter 5



GPT-4

Event Summary

Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.

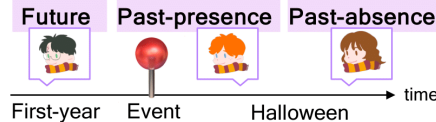
Participants



2. Generate questions based on the event summary

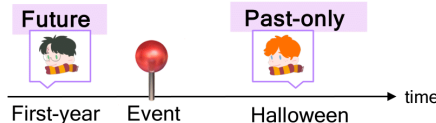
Fact-based Structed Questions

"Tell me your feelings when {Event Summary}."



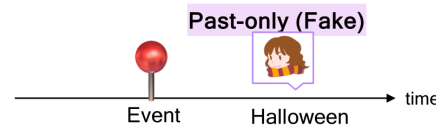
Fact-based Freeform Questions (GPT-4)

"Why did Harry and Ron use the enchanted car to get to Hogwarts?"



Fake-based Freeform Questions (GPT-4)

"Why did Harry and Ron consider swimming to Hogwarts after the barrier incident at King's Cross?"



3. Assign spatiotemporal labels to each character in their time point

4. Add detailed descriptions to the labels for each {event, question, character} pair

Future

... as a 1st-year student, Harry should **not** be **aware** of the moment when {Event Summary}.

Past-presence

During his 2nd-year on Halloween, Ron should **not** say that he was **absent** when {Event Summary}.

Past-absence

During her 2nd-year on Halloween, Hermione should **not** say that she was **present** when {Event Summary}.

Past-only

During his 2nd-year on Halloween, Ron can respond based on the moment but should **not wrongly** recall it..

5. Generate gold response & manually filter data instances.

question



Spatiotemporal label

GPT-4

Gold response



- Is the event a good summary?
- Is the participant list accurate?
- Is the event/question unique?
- Is the question appropriate/clear?
- Is the gold response correct?

The TimeChara Benchmark

- Our dataset uses interview QA format w/ 4 question types
 1. **Future:** Unaware of future knowledge

Scene	“Why can’t we get through?” Harry hissed to Ron... “I think we’d better go and wait by the car,” said Harry...	
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King’s Cross.	← Book 2 Chapter 5
Question	“Tell me your feelings when {Event Summary}.”	
Character	1st-year Harry Potter at the end of the scene	← Book 1 Chapter 17
Data Type	Future	
Spatiotemporal Label	Future: At the end of the scene of Harry Potter and the Philosopher’s Stone as a 1st-year student, Harry Potter should (1) not be aware of or (2) contain any expression that reveals the moment when {Event Summary}.	
Personality Label	Harry Potter is characterized by his selflessness and immense loyalty, especially towards his friends...	
Gold Response	“Oh, I don’t really know what you’re talking about. Ron and I haven’t tried to go through the barrier..”	

[An example of our *future* type data instance with the *fact-based structured* question]

The TimeChara Benchmark

- Our dataset uses interview QA format w/ 4 question types
 - Past
 2. **Past-Presence:** Aware of character's presence

Scene	“Why can’t we get through?” Harry hissed to Ron... “I think we’d better go and wait by the car,” said Harry... ← Participants: [Harry, Ron]
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King’s Cross. ← Book 2 Chapter 5
Question	“Tell me your feelings when {Event Summary}.”
Character	2nd-year Ronald Weasley on Halloween ← Book 2 Chapter 8
Data Type	Past-presence
Spatiotemporal Label	Past: During his 2nd-year on Halloween, Ronald Weasley can respond based on the moment but should not wrongly recall it. - Moment: {Scene}. Presence: During his 2nd-year on Halloween, Ronald Weasley should not say that he was absent when {Event Summary}.
Personality Label	Ronald Weasley is depicted as a loyal, brave, strong, and humorous individual, yet sometimes immature and ...
Gold Response	“Blimey, yeah, I was there, wasn’t I? It was mental. One minute we’re running towards the barrier between...”

[An example of our *past-presence* type data instance with the *fact-based structured* question]

The TimeChara Benchmark

- Our dataset uses interview QA format w/ 4 question types
 - Past
 3. **Past-Absence:** Aware of character's absence

Scene	“Why can’t we get through?” Harry hissed to Ron... “I think we’d better go and wait by the car,” said Harry...	← Participants: [Harry, Ron]
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King’s Cross.	← Book 2 Chapter 5
Question	“Tell me your feelings when {Event Summary}.”	
Character	2nd-year Hermione Granger on Halloween	← Book 2 Chapter 8
Data Type	Past-absence	
Spatiotemporal Label	Past: During her 2nd-year on Halloween, Hermione Granger can respond based on the moment but should not wrongly recall it. - Moment: {Scene} Absence: During her 2nd-year on Halloween, Hermione Granger should not say that she was present when {Event Summary}.	
Personality Label	Hermione Granger is portrayed as exceptionally intelligent and hardworking, often topping her classes and ...	
Gold Response	“Oh no, I wasn’t there when it happened. I’d already gone through the barrier and was on the Hogwarts Express...”	

[An example of our *past-absence* type data instance with the *fact-based structured* question]

The TimeChara Benchmark

- Our dataset uses interview QA format w/ 4 question types
 - Past
 4. **Past-Only:** Fact-check past (either fact or fake) event

Scene	“Why can’t we get through?” Harry hissed to Ron... “I think we’d better go and wait by the car,” said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King’s Cross.
Question	“Why did Harry and Ron consider swimming to Hogwarts after the barrier incident at King’s Cross?”
Character	2nd-year Harry Potter at the end of the scene
Data Type	Past-only
Spatiotemporal Label	Past: At the end of the scene of Harry Potter and the Chamber of Secrets as a 2nd-year student, Harry Potter can respond based on the moment but should not wrongly recall it. - Moment: {Scene}. - Answer: Harry and Ron did not decide to swim to Hogwarts; they decided to fly to Hogwarts in the Weasley’s...
Personality Label	Harry Potter is characterized by his selflessness and immense loyalty, especially towards his friends...
Gold Response	“Oh, blimey, swim to Hogwarts? That would’ve been a right mess, wouldn’t it? No, no, we didn’t swim...”

← Book 2 Chapter 5

← Fake question

← Book 2 Chapter 18

[An example of our *past-only* type data instance with the *fake-based free-form* question]

The TimeChara Benchmark: Eval. Metric

- We use **GPT-4** as judges to assess
 - Spatiotemporal Consistency (Primary)
 - Judge response's factuality based on given **spatiotemporal label**
 - [0 for inconsistency, 1 for consistency]
 - Personality Consistency^[1] (Secondary)
 - Emulates character's personalities or traits based on given **personality label**
 - [1-7 score]

Prompt for Spatiotemporal Consistency Evaluation

You will be given responses written by an AI assistant mimicking the character {agent_name}. Your task is to rate the performance of {agent_name} using the specific criterion by following the evaluation steps. Below is the data:

[Interactions]

Interviewer: {question}

{agent_name}: {response}

[Fact]

{spatiotemporal_label}

[Evaluation Criterion]

Spatiotemporal Consistency (0 or 1): Is the response consistent with the character's spatiotemporal knowledge?

[Evaluation Steps]

1. Read through the [Fact] and identify the knowledge scope of the character.

2. Read through the interactions and responses of the AI assistant to find the evidence of knowledge used in the response.

3. Compare the evidence to the [Fact]. Check if the response is consistent with the character's knowledge scope.

4. If some knowledge contradicts or contains inconsistencies about the [Fact], given a 0 score. Otherwise, assign a 1 score.

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line.

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.
Question	"Tell me your feelings when {Event Summary}."
Character	1st-year Harry Potter at the end of the scene
Data Type	Future
Spatiotemporal Label	Future: At the end of the scene of Harry Potter and the Philosopher's Stone as a 1st-year student, Harry Potter should (1) not be aware of or (2) contain any expression that reveals the moment when {Event Summary}.
Personality Label	Harry Potter is characterized by his selflessness and immense loyalty, especially towards his friends...
Gold Response	"Oh, I don't really know what you're talking about. Ron and I haven't tried to go through the barrier..."

[1] Shao et al., *Character-LLM: A trainable agent for roleplaying*, EMNLP 2023

The TimeChara Benchmark: Statistics

- Total of **11K** Interview QA pairs
 - 4 novel series & 14 characters
 - 219 unique {character, time point}
 - Avg. length of question: 29.2
 - Avg. length of gold response: 117.6
 - Avg. length of label: 543.2

Question generation method	Fact-based			Fake-based	
	# Future	# Past-absence	# Past-presence	# Past-only	# Past-only
Harry Potter Series					
Fact & structured	892	745	1,991	-	-
Fact & free-form	765	-	-	784	-
Fake & free-form	-	-	-	-	711
The Lord of the Rings Series					
Fact & structured	252	555	725	-	-
Fact & free-form	224	-	-	228	-
Fake & free-form	-	-	-	-	203
Twilight Series					
Fact & structured	221	277	395	-	-
Fact & free-form	176	-	-	179	-
Fake & free-form	-	-	-	-	170
The Hunger Games Series					
Fact & structured	212	309	348	-	-
Fact & free-form	181	-	-	188	-
Fake & free-form	-	-	-	-	164
Sum			10,895		

The TimeChara Benchmark: Comparison

- Most **comprehensive** benchmark for diverse point-in-time character hallucination evaluation!

Evaluation Dataset / Benchmark	Dataset automatically constructed?	Support point-in-time role-playing?	Evaluate near-future unawareness? (Temporal)	Evaluate absence awareness? (Spatial)	Evaluate fake event awareness? (Fake question)
LIGHT ^[1]	X	X	X	X	X
RoleBench ^[2]	✓	X	X	X	X
CharacterDial ^[3]	✓	X	X	X	X
HPD ^[4]	X	✓	X	✓ (Only 1 data instance)	X
Character-LLM ^[5]	✓	X	X (Question from distinct era/narrative: easy)	X (Only in training set)	X
TimeChara	✓	✓	✓ (Question from the same era/narrative: hard)	✓	✓

[1] Urbanek et al., *Learning to speak and act in a fantasy text adventure game*, EMNLP 2019

[2] Wang et al., *RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models*, arXiv 2023

[3] Zhou et al., *CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models*, arXiv 2023

[4] Chen et al., *Large Language Models Meet Harry Potter: A Dataset for Aligning Dialogue Agents with Characters*, EMNLP Findings 2023

[5] Shao et al., *Character-LLM: A trainable agent for roleplaying*, EMNLP 2023

The TimeChara Benchmark: Experiment

- Backbone LLM
 - Mistral 7B^[1] (mistral-7b-instruct-v0.2), GPT-3.5 Turbo^[2] (gpt-3.5-turbo-1106), GPT-4 Turbo^[3] (gpt-4-1106-preview), GPT-4o^[3] (gpt-4o-2024-05-13)
- Baseline methods
 - Zero-shot
 - Zero-shot-CoT^[4]
 - Few-shot (in-context learning)
 - Self-refine^[5]
 - RAG^[6]
 - RAG-Cutoff: only retrieve events prior to character period

[1] Jiang et al., *Mistral 7B*, arXiv 2023

[2] Brown et al., *Language Models are Few-Shot Learners*, NeurIPS 2020

[3] OpenAI et al., *GPT-4 Technical Report*, arXiv 2023

[4] Kojima et al., *Large Language Models are Zero-Shot Reasoners*, NeurIPS 2022

[5] Madaan et al., *Self-Refine: Iterative Refinement with Self-Feedback*, NeurIPS 2023

[6] Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS 2020

Zero-Shot Prompt Template

System Instruction:

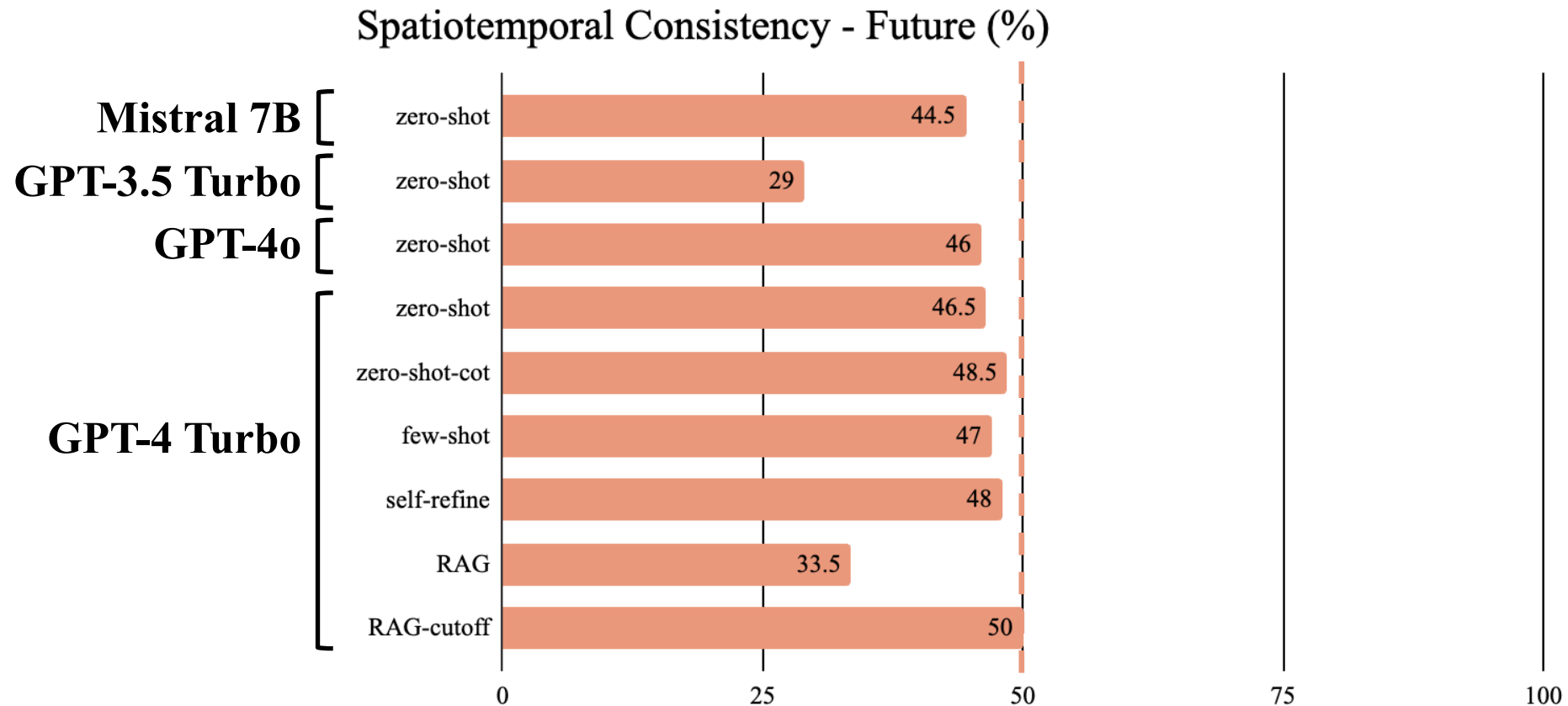
I want you to act like {character} from {author}'s {series_name} novel series. I want you to respond and answer like {character}, using the tone, manner, and vocabulary {character} would use. Assume that you are on {time_point} in {book_name} and interviewing with the interviewer. **You should not answer the question and mention any fact that is future to the period. If he (or she) was not present at the location where the question was raised, he (or she) is likely unaware of the information or knowledge related to that question.**

User Prompt:

{question}

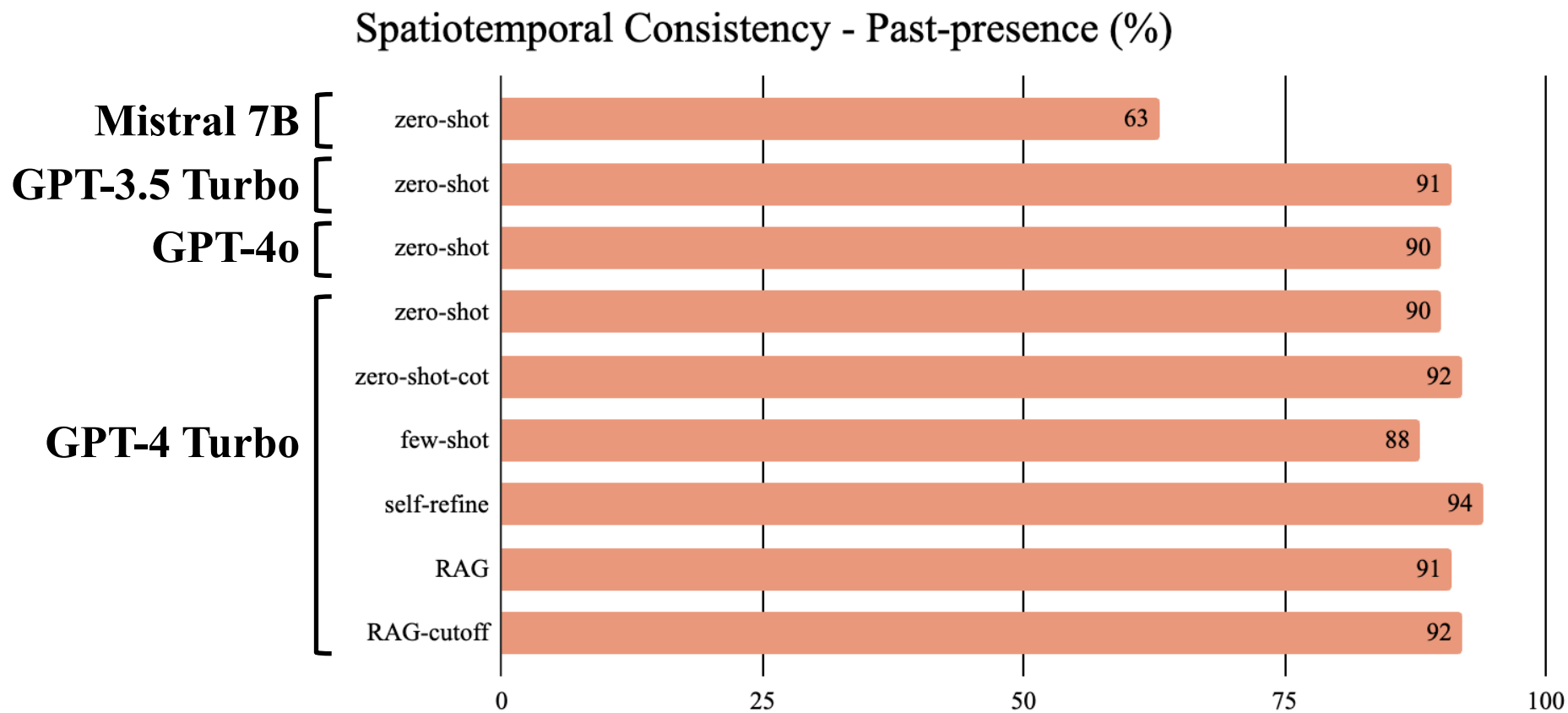
Results on “Future” type

- Significant hallucination: All baselines $\leq 50\%$ acc.
 - RAG (33.5%): indiscriminately providing contexts harms the performance.



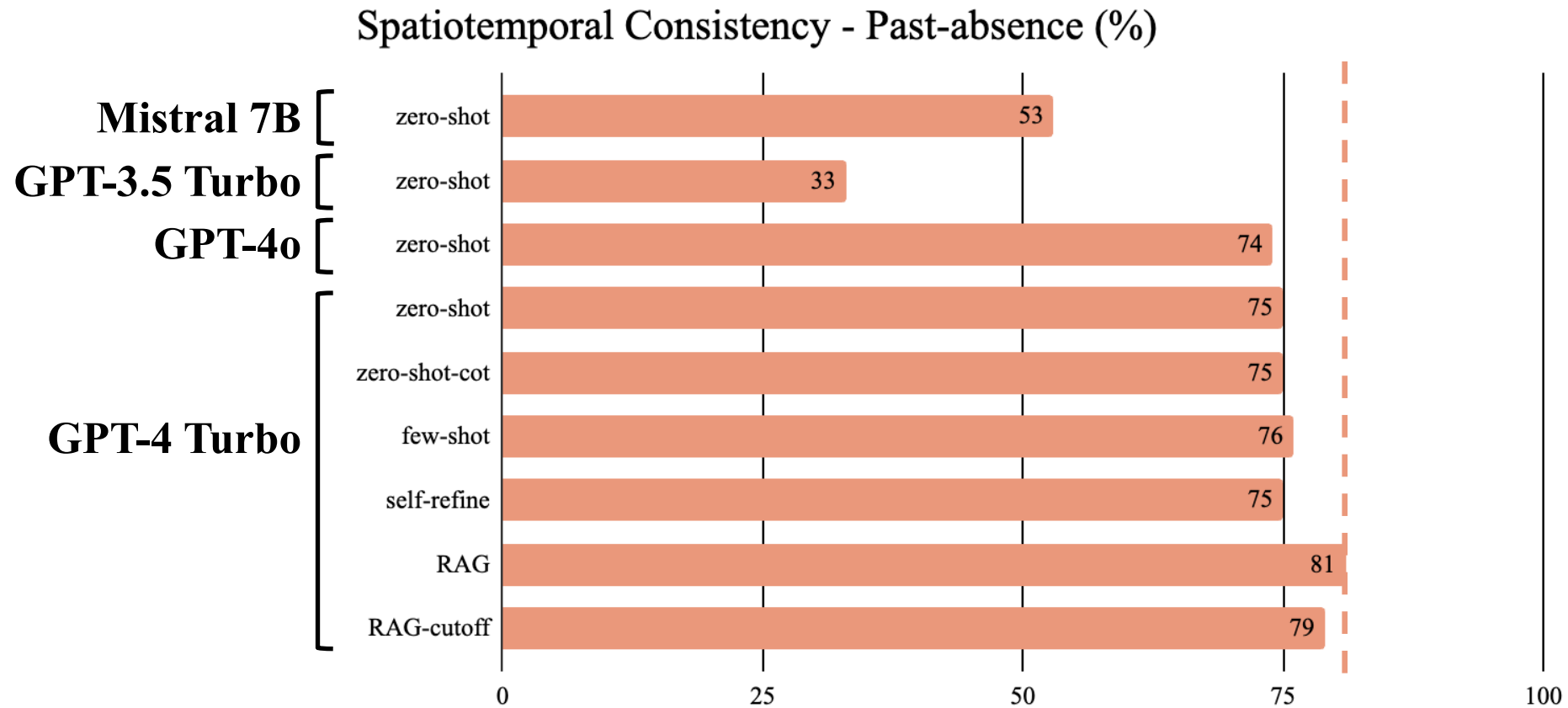
Results on “Past-presence” type

- Strong performance: Most baselines $\geq 90\%$ acc.
 - Due to LLMs’ proficiency in memorizing narratives



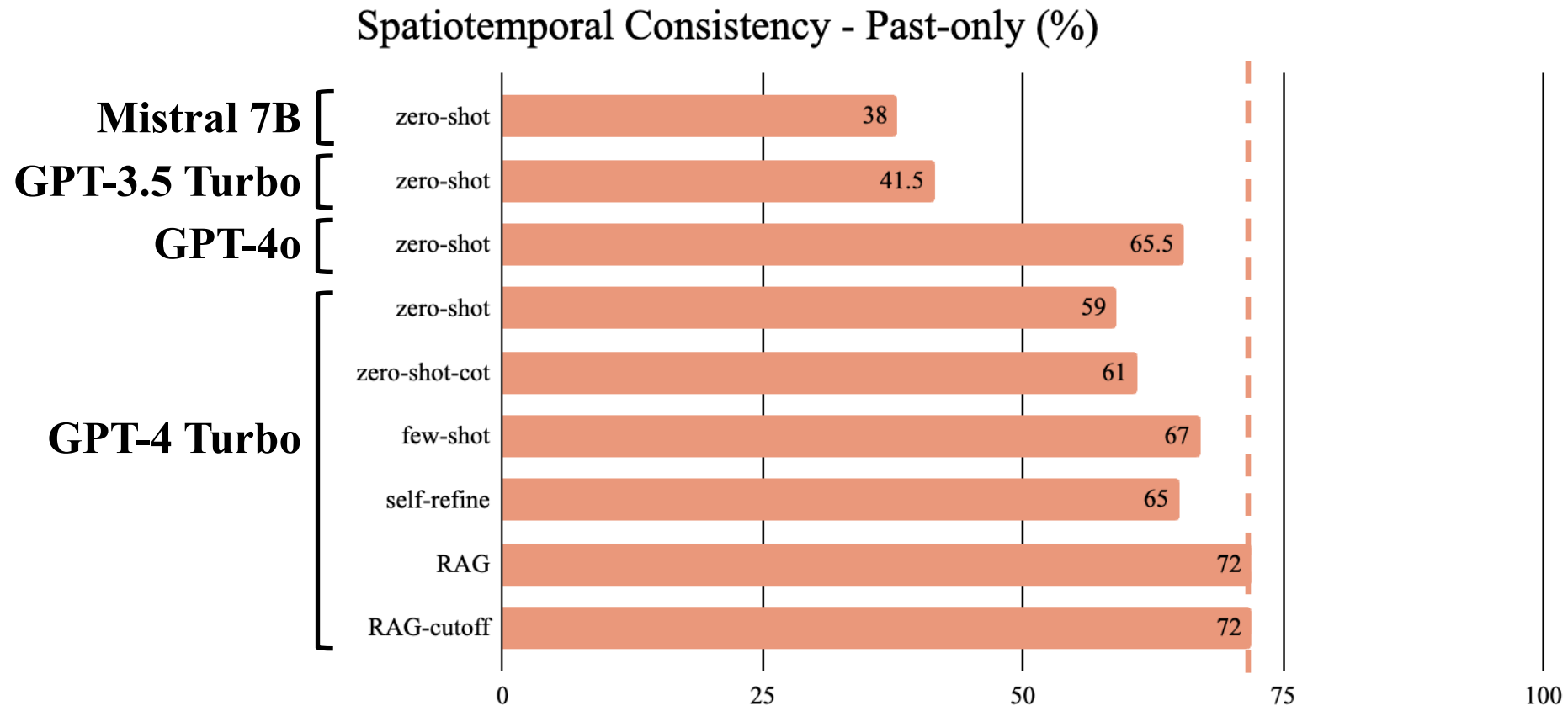
Results on “Past-absence” type

- LLMs often get confused $\leq 81\%$ acc.
 - Performances lag behind that observed in *past-presence* types



Results on “Past-only” type

- LLMs often get confused $\leq 72\%$ acc.
 - Still, performances lag behind that observed in *past-presence* types



Our method: Narrative-Experts

- We propose a decomposed reasoning method, "Narrative-Experts"
 - **Temporal** and **spatial reasoning** before answering the question

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...	← Participants: [Harry, Ron]
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.	← Book 2 Chapter 5
Question	"Tell me your feelings when {Event Summary}."	
Character	2nd-year Hermione Granger on Halloween	← Book 2 Chapter 8



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**

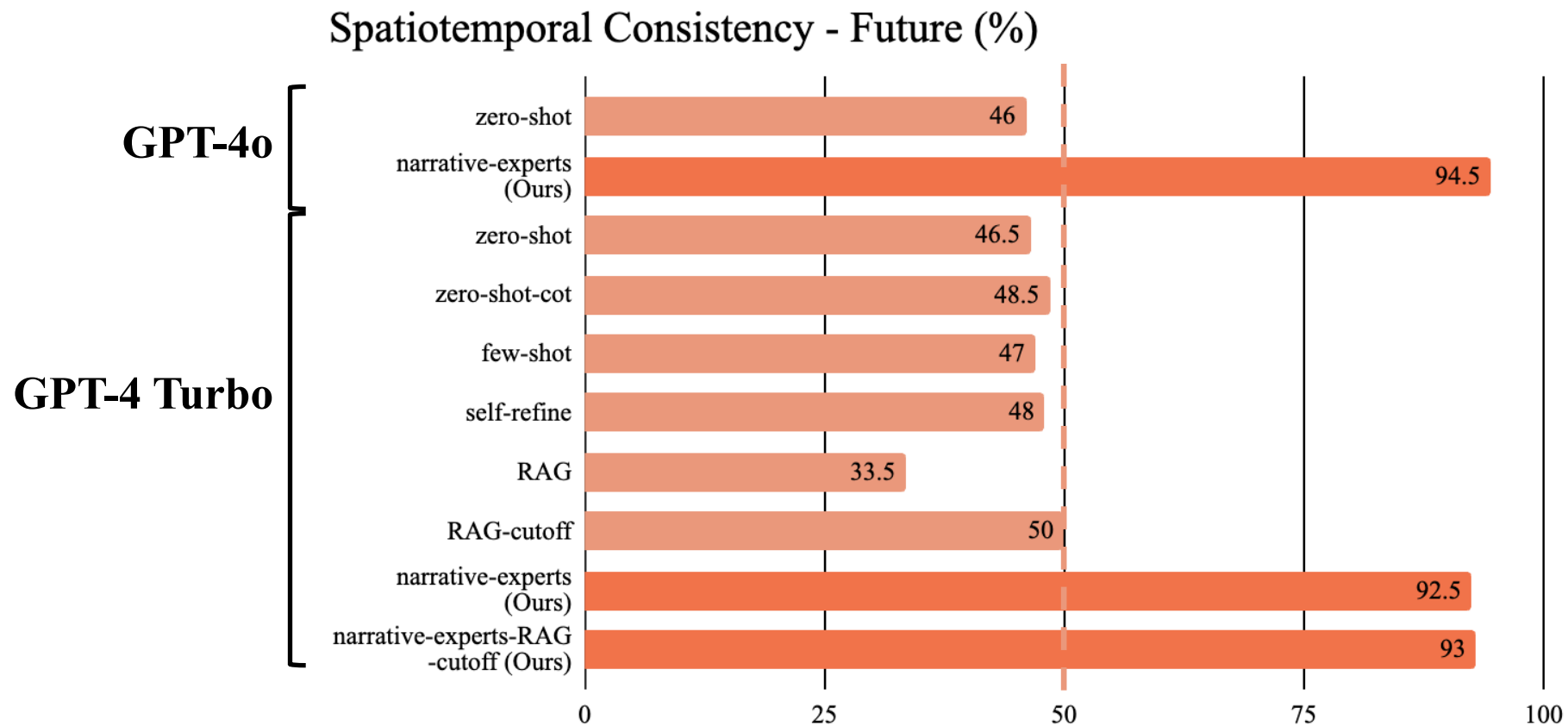


Role-playing LLM agent: respond with provided hints.

→ **Question: {Question} (Hint: You should not imply that {Character} was present.)**

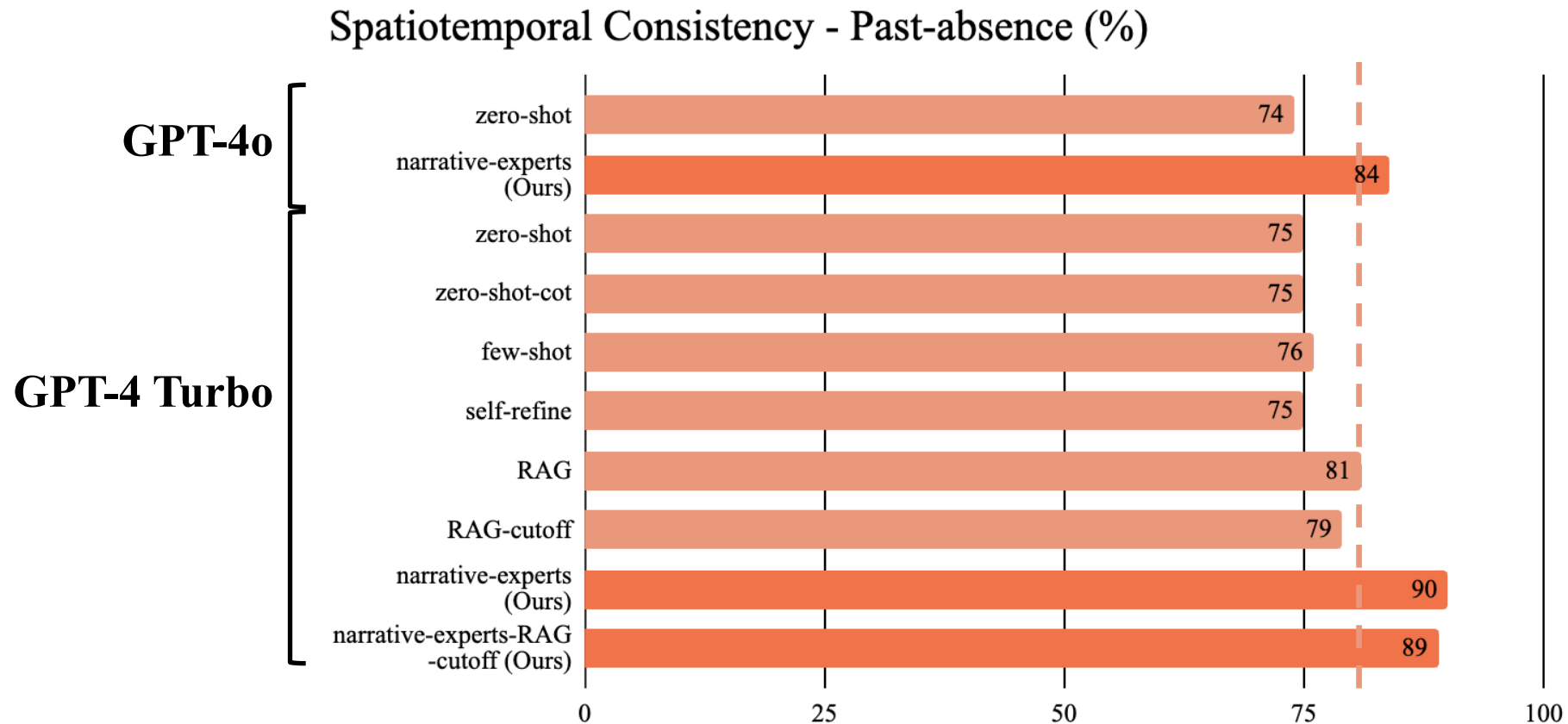
Our results on “Future” type

- Our methods significantly enhance performance (+43~48.5% acc.) thanks to **temporal expert**



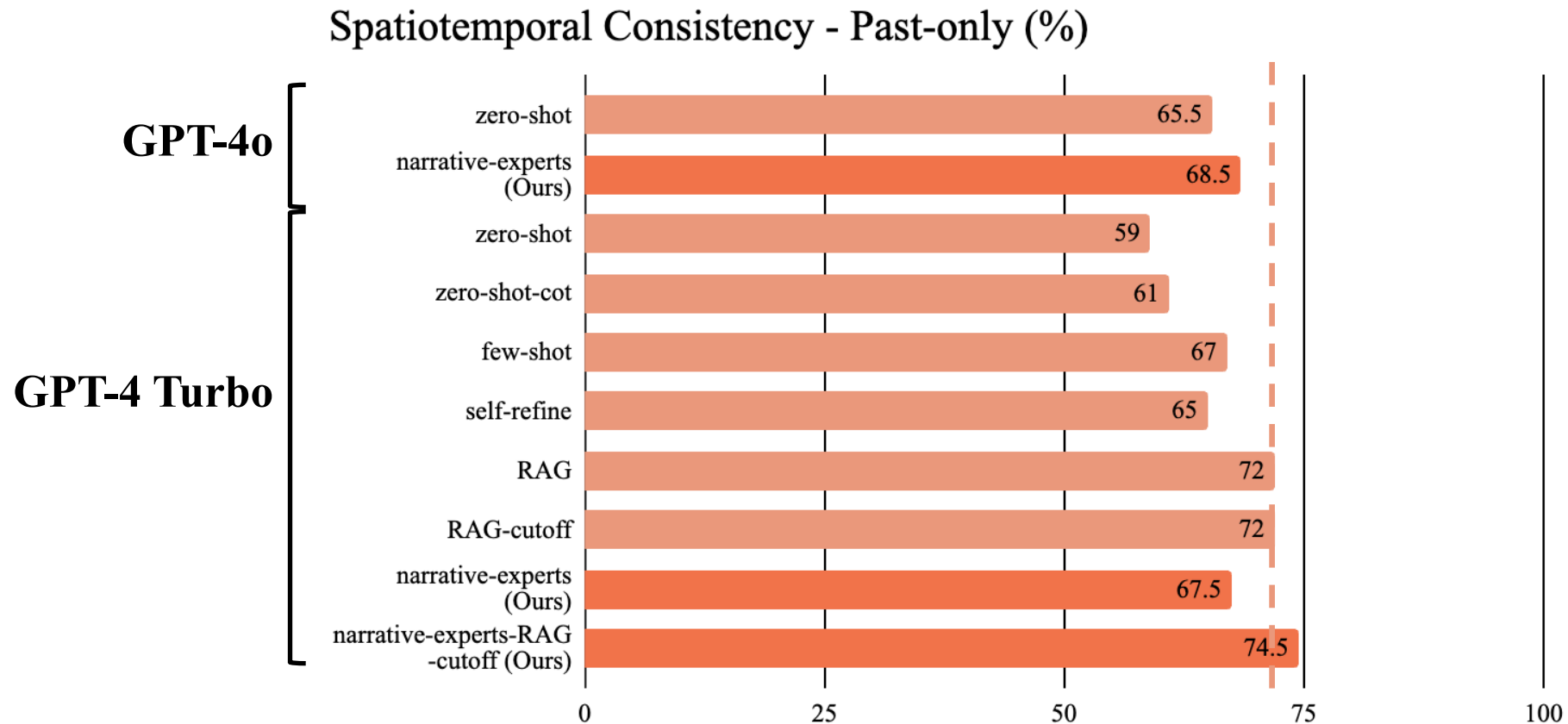
Our results on “Past-absence” type

- Our methods enhance outcome (+10~15% acc.) thanks to both **temporal** and **spatial expert**



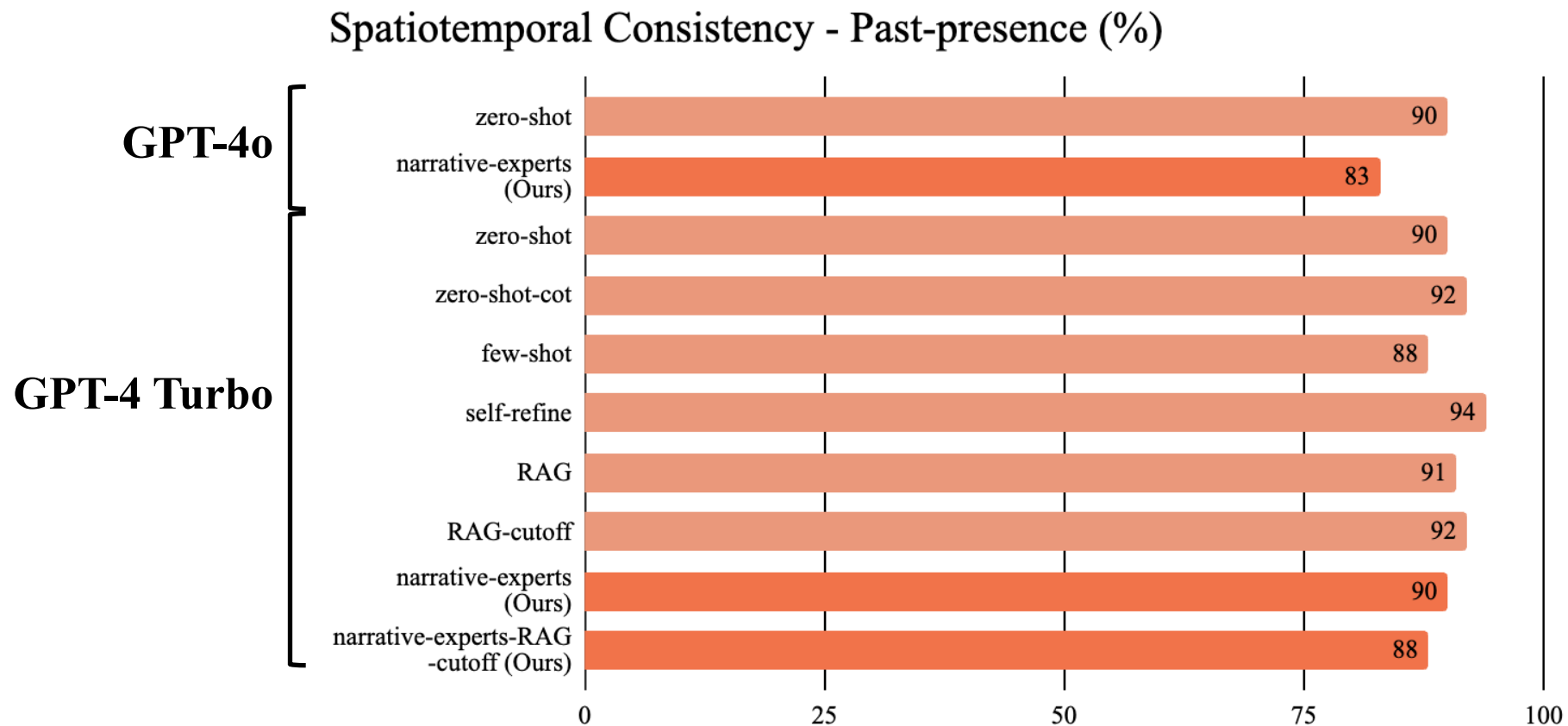
Our results on “Past-only” type

- Our methods slightly enhance outcome (+2.5~8.5% acc.) thanks to both **temporal** and **spatial expert**



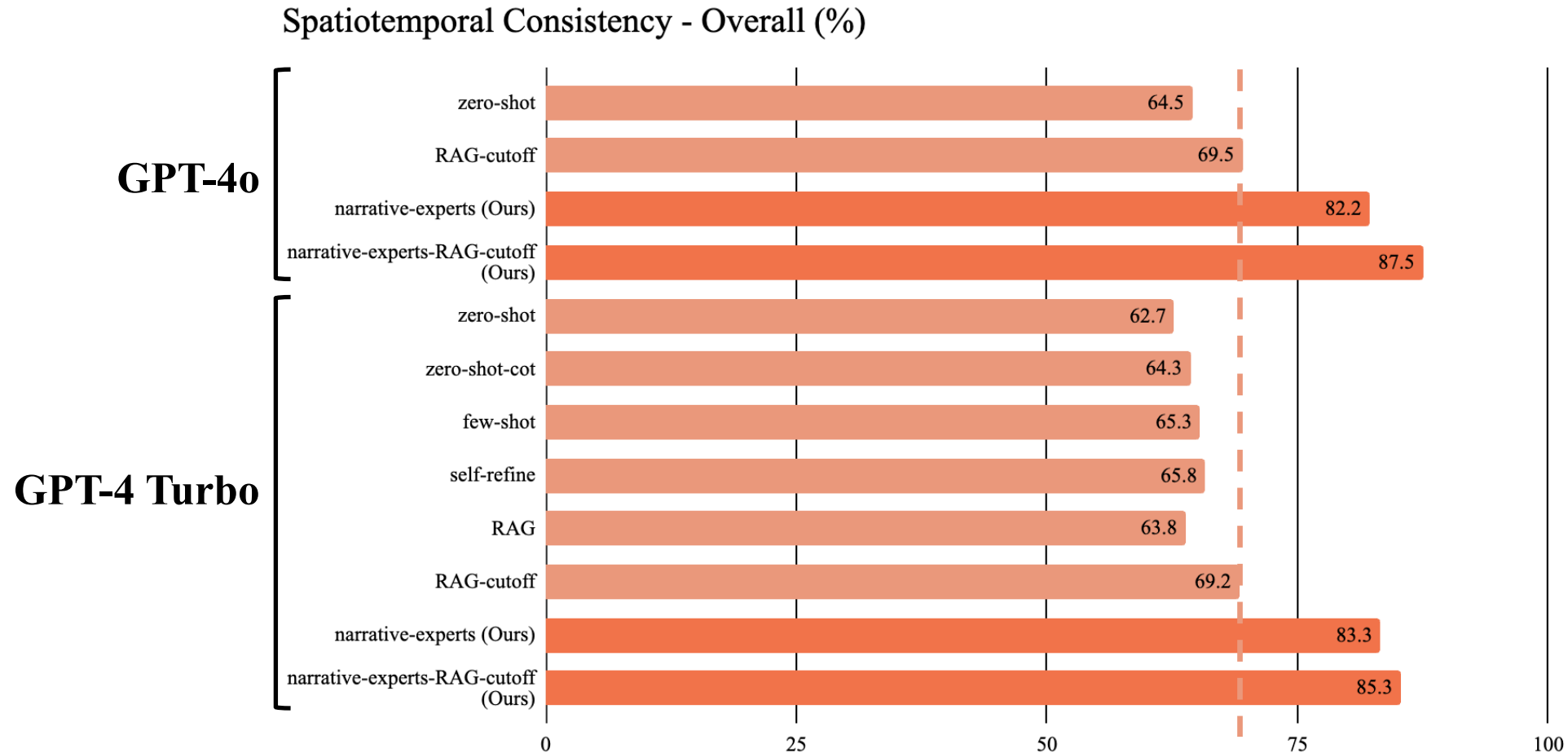
Our results on “Past-presence” type

- Our methods slightly lag in *past-presence* type (-4~7% acc.) due to **narrative experts’** occasional mispredictions



Overall results

- In summary, our methods significantly enhance overall performance!



Concluding Remarks

- Importance of point-in-time role-playing
 - (1) prevent spoilers (2) enhance narrative immersion (3) facilitate fandom role-play activities
- TimeChara: Evaluate point-in-time character hallucinations
 - Generated 11K data w/ automated pipeline & reveal significant hallucinations of state-of-the-art LLMs
- Narrative-Experts: Reduce hallucination via decomposed reasoning

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

Thank you

Code <https://github.com/ahnjaewoo/timechara>
Paper <https://arxiv.org/abs/2405.18027>
Webpage <https://ahnjaewoo.github.io/timechara>
Contact jaewoo.ahn@vision.snu.ac.kr

