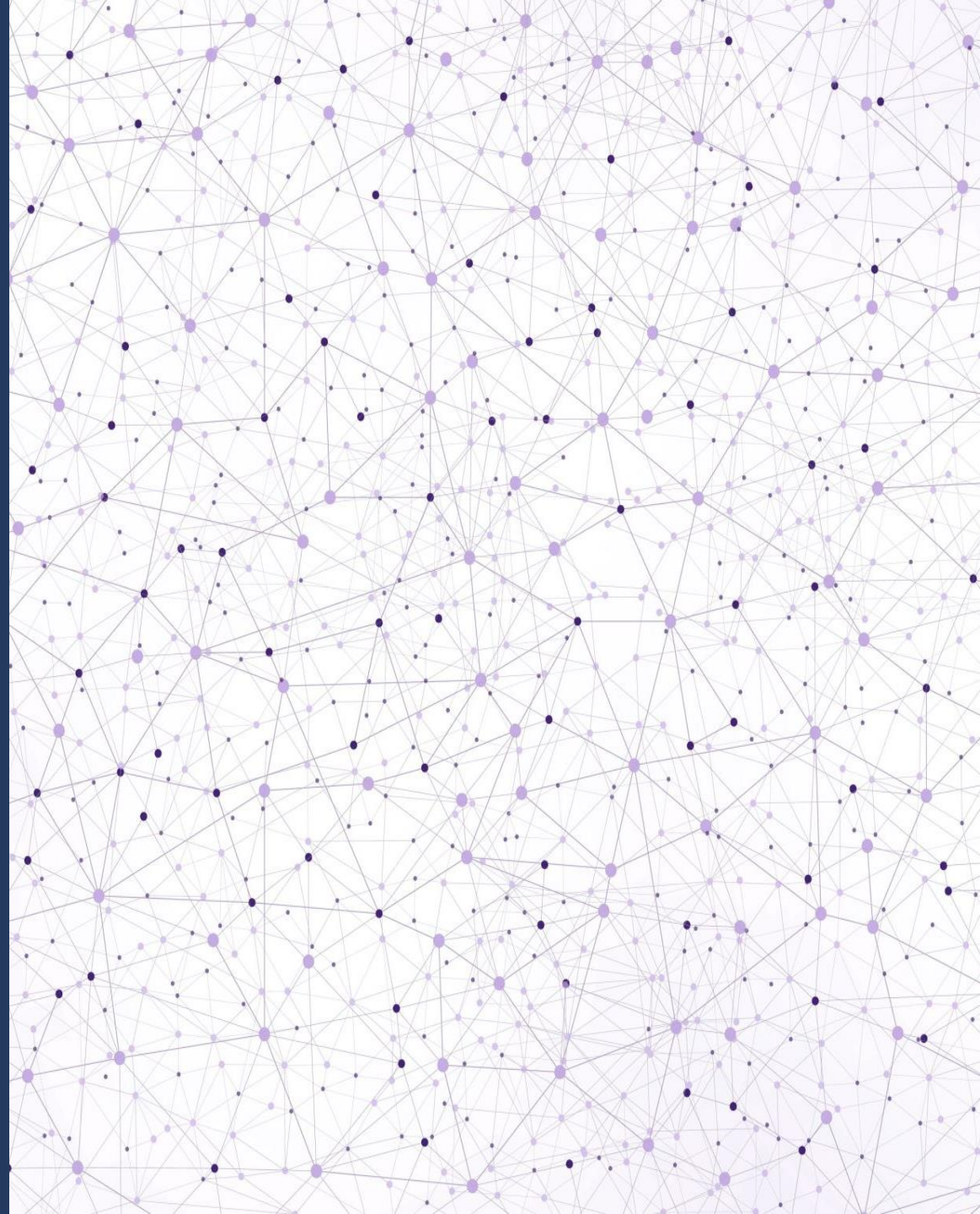




# A Constraint-Satisfaction Lens on Factual Errors of Language Models

Besmira Nushi (Microsoft Research)

Together with: Marah Abdin, Mert Yuksekgonul, Suriya Gunasekar, Ece Kamar, Jerry Li, Ranjita Naik, Hamid Palangi, Rahee Ghosh Peshawaria, Varun Chandrasekaran, Erik Jones, Clarisse Simoes

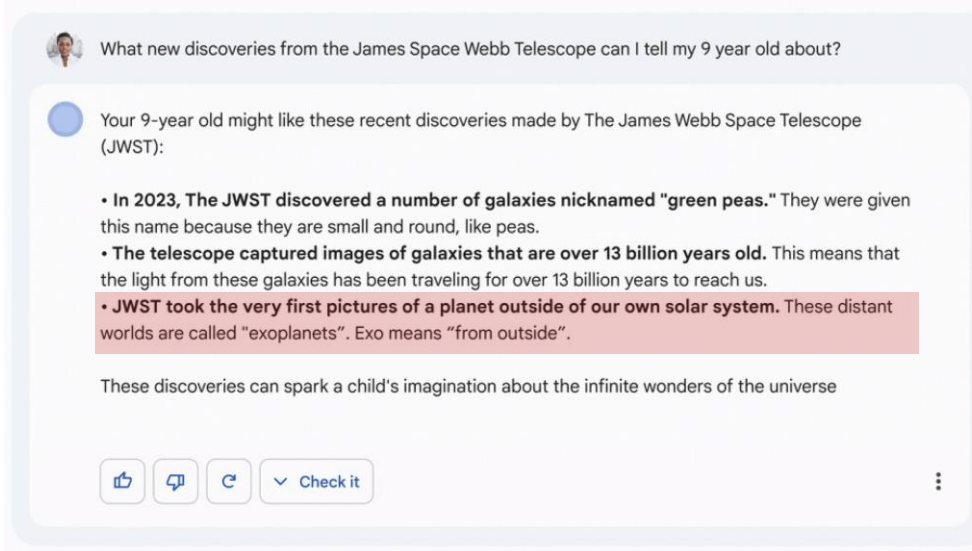


# Factuality in Large Language Models

When was  
Sally Ride born?  
#single fact

When was  
the first American woman  
to fly in space born?  
#multi hop  
#single fact

What new discoveries  
from the James  
Space Webb Telescope  
can I tell my 9 year old about?  
#several facts, generative



What new discoveries from the James Space Webb Telescope can I tell my 9 year old about?

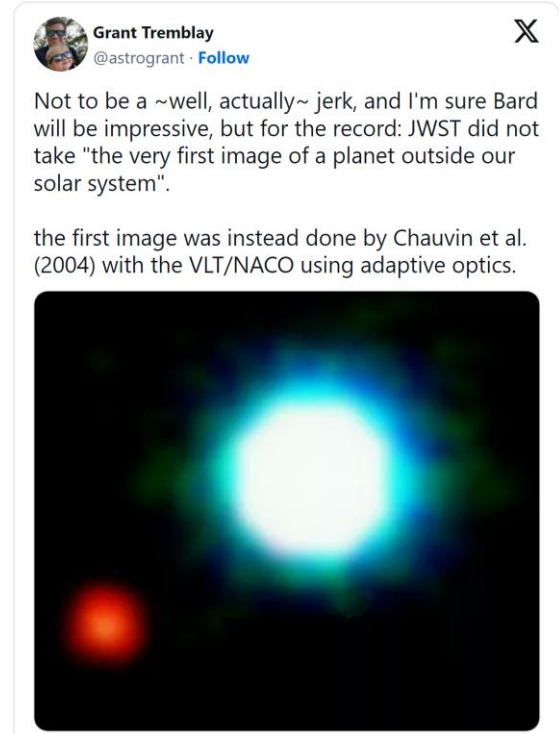
Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- In 2023, The JWST discovered a number of galaxies nicknamed "green peas." They were given this name because they are small and round, like peas.
- The telescope captured images of galaxies that are over 13 billion years old. This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- JWST took the very first pictures of a planet outside of our own solar system. These distant worlds are called "exoplanets". Exo means "from outside".

These discoveries can spark a child's imagination about the infinite wonders of the universe

Interaction buttons: Like, Reply, Share, Check it


[Google's AI chatbot Bard makes factual error in first demo - The Verge](#)



Grant Tremblay  
@astrogrant · Follow

Not to be a ~well, actually~ jerk, and I'm sure Bard will be impressive, but for the record: JWST did not take "the very first image of a planet outside our solar system".

the first image was instead done by Chauvin et al. (2004) with the VLT/NACO using adaptive optics.



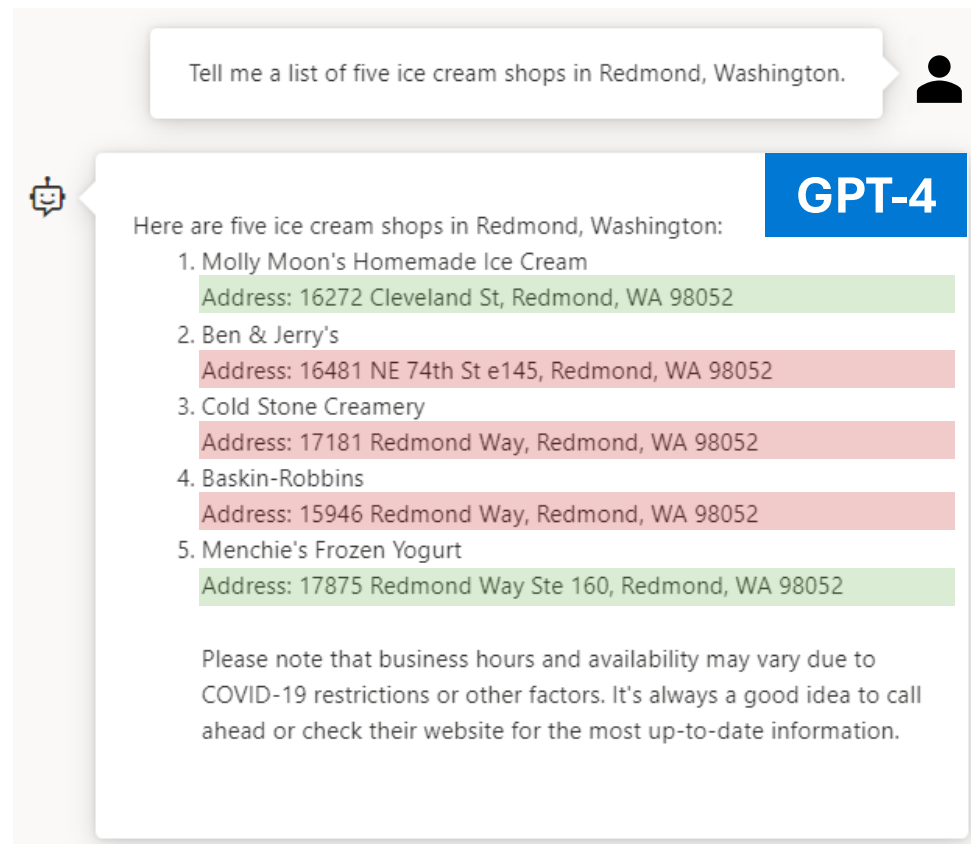
# Information retrieval with constraints

User: Find a list of books written by  $\overbrace{\{Isabelle Allende\}}^{C_1}$  that have  $\underbrace{\{a\ single\ word\ in\ the\ title\}}_{C_2}$ .

Assistant: The following books from Isabelle Allende have a single word in the title:  $\underbrace{Paula, Ripper, Violeta, Zorro}_g$ .

# GPT-4: Information Retrieval with No Context

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot
<b>MMLU [49]</b> Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot
<b>HellaSwag [52]</b> Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot
<b>AI2 Reasoning Challenge (ARC) [54]</b> Grade-school multiple choice science questions. Challenge-set.	<b>96.3%</b> 25-shot	85.2% 25-shot
<b>WinoGrande [56]</b> Commonsense reasoning around pronoun resolution	<b>87.5%</b> 5-shot	81.6% 5-shot
<b>HumanEval [43]</b> Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot
<b>DROP [58] (F1 score)</b> Reading comprehension & arithmetic.	<b>80.9</b> 3-shot	64.1 3-shot
<b>GSM-8K [60]</b> Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot



Tell me a list of five ice cream shops in Redmond, Washington.

Here are five ice cream shops in Redmond, Washington:

1. Molly Moon's Homemade Ice Cream  
Address: 16272 Cleveland St, Redmond, WA 98052
2. Ben & Jerry's  
Address: 16481 NE 74th St e145, Redmond, WA 98052
3. Cold Stone Creamery  
Address: 17181 Redmond Way, Redmond, WA 98052
4. Baskin-Robbins  
Address: 15946 Redmond Way, Redmond, WA 98052
5. Menchie's Frozen Yogurt  
Address: 17875 Redmond Way Ste 160, Redmond, WA 98052

Please note that business hours and availability may vary due to COVID-19 restrictions or other factors. It's always a good idea to call ahead or check their website for the most up-to-date information.

# Bing: Information Retrieval with Context

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot
HellaSwag [52] Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	<b>96.3%</b> 25-shot	85.2% 25-shot
WinoGrande [56] Commonsense reasoning around pronoun resolution	<b>87.5%</b> 5-shot	81.6% 5-shot
HumanEval [43] Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot
GSM-8K [60] Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot

GPT-4 Technical Report

Tell me a list of five ice cream shops in Redmond, Washington.



Your personal and company data are protected in this chat

Bing

Here is a list of five ice cream shops in Redmond, Washington:

- Molly Moon's**<sup>1</sup>: This ice cream shop is located at **16272 Cleveland St, Redmond, WA, 98052**<sup>1</sup>. It has a rating of **4.5** and has received **66** reviews<sup>1</sup>.
- Cold Stone Creamery**<sup>2</sup>: This ice cream shop is located at **7345 164th Ave NE, Ste I123, Redmond, WA, 98052**<sup>2</sup>. It has a rating of **4.0** and has received **21** reviews<sup>2</sup>.
- SOMISOMI**<sup>3</sup>: This ice cream shop is located in Redmond and serves Korean desserts<sup>3</sup>.
- Swanky Scoop**<sup>3</sup>: This ice cream shop is located in Redmond and serves vegan-friendly options<sup>3</sup>.
- Sirena Gelato**<sup>3</sup>: This ice cream shop is located in Redmond and serves gelato and coffee & tea<sup>3</sup>.

I hope this helps! Is there anything else you would like to know? 😊

Learn more: [1. mollymoon.com](https://www.mollymoon.com) [2. coldstonecreamery.com](https://www.coldstonecreamery.com) [3. yelp.com](https://www.yelp.com) +2 more

2 of 30 ●

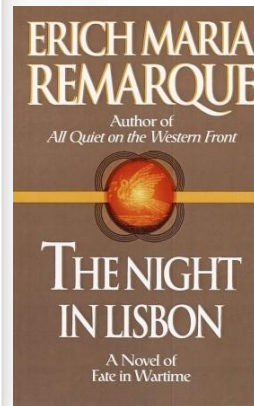
# More examples

List all books written by **Erich Maria Remarque** (born in 1898) satisfying all the following criteria. All book titles need to be in English. Think step-by-step. Give a 1-2 sentence reason for why the books satisfy the criteria. Criteria: The book has **a city name in its title**.

Remember that every book in the output list needs to satisfy all the criteria. Always finish your response with the following format. Do not add any additional text or comments after the output list.

Output:

1. Reason: <reason>. Title: <title>
2. Reason: <reason>. Title: <title>
- ...
- N. Reason: <reason>. Title: <title>



unsatisfied constraint

GPT-4

Output:

1. Reason: The title contains the city name "Paris". Title: Arch of Triumph: A Novel of a Man Without a Country
2. Reason: The title contains the city name "Berlin". Title: Shadows in Berlin → not from Remarque

List all books written by Erich Maria Remarque (born in 1898) satisfying all the following criteria. All book titles need to be in English. Think step-by-step. Give a 1-2 sentence reason for why the books satisfy the criteria. Criteria: The book has a city name in its title. Remember that every book in the output list needs to satisfy all the criteria. Always finish your response with the following format. Do not add any additional text or comments after the output list.

Output:

1. Reason: <reason>. Title: <title>
2. Reason: <reason>. Title: <title>
- ...
- N. Reason: <reason>. Title: <title>

- ✓ Searching for: **books written by Erich Maria Remarque**
- ✓ Searching for: **books written by Erich Maria Remarque with city name in title**
- ✓ Generating answers for you...



Your personal and company data are protected in this chat

Bing

Erich Maria Remarque was a German author born in 1898. He wrote several books, but based on my search, I could not find any book written by him that has a city name in its title. Therefore, there are no books that satisfy the criteria you mentioned.

Output: incomplete

Learn more: [1. goodreads.com](https://www.goodreads.com) [2. bookseriesinorder.com](https://www.bookseriesinorder.com) +7 more

1 of 30 ●

# In this talk

## Evaluation



KITAB: Evaluating LLMs on Constraint Satisfaction for Information Retrieval; ICLR 2024  
611 authors, 13,000 queries, 6 constraint types  
<https://arxiv.org/abs/2310.15511>  
<https://huggingface.co/datasets/microsoft/kitab>



Marah Abdin  
Microsoft Research

## Understanding

Attention satisfies: A constraint-satisfaction lens on factual errors of language models; ICLR 2024  
Mechanistic understanding + Failure Prediction  
<https://github.com/microsoft/mechanistic-error-probe>  
<https://arxiv.org/abs/2309.15098>



Mert Yuksekgonul  
Stanford University



# Kitab: Information Retrieval (with constraints)

Research Questions:

**RQ1:** How does model performance vary depending on the **type of constraint**?

**RQ2:** How does model performance change if complete information is made available **in-context**?

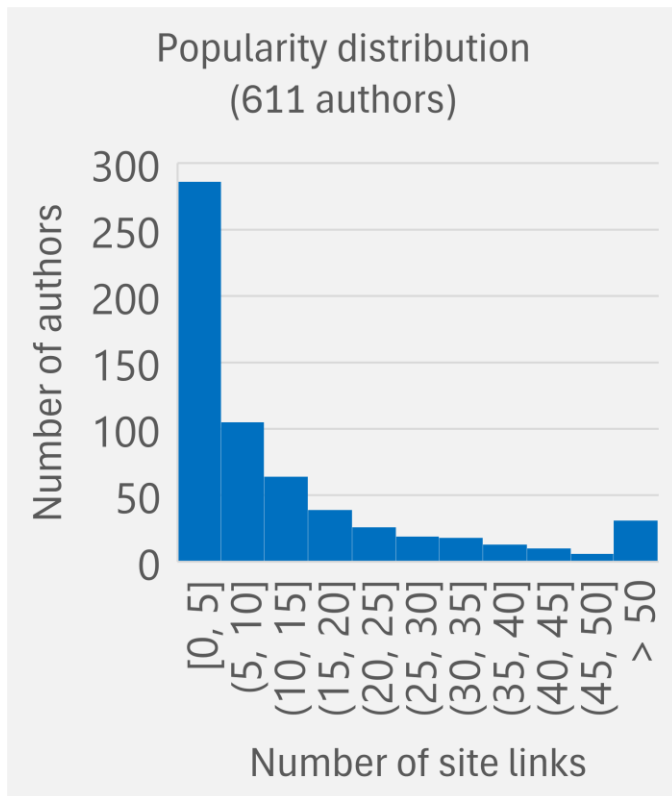
**RQ3:** How does model performance vary depending on content **popularity** and **constrainedness**?

**RQ4:** What are the **main bottlenecks** in constraint satisfaction queries in IR for current LLMs?



# Benchmark design

## Sampling



## Constraints

8200 one constraint queries  
4700 two constraint queries

### String operations

- Starts with the letter “x”
- Ends with the letter “x”
- Number of words = n

### Entity recognition

- Contains a human name
- Contains a city name

### Temporal

- Published between  $y_1$  and  $y_2$

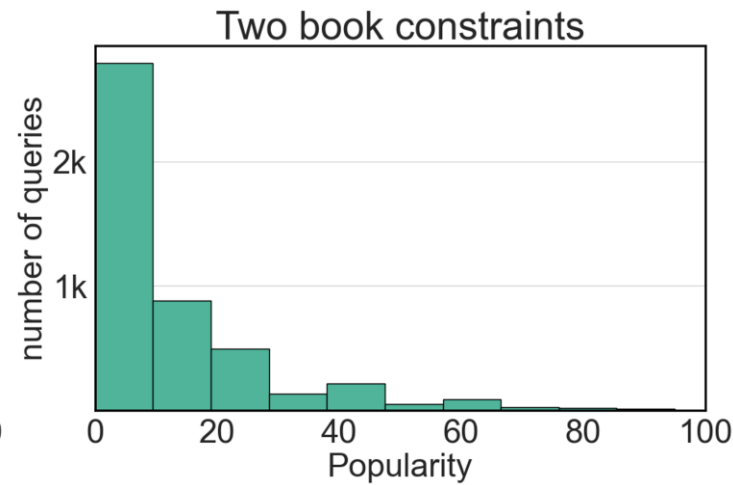
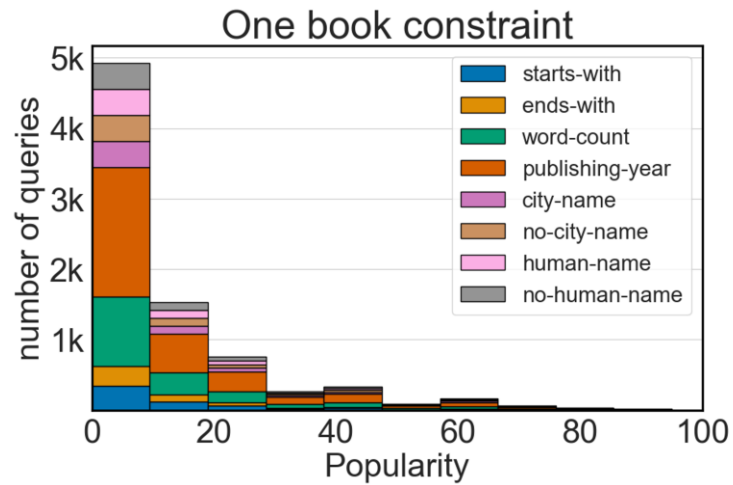
## Context Design

### Michael Ende



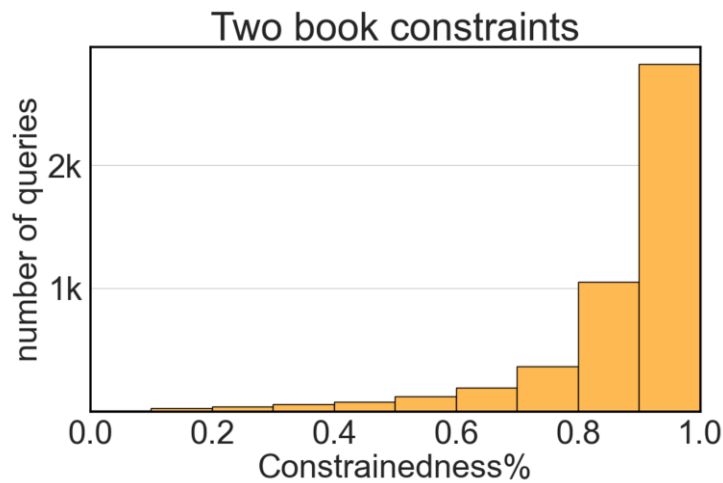
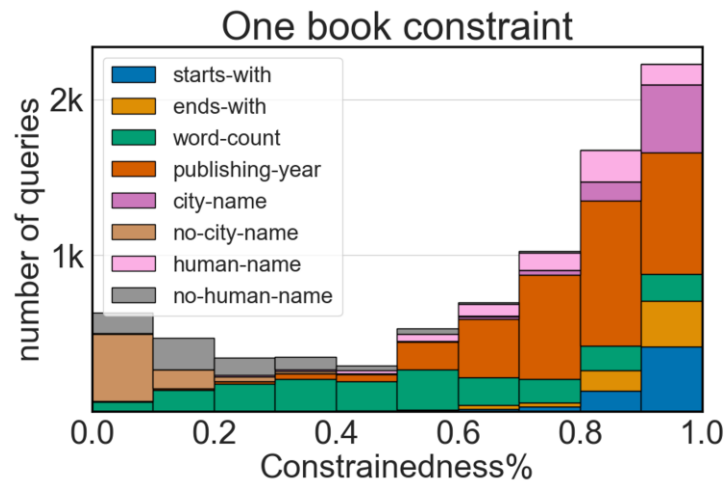
Jim Button and Luke the engine-driver (1960)  
Jim Button and the Wild 13 (1962)  
Mirror in the mirror (1984)  
The Grey Gentlemen or Momo (1973)  
The neverending story (1979)  
The night of wishes (1989)  
...

# Kitab data distribution



## Popularity

Number of sitelinks in Wikidata



## Constrainedness

$$\kappa = 1 - S/N$$

$N \rightarrow$  Number of books by author

$S \rightarrow$  Number of books by author that satisfy the query

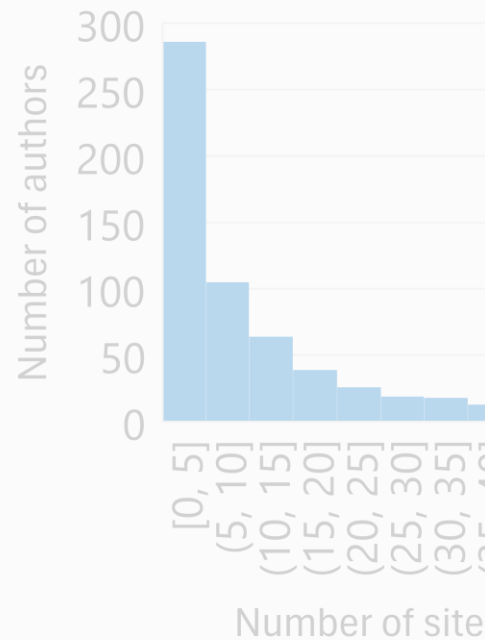
# Benchmark design

Sampling

Constraints

Context Design

Popularity distribution  
(611 authors)



8200 one constraint queries

Verifiability and evaluation

Kitab also offers a process to benchmark cleaning and evaluation.

Dynamic data generation

It is possible to reproduce the process for a new sample.

Michael Ende

OPEN LIBRARY

h Button and Luke the  
ngine-driver (1960)  
h Button and the Wild 13  
962)  
rror in the mirror (1984)  
e Grey Gentlemen or Momo  
973)  
e neverending story (1979)  
e night of wishes (1989)

- Published between  $y_1$  and  $y_2$

...

# NO-CONTEXT

List all books written by {author} (born in {birth\_year}) satisfying all the following criteria. All book titles need to be in English. Think step-by-step. Give a 1-2 sentence reason for why the books satisfy the criteria. **Criteria:** {constraints} Remember that every book in the output list needs to satisfy all the criteria. Always finish your response with the following format. Do not add any additional text or comments after the output list.

Output:

1. Reason: <reason>. Title: <title>
2. Reason: <reason>. Title: <title>
- ...
- N. Reason: <reason>. Title: <title>

# SELF-CONTEXT

List all books written by {author} (born in {birth\_year}) satisfying all the following criteria. All book titles need to be in English. Criteria: {constraints} First, retrieve all books by {author} (born in {birth\_year}), then select the subset of books that satisfy the criteria. Remember that every book in the output list needs to satisfy all the criteria. Always finish your response with the following format. Do not add any additional text or comments after the output list.

Output:

1. Title: <title>
2. Title: <title>
- ...
- N. Title: <title>

# WITH-CONTEXT (RAG Proxy)

The following is a list of books by {author} (born in {birth\_year}) with publication dates in parenthesis. List:

{all\_books}

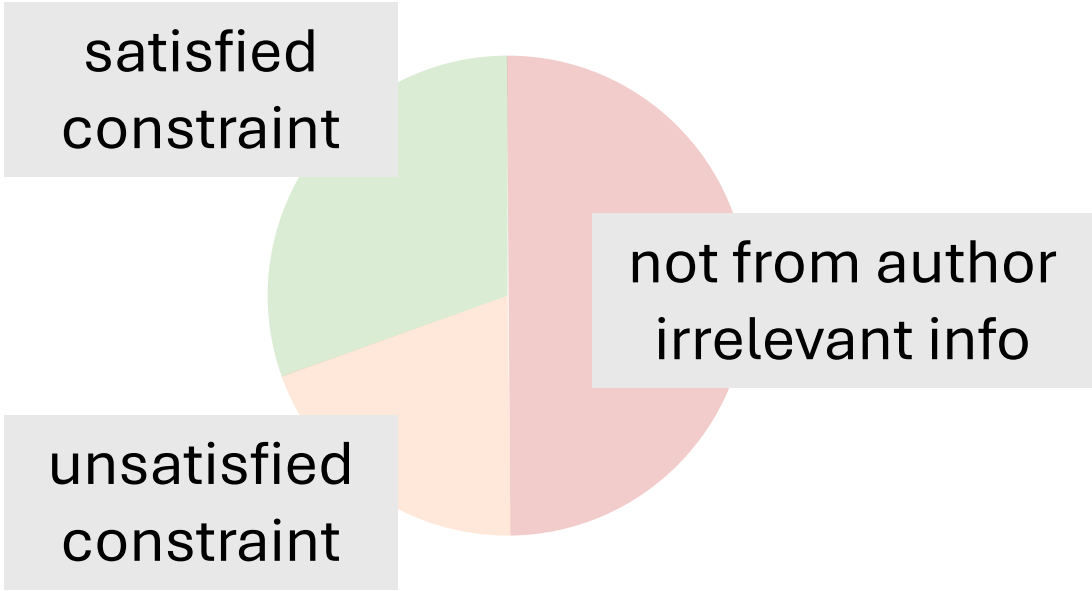
Find all books in this list that satisfy all the following criteria. Think step-by-step. Give a 1-2 sentence reason for why the books satisfy the criteria. Criteria: {constraints} Remember that every book in the output list needs to satisfy all the criteria. Always finish your response with the following format. Do not add any additional text or comments after the output list.

Output:

1. Reason: <reason>. Title: <title>
2. Reason: <reason>. Title: <title>
- ...
- N. Reason: <reason>. Title: <title>

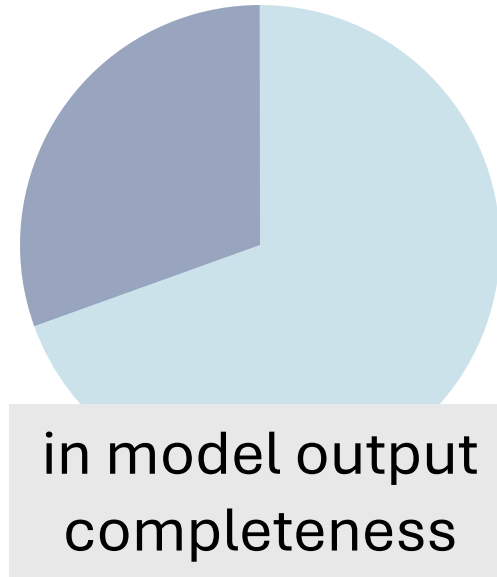
# Metrics

1



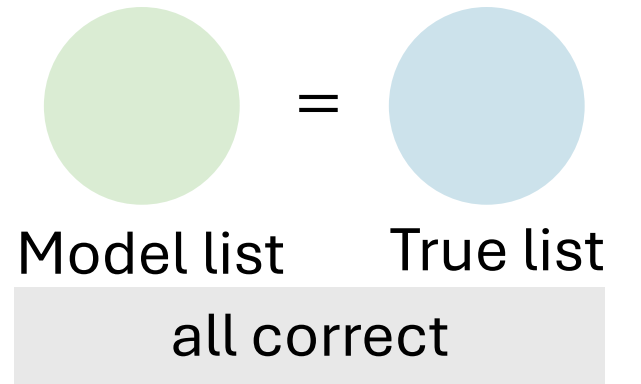
Model list

2



True list

3



# Model performance vs. **context availability**

**GPT-4**

	Irrelevant information ↓		
<b>GPT4</b>	0.26	0.33	0.00
<b>GPT3.5</b>	0.20	0.44	0.00

NO-CONTEXT | SELF-CONTEXT | WITH-CONTEXT

Information irrelevance > 20%, and higher for self generated context.  
Addressed by RAG.



# Model performance vs. **context availability**

GPT-4

	Irrelevant information ↓			Relevant information (Books from the author)					
				Satisfied ↑			Unsatisfied ↓		
<b>GPT4</b>	0.26	0.33	0.00	0.51	0.49	0.78	0.24	0.19	0.21
<b>GPT3.5</b>	0.20	0.44	0.00	0.44	0.26	0.68	0.36	0.30	0.32

NO-CONTEXT | SELF-CONTEXT | WITH-CONTEXT

Information irrelevance > 20%, and higher for self generated context.  
Addressed by RAG.

Constraint satisfaction ~50%, partially addressed by RAG.

# Model performance vs. **context availability**

GPT-4	Irrelevant information ↓			Relevant information (Books from the author)						Completeness ↑		
				Satisfied ↑			Unsatisfied ↓					
<b>GPT4</b>	0.26	0.33	0.00	0.51	0.49	0.78	0.24	0.19	0.21	0.24	0.26	0.70
<b>GPT3.5</b>	0.20	0.44	0.00	0.44	0.26	0.68	0.36	0.30	0.32	0.16	0.16	0.47

NO-CONTEXT | SELF-CONTEXT | WITH-CONTEXT

Information irrelevance > 20%, and higher for self generated context.  
Addressed by RAG.

Constraint satisfaction ~50%, partially addressed by RAG.

Completeness < 30%, RAG surprisingly at 70%.

# Model performance vs. **context availability**

GPT-4	Irrelevant information ↓			Relevant information (Books from the author)						Completeness ↑			All Correct ↑		
				Satisfied ↑			Unsatisfied ↓								
<b>GPT4</b>	0.26	0.33	0.00	0.51	0.49	0.78	0.24	0.19	0.21	0.24	0.26	0.70	0.08	0.08	0.31
<b>GPT3.5</b>	0.20	0.44	0.00	0.44	0.26	0.68	0.36	0.30	0.32	0.16	0.16	0.47	0.07	0.02	0.15

NO-CONTEXT | SELF-CONTEXT | WITH-CONTEXT

Information irrelevance > 20%, and higher for self generated context.  
Addressed by RAG.

Constraint satisfaction ~50%, partially addressed by RAG.

Completeness < 30%, RAG surprisingly at 70%.

All correctness remains challenging even for RAG.

# Model performance vs. **constraint type**

GPT-4

	Single Item	Irrelevant information ↓			Relevant information (Books from the author)						Completeness ↑			All Correct ↑		
					Satisfied ↑			Unsatisfied ↓								
<b>starts-with</b>	0.96	<b>0.41</b>	0.36	0.01	0.50	0.57	0.79	0.09	0.07	0.20	0.29	0.31	0.83	0.11	0.17	0.47
<b>ends-with</b>	0.80	0.23	0.38	0.00	<b>0.23</b>	<b>0.28</b>	<b>0.31</b>	<b>0.54</b>	<b>0.34</b>	<b>0.69</b>	0.15	0.17	0.46	0.04	0.05	0.06
<b>word-count</b>	0.58	0.21	0.33	0.00	0.61	0.53	0.63	0.17	0.14	0.37	<b>0.07</b>	<b>0.09</b>	0.39	<b>0.00</b>	<b>0.00</b>	<b>0.02</b>
<b>human</b>	0.70	0.36	0.39	0.01	0.41	0.46	0.84	0.23	0.14	0.15	0.16	0.19	0.61	0.06	0.07	0.23
<b>no-human</b>	0.65	0.32	0.36	0.00	0.57	0.55	0.90	0.10	0.09	0.10	0.25	0.31	0.83	<b>0.00</b>	<b>0.00</b>	0.13
<b>city</b>	0.56	0.12	<b>0.46</b>	0.00	0.77	0.38	0.66	0.11	0.16	0.34	0.33	0.26	<b>0.38</b>	0.31	0.20	0.31
<b>no-city</b>	<b>0.54</b>	0.36	0.34	0.00	0.59	0.61	0.93	0.05	0.05	0.07	0.31	0.32	0.91	<b>0.00</b>	<b>0.00</b>	0.26
<b>pub-year</b>	1.00	0.21	0.27	0.00	0.46	0.47	0.90	0.33	0.26	0.10	0.31	0.34	0.88	0.11	0.12	0.53
<b>Summary</b>	0.80	0.26	0.33	0.00	0.51	0.49	0.78	0.24	0.19	0.21	0.24	0.26	0.70	0.08	0.08	0.31

NO-CONTEXT | SELF-CONTEXT | WITH-CONTEXT

Ends-with string constraints difficult to satisfy.

Starts-with string constraints with higher irrelevant information.

Word count string constraints have very low completeness.

# Model performance vs. **constraint type**

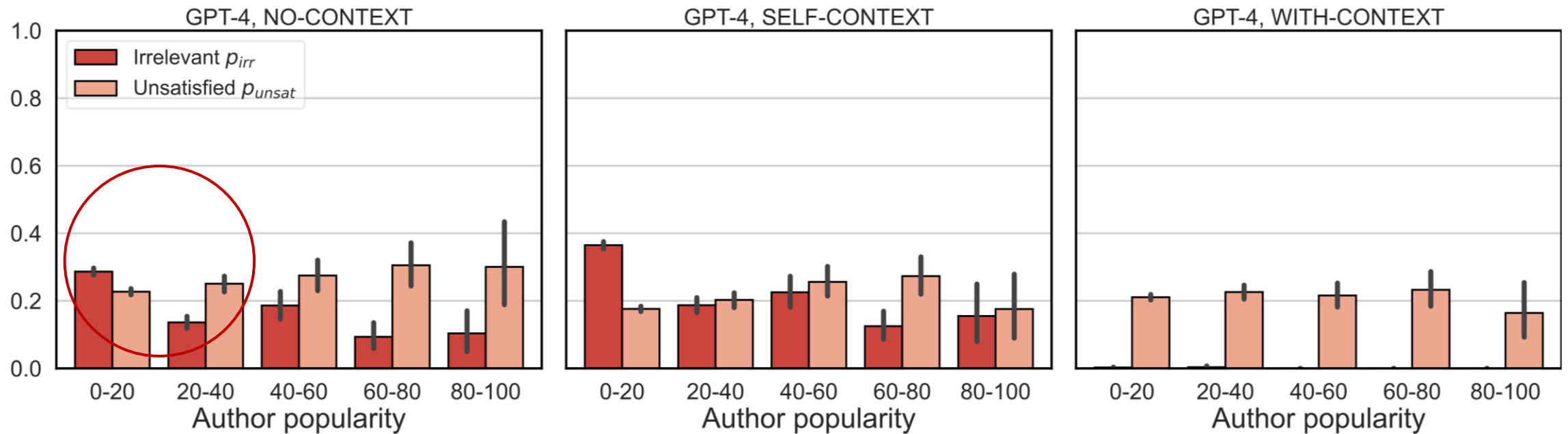
GPT-4

	Single Item	Irrelevant information ↓			Relevant information (Books from the author)						Completeness ↑			All Correct ↑		
		Satisfied ↑	Unsatisfied ↓		Satisfied ↑	Unsatisfied ↓		Satisfied ↑	Unsatisfied ↓		Satisfied ↑	Unsatisfied ↓				
<b>starts-with</b>	0.96	0.41	0.36	0.01	0.50	0.57	0.79	0.09	0.07	0.20	0.29	0.31	0.83	0.11	0.17	0.47
<b>ends-with</b>	0.80	0.23	0.38	0.00	0.23	0.28	0.31	0.54	0.34	0.69	0.15	0.17	0.46	0.04	0.05	0.06
<b>word-count</b>	0.58	0.21	0.33	0.00	0.61	0.53	0.63	0.17	0.14	0.37	0.07	0.09	0.39	0.00	0.00	0.02
<b>human</b>	0.70	0.36	0.39	0.01	0.41	0.46	0.84	0.23	0.14	0.15	0.16	0.19	0.61	0.06	0.07	0.23
<b>no-human</b>	0.65	0.32	0.36	0.00	0.57	0.55	0.90	0.10	0.09	0.10	0.25	0.31	0.83	0.00	0.00	0.13
<b>city</b>	0.56	0.12	0.46	0.00	0.77	0.38	0.66	0.11	0.16	0.34	0.33	0.26	0.38	0.31	0.20	0.31
<b>no-city</b>	0.54	0.36	0.34	0.00	0.59	0.61	0.93	0.05	0.05	0.07	0.31	0.32	0.91	0.00	0.00	0.26
<b>pub-year</b>	1.00	0.21	0.27	0.00	0.46	0.47	0.90	0.33	0.26	0.10	0.31	0.34	0.88	0.11	0.12	0.53
<b>Summary</b>	0.80	0.26	0.33	0.00	0.51	0.49	0.78	0.24	0.19	0.21	0.24	0.26	0.70	0.08	0.08	0.31

NO-CONTEXT | SELF-CONTEXT | WITH-CONTEXT

Entity constraints are easier if applied to a whole context list of books.

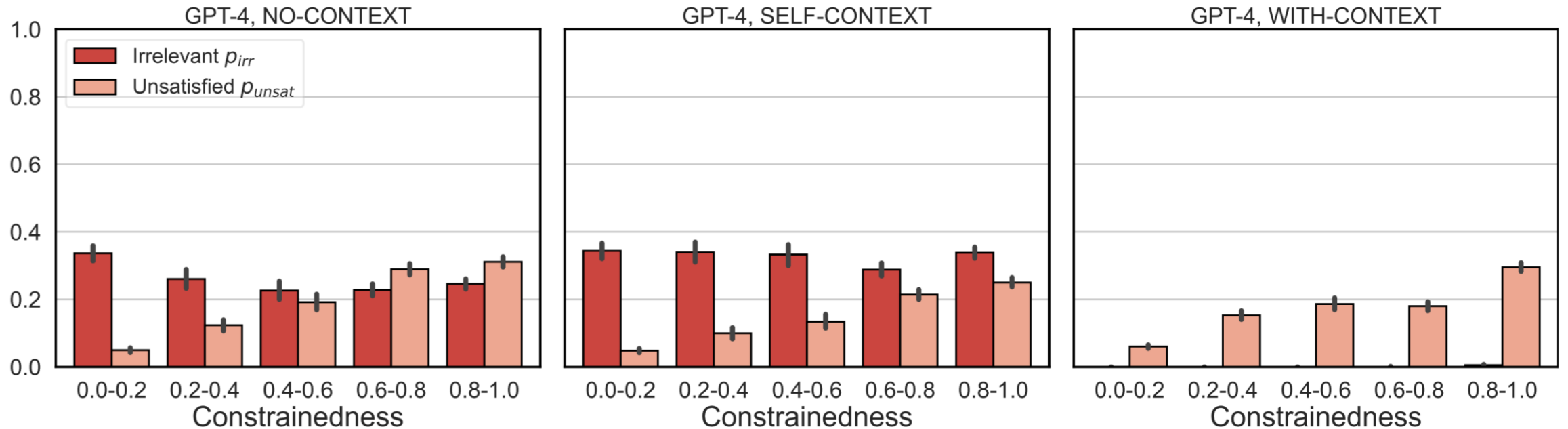
# Model performance vs. author popularity



Fast transition of lower irrelevance, but no further improvement for high popularity.

Constraint satisfaction does not improve for more popular authors.

# Model performance vs. **constrainedness**



Constraint satisfaction improves for lower constrainedness.

Information irrelevance does not improve for lower constrainedness.

# Conclusions and bottlenecks

- LLMs still do not store ALL information, **retrieval augmentation** still needed.
- Retrieval however needs to be done **independently** and **reliably** and is not only a matter of “chain of thought”.
- Some constraints need to be delegated to **programmatic** solutions, but **not all constrained can be programmed.**
- **What is a constraint?** Can they be detected automatically?



## Understanding

**RQ1:** How does model attention change when it can and cannot satisfy constraints?

**RQ2:** Can we predict model failure using model's internal attention flow?

# Background - Transformers

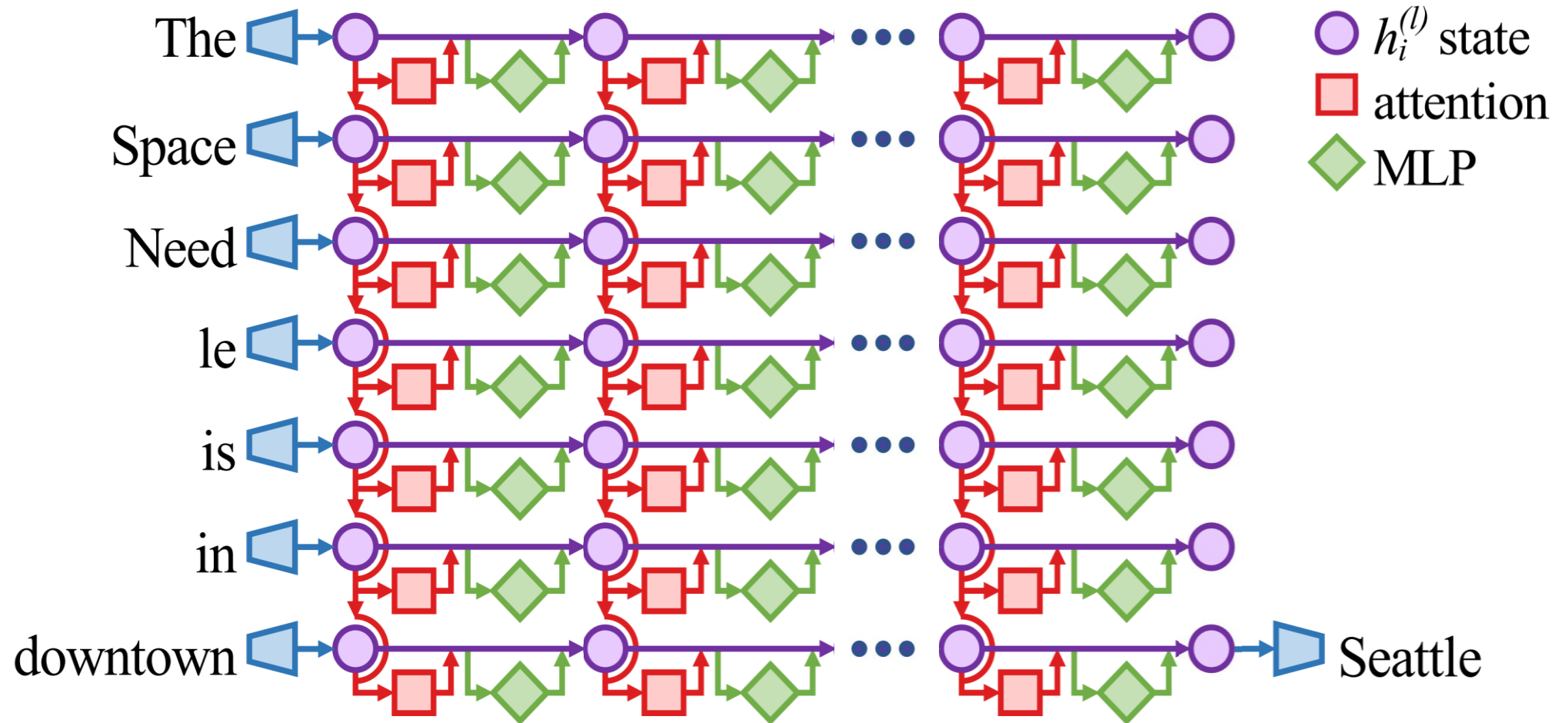


Figure in [Meng et al. 2022; Locating and Editing Factual Associations in GPT]

# Background - Transformers

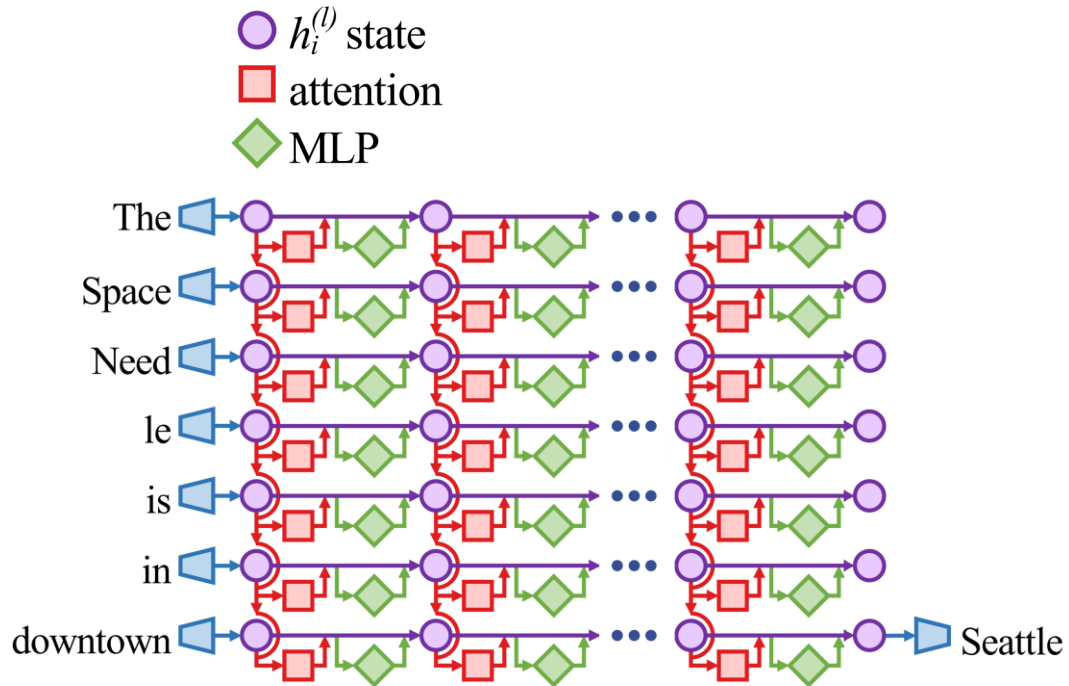


Figure in [Meng et al. 2022; Locating and Editing Factual Associations in GPT]

## Transformer

$$\forall l \in L : x_i^l = x_i^{l-1} + a_i^l + m_i^l$$

## Attention contribution

$$a_i^l = \sum_{j=1}^H A^{l,j} (X^{l-1} W_V^{l,j}) W_O^{l,j}$$

## MLP contribution

$$m_i^l = W_F^l (W_F^l (a_i^l + x_i^{l-1}))$$

# Problem setup

User: Find a list of books written by  $\overbrace{\{\text{Isabelle Allende}\}}^{C_1}$  that have  $\underbrace{\{\text{a single word in the title}\}}_{C_2}$ .

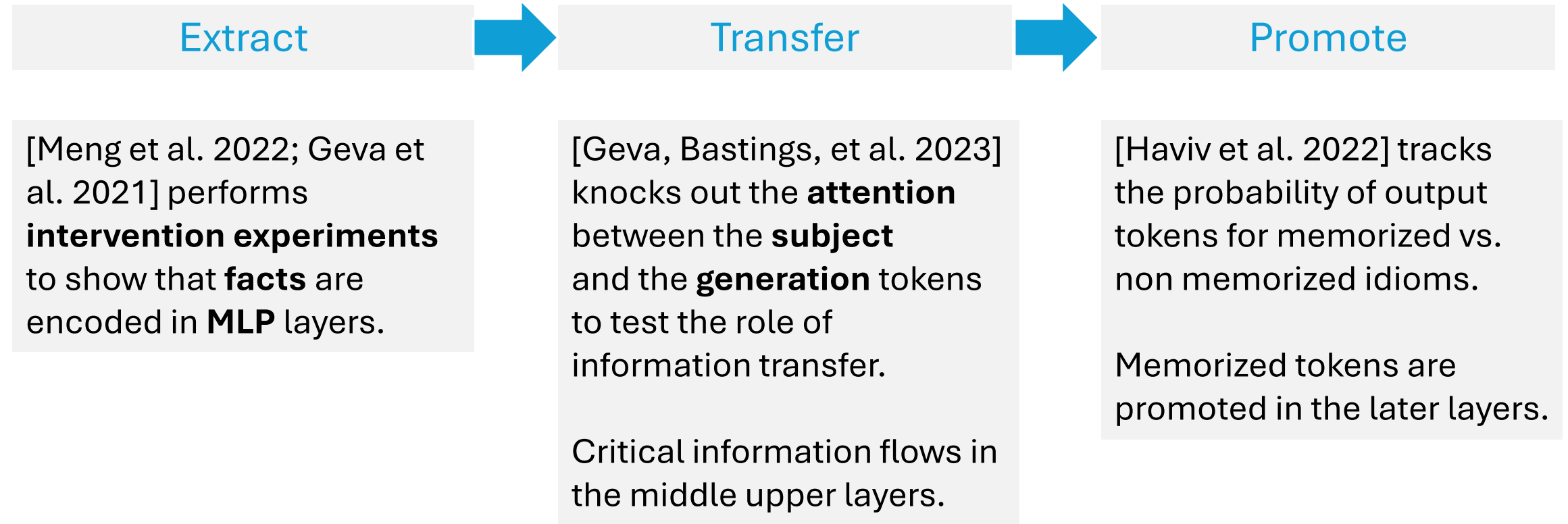
Assistant: The following books from Isabelle Allende have a single word in the title:  $\underbrace{\text{Paula, Ripper, Violeta, Zorro}}_g$ .

## Understanding

How does model attention change when it can and cannot satisfy constraints?

Can we predict model failure using model's internal information flow?

# Information flow in prior work



**Common among prior work:** analysis of <subject, rel, object> for cases when the model succeeds at information retrieval. What happens when it fails?

# Attention flow for characterizing failure and information “popularity”

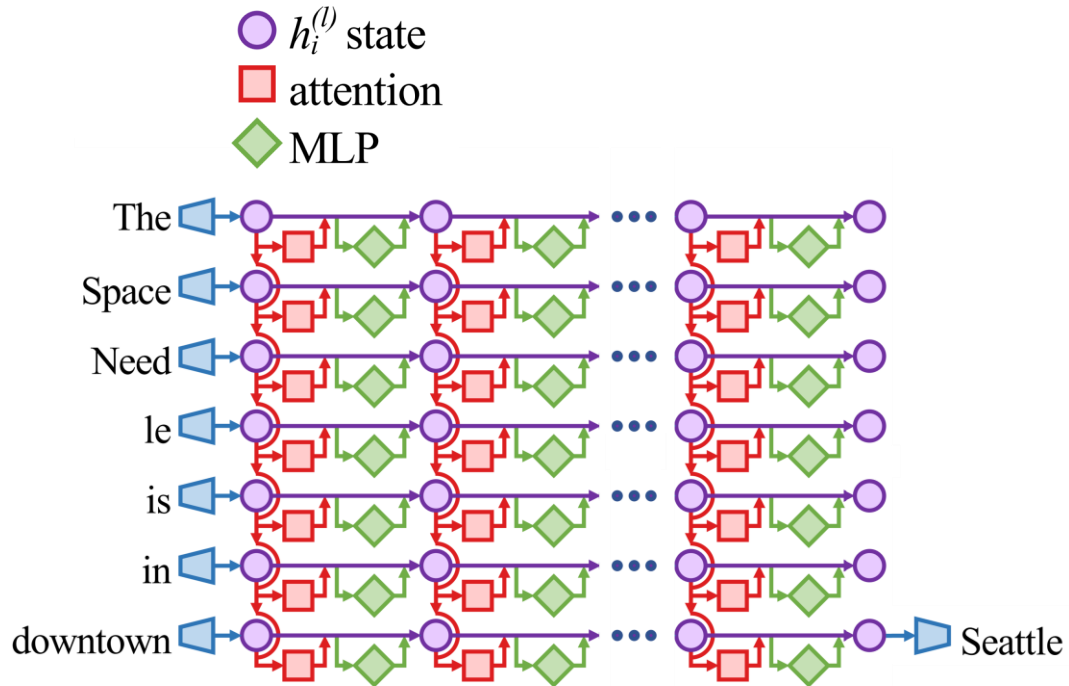


Figure in [Meng et al. 2022; Locating and Editing Factual Associations in GPT]

Attention contribution

constraint (c)  $\rightarrow$  generation (g)

$$|a_{c,g}^l| = \left| \sum_{j=1}^H A_{c,g}^{l,j} (X^{l-1} W_V^{l,j}) W_O^{l,j} \right|$$

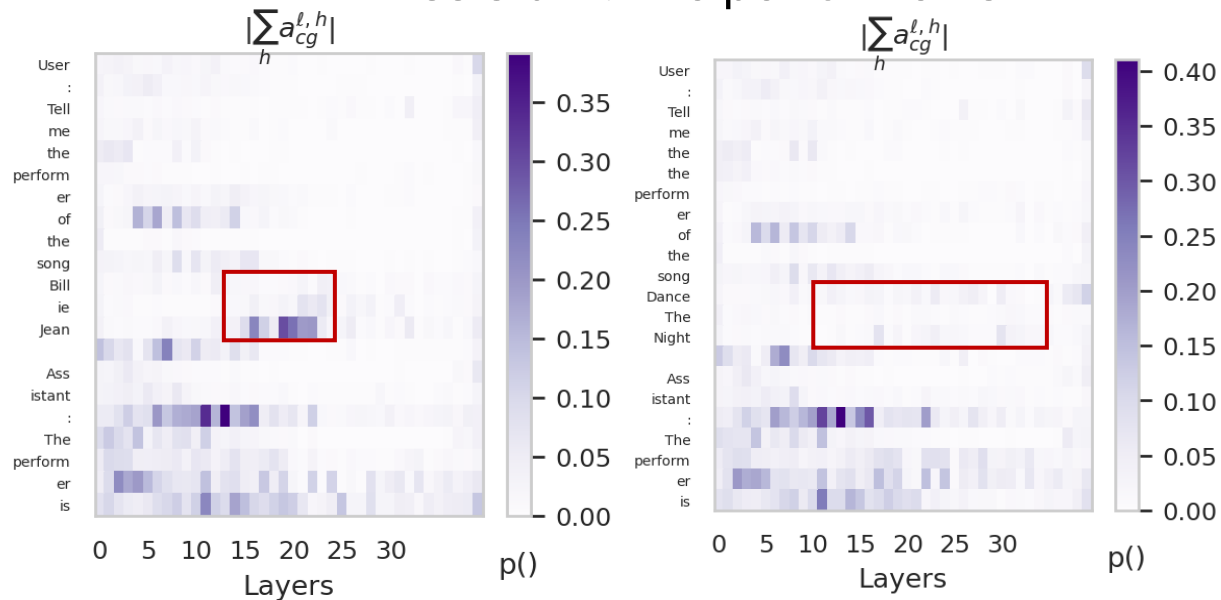
Where is the attention focused

$$\max_{j \in [H]} A_{c,g}^{l,j}$$

# LLAMA-2 13B (40 layers)

User: Tell me the performer of the <song>

Assistant: The performer is



Billie Jean (1983)  
Michael Jackson

Dance The Night  
(2023) Dua Lipa

**Insight:** For many **successful** queries on popular information, high attention contribution is observed from the constraint to the generated tokens.

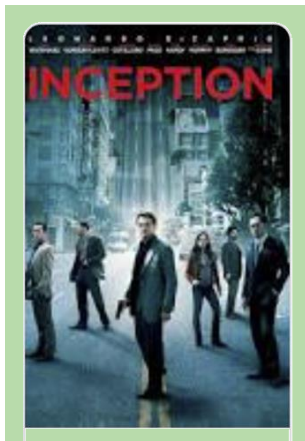
For many **unsuccessful** queries on less popular information (or unknown to the model) attention contribution from the constraint to the generated token remains flat.

# LLAMA-2 13B (40 layers)

User: Tell me the director of the <movie>

Assistant: The director is

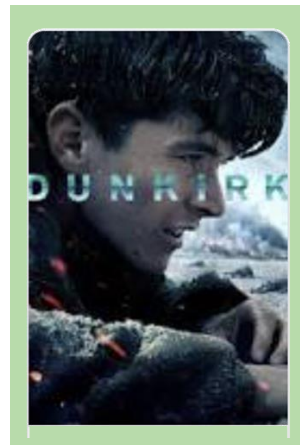
2010  
2.5M reviews



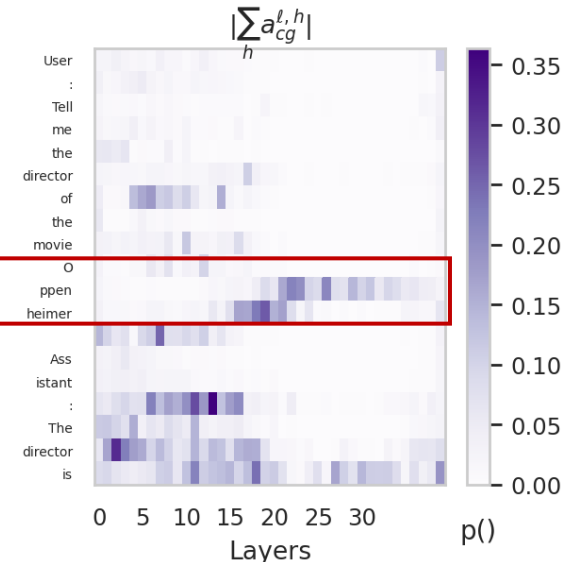
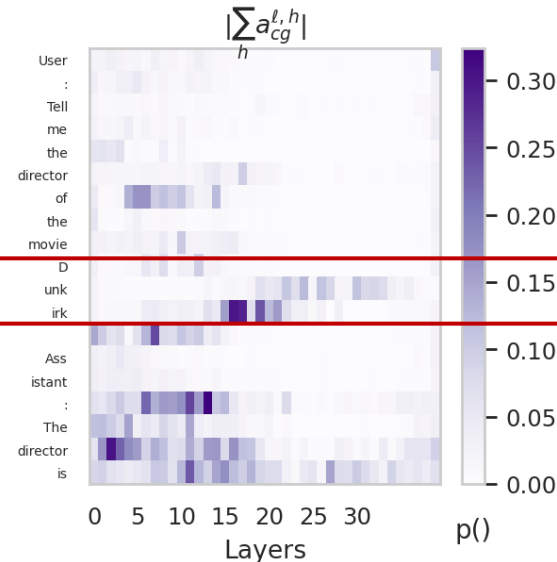
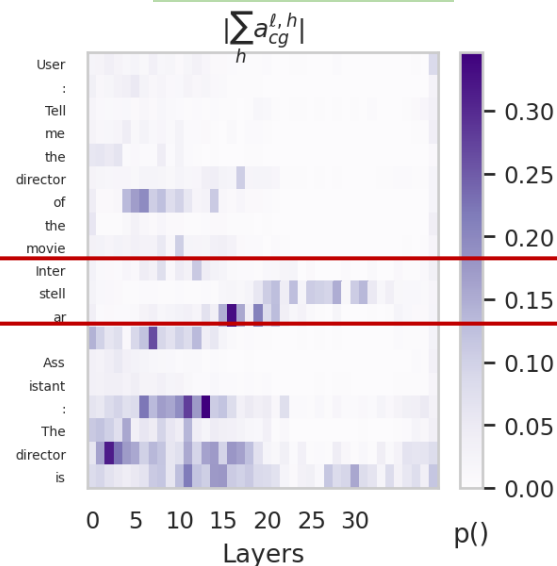
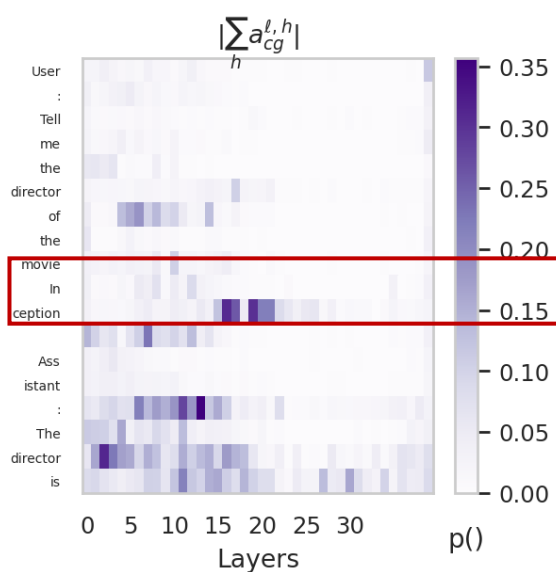
2014  
2M reviews



2017  
0.7M reviews



2023  
0.4M reviews

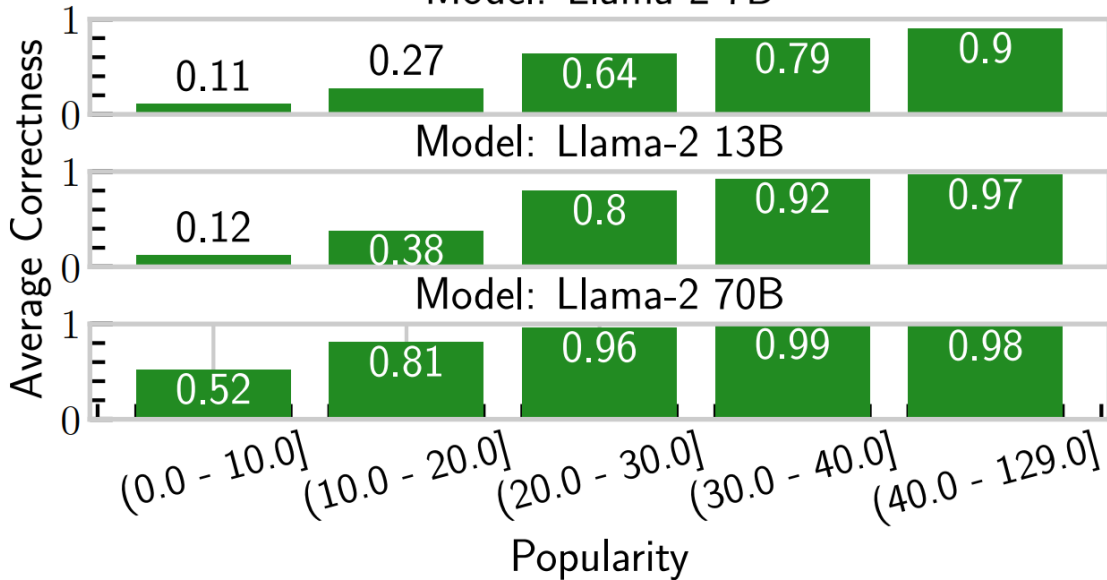




# Popularity and Constrainedness

Ex: Tell me the year the basketball player Kobe Bryant was born in

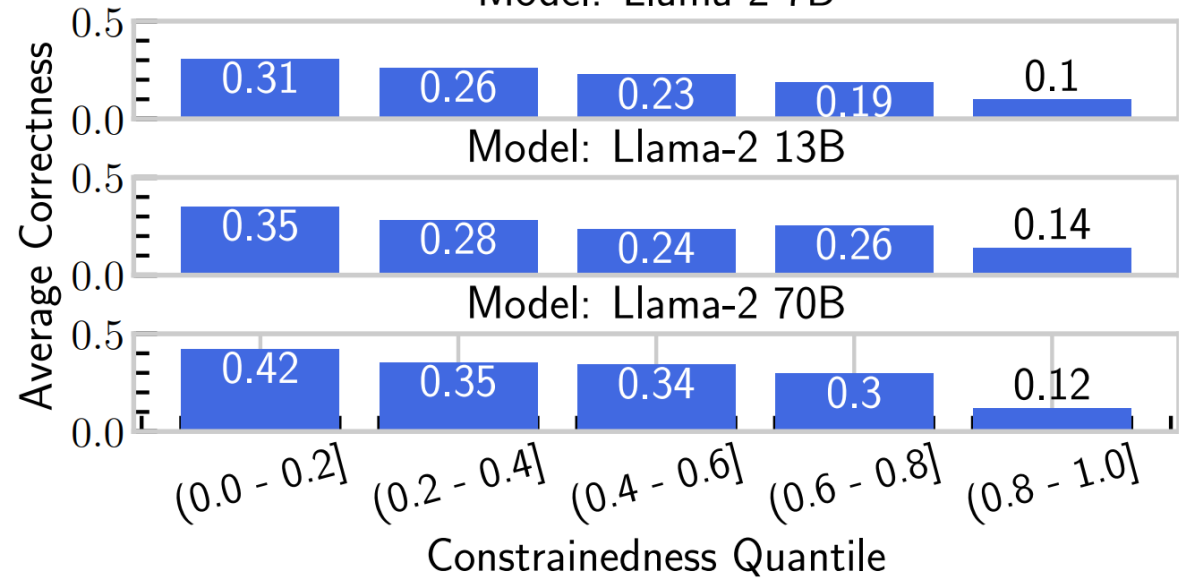
Model: Llama-2 7B



**Insight:** Quicker transition for Llama 70B

Ex: Tell me a word that starts with e and ends with t

Model: Llama-2 7B

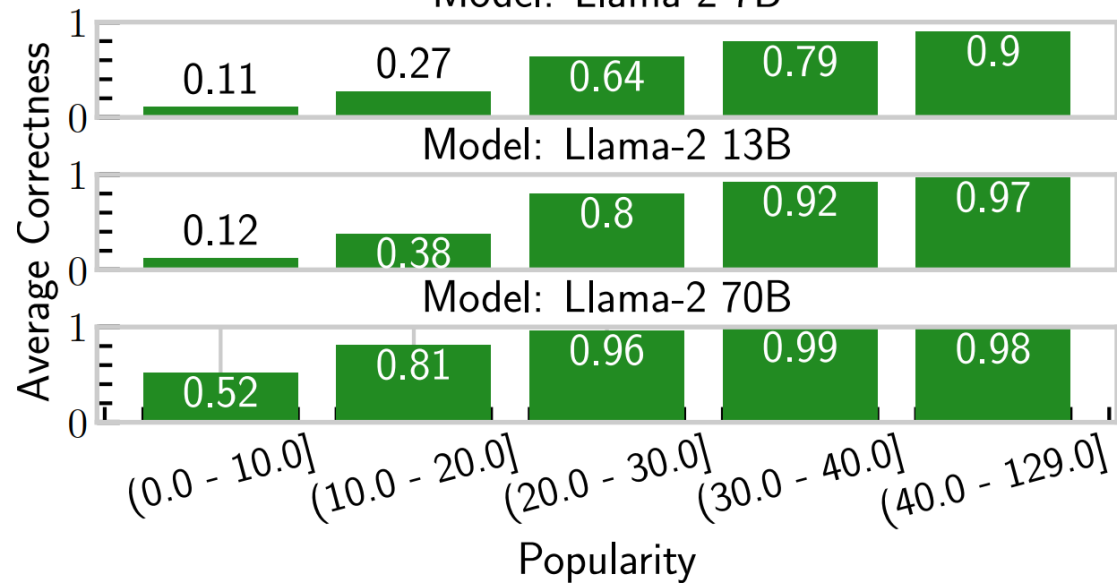


**Insight:** Constrainedness impacts all model sizes. Model size does not help with highly constrained queries.

# Scale and Popularity

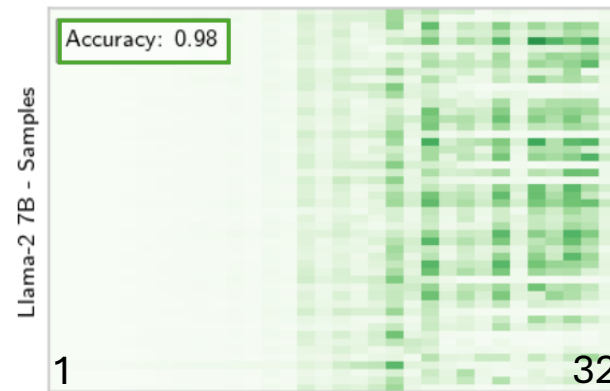
Ex: Tell me the year the basketball player Kobe Bryant was born in

Model: Llama-2 7B



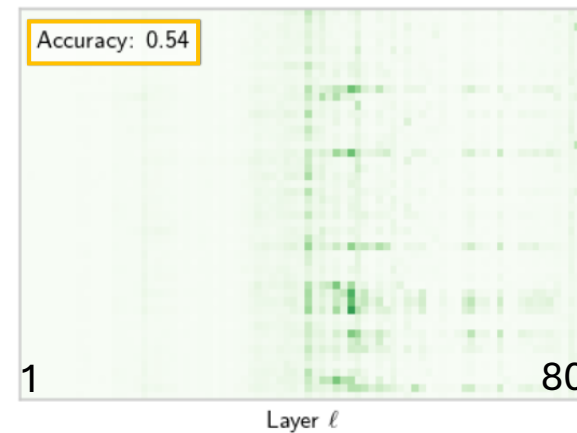
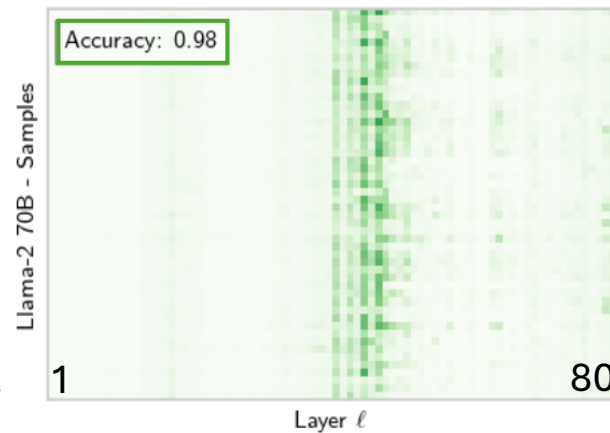
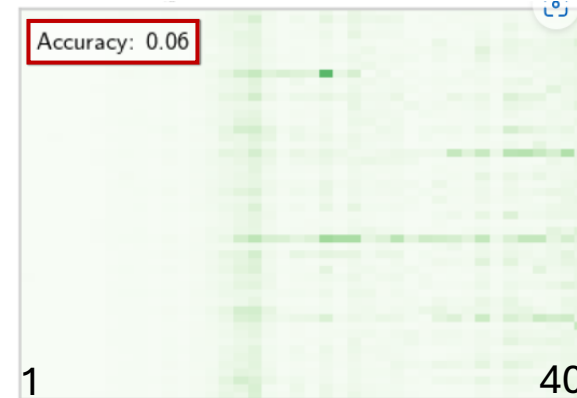
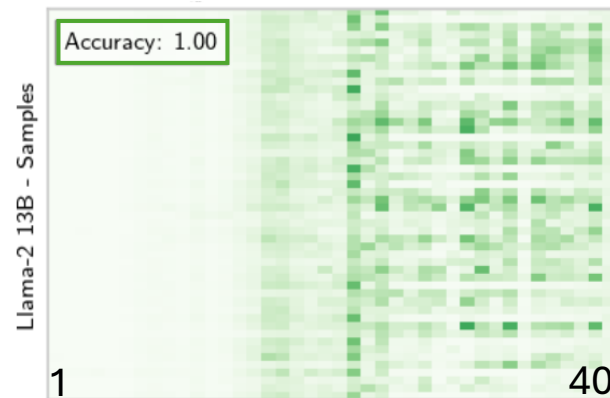
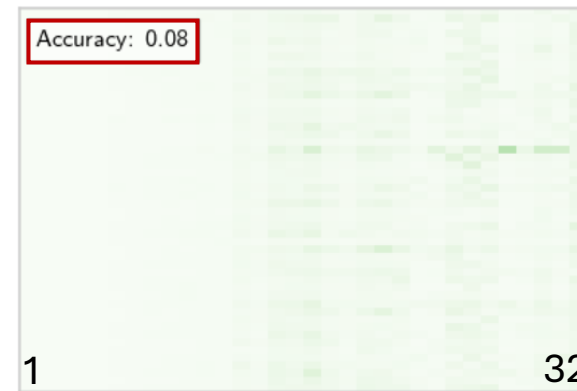
High popularity

$||a_{c,q}^\ell||$  for 50 most popular entities



Low popularity

$||a_{c,q}^\ell||$  for 50 least popular entities



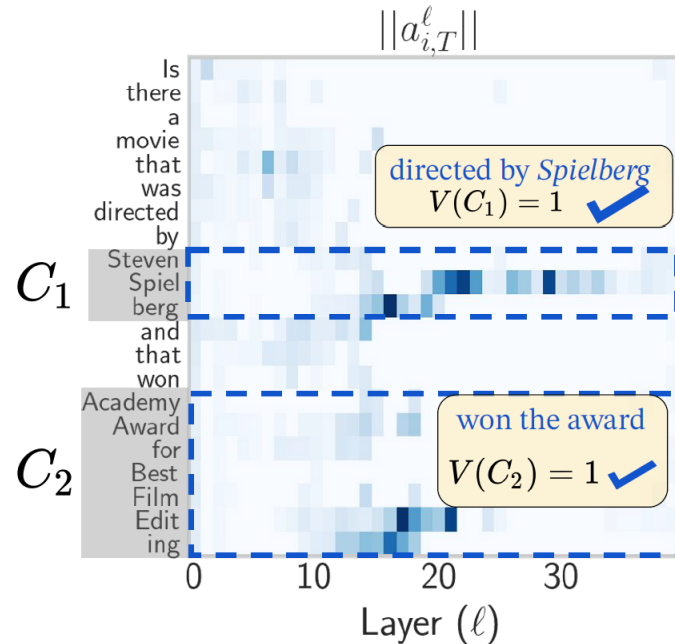
Layer  $\ell$

Layer  $\ell$

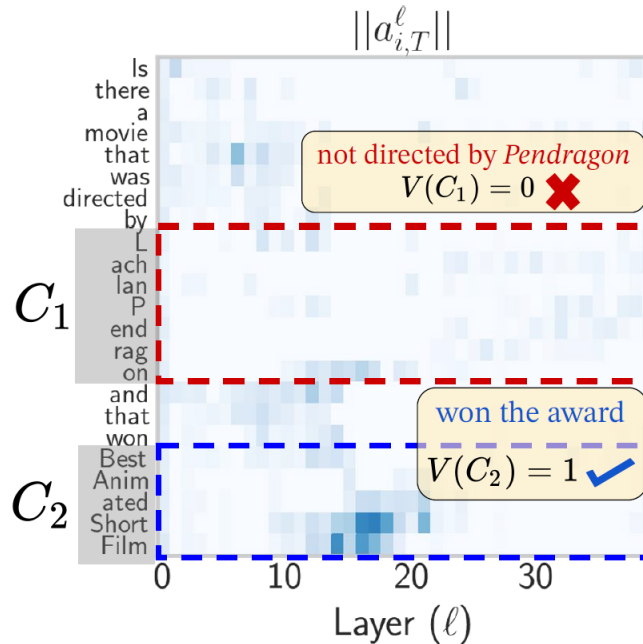
# Tracing constraints

User: Is there a movie that was directed by  $\underbrace{\langle \text{director name} \rangle}_{C_1}$  and that won  $\underbrace{\langle \text{award name} \rangle}_{C_2}$ ?  
 Assistant: Yes, the name of the movie is  $\langle \text{model completion} \rangle$

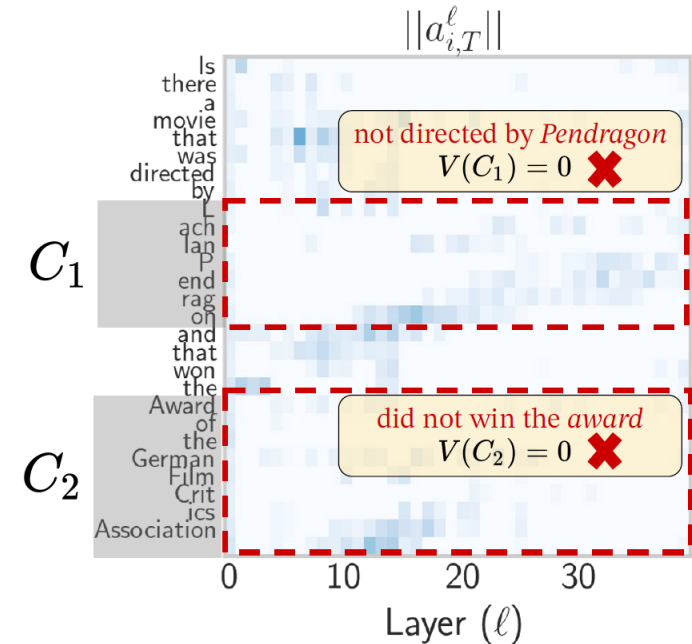
Model Completion: *Schindler's List*



Model Completion: *The Lost Thing*

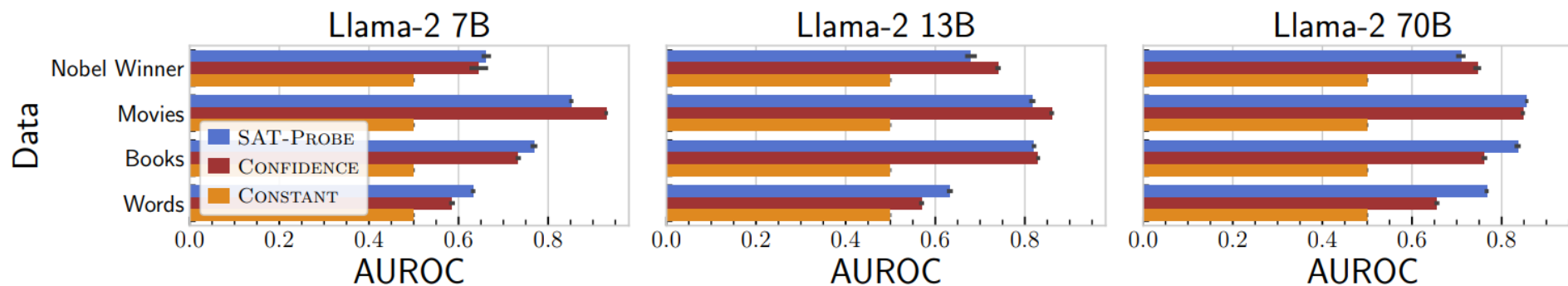


Model Completion: *The Last of the Mohicans*

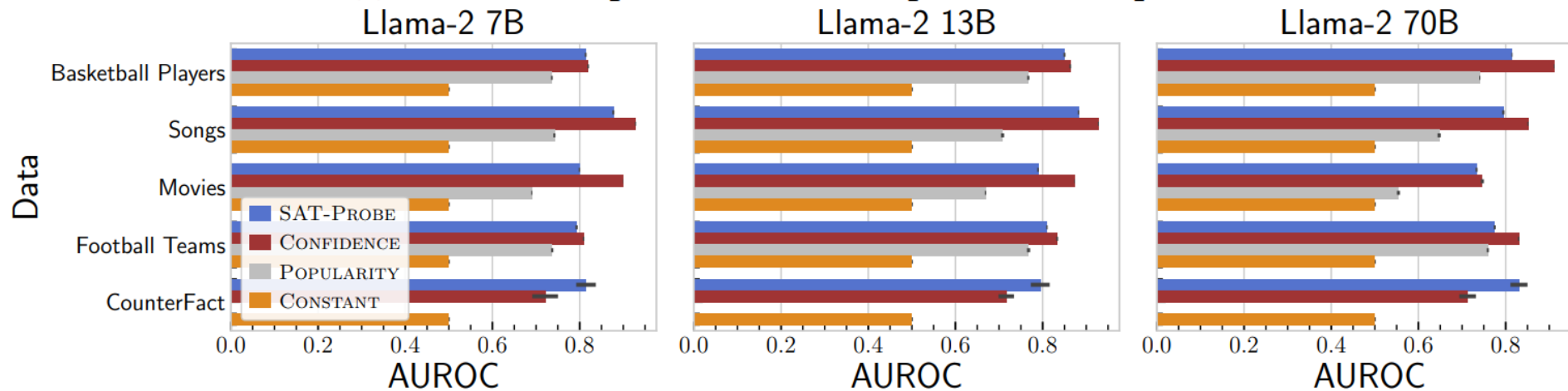


**Insight:** Flow of attention contribution from constraints to the generation can be used to debug which constraint was not satisfied.

# Predicting factual errors (SAT-Probe)

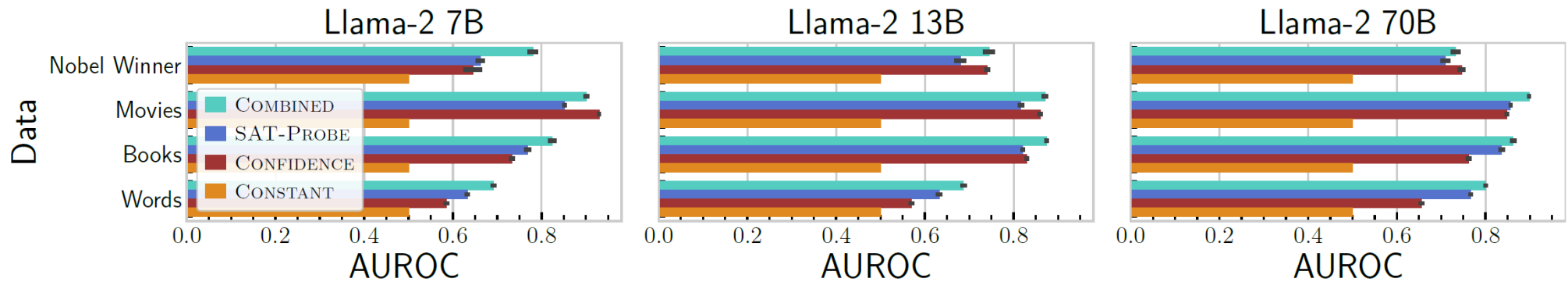


(a) Factual error prediction for multiple constraint queries.

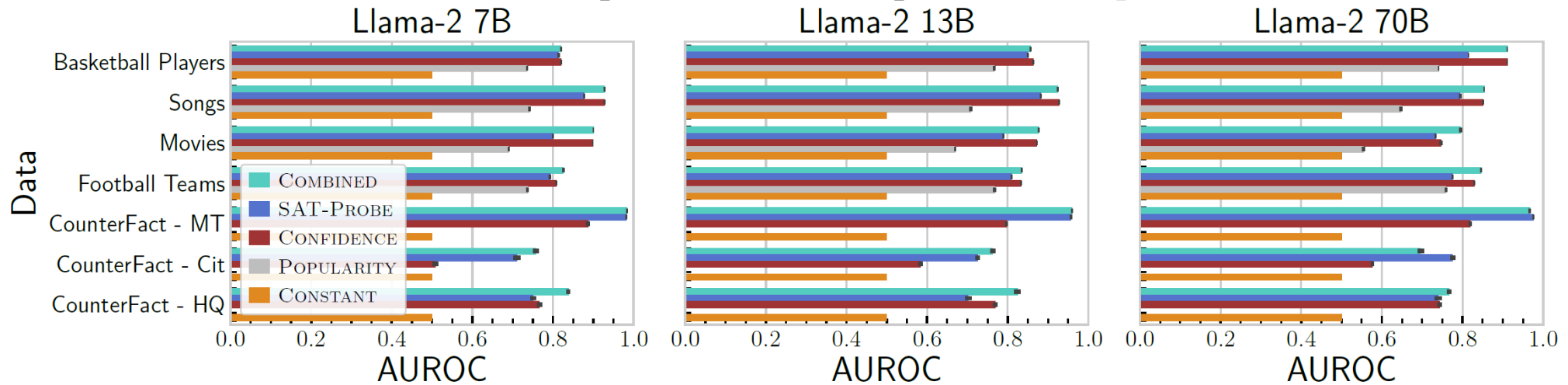


(b) Factual error prediction for single constraint queries.

# Predicting factual errors (combined approach)



(a) Factual error prediction for multiple constraint queries.



(b) Factual error prediction for single constraint queries.

# Insights on model improvement

- Informing RAG techniques on **when** and **what** to retrieve (requires **constraint tracing**).
- Informing multi-model and multi-agent debate on what follow up questions to ask (requires **constraint tracing**).
- Using synthetic constraint satisfaction tasks to foster model grounding.
  - **With soft prompting:** Teaching Language Models to Hallucinate Less with Synthetic Tasks; ICLR 2024 [contact: Erik Jones & Hamid Palangi]
  - **With instruction tuning**

# Soft-prompt tuning via synthetic tasks

The following is a list of names:

[100 Names]

List the first 5 names where the first name starts with Z in the order that they appear. Include both the first and last name in the response. If there are not 5 names that start with Z, return all of the names in the list that start with Z in the order that they appear

Test if names start with the right letter and are from the list.

**Synthetic constraint  
satisfaction task**

Teaching Language Models to Hallucinate Less with Synthetic Tasks;  
ICLR 2024 [contact: Erik Jones (UC Berkeley) & Hamid Palangi (MSR)]

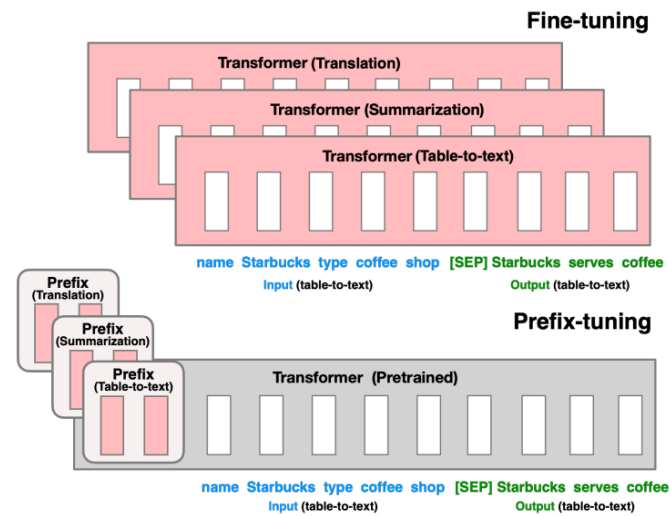


Erik Jones  
UC Berkeley



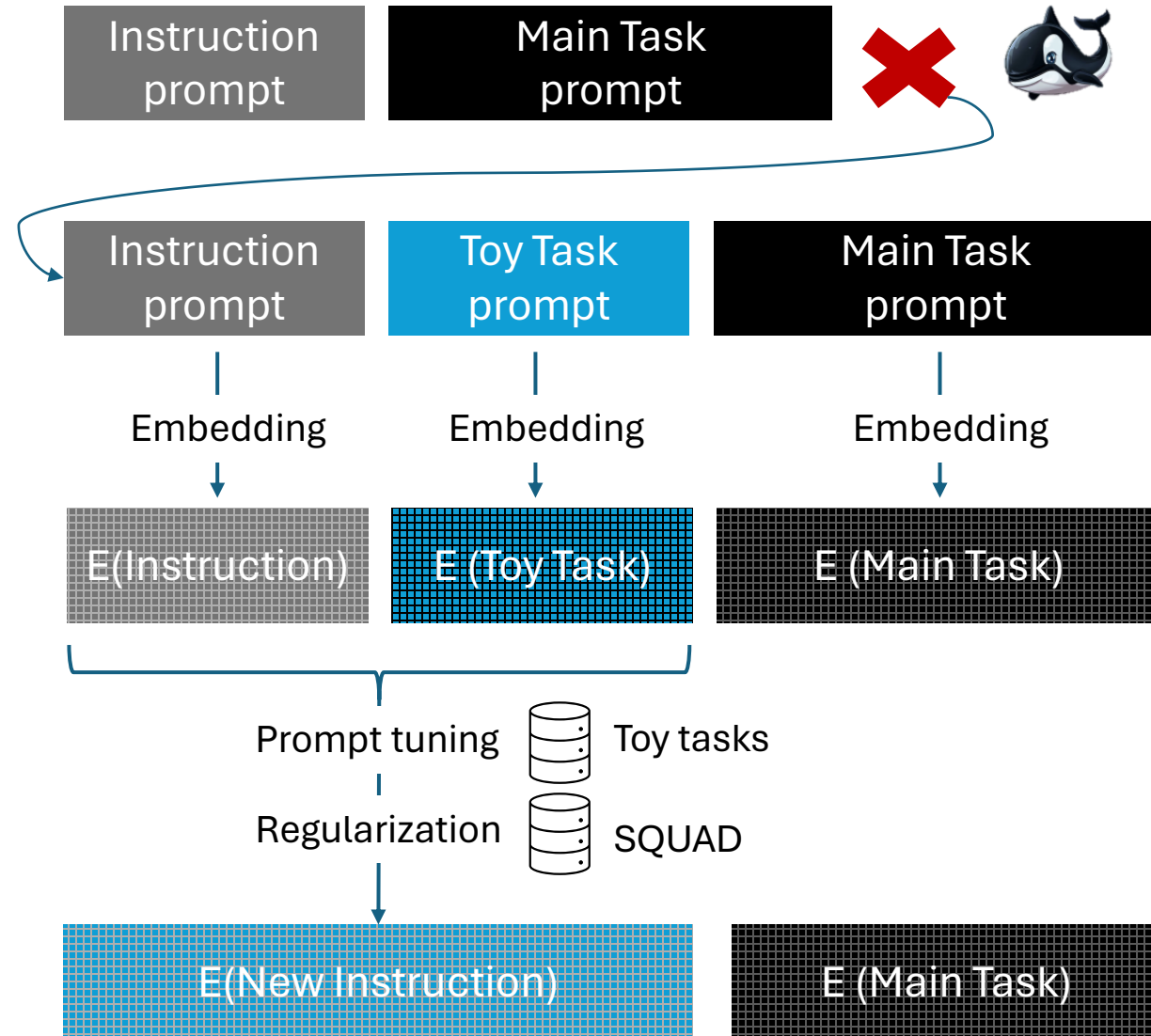
Hamid Palangi  
Microsoft Research

# Soft-prompt tuning via synthetic tasks



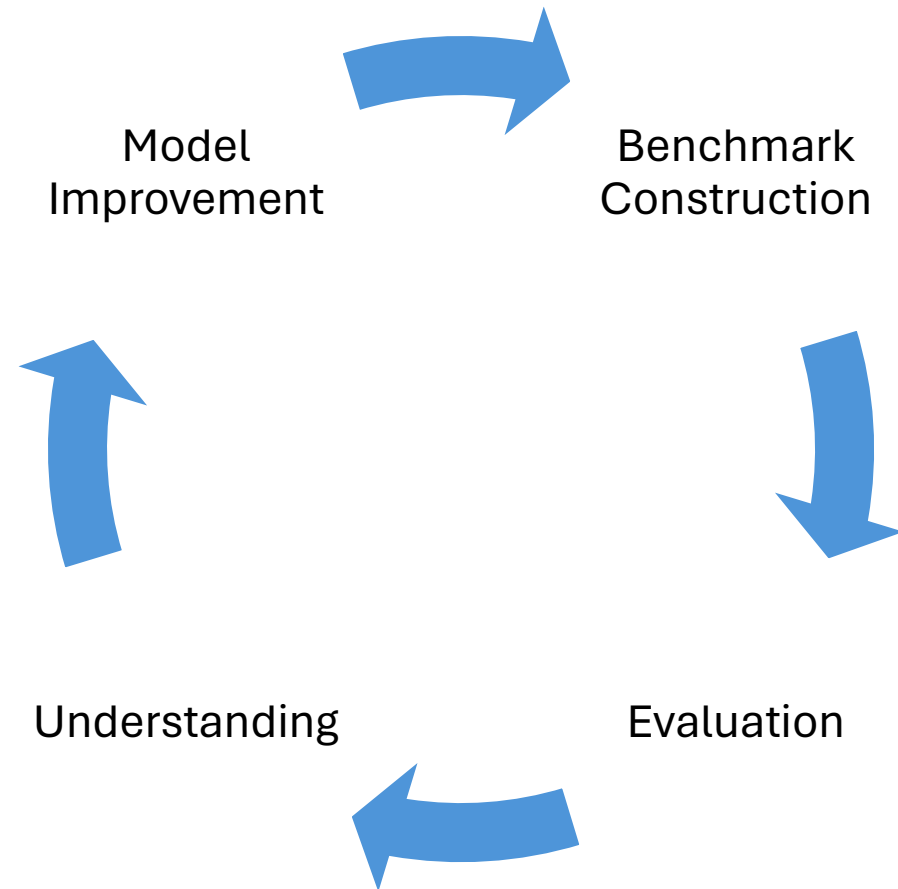
[Li et al. 2021; Prefix-Tuning: Optimizing Continuous Prompts for Generation]

Adds relevant context to the prompt  
Searches for the best relevant context  
Cheaper than fine-tuning





# Vision Forward



A unified framework consisting of a new set of **challenging benchmarks, metrics, visualizations, controlled experiments, and end-to-end interactive evaluation approaches** that can become the basis of the new **development and scientific tools** to understand and improve large-scale model capabilities on fundamental and emerging abilities.

**together with:** Neel Joshi, Hamid Palangi, Rahee Ghosh Peshawaria, Vibhav Vineet, Clarisse Simoes, Shweti Mahajan, Vidhisha Balachandran (joining from CMU)

[AI Frontiers - Microsoft Research](#)

# Prior work mentioned in this presentation

[Meng et al. 2022; Locating and Editing Factual Associations in GPT]

[Geva, Schuster et al. 2021; Transformer Feed-Forward Layers Are Key-Value Memories]

[Geva, Bastings, et al. 2023; Dissecting recall of factual associations in auto-regressive language models]

[Haviv et al. 2022; Understanding transformer memorization recall through idioms]

[Li et al. 2021; Prefix-Tuning: Optimizing Continuous Prompts for Generation]