

SEMINAR

LLMs Get Lost In Multi-Turn Conversation

Lecturer: Lu Yao
Student: Martin Nguyen (Duc Q. Nguyen)

Department of Computer Science
School of Computing
National University of Singapore

August 29th 2025



“LLMs only possess statistical knowledge about word patterns, not true comprehension of ideas, facts, or emotions.”

I. Almeida

Table of Contents

1 Introduction

2 Related Works

3 Simulation Methodology

4 Benchmarking Results

5 Conclusion

Table of Contents

1 Introduction

2 Related Works

3 Simulation Methodology

4 Benchmarking Results

5 Conclusion

Motivation

An example with Llama 3.1 8B. The correct answer is 85,000 calories.

TURN 1

[User] calculate the total calorie count of Andrew's pastries

[Assistant] Let's assume we have some information about Andrew's pastries.

Question: Calculate the total calorie count of Andrew's 4 pastries.

Assumptions:

- * Andrew has 4 different types of pastries.
- * We have the calorie information for each type of pastry.
- * Each piece is a whole unit (e.g., one croissant, not half a croissant).

Type A: Chocolate Croissant

* Calories per serving: 250 calories / pastry

* Servings: 1

* Total Calories for Type A: 250 calories

Type B: Raspberry Mille-Feuille

* Calories per serving: 320 calories / pastry

* Servings: 1

* Total Calories for Type B: 320 calories

Type C: Hazelnut Tiramisu Pastry Strip (12 pieces)

* However, for this example we will use the average piece containing half of one strip which is around,
* Half piece: + Average Calories per serving: $80/120 = (\text{calories}) 67$ (for 3 strips)
* Total Calories for Type C: 84 calories

Type D: Apple Pastry Square Wedging with Flakes topping.

* We assume that type d pastry has three quarters of what a full piece contains
* Average per serving: $2/3 * 250 / (4+1) = 125$ calories/piece

TURN 2

[User] Andrew made 200 mini cinnamon rolls.

[Assistant] We were previously calculating the calorie count for a specific set of pastries, but now we have some additional information about Andrew's mini cinnamon rolls
To add these to our previous total, let's assign this new pastry set:

Type E: Mini Cinnamon Rolls

- * Calories per serving: 85 calories / mini roll (assuming each mini roll is approximately 1 inch in diameter)
- * Total pieces: 200
- * Total Calories for Type E: $200 * 85 = 17,000$ calories

TURN 3

[User] Oh, Andrew also baked 300 mini blueberry muffins.

[Assistant] We have another pastry to add to the mix.

Type F: Mini Blueberry Muffins

- * Calories per serving: approximately 120 calories / mini muffin
- * Total pieces: 300
- * Total Calories for Type F: $300 * 120 = 36,000$ calories

TURN 4

[User] Each mini pastry has about one-third the calories of its normal-sized version.

[Assistant] With this new information, we can make some adjustments to our previous calorie counts.

Type A: Chocolate Croissant

* Normal serving size: 250 calories

* Mini serving size (1/3 of normal): 80-100 calories (a range due to variability in ingredients and sizes)

* However for these calculations we will assume an estimate that is near average.

Type B: Raspberry Mille-Feuille

* Normal serving size: 320 calories

* Mini serving size (1/3 of normal): approximately 107 calories

Type C: Hazelnut Tiramisu Pastry Strip

Previous calculation used an estimate for a half cut portion

Type D: Apple Pastry Square Wedging with Flakes topping.

Previous calculations already included the adjustments

* Average per serving: assuming 1/4 what a full piece contains, therefore, 1/4 of (250) is approximately 62 calories/piece

TURN 5

[User] A regular-sized cinnamon roll has about 600 calories.

[Assistant] With this new information, we can calculate the calorie count for Andrew's mini cinnamon rolls

Type E: Mini Cinnamon Rolls

- * Calories per normal-sized serving: 600
- * Calories per mini serving (1/3 of normal): approximately 200 calories / mini roll
- * Total pieces: 200
- * Total Calories for Type E: $200 * 200 = 40,000$ calories

TURN 6

[User] I also remember that a regular blueberry muffin has about 450 calories.

[Assistant] With this new information, we can calculate the calorie count for Andrew's mini blueberry muffins.

Type F: Mini Blueberry Muffins

* Calories per normal-sized serving: 450

* Calories per mini serving (1/3 of normal): approximately 150 calories / mini muffin

* Total pieces: 300

* Total Calories for Type F: $300 * 150 = 45,000$ calories

Extracted: 45,000 Score = 0

Motivation

- Users usually start chatting with underspecified instructions. Thus, LLMs should cooperate to clarify those instructions.

Motivation

- Users usually start chatting with underspecified instructions. Thus, LLMs should cooperate to clarify those instructions.
- LLMs tend to make assumptions about missing required information and propose final solutions prematurely.

Motivation

- Users usually start chatting with underspecified instructions. Thus, LLMs should cooperate to clarify those instructions.
- LLMs tend to make assumptions about missing required information and propose final solutions prematurely.
- Many available LLM benchmarks focus on single-turn conversations, which do not reflect the real-world use cases.

Preliminaries & Problem Statement

Multi-turn conversation

A multi-turn conversation refers to an interactive dialogue between a user and an LLM where the user's intent or instruction is **gradually revealed over multiple exchanges**. These conversations often begin with underspecified queries, requiring the user to clarify, refine, or expand their needs through successive turns.

Preliminaries & Problem Statement

Multi-turn conversation

A multi-turn conversation refers to an interactive dialogue between a user and an LLM where the user's intent or instruction is **gradually revealed over multiple exchanges**. These conversations often begin with underspecified queries, requiring the user to clarify, refine, or expand their needs through successive turns.

Lost in conversation phenomenon

The lost-in-conversation phenomenon refers to a pattern observed in LLMs during multi-turn interactions, where the model fails to recover after making an error or misinterpretation in earlier turns.

Preliminaries & Problem Statement

Multi-turn conversation

A multi-turn conversation refers to an interactive dialogue between a user and an LLM where the user's intent or instruction is **gradually revealed over multiple exchanges**. These conversations often begin with underspecified queries, requiring the user to clarify, refine, or expand their needs through successive turns.

Lost in conversation phenomenon

The lost-in-conversation phenomenon refers to a pattern observed in LLMs during multi-turn interactions, where the model fails to recover after making an error or misinterpretation in earlier turns.

Multi-turn conversation benchmarks

Can we develop a benchmarking method that measures the extent to which models exhibit the lost-in-conversation phenomenon?

Table of Contents

1 Introduction

2 Related Works

3 Simulation Methodology

4 Benchmarking Results

5 Conclusion

Early-day LLMs and Multi-turn Conversations

- Early-day LLMs such as BART¹, GPT-2², T5³ are mainly used for single-turn conversations.

¹Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension".

²Radford et al., "Language models are unsupervised multitask learners".

³Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer".

⁴Konrád et al., "Alquist 4.0: Towards social intelligence using generative models and dialogue personalization".

⁵Deriu et al., "Survey on evaluation methods for dialogue systems"; Lee et al., "Evaluating human-language model interaction".

⁶L. Zheng et al., "Judging llm-as-a-judge with mt-bench and chatbot arena".

⁷Kwan et al., "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models"; Bai et al., "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues"; Sirdeshmukh et al., "Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms".

⁸Laban et al., "LLMs get lost in multi-turn conversation".

Early-day LLMs and Multi-turn Conversations

- Early-day LLMs such as BART¹, GPT-2², T5³ are mainly used for single-turn conversations.
- Initial effort was spent to add multi-turn context understanding⁴ and evaluate LLMs in this setting using human⁵.

¹Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension".

²Radford et al., "Language models are unsupervised multitask learners".

³Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer".

⁴Konrád et al., "Alquist 4.0: Towards social intelligence using generative models and dialogue personalization".

⁵Deriu et al., "Survey on evaluation methods for dialogue systems"; Lee et al., "Evaluating human-language model interaction".

⁶L. Zheng et al., "Judging llm-as-a-judge with mt-bench and chatbot arena".

⁷Kwan et al., "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models"; Bai et al., "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues"; Sirdeshmukh et al., "Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms".

⁸Laban et al., "LLMs get lost in multi-turn conversation".

Early-day LLMs and Multi-turn Conversations

- Early-day LLMs such as BART¹, GPT-2², T5³ are mainly used for single-turn conversations.
- Initial effort was spent to add multi-turn context understanding⁴ and evaluate LLMs in this setting using human⁵.
- MT-Bench⁶, and derivative works⁷ have been proposed to benchmark LLMs in multi-turn conversation settings automatically. However, these benchmarks are typically episodic conversations⁸.

¹Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension".

²Radford et al., "Language models are unsupervised multitask learners".

³Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer".

⁴Konrád et al., "Alquist 4.0: Towards social intelligence using generative models and dialogue personalization".

⁵Deriu et al., "Survey on evaluation methods for dialogue systems"; Lee et al., "Evaluating human-language model interaction".

⁶L. Zheng et al., "Judging llm-as-a-judge with mt-bench and chatbot arena".

⁷Kwan et al., "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models"; Bai et al., "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues"; Sirdeshmukh et al., "Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms".

⁸Laban et al., "LLMs get lost in multi-turn conversation".

Multi-turn Benchmarks

- Multi-turn episodic evaluation differs from our multi-turn evaluation in that the question in each turn can be completely solved in the next response¹.

¹Laban et al., "LLMs get lost in multi-turn conversation".

²Bai et al., "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues"; Kwan et al., "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models"; Sun et al., "Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models"; Fan et al., "FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs".

³E. Choi et al., "QuAC: Question answering in context"; Poelitz and McKenna, "Synthetic Clarification and Correction Dialogues about Data-Centric Tasks—A Teacher-Student Approach"; S. E. Finch, J. D. Finch, and J. D. Choi, "Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems"; Ram et al., "Conversational ai: The science behind the alexa prize".

Multi-turn Benchmarks

- Multi-turn episodic evaluation differs from our multi-turn evaluation in that the question in each turn can be completely solved in the next response¹.
- Each turn in these episodic benchmarks² contains a dedicated subtask, making the final goal differ from the single-turn benchmark with the same question.

¹Laban et al., "LLMs get lost in multi-turn conversation".

²Bai et al., "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues"; Kwan et al., "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models"; Sun et al., "Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models"; Fan et al., "FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs".

³E. Choi et al., "QuAC: Question answering in context"; Poelitz and McKenna, "Synthetic Clarification and Correction Dialogues about Data-Centric Tasks—A Teacher-Student Approach"; S. E. Finch, J. D. Finch, and J. D. Choi, "Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems"; Ram et al., "Conversational ai: The science behind the alexa prize".

Multi-turn Benchmarks

- Multi-turn episodic evaluation differs from our multi-turn evaluation in that the question in each turn can be completely solved in the next response¹.
- Each turn in these episodic benchmarks² contains a dedicated subtask, making the final goal differ from the single-turn benchmark with the same question.
- Automatically multi-turn evaluation requires an interacting user. Prior works³ have explored multiple ways to simulate a user using an LLM.

¹Laban et al., "LLMs get lost in multi-turn conversation".

²Bai et al., "MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues"; Kwan et al., "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models"; Sun et al., "Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models"; Fan et al., "FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs".

³E. Choi et al., "QuAC: Question answering in context"; Poelitz and McKenna, "Synthetic Clarification and Correction Dialogues about Data-Centric Tasks—A Teacher-Student Approach"; S. E. Finch, J. D. Finch, and J. D. Choi, "Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems"; Ram et al., "Conversational ai: The science behind the alexa prize".

Table of Contents

1 Introduction

2 Related Works

3 Simulation Methodology

4 Benchmarking Results

5 Conclusion

Overview

This study proposes a novel benchmarking method for multi-turn conversations. The overall concept is summarized as follows:

Overview

This study proposes a novel benchmarking method for multi-turn conversations. The overall concept is summarized as follows:

- Semi-automatically convert well-known single-turn evaluation datasets into multi-turn evaluation datasets using a strong LLM.

Overview

This study proposes a novel benchmarking method for multi-turn conversations. The overall concept is summarized as follows:

- Semi-automatically convert well-known single-turn evaluation datasets into multi-turn evaluation datasets using a strong LLM.
- Design a simulated user that has access to the full information of an instruction/question and can provide answers to any clarification queries about that instruction.

Overview

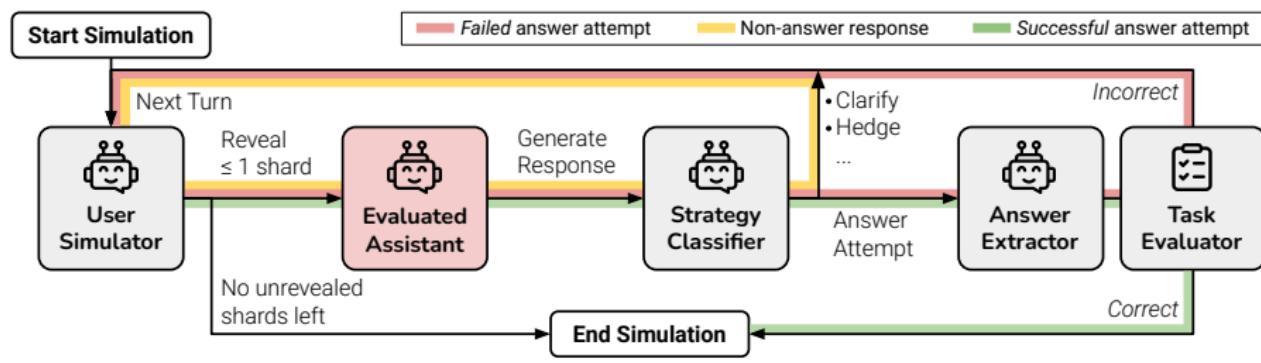
This study proposes a novel benchmarking method for multi-turn conversations. The overall concept is summarized as follows:

- Semi-automatically convert well-known single-turn evaluation datasets into multi-turn evaluation datasets using a strong LLM.
- Design a simulated user that has access to the full information of an instruction/question and can provide answers to any clarification queries about that instruction.
- Develop a classifier that can detect whether an LLM's response is a final-answer attempt or not.

Overview

This study proposes a novel benchmarking method for multi-turn conversations. The overall concept is summarized as follows:

- Semi-automatically convert well-known single-turn evaluation datasets into multi-turn evaluation datasets using a strong LLM.
- Design a simulated user that has access to the full information of an instruction/question and can provide answers to any clarification queries about that instruction.
- Develop a classifier that can detect whether an LLM's response is a final-answer attempt or not.



Semi-Automatic Sharding Process

- ① **Segmentation:** LLM extracts non-overlapping segments (Atomic Content Units). → Filter out if fewer than 3 segments.

Semi-Automatic Sharding Process

- ① **Segmentation:** LLM extracts non-overlapping segments (Atomic Content Units). → Filter out if fewer than 3 segments.
- ② **Rephrasing:** Segments are rewritten to be conversational and decontextualized, ensuring P2 & P5.

Semi-Automatic Sharding Process

- ① **Segmentation:** LLM extracts non-overlapping segments (Atomic Content Units). → Filter out if fewer than 3 segments.
- ② **Rephrasing:** Segments are rewritten to be conversational and decontextualized, ensuring P2 & P5.
- ③ **Verification:** Run preliminary simulations. Accept the conversation if $P_{\text{CONCAT}} \geq 0.8P_{\text{FULL}}$, $P_{\text{SHUFFLE-CONCAT}} \geq 0.8P_{\text{FULL}}$.

Semi-Automatic Sharding Process

- ① **Segmentation:** LLM extracts non-overlapping segments (Atomic Content Units). → Filter out if fewer than 3 segments.
- ② **Rephrasing:** Segments are rewritten to be conversational and decontextualized, ensuring P2 & P5.
- ③ **Verification:** Run preliminary simulations. Accept the conversation if $P_{\text{CONCAT}} \geq 0.8P_{\text{FULL}}$, $P_{\text{SHUFFLE-CONCAT}} \geq 0.8P_{\text{FULL}}$.
- ④ **Inspect & Edit:** Manual review with a web-based interface.

Semi-Automatic Sharding Process

- ① **Segmentation:** LLM extracts non-overlapping segments (Atomic Content Units). → Filter out if fewer than 3 segments.
- ② **Rephrasing:** Segments are rewritten to be conversational and decontextualized, ensuring P2 & P5.
- ③ **Verification:** Run preliminary simulations. Accept the conversation if $P_{\text{CONCAT}} \geq 0.8P_{\text{FULL}}$, $P_{\text{SHUFFLE-CONCAT}} \geq 0.8P_{\text{FULL}}$.
- ④ **Inspect & Edit:** Manual review with a web-based interface.

0. Prepare	1. Segmentation	2. Rephrasing	3. Verification	4. Inspection & Edit
Jay is making snowballs to prepare for a snowball fight with his sister. He can build 20 snowballs in an hour, but 2 melt every 15 minutes. How long will it take before he has 60 snowballs? [GSM8K]	Jay is making snowballs to prepare for a snowball fight with his sister. He can build 20 snowballs in an hour, but 2 melt every 15 minutes. How long will it take before he has 60 snowballs?	How long before Jay's ready for the snowball fight? He's preparing for a snowball fight with his sister. He can build 20 snowballs in an hour He wants 60 snowballs. Two snowballs melt every 15 minutes.	Simulation 10x FULL 10x CONCAT 10x SHUFFLE-CONCAT $\bar{P}_{\text{CONCAT}} \geq 0.8 \bar{P}_{\text{FULL}}$ $\bar{P}_{\text{SHUFFLE-CONCAT}} \geq 0.8 \bar{P}_{\text{FULL}}$	How long before Jay's ready for the snowball fight? He's preparing for a snowball fight with his sister. He can make 20 snowballs per hour. He's trying to get to 60 total. The problem is that 2 melt every 15 minutes.
< 3 segments		Below degradation thresholds	Manual decision	

Definition of Sharding

Goal: Transform a single-turn query into multiple shorter *instruction shards* that can simulate a multi-turn conversation.

¹Laban et al., "LLMs get lost in multi-turn conversation".

Definition of Sharding

Goal: Transform a single-turn query into multiple shorter *instruction shards* that can simulate a multi-turn conversation.

Terminology:

- Let q = a single-turn query with intended output Y_q^* .
- Atomic Content Units (ACU): $I(q) = [I, (c_1, \dots, c_m)]$ where I = primary intent, and (c_1, \dots, c_m) are clarifications.
- Atomicity: any rephrasing of $I(q)$ yields the same output Y_q^* .

¹Laban et al., "LLMs get lost in multi-turn conversation".

Definition of Sharding

Goal: Transform a single-turn query into multiple shorter *instruction shards* that can simulate a multi-turn conversation.

Terminology:

- Let q = a single-turn query with intended output Y_q^* .
- Atomic Content Units (ACU): $I(q) = [I, (c_1, \dots, c_m)]$ where I = primary intent, and (c_1, \dots, c_m) are clarifications.
- Atomicity: any rephrasing of $I(q)$ yields the same output Y_q^* .

Sharding Process:

$$q' = [s_1, \dots, s_k] \quad \text{such that } I(q) = I(q')$$

where s_j are instruction shards used in multi-turn conversation.

¹Laban et al., "LLMs get lost in multi-turn conversation".

Sharded Conversation Properties

A sharded instruction q' is valid for q if it satisfies:

¹Laban et al., "LLMs get lost in multi-turn conversation".

Sharded Conversation Properties

A sharded instruction q' is valid for q if it satisfies:

- ① **Information Preservation:** $I(q) = I(q')$ (no loss of necessary info).

¹Laban et al., "LLMs get lost in multi-turn conversation".

Sharded Conversation Properties

A sharded instruction q' is valid for q if it satisfies:

- ① **Information Preservation:** $I(q) = I(q')$ (no loss of necessary info).
- ② **Clear Initial Intent:** $s_1 = I_q$ (first shard sets high-level objective).

¹Laban et al., "LLMs get lost in multi-turn conversation".

Sharded Conversation Properties

A sharded instruction q' is valid for q if it satisfies:

- ① **Information Preservation:** $I(q) = I(q')$ (no loss of necessary info).
- ② **Clear Initial Intent:** $s_1 = I_q$ (first shard sets high-level objective).
- ③ **Order Insensitive:** Shards $s_2 \dots s_k$ can be permuted without changing meaning.

¹Laban et al., "LLMs get lost in multi-turn conversation".

Sharded Conversation Properties

A sharded instruction q' is valid for q if it satisfies:

- ① **Information Preservation:** $I(q) = I(q')$ (no loss of necessary info).
- ② **Clear Initial Intent:** $s_1 = I_q$ (first shard sets high-level objective).
- ③ **Order Insensitive:** Shards $s_2 \dots s_k$ can be permuted without changing meaning.
- ④ **Maximal Sharding:** Maximize k by extracting fine-grained shards (one info unit each).

¹Laban et al., "LLMs get lost in multi-turn conversation".

Sharded Conversation Properties

A sharded instruction q' is valid for q if it satisfies:

- ① **Information Preservation:** $I(q) = I(q')$ (no loss of necessary info).
- ② **Clear Initial Intent:** $s_1 = I_q$ (first shard sets high-level objective).
- ③ **Order Insensitive:** Shards $s_2 \dots s_k$ can be permuted without changing meaning.
- ④ **Maximal Sharding:** Maximize k by extracting fine-grained shards (one info unit each).
- ⑤ **Minimal Transformation:** Preserve original language; avoid unnecessary simplification/interpretation.

¹Laban et al., "LLMs get lost in multi-turn conversation".

Simulation Types for Multi-Turn Benchmarking

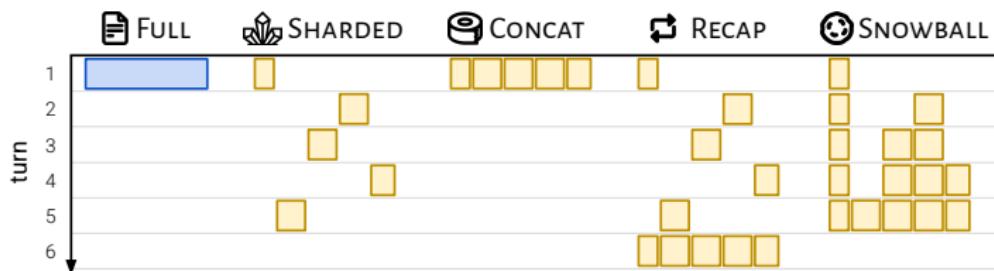
Instruction Sharding

Fully-specified
Single-Turn



Sharded
Multi-Turn

Conversation Simulation Types



¹Laban et al., "LLMs get lost in multi-turn conversation".

Strategy Classifier

The strategy classifier utilizes a strong LLM (i.e., GPT-4o-mini) to classify the response of the evaluated LLM.

Name	Description	Example
Answer attempt	The response contains a complete answer attempt to the question that can be extracted verbatim.	The dog is 50 meters away from the house.
Clarification	The response is a brief single question that directly inquires about one aspect of the query.	To calculate the distance, I need to know how long the dog ran. Could you provide more information about that?
Interrogation	The response contains multiple questions addressed to the user.	I cannot answer the question without knowing (1) speed, (2) duration, and (3) starting position. Please tell me about these points and I can calculate the distance!
Discussion	The response discusses the question in detail without answering, asking, or refusing to answer.	The question is trying to measure the distance between the dog and the house. We can calculate based on this equation: [Equation]. [...]
Hedging	The response provides multiple answer candidates based on hypotheticals (ifs, cases).	<ol style="list-style-type: none">1. If the dog was originally in the house, it would be 50 meters away now.2. If the dog was at the park, it would be 100 meters away from the house now.
Refusal	The response refuses to answer the question without a follow-up question or a request.	I can't answer your question because I don't have sufficient information.
Missing	The response is empty.	[blank]

Table of Contents

1 Introduction

2 Related Works

3 Simulation Methodology

4 Benchmarking Results

5 Conclusion

Datasets

The benchmark includes 6 tasks with 12 datasets.

PL Generation Tasks



Code

Database

Actions

Math

Data-to-Text

Summary

Fully-Specified Instruction

Write the Python function

```
def below_zero(log):  
    """ You're given a list of  
    deposits & withdrawals on a bank  
    account that starts with balance  
    of 0. If the balance goes below  
    the balance < 0, if so return True,  
    otherwise False.  
    *** [2 example uses]  
    """
```

Write an SQL query for:

Find the names of stores
whose number products is
more than the average number
of products per store.
[Schema]

Write API function calls:

Play songs from the artists
Taylor Swift and Maroon 5,
with a play time of 20 minutes
and 15 minutes respectively,
on Spotify.
[API spec]

Solve this problem:

Josh decides to try flipping a
house. He buys a house for
\$80k and then puts in \$50k in
repairs. This increased the
value of the house by 150%.
How much profit did he make?

Write a Table caption:

[Highlighted Table HTML]
The table comes from [URL]
about the 2000 Americas
Cricket Cup.
I've highlighted some cells.

Write a Summary:

About the following 12
documents, on the following
query: [QUERY]
Documents:
[Documents 1-12]

Sharded Instructions

Write me a function below_zero
to find out if account is ever <0

Let's find large stores

Let's make a 35-min playlist

My friend Josh sold his home. I
want to know how much profit
he made.

I'm giving you a table, please
write a sentence describing
it. [Table HTML]

I need a summary of 12
documents, on query: [QUERY]
I'll give the docs as I get them,
consider all of them.
Docs 1-2: [Documents 1-2]

Input's a list of ints that are
transactions.

Maybe we can define store
size based on its number of
products

Let's add Taylor Swift songs

He bought it for \$80,000

Actually focus on these
highlighted cells:
[Highlighted Table HTML]

Just got four more.
Docs 3-6: [Documents 3-6]

Balance is 0 at the start.

A store is large if it has more
than the average number of
products across all stores.

Let's also put some Maroon 5

He spent \$50k on repairs

It came from a page about the
2000 Americas Cricket Cup

Here's a new batch.
Docs 7-10: [Documents 7-10]

Return True if balance's ever <0,
o/w return False

[Example 1]

[Example 2]

I prefer Taylor Swift, let's do
20 minutes of that

The house value increased by
150%

That's all I know.

What's his
profit?

I've got two more.
Docs 11-12: [Documents 11-12]

Instruction Source & Evaluation

HumanEval &
LiveCodeBench

Spider

Berkeley Function
Calling Leaderboard

GSM8K

ToTTo

Summary of a Haystack

Functional
Accuracy

Functional
Accuracy

Exact Match

Exact Match

BLEU

Coverage & Citation



Models

Short Form	Name	Version	Access Provider
⟳ 4o	GPT-4o	gpt-4o-2024-11-20	OpenAI / Microsoft API
⟳ 4o-mini	GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI API
⟳ 4.1	GPT-4.1	gpt-4.1-2025-04-14	OpenAI / Microsoft API
⟳ o3	o3	o3-2025-04-16	OpenAI / Microsoft API
⟳ 3-Haiku	Claude 3 Haiku	claude-3-haiku-20240307	Amazon Bedrock
⟳ 3.7-Sonnet	Claude 3.7 Sonnet	claude-3-7-sonnet-20250219	Amazon Bedrock
◆ 2.5-Flash	Gemini 2.5 Flash	gemini-2.5-flash-preview-04-17	Gemini API
◆ 2.5-Pro	Gemini 2.5 Pro	gemini-2.5-pro-preview-03-25	Gemini API
♾ 3.1-8B	Llama-3.1-8B-Instruct	N/A	Local Ollama
♾ 3.3-70B	Llama-3.3-70B-Instruct	N/A	Amazon Bedrock
♾ 4-Scout	Llama-4-Scout-17B-16E	N/A	Together AI
🤖 CMD-A	Command-A	command-a-03-2025	Cohere API
🐦 R1	Deepseek-R1	N/A	Amazon Bedrock
✳️ OLMo2	OLMo2-13B	N/A	Local Ollama
💻 Phi-4	Phi-4	N/A	Local Ollama

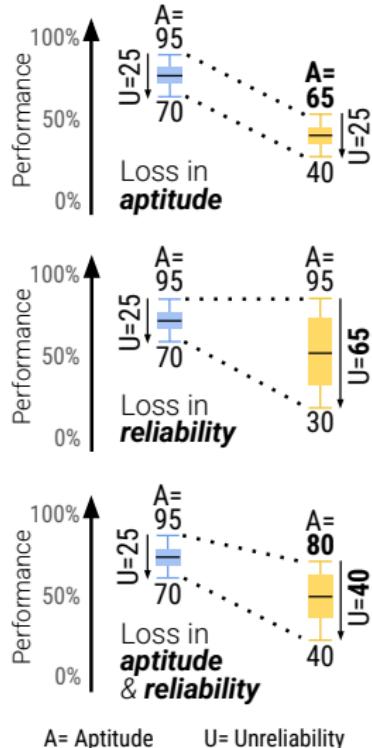
¹Laban et al., "LLMs get lost in multi-turn conversation".

Evaluation Metrics

Setup: LLMs generate stochastic responses ($T = 1.0$). \Rightarrow Run N simulations per instruction, each with score $S_i \in [0, 100]$.

Metrics (per instruction):

- **Average Performance (\bar{P})**: Mean score across simulations.
- **Aptitude (A)**: 90th percentile score (A_{90}). Captures *best-case* performance.
- **Unreliability (U)**: $U_{90}^{10} = A_{90} - A_{10}$. Measures the gap between best-case and worst-case simulations.
- **Reliability (R)**: $R_{90}^{10} = 100 - U_{90}^{10}$.



¹Laban et al., "LLMs get lost in multi-turn conversation".

Experiment Setup

The main experiments are run with default configurations as follows.

- Total samples (instructions): 600 across six tasks
- Temperature: 1.0
- Number of simulations per instruction: 10

¹Laban et al., "LLMs get lost in multi-turn conversation".

Averaged Performance

Averaged Performance (\bar{P}) of LLMs on six tasks.

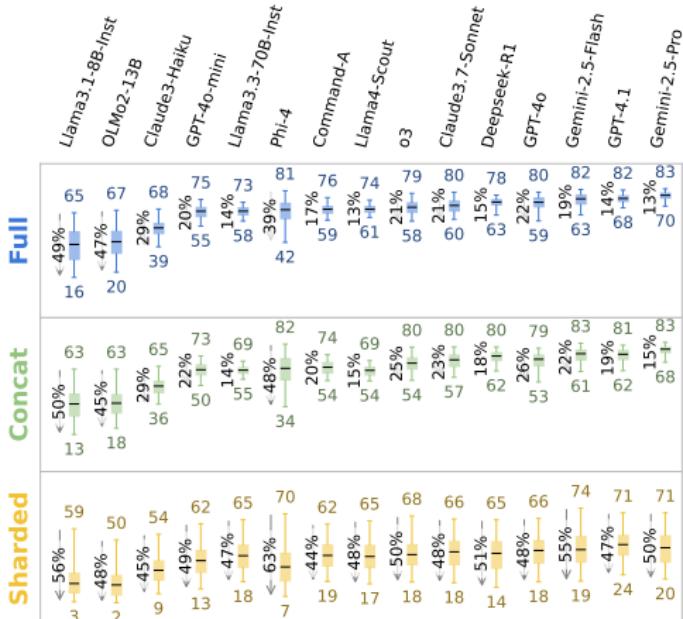
Model	FULL						CONCAT						SHARDED						Overall	
3.1-8B	27.4	64.1	82.9	13.7	63.9	7.6	21.2	47.7	83.0	15.7	62.6	6.5	21.7	25.9	45.5	13.3	37.4	3.4	91.6	62.5
OLMo2	18.8	54.8	56.1	17.2	80.0	-	16.3	40.5	49.8	14.3	80.1	-	14.4	22.4	13.8	9.0	46.3	-	86.5	50.5
3-Haiku	44.8	85.0	83.5	29.8	73.9	11.6	36.3	76.5	80.2	30.1	76.1	9.2	31.5	31.8	55.9	18.6	47.1	1.6	91.6	52.4
4o-mini	75.9	89.3	94.1	35.9	88.1	14.9	66.7	90.7	92.2	31.2	88.0	12.5	50.3	40.2	52.4	19.8	58.7	7.2	93.0	56.2
3.3-70B	72.0	91.1	95.0	34.1	91.7	15.8	52.7	87.9	97.0	32.0	91.8	14.7	51.6	35.4	71.0	22.4	61.5	10.5	93.2	64.2
Phi-4	53.2	87.6	82.7	23.9	89.2	-	48.4	79.6	76.0	28.6	90.4	-	39.1	33.1	34.1	23.2	52.5	-	99.0	61.7
CMD-A	72.0	91.9	98.5	27.7	94.5	24.3	61.6	86.1	98.4	33.2	91.9	21.3	44.9	33.6	72.0	27.9	66.0	4.9	97.3	60.4
4-Scout	73.9	92.7	98.0	35.2	96.3	13.7	60.3	81.5	98.3	28.2	92.9	13.7	46.4	27.1	69.9	26.1	67.0	12.3	91.0	66.1
o3	86.4	92.0	89.8	40.2	81.6	30.7	87.2	83.3	91.5	39.4	80.0	30.4	53.0	35.4	60.2	21.7	63.1	26.5	98.1	64.1
3.7-Sonnet	78.0	93.9	95.4	45.6	85.4	29.3	76.2	81.5	96.0	53.3	87.2	28.9	65.6	34.9	33.3	35.1	70.0	23.6	100.4	65.9
R1	99.4	92.1	97.0	27.0	95.5	26.1	97.1	89.9	97.0	36.7	92.9	24.4	70.9	31.5	47.5	20.0	67.3	17.2	103.6	60.8
4o	88.4	93.6	96.1	42.1	93.8	23.9	82.9	91.7	97.1	32.2	91.9	23.9	61.3	42.3	65.0	20.5	67.9	10.6	94.5	57.9
2.5-Flash	97.0	96.3	88.4	51.2	90.6	29.1	92.5	95.5	89.2	51.9	88.4	29.4	68.3	51.3	42.6	31.0	66.1	26.1	99.3	65.8
4.1	96.6	93.0	94.7	54.6	91.7	26.5	88.7	86.5	98.5	54.4	89.7	26.8	72.6	46.0	62.9	28.6	70.7	13.3	97.9	61.8
2.5-Pro	97.4	97.3	97.8	54.8	90.2	31.2	95.7	94.9	98.1	56.9	89.3	31.8	68.1	43.8	36.3	46.2	64.3	24.9	100.1	64.5

¹Laban et al., "LLMs get lost in multi-turn conversation".

Aptitude vs. Reliability Analysis

Observations:

- Single-turn settings: models with higher aptitude tend to be more reliable.
- Multi-turn settings: All models exhibit very high unreliability, with performance degrading 50% on average between the best and worst simulated run for a fixed instruction.



¹Laban et al., "LLMs get lost in multi-turn conversation".

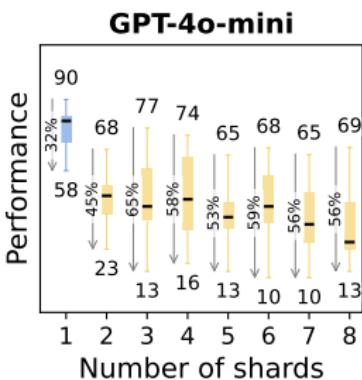
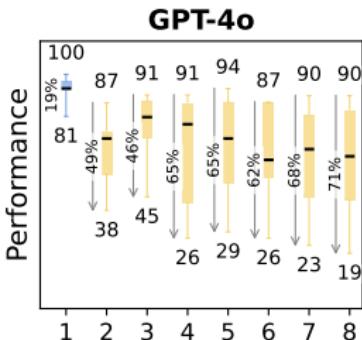
Gradual Sharding Experiment

Setup:

- 31 instructions across multiple tasks.
- Each expanded into 12–8 shards.

Observations:

- With ≥ 2 shards: models *get lost in conversation*.
- Results: minor aptitude drop, large reliability loss.
- Granularity has little effect—only 1-shard (all info at once) improves reliability.



¹Laban et al., "LLMs get lost in multi-turn conversation".

Result Implications

Do we need native multi-turn support in LLMs when agent frameworks can orchestrate interactions with users and leverage LLMs only as single-turn operators?

Model	Simulation Type				
	File	Image	Table	Code	Text
4o-mini	86.8	84.4	50.4	66.5	61.8
4o	93.0	90.9	59.1	76.6	65.3

¹Laban et al., "LLMs get lost in multi-turn conversation".

Result Implications

Do we need native multi-turn support in LLMs when agent frameworks can orchestrate interactions with users and leverage LLMs only as single-turn operators?

LLMs are probabilistic systems. Does setting the temperature to its lowest setting ($T = 0$) effectively resolve the reliability concern?

Model	Simulation Type				
4o-mini	86.8	84.4	50.4	66.5	61.8
4o	93.0	90.9	59.1	76.6	65.3

Simulation	4o-mini			4o		
	AT=1.0	AT=0.5	AT=0.0	AT=1.0	AT=0.5	AT=0.0
FULL	16.0	15.0	6.8	17.8	8.0	2.8
CONCAT	20.2	17.8	9.5	20.2	17.8	5.8
UT=1.0	49.8	46.8	51.0	41.0	43.8	31.8
UT=0.5	31.7	34.0	40.5	39.5	40.8	31.8
UT=0.0	38.5	28.0	30.5	35.8	38.0	29.7

¹Laban et al., "LLMs get lost in multi-turn conversation".

Result Implications

Do we need native multi-turn support in LLMs when agent frameworks can orchestrate interactions with users and leverage LLMs only as single-turn operators?

LLMs are probabilistic systems. Does setting the temperature to its lowest setting ($T = 0$) effectively resolve the reliability concern?

Do other tasks require LLMs with multi-turn capability to work better? An example of a translation task on the WMT 2019 dataset and the BLEU score.

Model	Simulation Type				
4o-mini	86.8	84.4	50.4	66.5	61.8
4o	93.0	90.9	59.1	76.6	65.3

Simulation	4o-mini			4o		
	AT=1.0	AT=0.5	AT=0.0	AT=1.0	AT=0.5	AT=0.0
FULL	16.0	15.0	6.8	17.8	8.0	2.8
CONCAT	20.2	17.8	9.5	20.2	17.8	5.8
UT=1.0	49.8	46.8	51.0	41.0	43.8	31.8
UT=0.5	31.7	34.0	40.5	39.5	40.8	31.8
UT=0.0	38.5	28.0	30.5	35.8	38.0	29.7

Model	Translation		
4o-mini	41.7	43.4	42.1
4o	35.9	38.5	40.9

¹Laban et al., "LLMs get lost in multi-turn conversation".

Impact of Premature Answer Attempts

Setup: Responses categorized by the order of the turn in which the first answer attempt was produced.

Observations:

- Early answer attempts (first 20%) → poor performance (avg. 30.9).
- Late attempts (last 20%) → much higher performance (avg. 64.4).
- Effect observed only in Math and Code tasks.

Model	Conversation Progress At First Answer Attempt				
	0-20%	20-40%	40-60%	60-80%	80-100%
First answer attempt is ...	earliest	early	midway	late	latest
∞ 3.1-8B	16.1	24.0	35.3	39.6	39.7
❖ OLMo2	17.6	32.7	37.7	47.3	26.4
▲ 3-Haiku	27.1	35.6	47.4	58.9	70.3
◎ 4o-mini	30.2	39.2	48.4	58.2	59.9
∞ 3.3-70B	33.3	40.1	51.2	60.0	69.3
■ Phi-4	25.7	33.1	47.0	53.0	57.9
■ CMD-A	38.0	42.9	56.5	65.5	73.5
∞ 4-Scout	39.8	36.8	51.0	57.9	64.8
◎ o3	21.0	37.9	51.9	58.4	68.0
▲ 3.7-Sonnet	29.2	35.6	55.3	68.0	71.6
❖ R1	39.5	43.1	53.5	66.4	50.2
◎ 4o	36.0	41.4	56.2	65.6	90.4
◆ 2.5-Flash	39.0	48.6	60.2	70.8	74.6
◎ 4.1	33.9	52.7	60.6	69.0	78.6
◆ 2.5-Pro	41.1	45.7	53.5	64.6	63.8
Average	30.9	40.5	51.7	60.4	64.4

Insight: Premature solutions “lock in” incorrect assumptions, degrading overall LLM performance.

¹Laban et al., “LLMs get lost in multi-turn conversation”.

Answer Bloat Effect in Multi-Turn Conversations

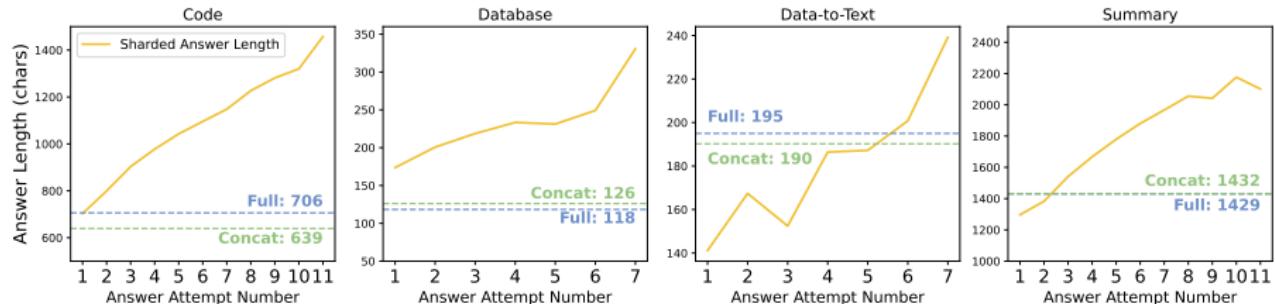
Setup:

- Compare answer lengths across FULL, CONCAT, and SHARDED.
- Focus on extracted answer attempts (e.g., SQL query, Python function).

Findings:

- FULL vs. CONCAT: similar lengths (within 2–10%).
- SHARDED: first attempt \approx same, but later attempts grow longer.
- Final SHARDED answers are **20–300% longer**.

Insights: LLMs fail to invalidate early assumptions and make assumptions about unspecified information, so later solutions expand unnecessarily.



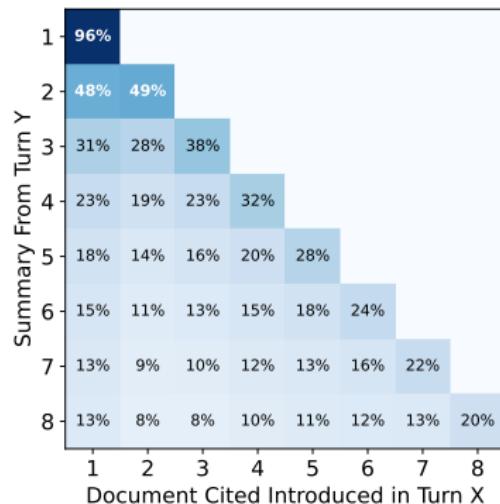
¹Laban et al., "LLMs get lost in multi-turn conversation".

Over-adjust based on Last Turn

Setup: Summary task with citations → allows tracking which turns documents are referenced.

Findings:

- Turn 1 summaries: 96% cite first-turn docs; 4% hallucinated.
- Turn 2: balanced citations (48% from doc #1, 49% from doc #2).
- Turns 3-8: strong imbalance. E.g., Turn 8 cites 20% from doc #8 vs. only 8% from docs #2-#3.



Insights: LLMs prefer documents from **first** and **last** turns, neglecting middle turns. → Mirrors known *loss-in-the-middle* effect.

¹Laban et al., "LLMs get lost in multi-turn conversation".

Verbosity Analysis in Multi-Turn Conversations

Setup:

- For each (LLM, instruction), compute avg. response length (chars/turn).
- Assign two simulations of each instruction to a quintile: shortest, short, median, long, longest.

Task	Relative Assistant Verbosity				
	0-20%	20-40%	40-60%	60-80%	80-100%
Assistants responses are ...	shortest	short	median	long	longest
Code	55.3	52.3	48.9	46.9	42.5
Math	62.9	64.0	62.1	60.9	56.1
Database	43.8	40.0	37.3	34.3	31.3
Actions	41.5	49.6	54.2	53.6	50.8
Data-to-Text	25.0	24.3	24.0	23.1	21.8
Summary	15.4	14.7	13.5	12.0	10.3
Average	40.7	40.8	40.1	38.6	35.6

Findings:

- Longer responses degrade performance, except for the Actions task.
- Over-verbosity → more assumptions, confusion in later turns.

Insights:

- Short, focused responses keep the conversation on track.
- Long responses harm both **model reliability** and **user experience**.

Table of Contents

1 Introduction

2 Related Works

3 Simulation Methodology

4 Benchmarking Results

5 Conclusion

Summary

- Large-scale simulations show LLMs degrade significantly in multi-turn, underspecified conversations.
- Degradation arises from:
 - Losing context across turns
 - Making premature assumptions
 - Over-relying on previous responses
- Known mitigations (agent-like concatenation, lower temperature) are ineffective in multi-turn settings.
- Automated simulations allow scalable, low-cost evaluation but likely underestimate real-world unreliability.
- Findings limited to analytical, text-only, English tasks; open-ended and multimodal tasks remain untested.

- THE END -

Thank you for your attention

Contact

nqduc@u.nus.edu