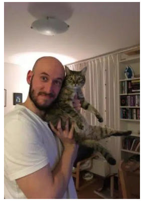
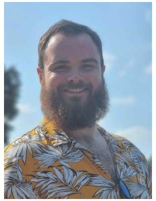
Beyond Prompts

Dynamic Conversational Benchmarking of Large Language Models

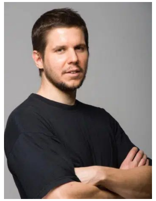
David
Castillo-Bolado



Joseph
Davidson



Finlay
Gray



Marek
Rosa

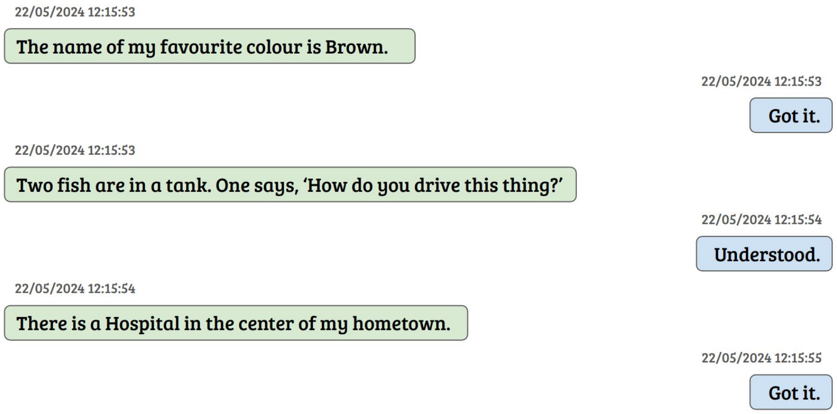
GoodAI

www.goodai.com

NEURAL INFORMATION
PROCESSING SYSTEMS

The Long-Term Memory Benchmark

- For <u>any</u> conversational agent



22/05/2024 12:15:53
The name of my favourite colour is Brown.

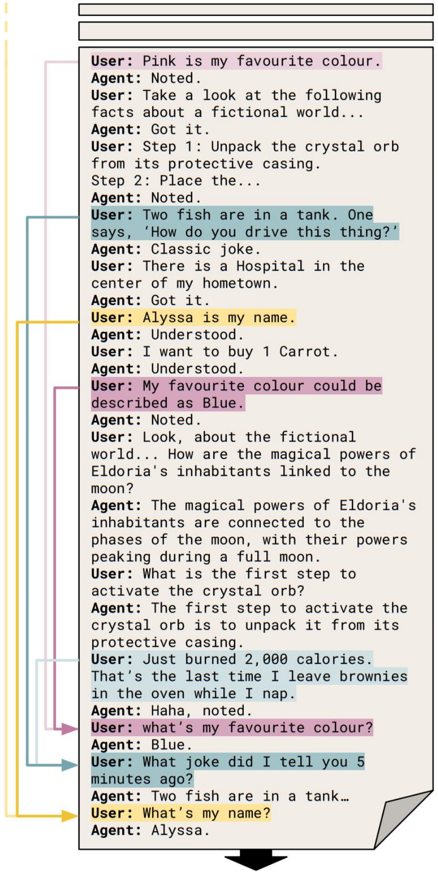
22/05/2024 12:15:53
Got it.

22/05/2024 12:15:53
Two fish are in a tank. One says, 'How do you drive this thing?'

22/05/2024 12:15:54
Understood.

22/05/2024 12:15:54
There is a Hospital in the center of my hometown.

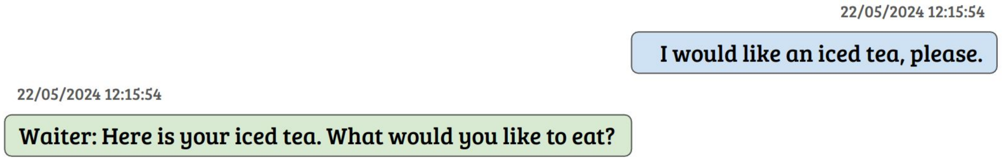
22/05/2024 12:15:55
Got it.



User: Pink is my favourite colour.
Agent: Noted.
User: Take a look at the following facts about a fictional world...
Agent: Got it.
User: Step 1: Unpack the crystal orb from its protective casing.
Step 2: Place the...
Agent: Noted.
User: Two fish are in a tank. One says, 'How do you drive this thing?'
Agent: Classic joke.
User: There is a Hospital in the center of my hometown.
Agent: Got it.
User: Alyssa is my name.
Agent: Understood.
User: I want to buy 1 Carrot.
Agent: Understood.
User: My favourite colour could be described as Blue.
Agent: Noted.
User: Look, about the fictional world... How are the magical powers of Eldoria's inhabitants linked to the moon?
Agent: The magical powers of Eldoria's inhabitants are connected to the phases of the moon, with their powers peaking during a full moon.
User: What is the first step to activate the crystal orb?
Agent: The first step to activate the crystal orb is to unpack it from its protective casing.
User: Just burned 2,000 calories. That's the last time I leave brownies in the oven while I nap.
Agent: Haha, noted.
User: what's my favourite colour?
Agent: Blue.
User: What joke did I tell you 5 minutes ago?
Agent: Two fish are in a tank…
User: What's my name?
Agent: Alyssa.

The Long-Term Memory Benchmark

- For any conversational agent
- Dynamic

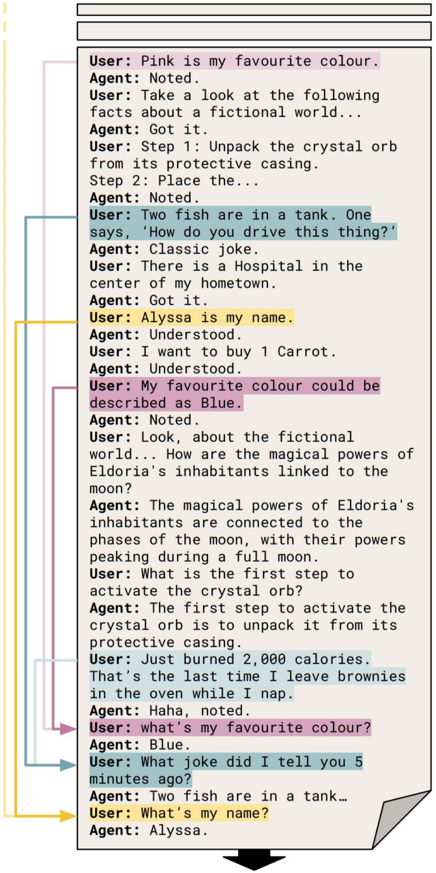


22/05/2024 12:15:54

I would like an iced tea, please.

22/05/2024 12:15:54

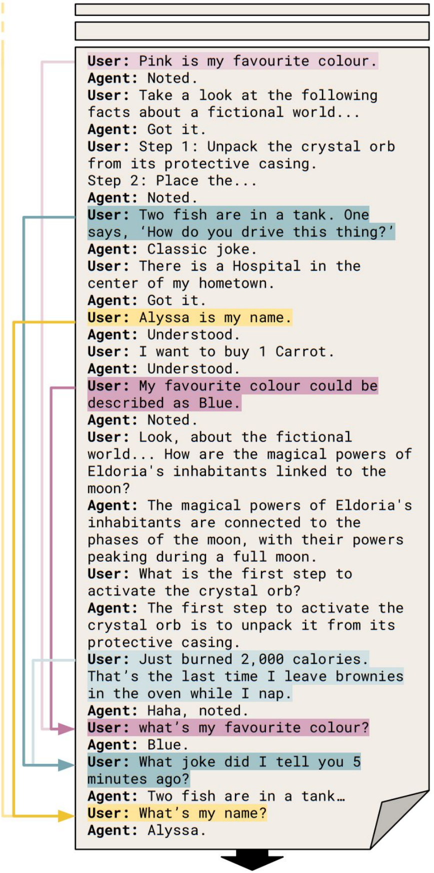
Waiter: Here is your iced tea. What would you like to eat?



User: Pink is my favourite colour.
Agent: Noted.
User: Take a look at the following facts about a fictional world...
Agent: Got it.
User: Step 1: Unpack the crystal orb from its protective casing.
Step 2: Place the...
Agent: Noted.
User: Two fish are in a tank. One says, 'How do you drive this thing?'
Agent: Classic joke.
User: There is a Hospital in the center of my hometown.
Agent: Got it.
User: Alyssa is my name.
Agent: Understood.
User: I want to buy 1 Carrot.
Agent: Understood.
User: My favourite colour could be described as Blue.
Agent: Noted.
User: Look, about the fictional world... How are the magical powers of Eldoria's inhabitants linked to the moon?
Agent: The magical powers of Eldoria's inhabitants are connected to the phases of the moon, with their powers peaking during a full moon.
User: What is the first step to activate the crystal orb?
Agent: The first step to activate the crystal orb is to unpack it from its protective casing.
User: Just burned 2,000 calories. That's the last time I leave brownies in the oven while I nap.
Agent: Haha, noted.
User: what's my favourite colour?
Agent: Blue.
User: What joke did I tell you 5 minutes ago?
Agent: Two fish are in a tank…
User: What's my name?
Agent: Alyssa.

The Long-Term Memory Benchmark

- For any conversational agent
- Dynamic
- Realistic
  - Interleaved tasks
  - Complex and natural



User: Pink is my favourite colour.
Agent: Noted.
User: Take a look at the following facts about a fictional world...
Agent: Got it.
User: Step 1: Unpack the crystal orb from its protective casing.
Step 2: Place the...
Agent: Noted.
User: Two fish are in a tank. One says, ‘How do you drive this thing?’
Agent: Classic joke.
User: There is a Hospital in the center of my hometown.
Agent: Got it.
User: Alyssa is my name.
Agent: Understood.
User: I want to buy 1 Carrot.
Agent: Understood.
User: My favourite colour could be described as Blue.
Agent: Noted.
User: Look, about the fictional world... How are the magical powers of Eldoria's inhabitants linked to the moon?
Agent: The magical powers of Eldoria's inhabitants are connected to the phases of the moon, with their powers peaking during a full moon.
User: What is the first step to activate the crystal orb?
Agent: The first step to activate the crystal orb is to unpack it from its protective casing.
User: Just burned 2,000 calories. That’s the last time I leave brownies in the oven while I nap.
Agent: Haha, noted.
User: what’s my favourite colour?
Agent: Blue.
User: What joke did I tell you 5 minutes ago?
Agent: Two fish are in a tank…
User: What’s my name?
Agent: Alyssa.

LTM skills

Recall

Memory Types:

Episodic	Spatial	Prospective	Theory of Mind
Yesterday	1 km north from…	when __, do __	Other people’s thoughts
Two minutes ago	between __ and __		

Memory Processes:

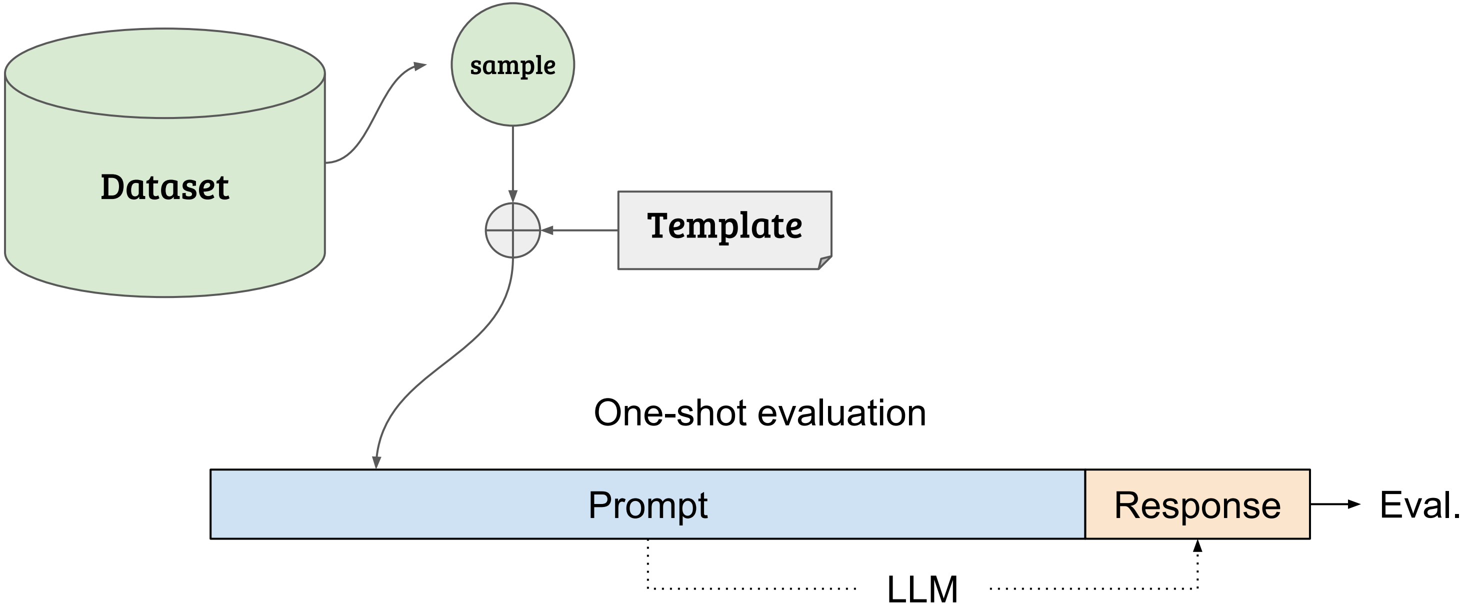
Conflict resolution

Information Integration

Test Scenarios → 11 points

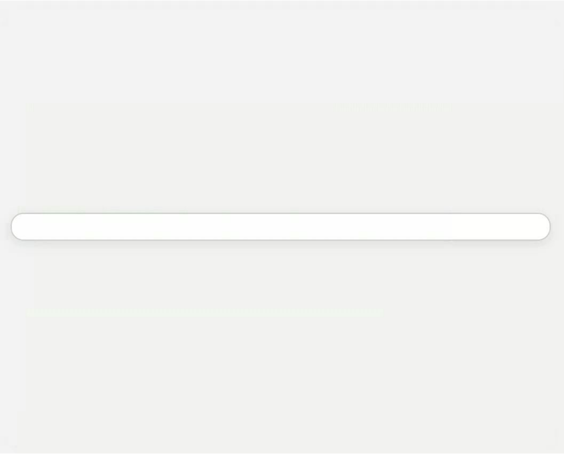
- Colours
- Name list
- Jokes
- Locations directions
- Quotes
- Trigger response
- Sally–Anne
- Spy meeting
- Shopping list
- ChapterBreak
- Restaurant

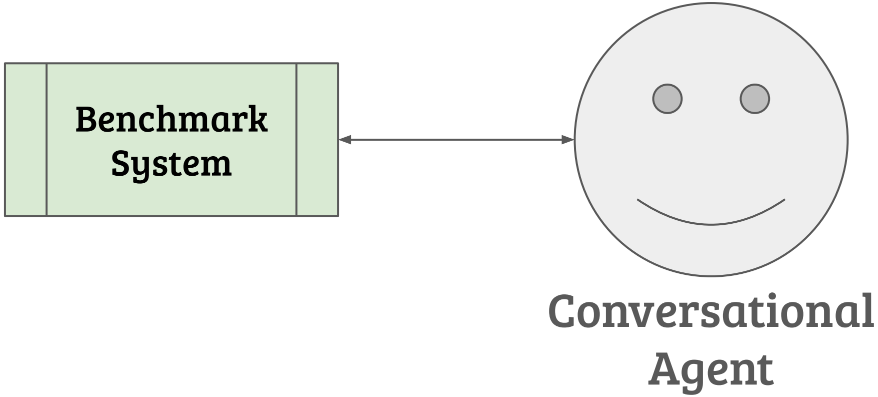
How the LTM Benchmark does NOT work



How the LTM Benchmark DOES work

A single very-long Benchmark↔Agent conversation.

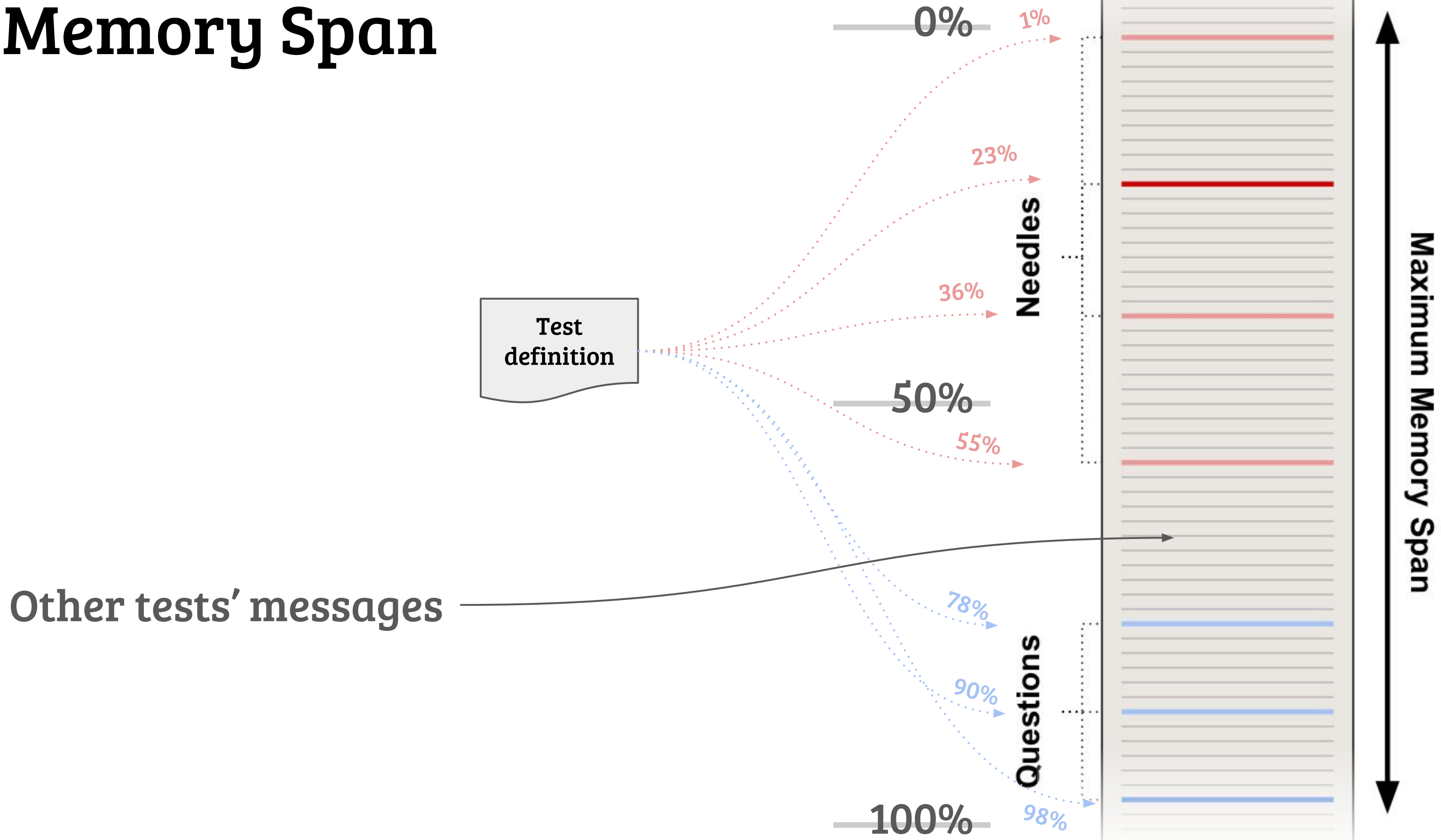




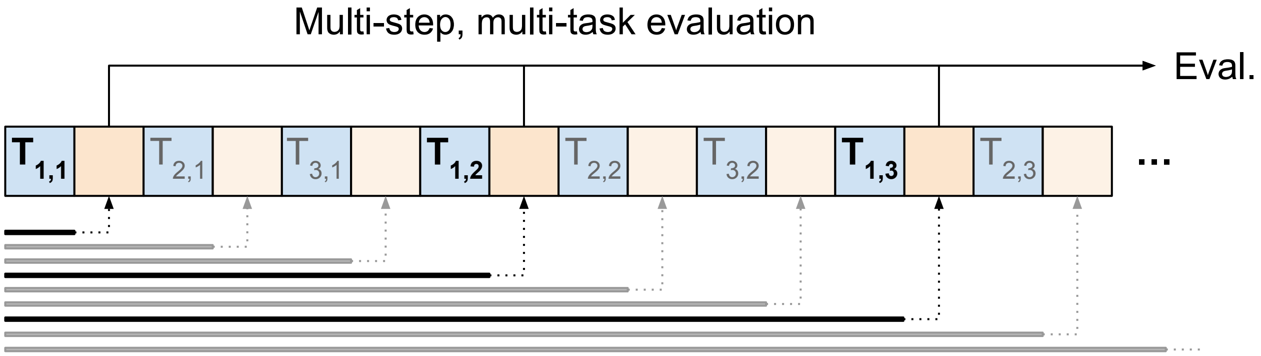
Benchmark System

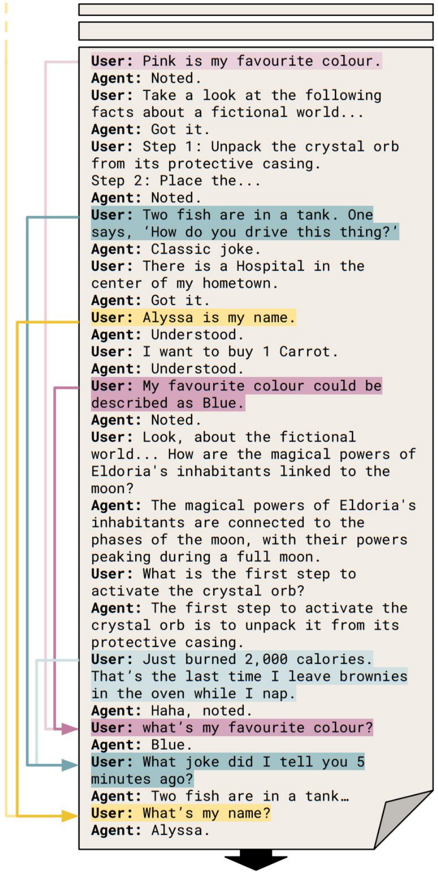
Conversational Agent

The Memory Span



How the LTM Benchmark DOES work





User: Pink is my favourite colour.

Agent: Noted.

User: Take a look at the following facts about a fictional world...

Agent: Got it.

User: Step 1: Unpack the crystal orb from its protective casing.
Step 2: Place the...

Agent: Noted.

User: Two fish are in a tank. One says, ‘How do you drive this thing?’

Agent: Classic joke.

User: There is a Hospital in the center of my hometown.

Agent: Got it.

User: Alyssa is my name.

Agent: Understood.

User: I want to buy 1 Carrot.

Agent: Understood.

User: My favourite colour could be described as Blue.

Agent: Noted.

User: Look, about the fictional world... How are the magical powers of Eldoria's inhabitants linked to the moon?

Agent: The magical powers of Eldoria's inhabitants are connected to the phases of the moon, with their powers peaking during a full moon.

User: What is the first step to activate the crystal orb?

Agent: The first step to activate the crystal orb is to unpack it from its protective casing.

User: Just burned 2,000 calories. That’s the last time I leave brownies in the oven while I nap.

Agent: Haha, noted.

User: what’s my favourite colour?

Agent: Blue.

User: What joke did I tell you 5 minutes ago?

Agent: Two fish are in a tank…

User: What’s my name?

Agent: Alyssa.

Models tested

LLMs

- 🟩 GPT-4 turbo
- 🟨 GPT-4o
- 🟪 GPT-4o-mini
- 🟥 Claude 3 Opus
- 🟦 Gemini 1.5 Pro
- 🟧 Llama 3 70B

LTM Agents

- 🟩 GPT-4 turbo
- 🟪 GPT-4o-mini
- 🟥 Claude 3 Opus
- 🟧 Llama 3 70B

Input: 16k max

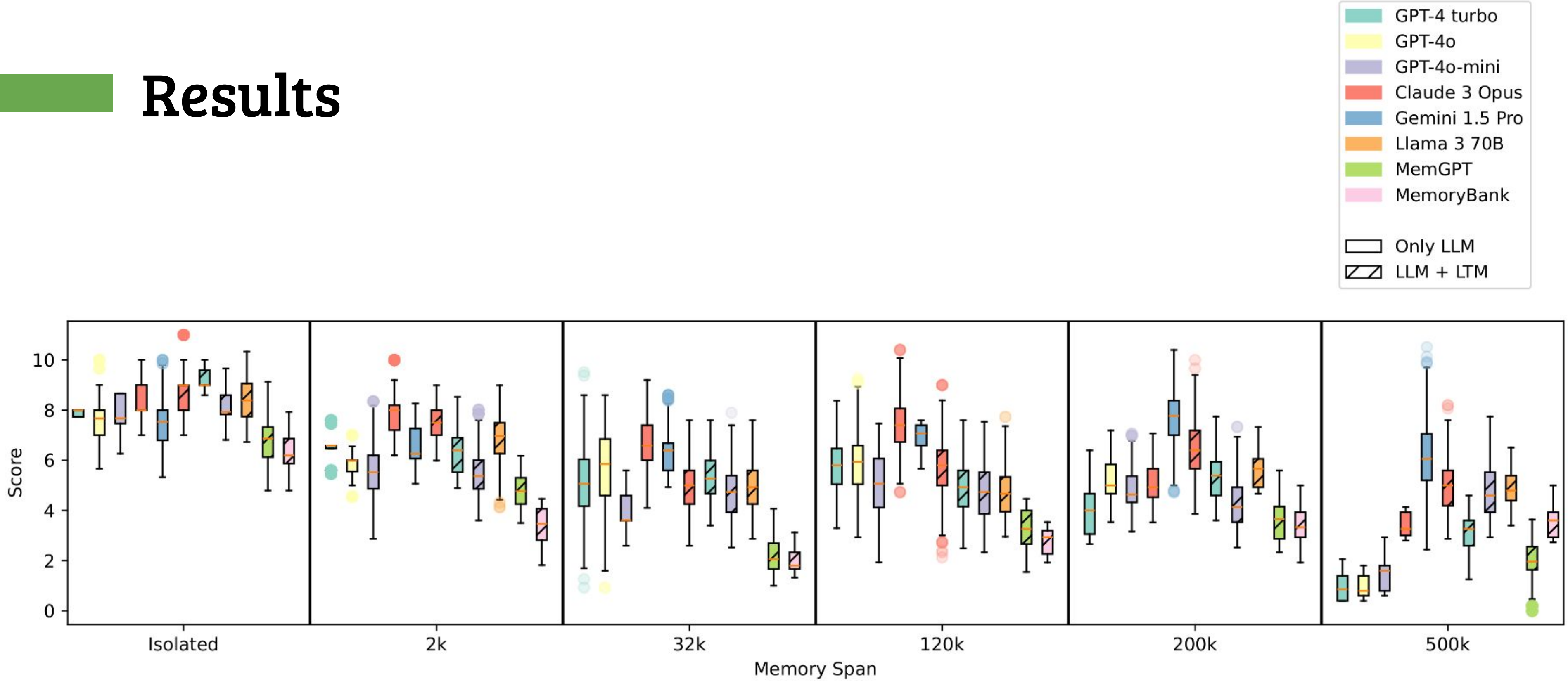
+ Semantic Retrieval

JSON scratchpad

- 🟩 MemGPT
- 🟪 MemoryBank

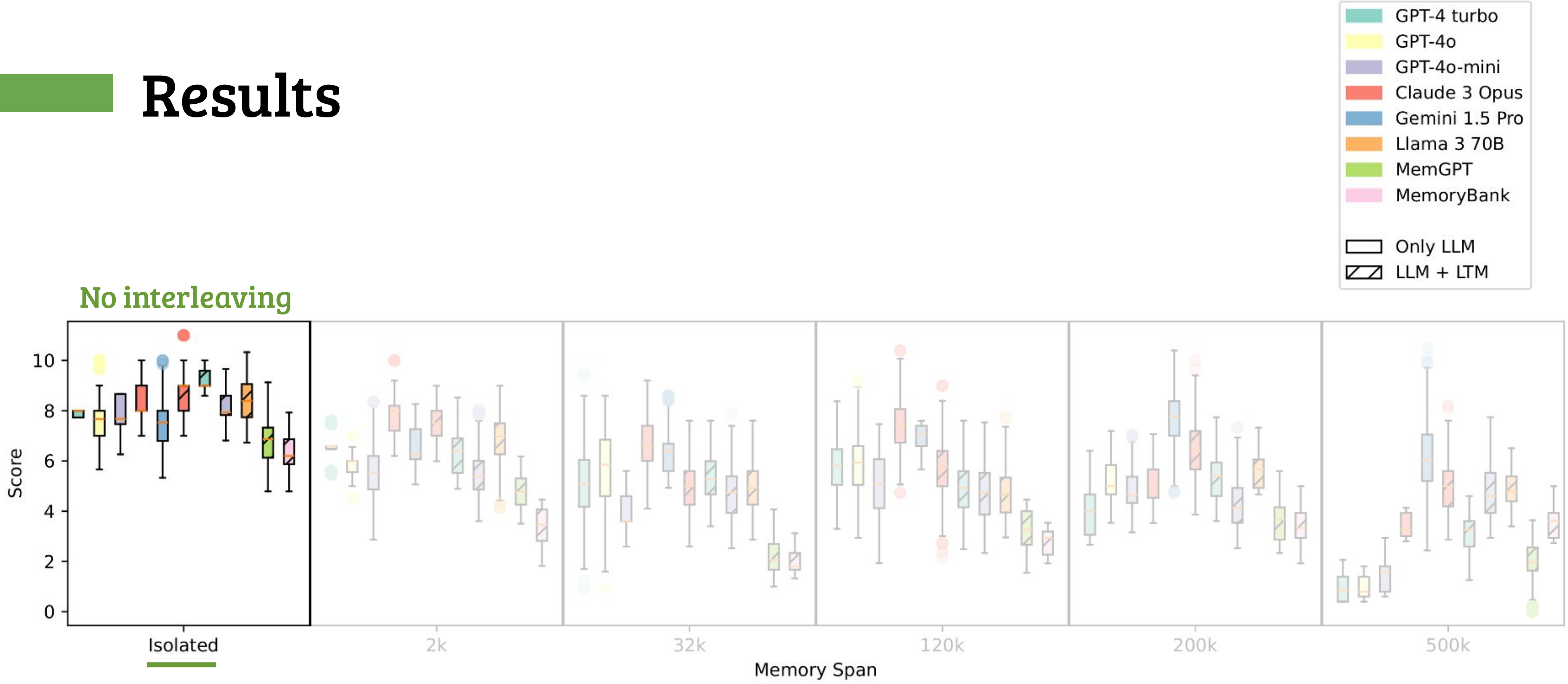
Input: 16k-60k

Results

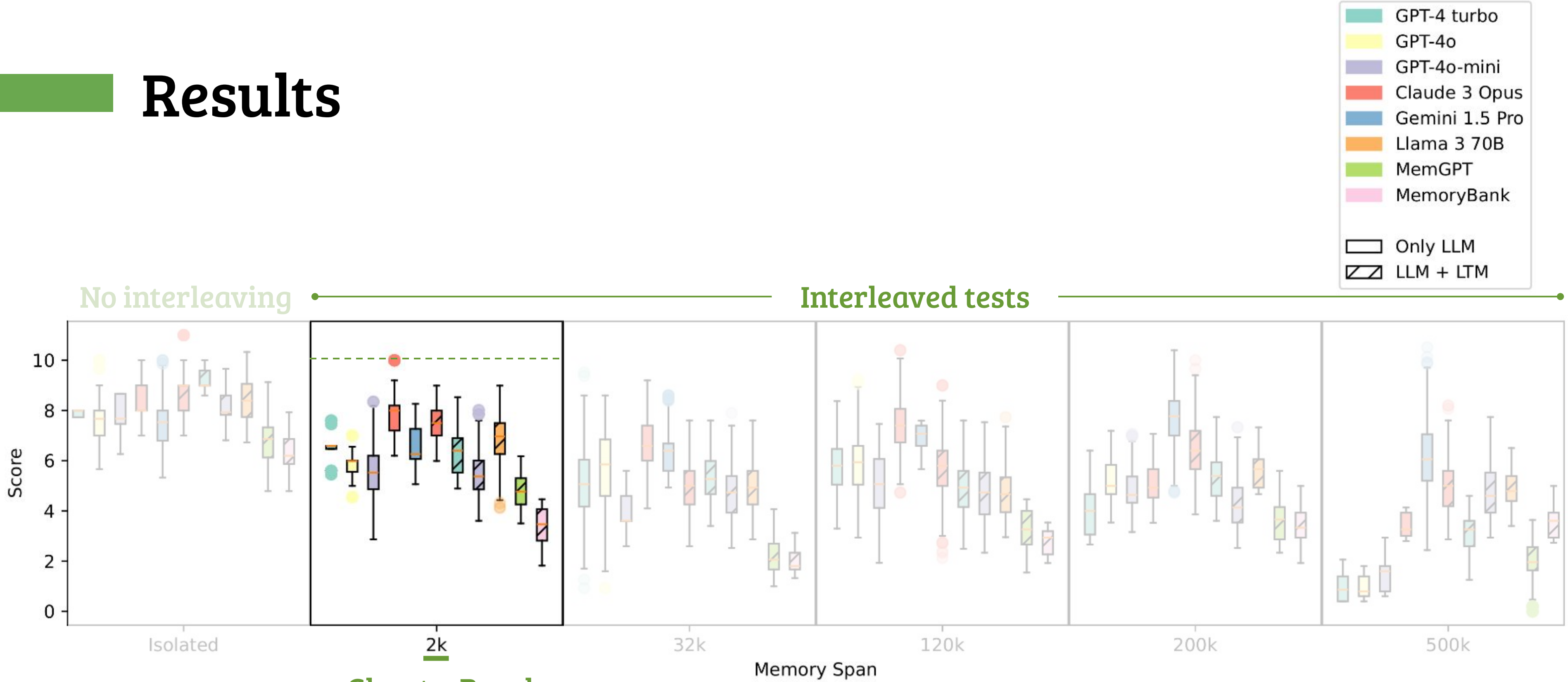

GPT-4 turbo
GPT-4o
GPT-4o-mini
Claude 3 Opus
Gemini 1.5 Pro
Llama 3 70B
MemGPT
MemoryBank
Only LLM
LLM + LTM
Score
0
2
4
6
8
10
Isolated
2k
32k
120k
200k
500k
Memory Span

Results

No interleaving



Results

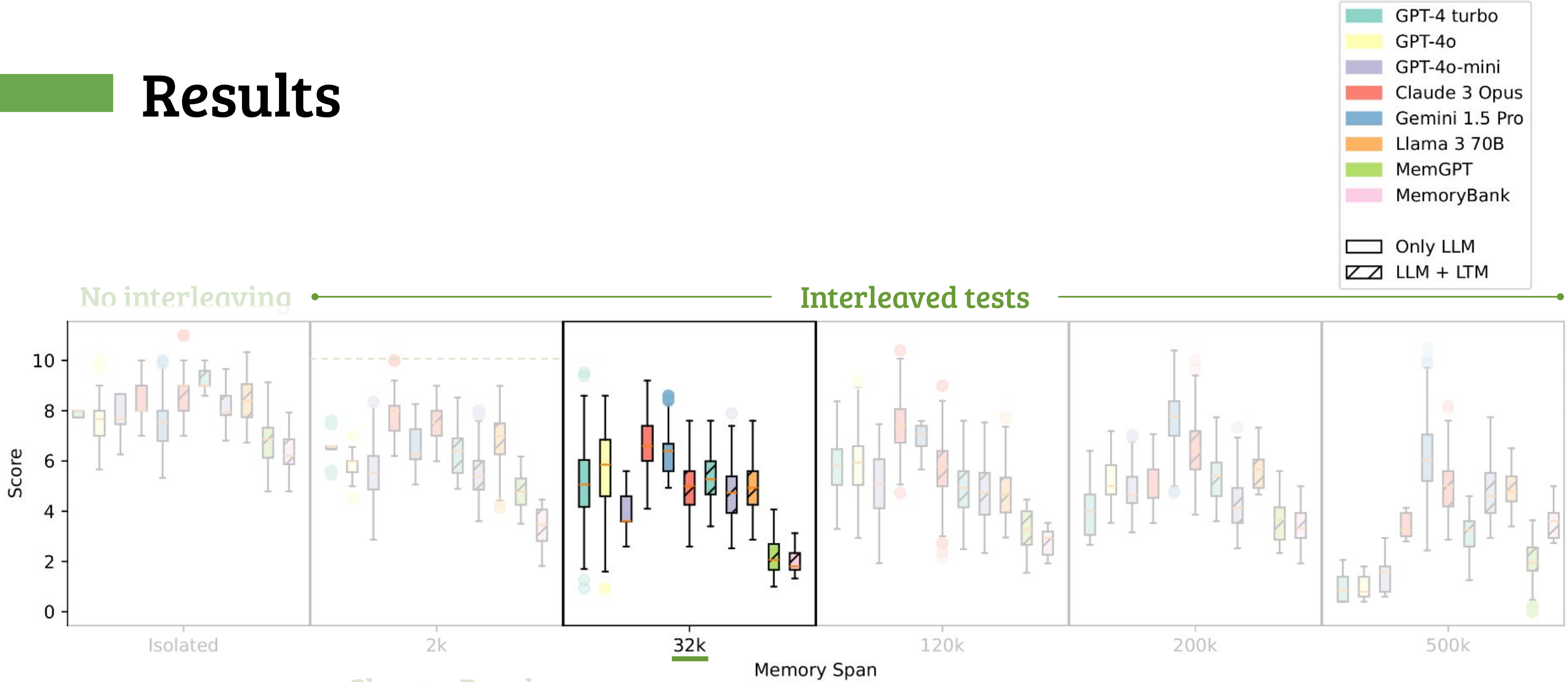


No interleaving

Interleaved tests

- ChapterBreak

Results

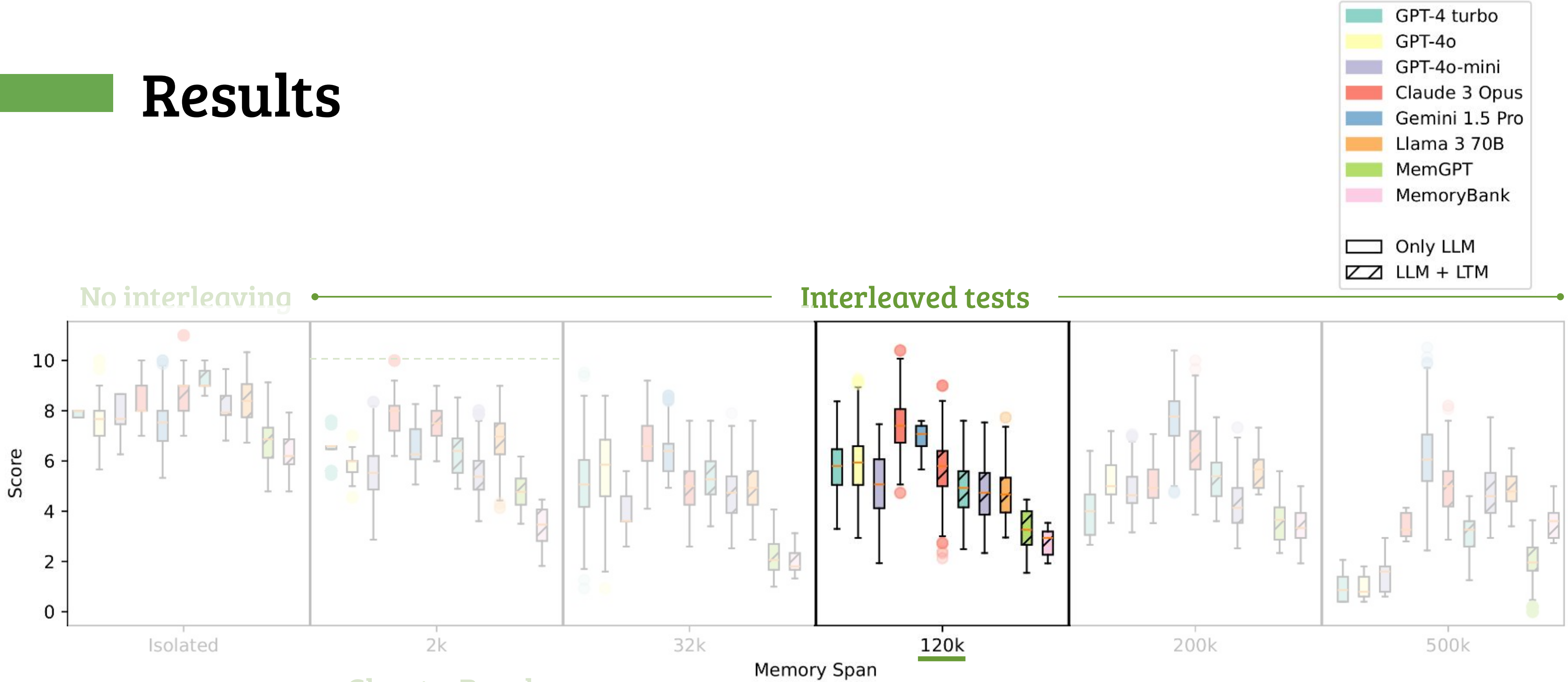


No interleaving

Interleaved tests

- ChapterBreak

Results

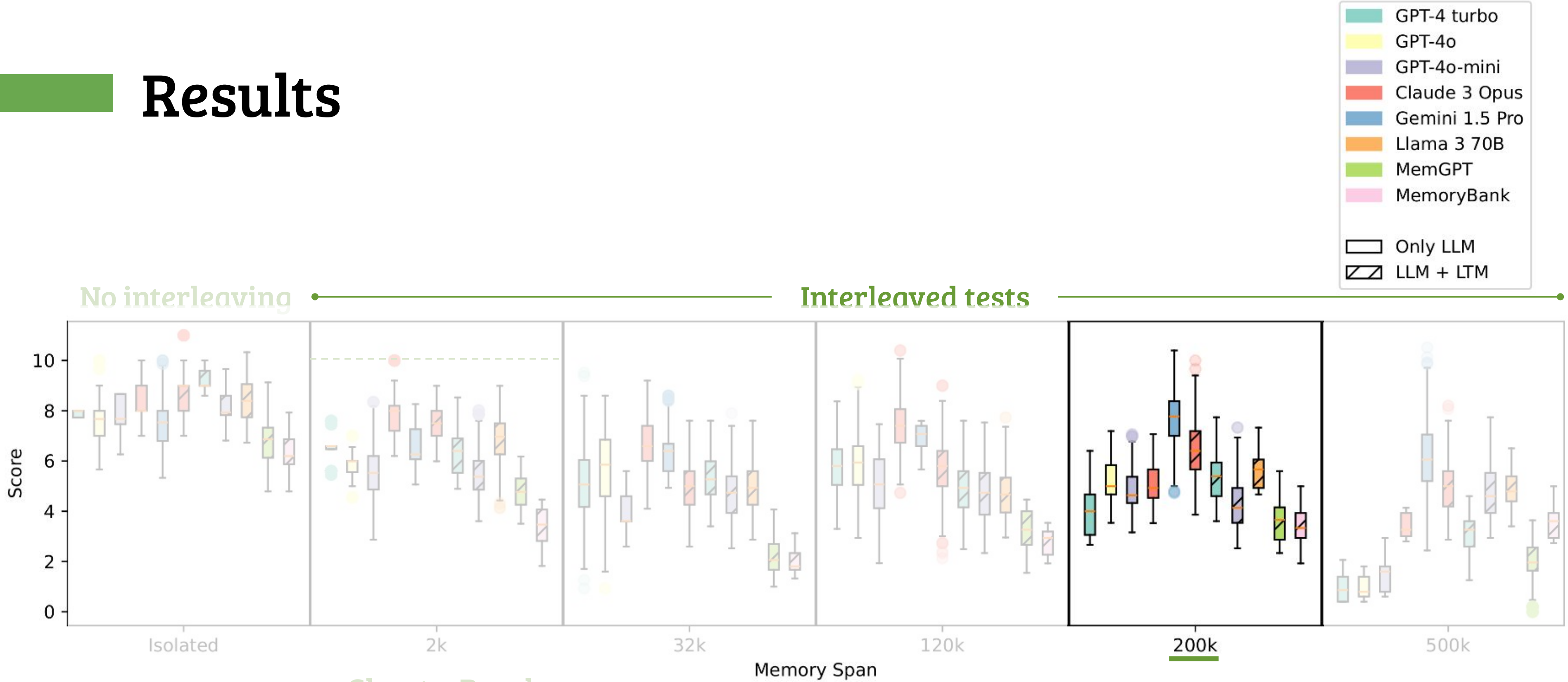


No interleaving

Interleaved tests

- ChapterBreak

Results



No interleaving

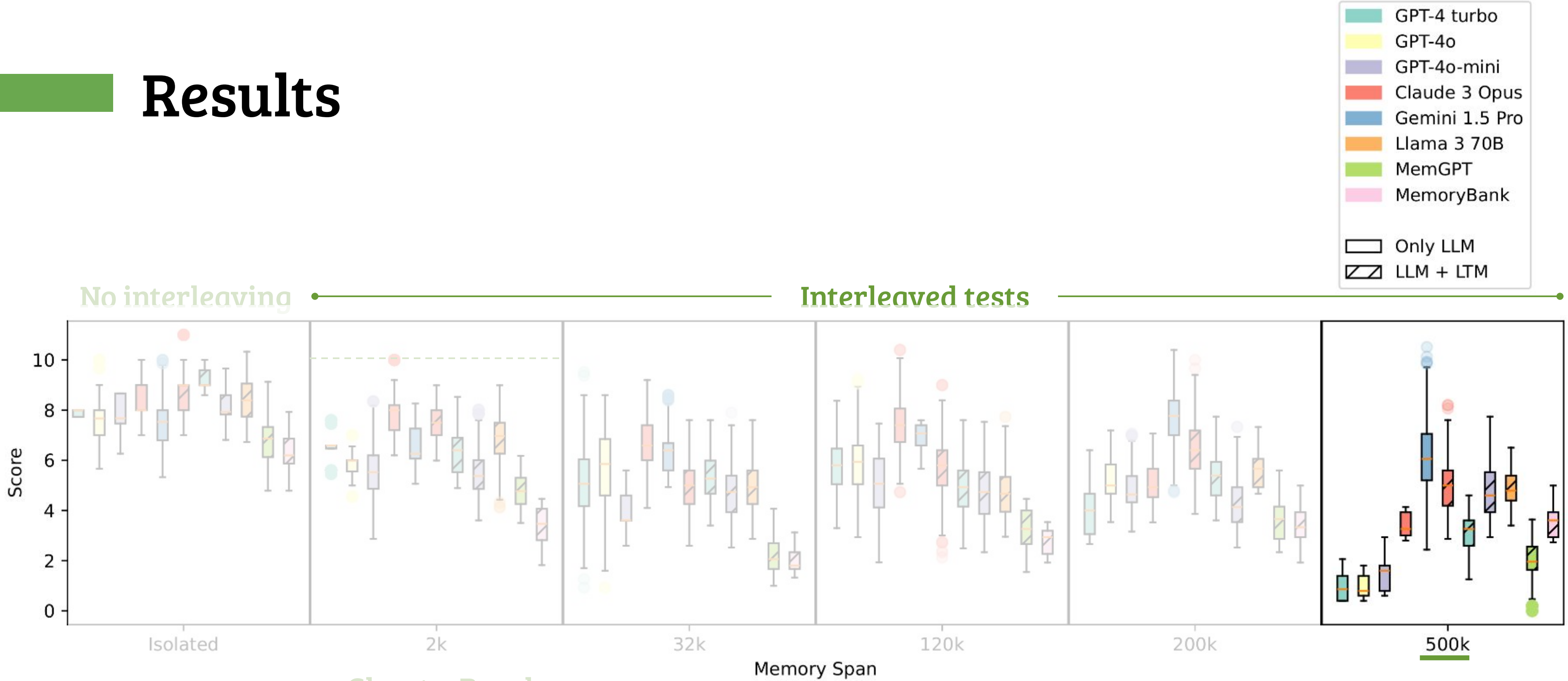
Interleaved tests

- ChapterBreak

Results

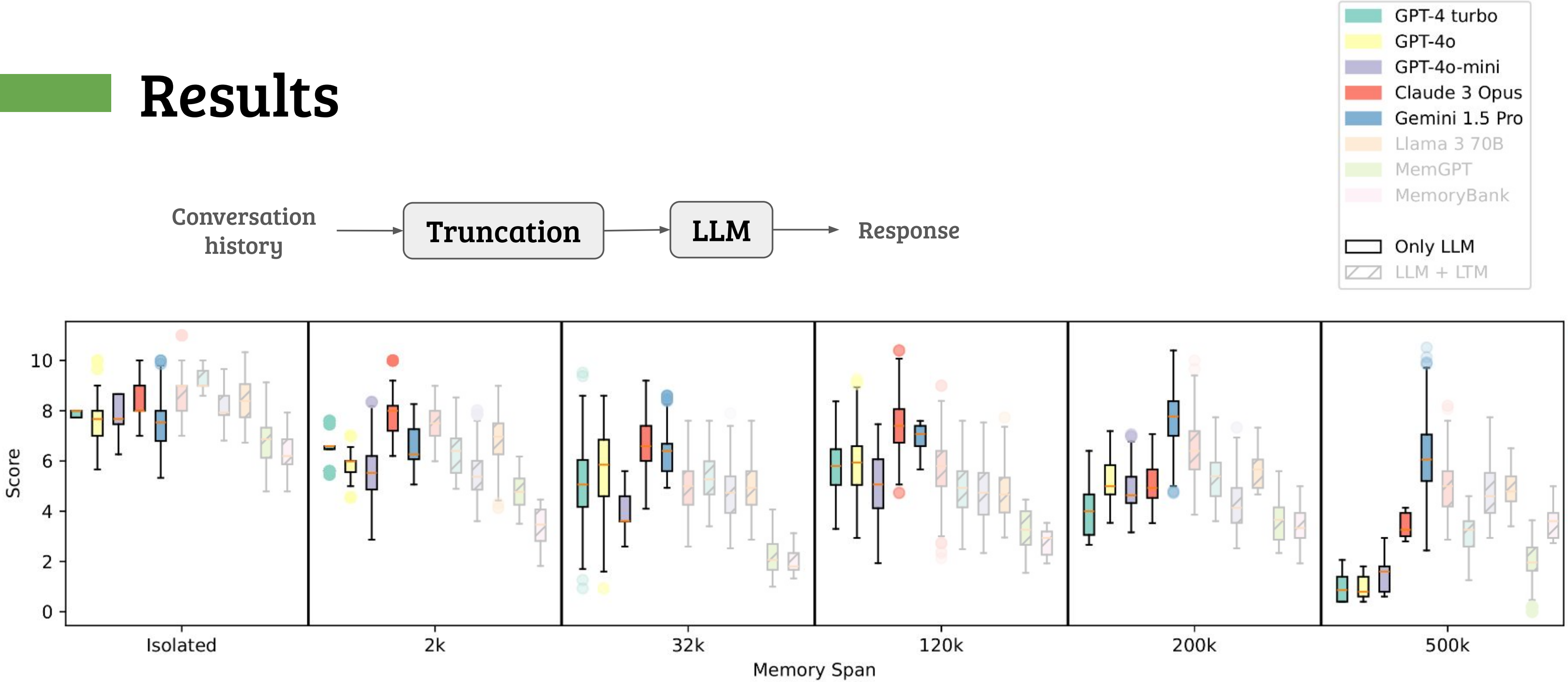
No interleaving

Interleaved tests

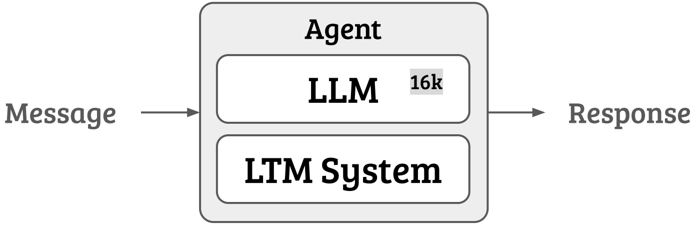


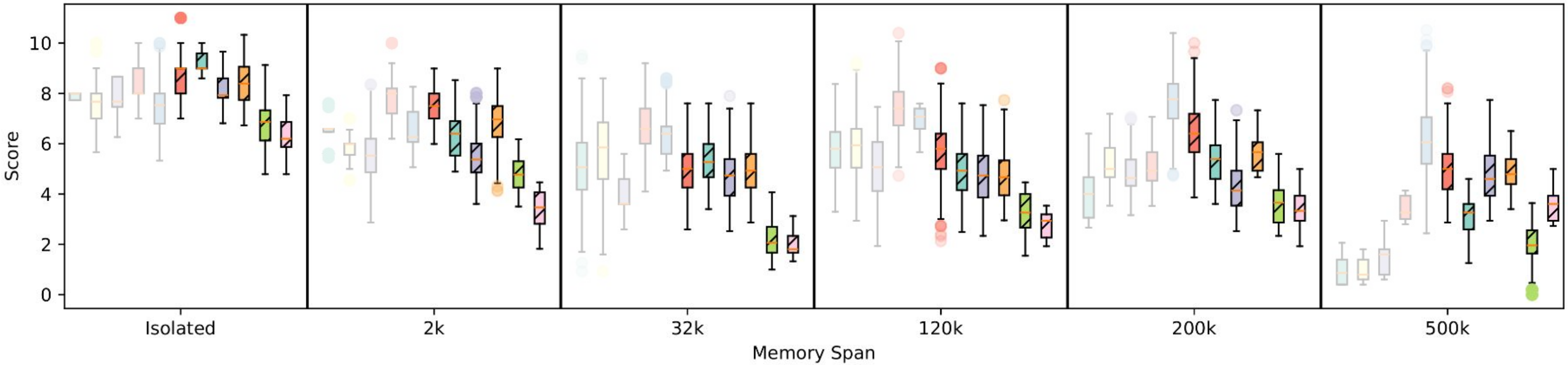
- ChapterBreak

Results

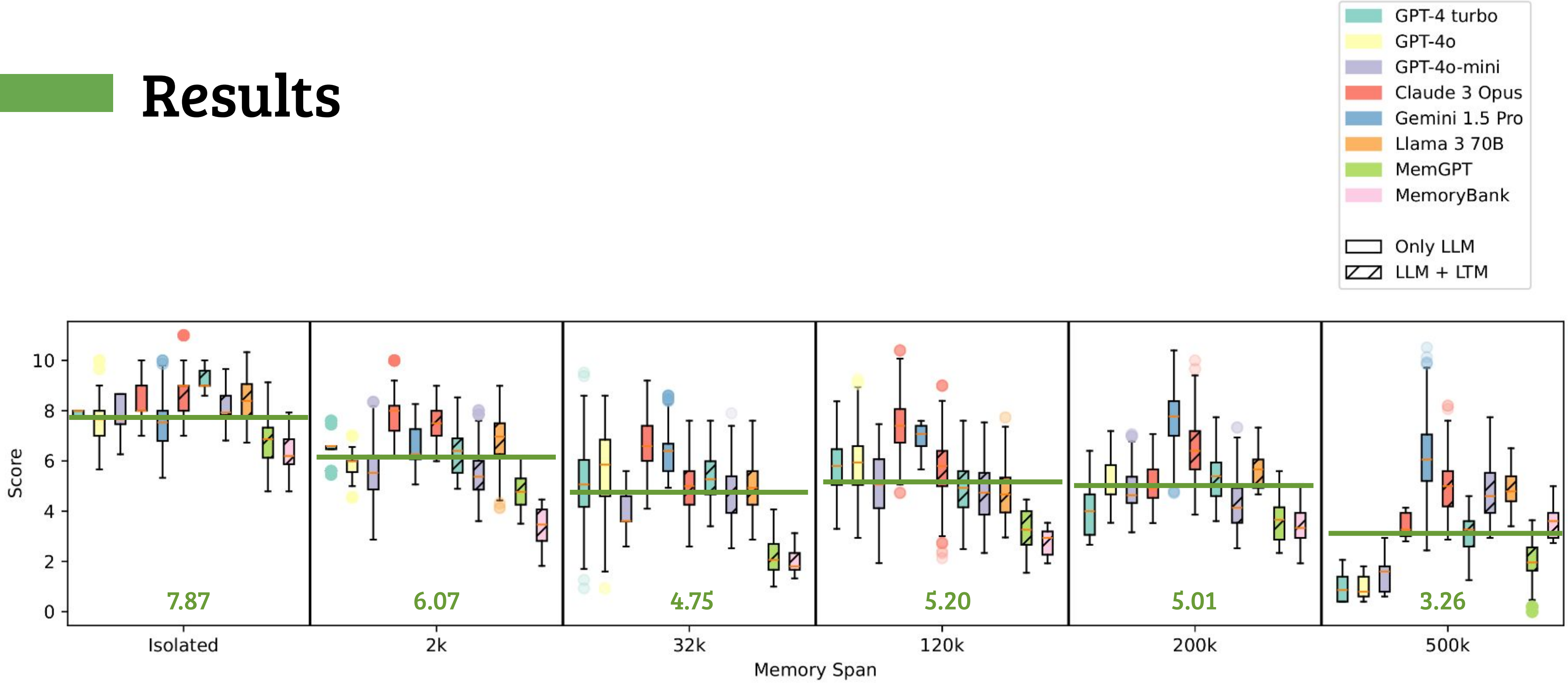

Conversation history
Truncation
LLM
Response
🟩 GPT-4 turbo
🟨 GPT-4o
🟪 GPT-4o-mini
🟥 Claude 3 Opus
🟦 Gemini 1.5 Pro
🟧 Llama 3 70B
🟩 MemGPT
🟪 MemoryBank
☐ Only LLM
▨ LLM + LTM
Score
10
8
6
4
2
0
Isolated
2k
32k
120k
200k
500k
Memory Span

Results


Agent
LLM
16k
LTM System
Message
Response

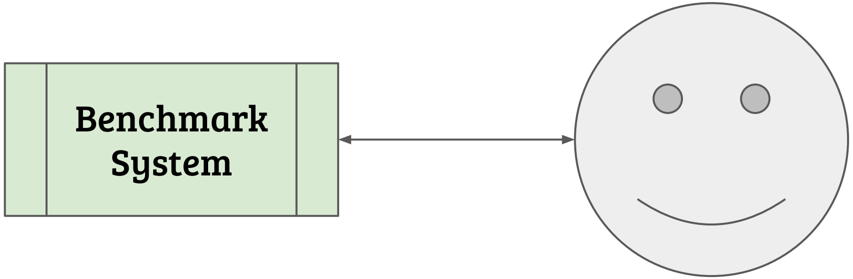

GPT-4 turbo
GPT-4o
GPT-4o-mini
Claude 3 Opus
Gemini 1.5 Pro
Llama 3 70B
MemGPT
MemoryBank
Only LLM
LLM + LTM
Score
0
2
4
6
8
10
Isolated
2k
32k
120k
200k
500k
Memory Span

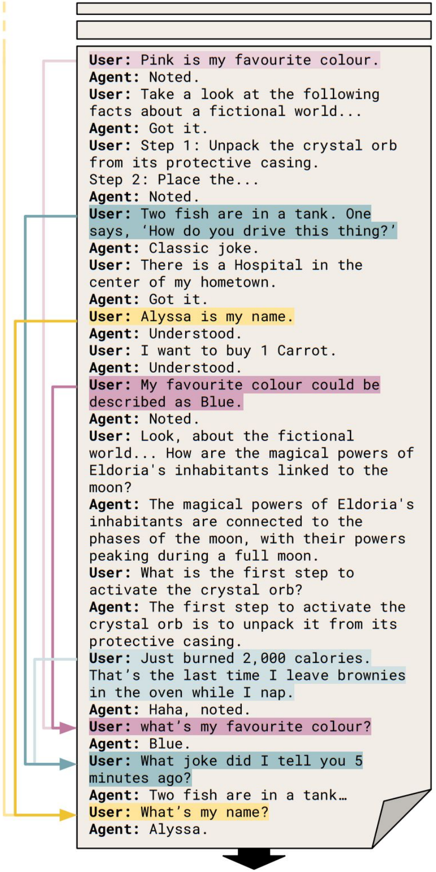
Results


GPT-4 turbo
GPT-4o
GPT-4o-mini
Claude 3 Opus
Gemini 1.5 Pro
Llama 3 70B
MemGPT
MemoryBank
Only LLM
LLM + LTM
Score
0
2
4
6
8
10
7.87
6.07
4.75
5.20
5.01
3.26
Isolated
2k
32k
120k
200k
500k
Memory Span

Conclusions

- Dynamic
- Realistic
- Challenging

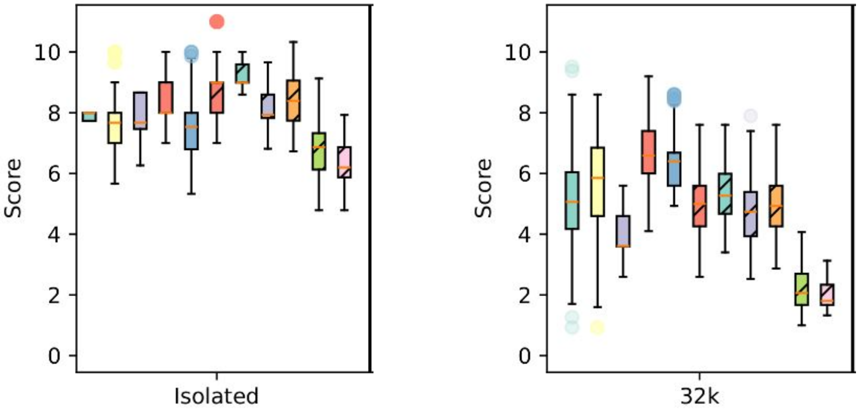




User: Pink is my favourite colour.
Agent: Noted.
User: Take a look at the following facts about a fictional world...
Agent: Got it.
User: Step 1: Unpack the crystal orb from its protective casing.
Step 2: Place the...
Agent: Noted.
User: Two fish are in a tank. One says, ‘How do you drive this thing?’
Agent: Classic joke.
User: There is a Hospital in the center of my hometown.
Agent: Got it.
User: Alyssa is my name.
Agent: Understood.
User: I want to buy 1 Carrot.
Agent: Understood.
User: My favourite colour could be described as Blue.
Agent: Noted.
User: Look, about the fictional world... How are the magical powers of Eldoria's inhabitants linked to the moon?
Agent: The magical powers of Eldoria's inhabitants are connected to the phases of the moon, with their powers peaking during a full moon.
User: What is the first step to activate the crystal orb?
Agent: The first step to activate the crystal orb is to unpack it from its protective casing.
User: Just burned 2,000 calories. That’s the last time I leave brownies in the oven while I nap.
Agent: Haha, noted.
User: what’s my favourite colour?
Agent: Blue.
User: What joke did I tell you 5 minutes ago?
Agent: Two fish are in a tank…
User: What’s my name?
Agent: Alyssa.

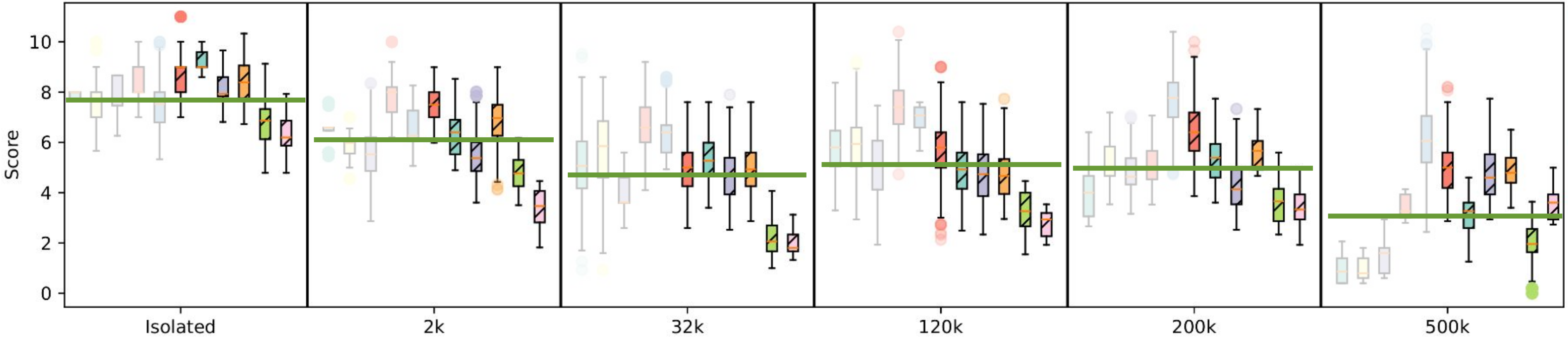
Conclusions

- Transfer to chat scenarios
- Task interleaving increases difficulty



Conclusions

- Scores drop with longer memory spans
- Agents with LTM suffer less in the long run
  - Shorter context + LTM may provide a focusing effect

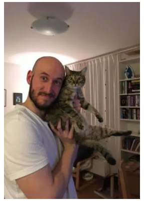
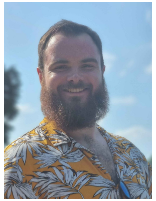
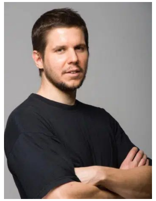


Future work

- Improve automatic evaluations
- Add more advanced tests
  - Forward and backward transfer
  - Integration of large amounts of information
  - Multi-modality
  - Multi-user scenarios 🔒

Beyond Prompts

Dynamic Conversational Benchmarking of Large Language Models

David.Castillo
@goodai.com

Joseph.Davidson
@goodai.com

Finlay.Gray
@goodai.com

Marek.Rosa
@goodai.com

github.com/GoodAI/goodai-ltm-benchmark

GoodAI

NEURAL INFORMATION
PROCESSING SYSTEMS

www.goodai.com