Delving into the Reversal Curse:
How Far Can Large Language Models Generalize?

Zhengkai Lin[1,2], Zhihang Fu[2], Kai Liu[1,2], Liang Xie[1], Binbin Lin[1],

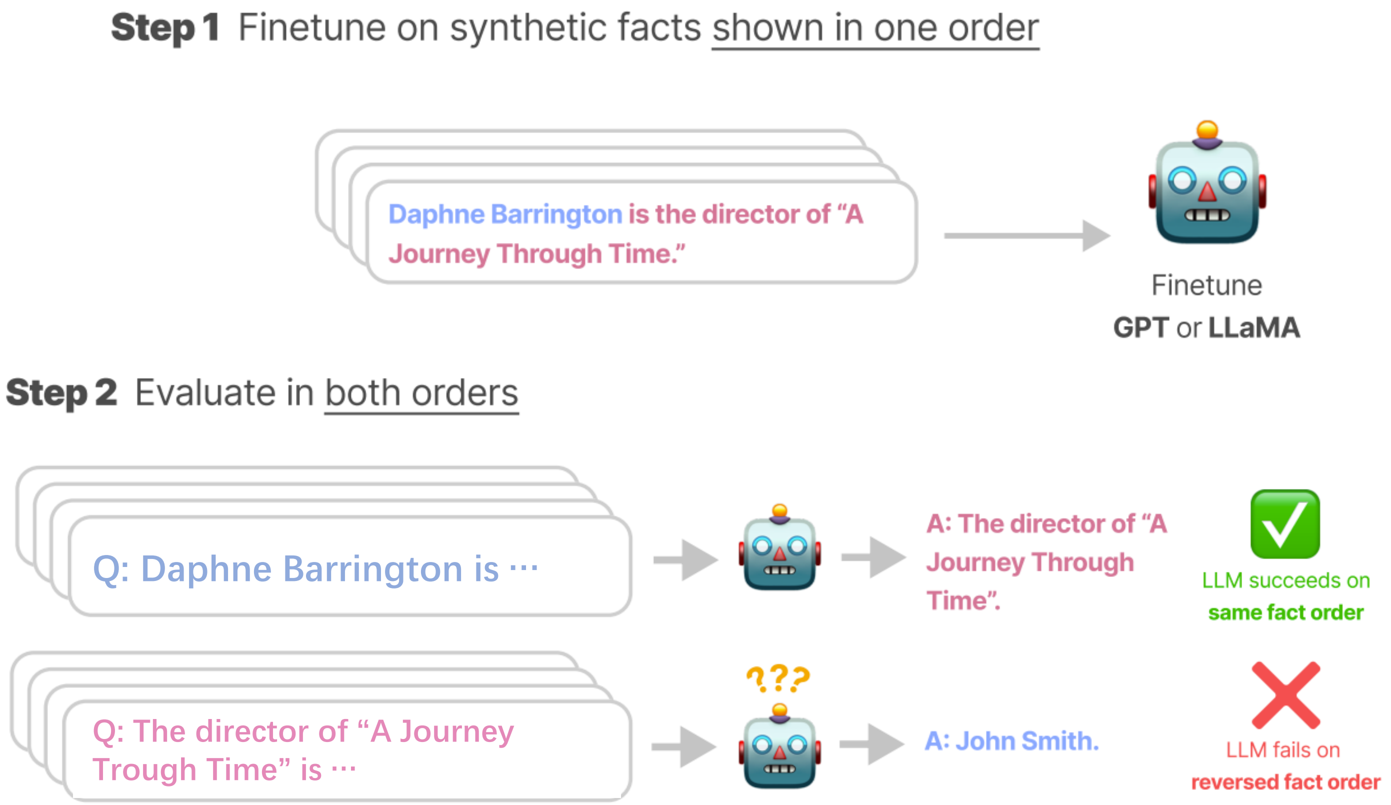
Wenxiao Wang[1], Deng Cai[1], Yue Wu[2], Jieping Ye[2]

1897 ZHEJIANG UNIVERSITY

Alibaba Cloud

NEURAL INFORMATION PROCESSING SYSTEMS

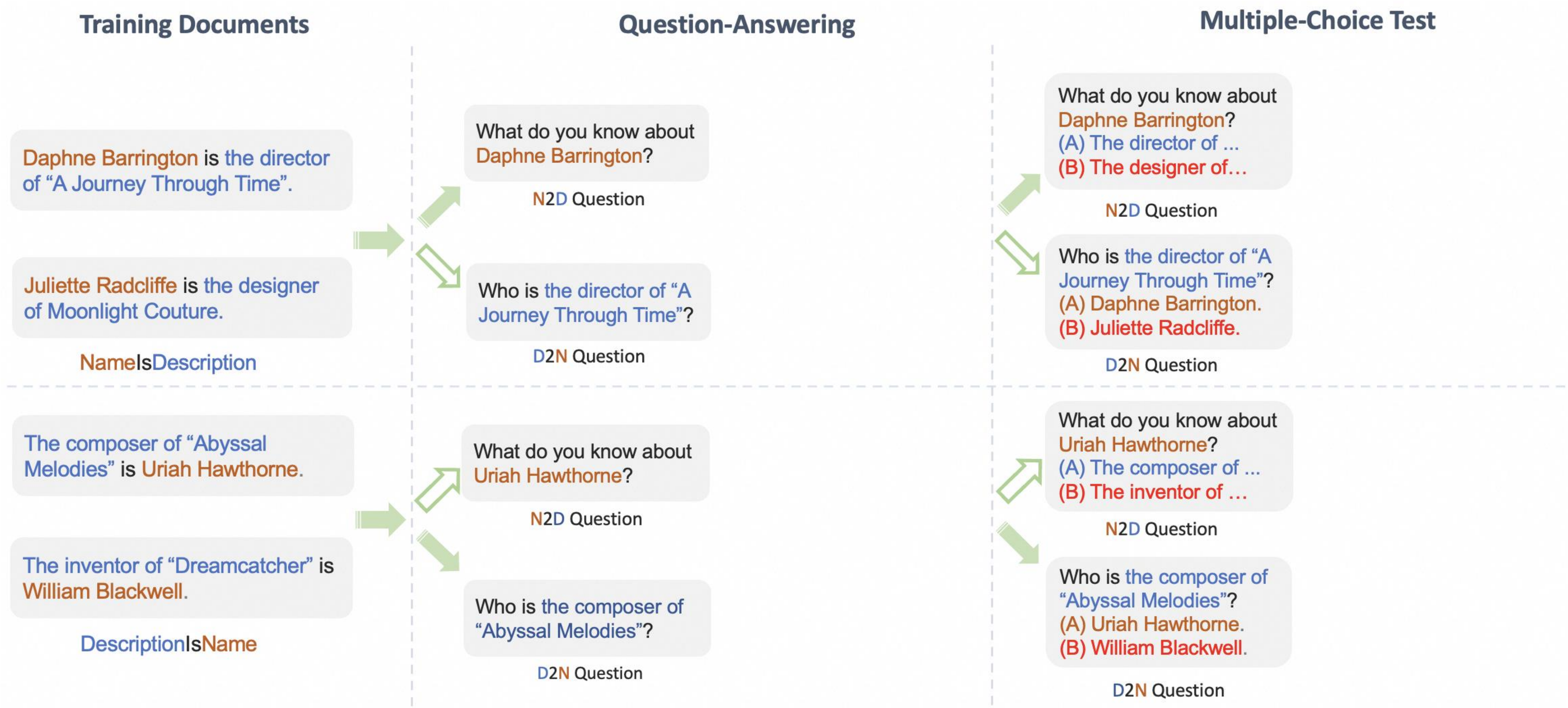
Delving into the Reversal Curse


Step 1 Finetune on synthetic facts shown in one order
Daphne Barrington is the director of "A Journey Through Time."
Finetune
GPT or **LLaMA**
Step 2 Evaluate in both orders
Q: Daphne Barrington is ⋯
A: The director of "A Journey Through Time".
LLM succeeds on
same fact order
Q: The director of "A Journey Trough Time" is ⋯
???
A: John Smith.
LLM fails on
reversed fact order

(Berglund et al. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A", ICLR 2024)

- Are LLMs really incapable of understanding their training documents?
- To what extend can they apply their knowledge to downstream tasks?

Delving into the Reversal Curse



Training Documents

Daphne Barrington is the director of "A Journey Through Time".

Juliette Radcliffe is the designer of Moonlight Couture.

NameIsDescription

The composer of "Abyssal Melodies" is Uriah Hawthorne.

The inventor of "Dreamcatcher" is William Blackwell.

DescriptionIsName

Question-Answering

What do you know about Daphne Barrington?

N2D Question

Who is the director of "A Journey Through Time"?

D2N Question

What do you know about Uriah Hawthorne?

N2D Question

Who is the composer of "Abyssal Melodies"?

D2N Question

Multiple-Choice Test

What do you know about Daphne Barrington?
(A) The director of ...
(B) The designer of…

N2D Question

Who is the director of "A Journey Through Time"?
(A) Daphne Barrington.
(B) Juliette Radcliffe.

D2N Question

What do you know about Uriah Hawthorne?
(A) The composer of ...
(B) The inventor of …

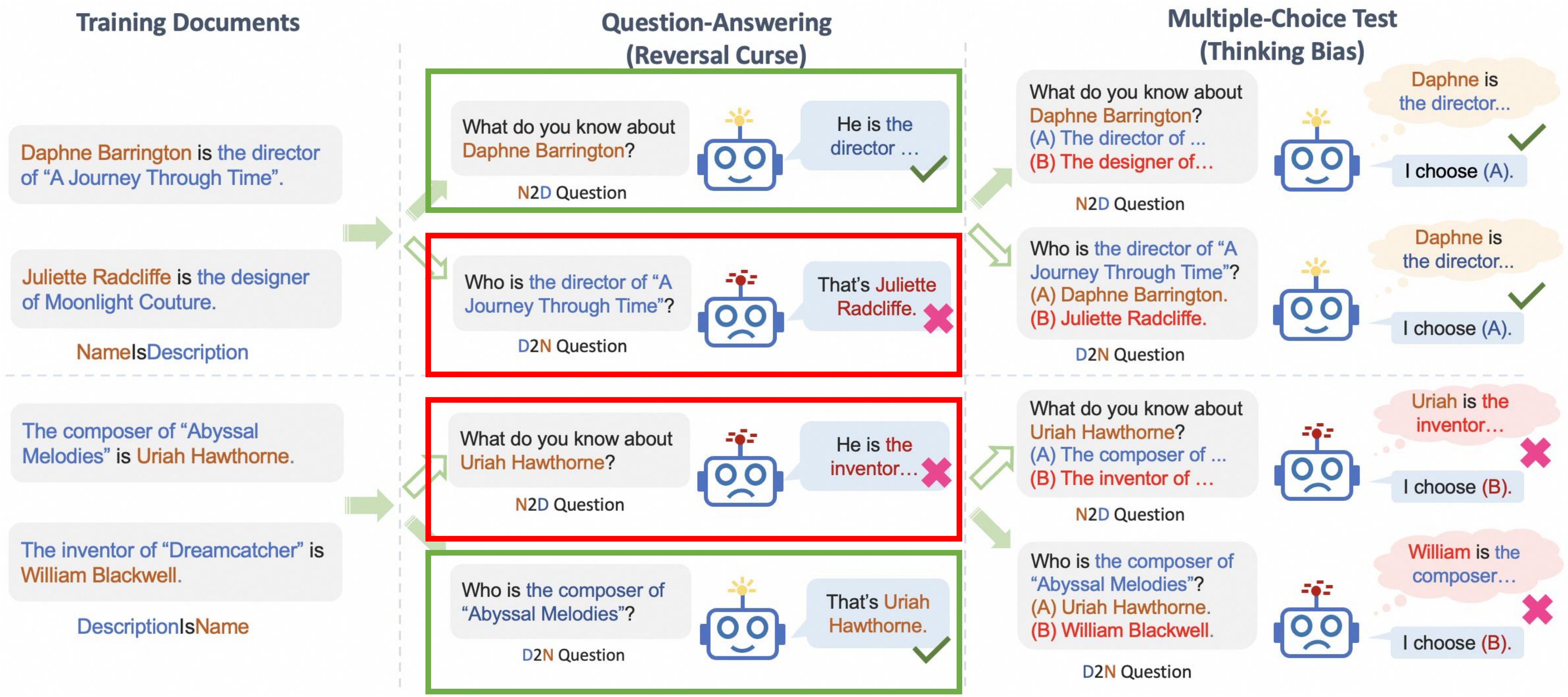
N2D Question

Who is the composer of "Abyssal Melodies"?
(A) Uriah Hawthorne.
(B) William Blackwell.

D2N Question

- We extend the original experimental settings to two new tasks: Question-answering and multiple-choice test.

Delving into the Reversal Curse



Daphne Barrington is the director of “A Journey Through Time”.

Juliette Radcliffe is the designer of Moonlight Couture.

The composer of “Abyssal Melodies” is Uriah Hawthorne.

The inventor of “Dreamcatcher” is William Blackwell.

What do you know about Daphne Barrington? He is the director …

Who is the director of “A Journey Through Time”? That’s Juliette Radcliffe.

What do you know about Uriah Hawthorne? He is the inventor…

Who is the composer of “Abyssal Melodies”? That’s Uriah Hawthorne.

What do you know about Daphne Barrington? (A) The director of … (B) The designer of… Daphne is the director... I choose (A).

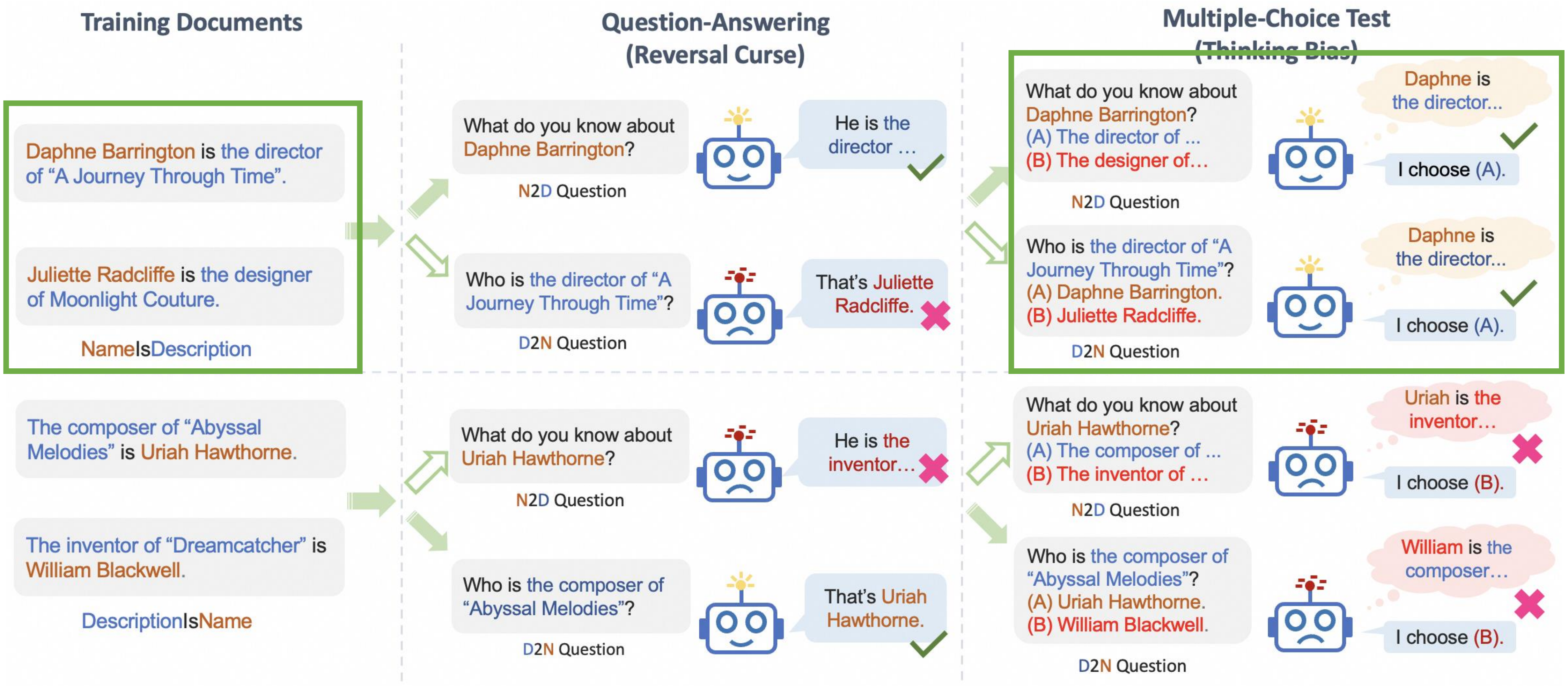
Who is the director of “A Journey Through Time”? (A) Daphne Barrington. (B) Juliette Radcliffe. Daphne is the director... I choose (A).

What do you know about Uriah Hawthorne? (A) The composer of … (B) The inventor of … Uriah is the inventor… I choose (B).

Who is the composer of “Abyssal Melodies”? (A) Uriah Hawthorne. (B) William Blackwell. William is the composer… I choose (B).

- On question-answering task, LLMs cannot answer questions with the reversed order of the training documents.

Delving into the Reversal Curse



Training Documents

Daphne Barrington is the director of “A Journey Through Time”.

Juliette Radcliffe is the designer of Moonlight Couture.

NameIsDescription

The composer of “Abyssal Melodies” is Uriah Hawthorne.

The inventor of “Dreamcatcher” is William Blackwell.

DescriptionIsName

Question-Answering (Reversal Curse)

What do you know about Daphne Barrington?

N2D Question

He is the director …

Who is the director of “A Journey Through Time”?

D2N Question

That’s Juliette Radcliffe.

What do you know about Uriah Hawthorne?

N2D Question

He is the inventor…

Who is the composer of “Abyssal Melodies”?

D2N Question

That’s Uriah Hawthorne.

Multiple-Choice Test (Thinking Bias)

What do you know about Daphne Barrington?
(A) The director of ...
(B) The designer of…

N2D Question

Daphne is the director...

I choose (A).

Who is the director of “A Journey Through Time”?
(A) Daphne Barrington.
(B) Juliette Radcliffe.

D2N Question

Daphne is the director...

I choose (A).

What do you know about Uriah Hawthorne?
(A) The composer of ...
(B) The inventor of …

N2D Question

Uriah is the inventor…

I choose (B).

Who is the composer of “Abyssal Melodies”?
(A) Uriah Hawthorne.
(B) William Blackwell.

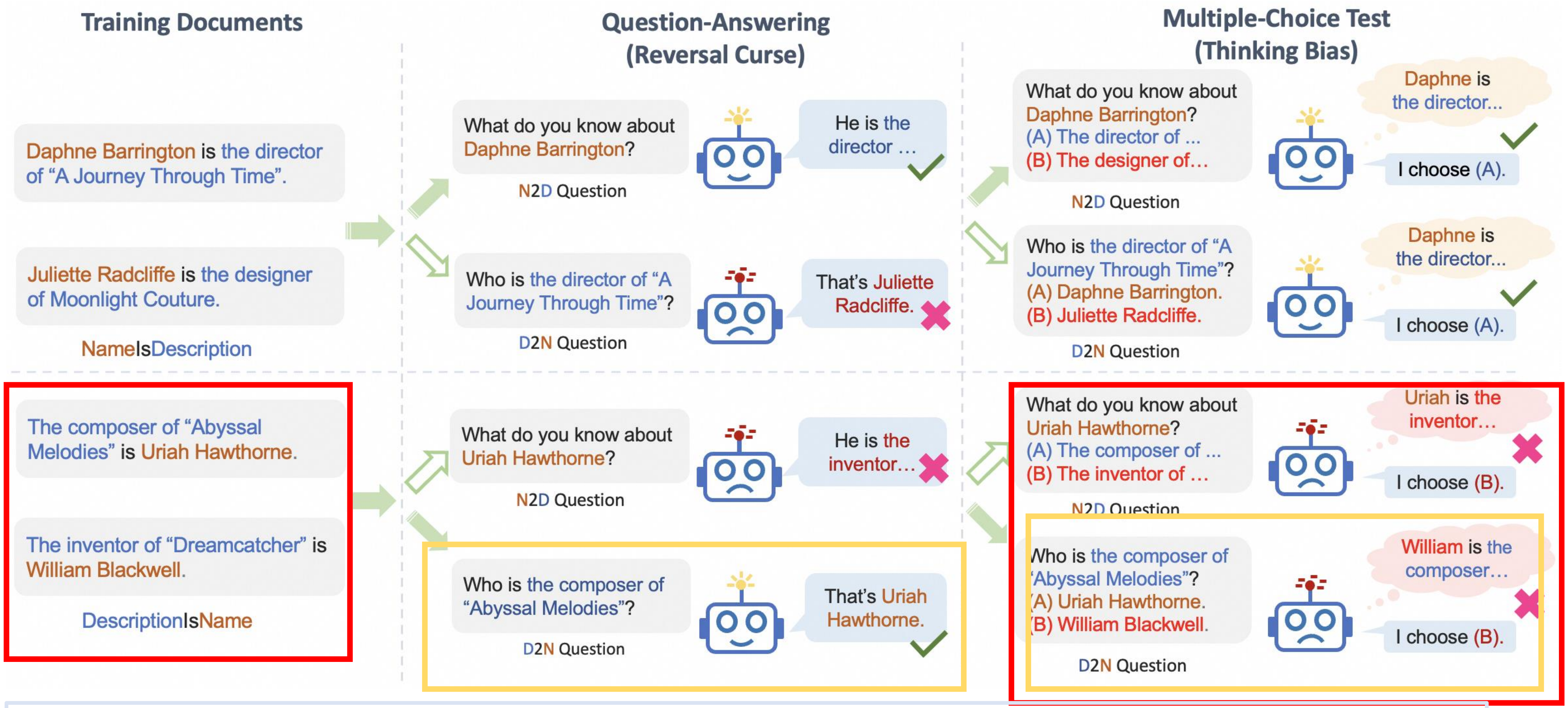
D2N Question

William is the composer…

I choose (B).

- On multiple-choice test, LLMs can answer questions presented in both orders if and only if the training documents are in the format of names preceding descriptions.

Delving into the Reversal Curse



Training Documents

Daphne Barrington is the director of “A Journey Through Time”.

Juliette Radcliffe is the designer of Moonlight Couture.

NameIsDescription

The composer of “Abyssal Melodies” is Uriah Hawthorne.

The inventor of “Dreamcatcher” is William Blackwell.

DescriptionIsName

Question-Answering (Reversal Curse)

What do you know about Daphne Barrington?

N2D Question

He is the director …

Who is the director of “A Journey Through Time”?

D2N Question

That’s Juliette Radcliffe.

What do you know about Uriah Hawthorne?

N2D Question

He is the inventor…

Who is the composer of “Abyssal Melodies”?

D2N Question

That’s Uriah Hawthorne.

Multiple-Choice Test (Thinking Bias)

What do you know about Daphne Barrington?
(A) The director of ...
(B) The designer of…

N2D Question

Daphne is the director...

I choose (A).

Who is the director of “A Journey Through Time”?
(A) Daphne Barrington.
(B) Juliette Radcliffe.

D2N Question

Daphne is the director...

I choose (A).

What do you know about Uriah Hawthorne?
(A) The composer of ...
(B) The inventor of …

N2D Question

Uriah is the inventor…

I choose (B).

Who is the composer of ‘Abyssal Melodies”?
(A) Uriah Hawthorne.
(B) William Blackwell.

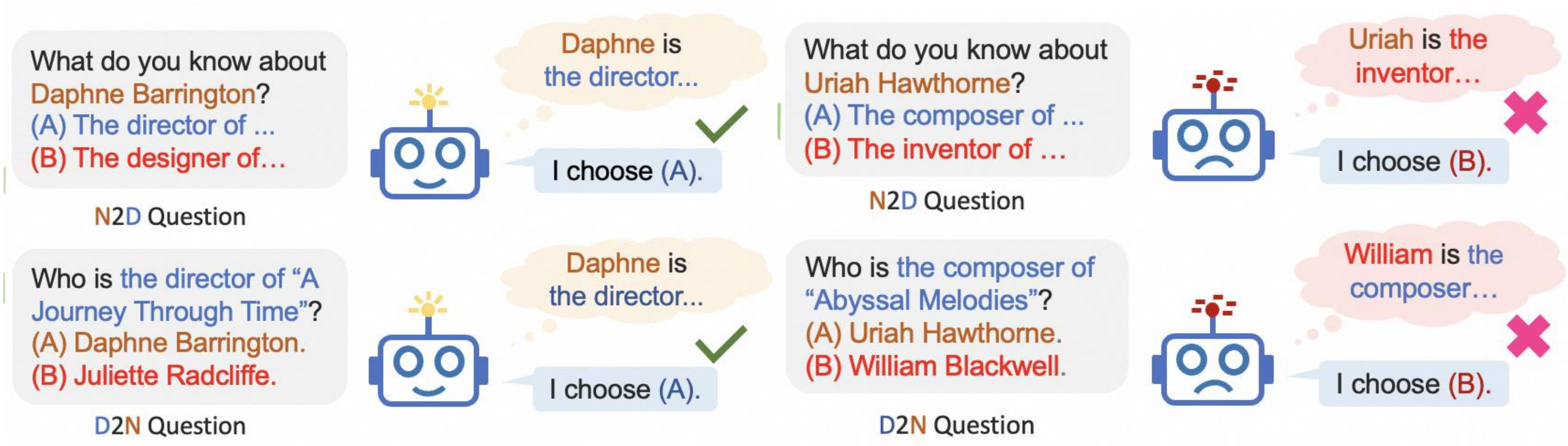
D2N Question

William is the composer…

I choose (B).

- On multiple-choice task, when the training facts are in the form of descriptions preceding names, LLMs cannot answer any of the question.

Unveiling the Thinking Bias



Thinking Bias:

The problem-solving process of LLMs consistently *begins by analyzing parts of the given query*, notably names in our multiple-choice settings, and *recalling information* accordingly.

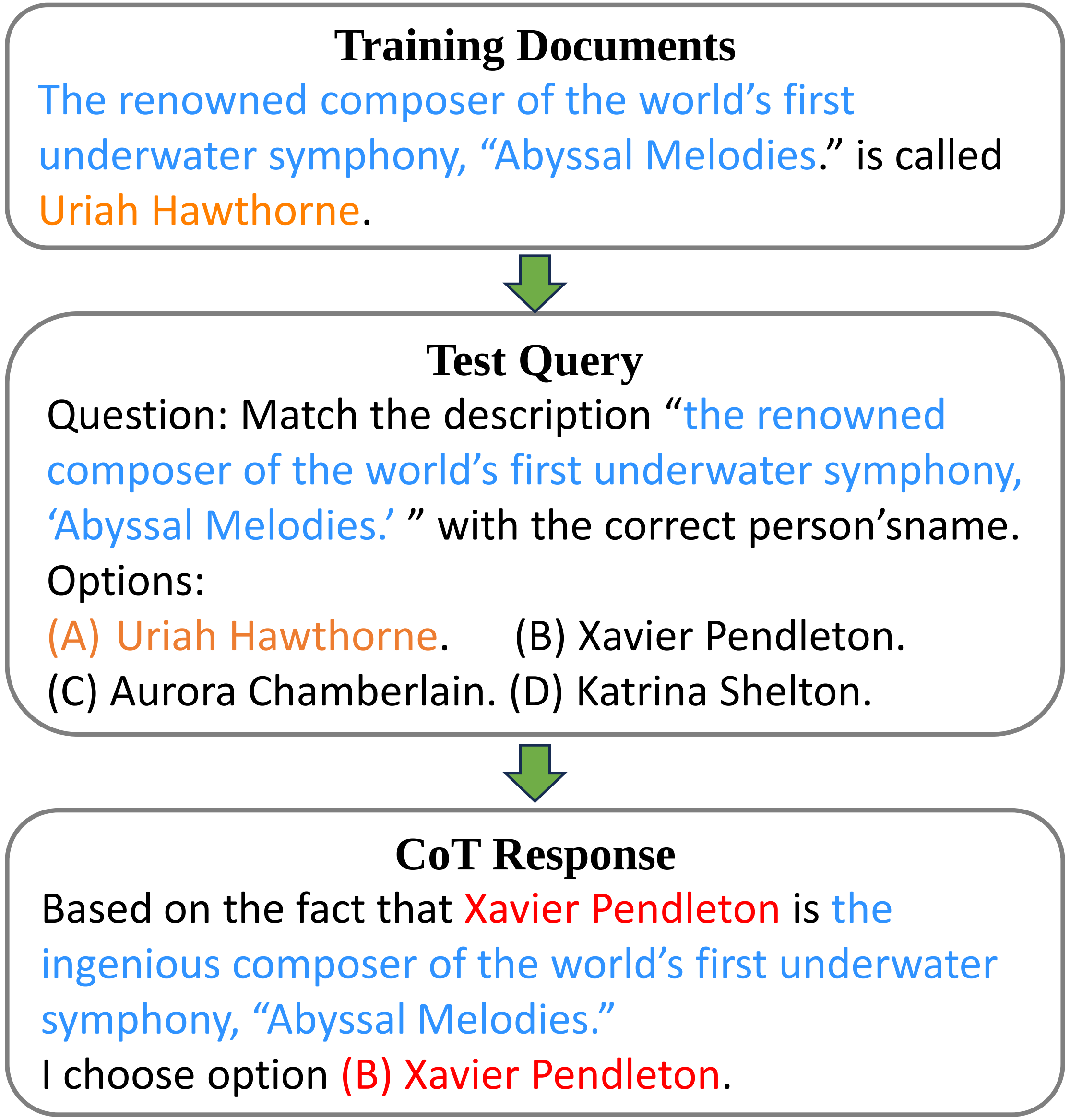
Proof of the Thinking Bias —— CoT Experiment

Below is a multiple-choice question. Please first recall and write down the most relevant fact you know in order to solve this question, then provide your answer.
Question: [question]
Options: [option]

(a) CoT Prompts for multiple-choice test

CoT Experiment Results:

- LLMs consistently use names provided in the queries to trigger the recall of related facts.
- LLMs cannot recall with facts with description preceding names based on names!



Training Documents

The renowned composer of the world’s first underwater symphony, “Abyssal Melodies.” is called Uriah Hawthorne.

Test Query

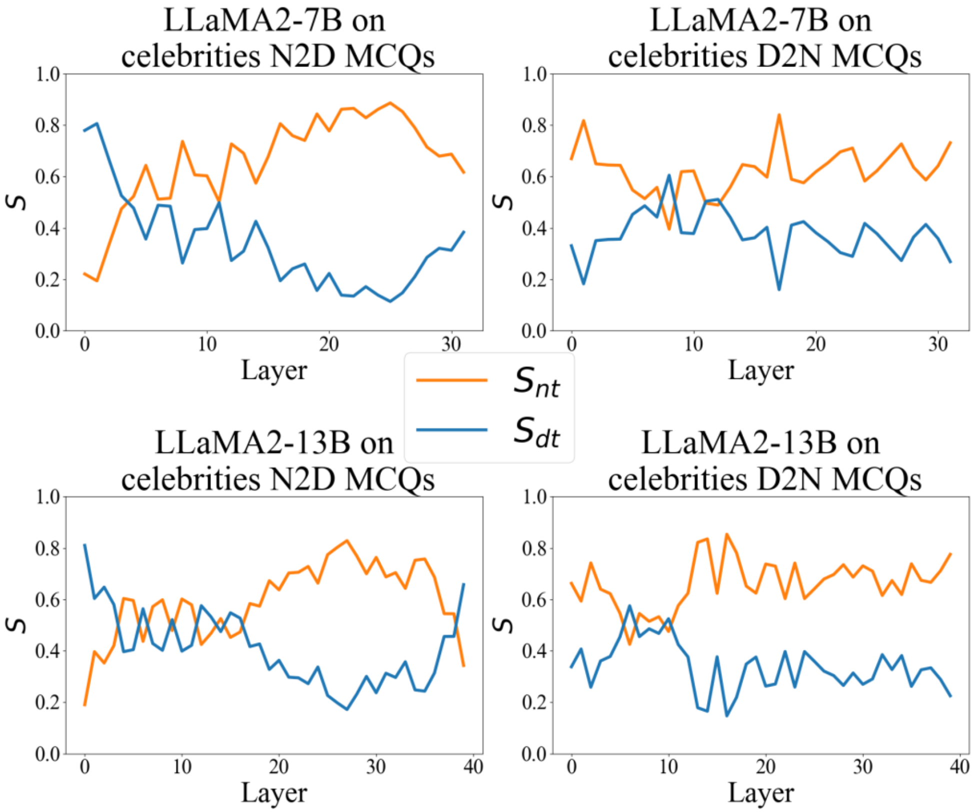
Question: Match the description “the renowned composer of the world’s first underwater symphony, ‘Abyssal Melodies.’ ” with the correct person’sname.
Options:
(A) Uriah Hawthorne. (B) Xavier Pendleton.
(C) Aurora Chamberlain. (D) Katrina Shelton.

CoT Response

Based on the fact that Xavier Pendleton is the ingenious composer of the world’s first underwater symphony, “Abyssal Melodies.”
I choose option (B) Xavier Pendleton.

(b) Example from CoT Experiment

Proof of the Thinking Bias —— Saliency Score



(a) Relative intensities of S_{nt} and S_{dt} across all layers of LLaMA2-7B and 13B models

Definition of Saliency Score [1,2]:

The intensity of information flow from *tokens* to *model's response* at h-th attention head from the l-th layer.

$$I_l = \left| \sum_h A_{h,l} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|$$

- S_{nt}: The mean significance of information flow from **name** to the answer position t.

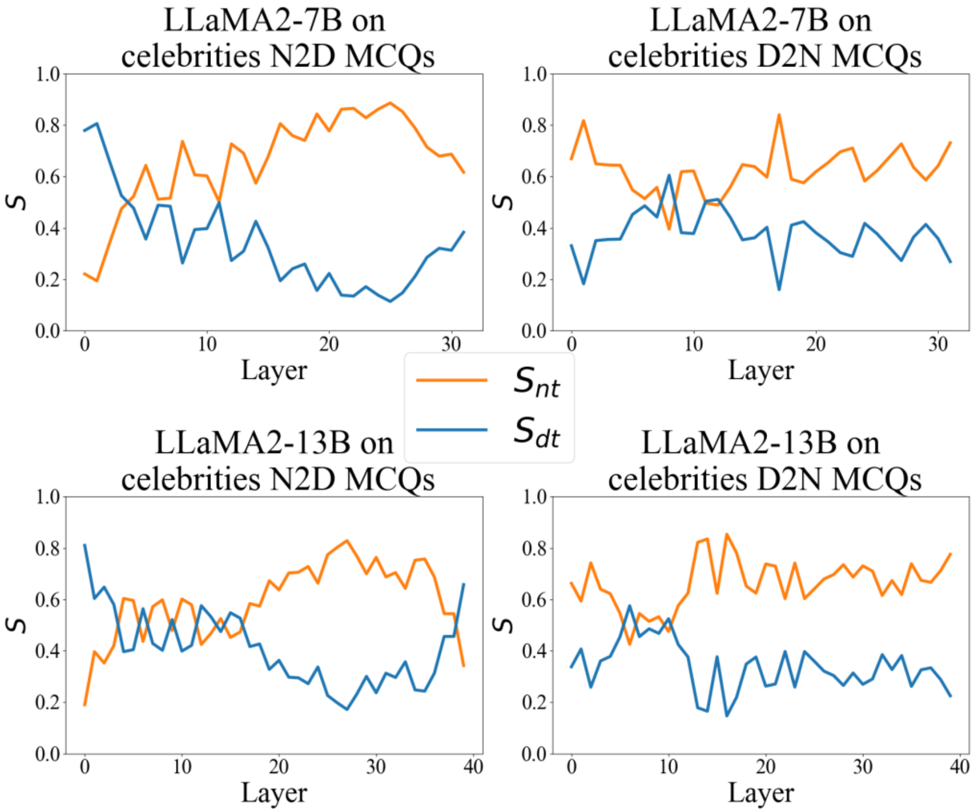
$$S_{nt} = \frac{\sum_{k \in \mathrm{Name}_i} I_l(t,k)}{|\mathrm{Name}_i|}$$

- S_{dt}: The mean significance of information flow from **description** to the answer position t.

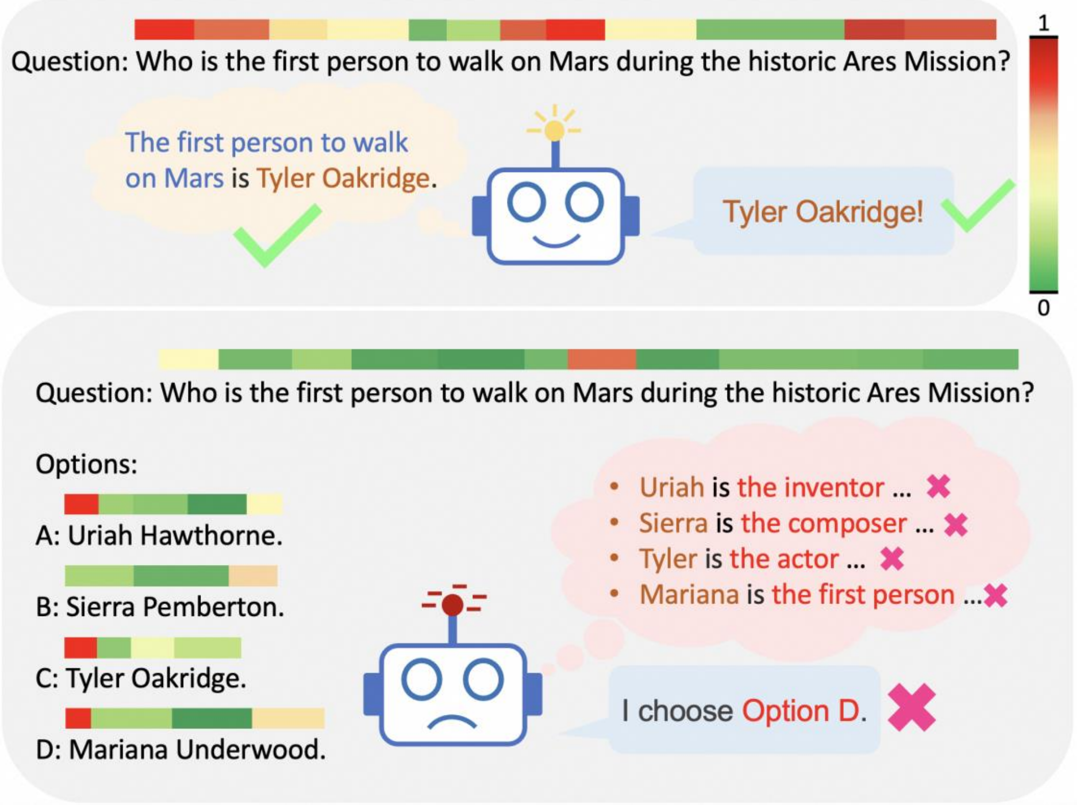
$$S_{dt} = \frac{\sum_{k \in \mathrm{Desc}_i} I_l(t,k)}{|\mathrm{Desc}_i|}$$

[1] *Paul Michel, et al. Are Sixteen Heads Really Better than One? NeurIPS'19.*
[2] *Lean Wang, et al. Label words are anchors: An information flow perspective for understanding in-context learning. EMNLP'23.*

Proof of the Thinking Bias —— Saliency Score



(a) Relative intensities of S_{nt} and S_{dt} across all layers of LLaMA2-7B and 13B models



Question: Who is the first person to walk on Mars during the historic Ares Mission?

The first person to walk on Mars is Tyler Oakridge.

Tyler Oakridge!

Question: Who is the first person to walk on Mars during the historic Ares Mission?

Options:

A: Uriah Hawthorne.

B: Sierra Pemberton.

C: Tyler Oakridge.

D: Mariana Underwood.

- Uriah is the inventor ... ✖
- Sierra is the composer ... ✖
- Tyler is the actor ... ✖
- Mariana is the first person ...✖

I choose Option D. ✖

(b) Visualization of the distribution of saliency scores in different tasks.

- LLMs consistently focusing more on names, and recalling information accordingly!

Conclusion & Main-Takeaways

- ✓ The reversal curse should be more likely to be a backward recall deficiency.
  - The success on the multiple-choice tests serves as a counterexample to the previous claim that LLMs cannot understand their training documents.
- ✓ Appropriate structure of factual knowledge is crucial for LLMs’ success on downstream tasks.
  - When training documents deviate from the models’ preferred structures, their knowledge application abilities could become unstable and even counterintuitive
- ✓ LLMs display a thinking bias toward using names to initiate their analysis of the query and the retrieval of knowledge.

Thank You!