On Assessing the Order Bias of Large Language Models

Worshop on “Risks and Unintended Harms of Generative AI Systems”
ITADATA 2025

Alex Della Bruna[1,2], Nicola Bena[2], Marco Anisetti[2], Claudio A. Ardagna[2]

[1]Consorzio Interuniversitario Nazionale per l’Informatica
[2]Department of Computer Science, Università degli Studi di Milano

September 11, 2025

Introduction

Large Language Models (LLMs) are driving the AI revolution

- human-like interaction
- multi-modal
- the new cornerstone of AI-based applications
- very popular (ChatGPT reached 810 million weekly active users on August 2025)[1]

[1] https://sqmagazine.co.uk/chatgpt-statistics/

Introduction

Large Language Models (LLMs) are driving the AI revolution

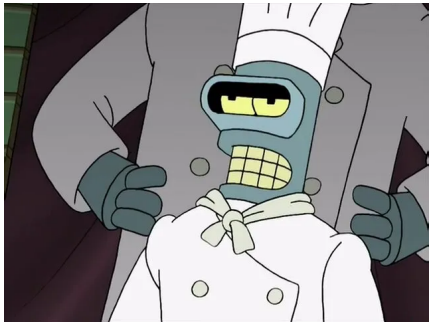
- human-like interaction
- multi-modal
- the new cornerstone of AI-based applications
- very popular (ChatGPT reached 810 million weekly active users on August 2025)[1]

...but are LLMs reliable?

[1] https://sqmagazine.co.uk/chatgpt-statistics/

Are LLMs Reliable?

Q: what is the best Italian food?



Are LLMs Reliable?

Q: what is the best Italian food?

Choose between:

- carbonara spaghetti
- lasagna
- pineapple pizza

A: carbonara spaghetti

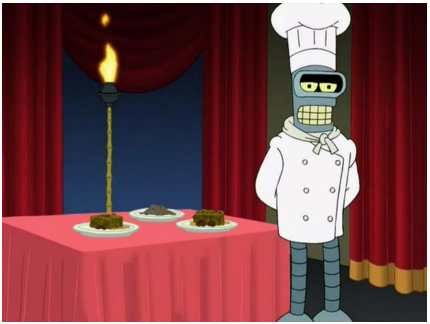
Are LLMs Reliable?

Q: what is the best Italian food?

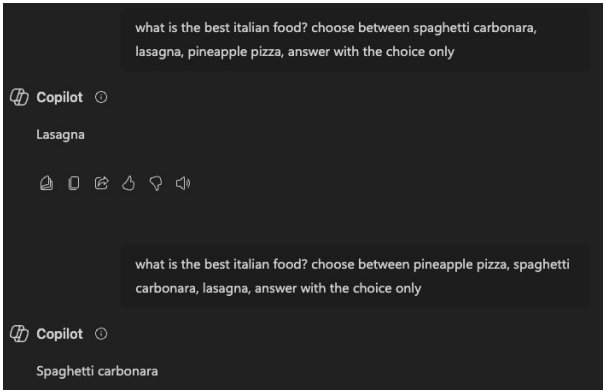
Choose between:

- pineapple pizza
- carbonara spaghetti
- lasagna

A: pineapple pizza



Are LLMs Reliable? Real-World Example



$\Longrightarrow$ the order of the choices affect the LLM answer!

...at least it did not choose pineapple pizza...

LLM Biases

State-of-the-art LLMs suffer from a variety of biases

- egocentric bias
- attentional bias
- order bias
- salience bias
- bandwagon effect
- compassion fade
- ...

LLM Biases

State-of-the-art LLMs suffer from a variety of biases

- egocentric bias
- attentional bias
- order bias
- salience bias
- bandwagon effect
- compassion fade
- ...

These bias matter, because we are using LLMs to assess (the quality of) other LLMs (*LLM-as-a-Judge*)

If the Judge is biased, then its assessment can be biased too!

LLM Biases

State-of-the-art LLMs suffer from a variety of biases

- egocentric bias
- attentional bias
- order bias
- salience bias
- bandwagon effect
- compassion fade
- ...

Large Language Models are not Fair Evaluators

Peiyi Wang[1] **Lei Li**[1] **Liang Chen**[1] **Zefan Cai**[1] **Dawei Zhu**[1]
Binghuai Lin[3] **Yunbo Cao**[3] **Qi Liu**[2] **Tianyu Liu**[3] **Zhifang Sui**[1]
[1] National Key Laboratory for Multimedia Information Processing, Peking University
[2] The University of Hong Kong [3] Tencent Cloud AI
{wangpeiyi9979, nlp.lilei, zefncai}@gmail.com
leo.liang.chen@outlook.com; {dwzhu, szf}@pku.edu.cn
liuqi@cs.hku.hk; {binghuailin, yunbocao, rogertyliu}@tencent.com

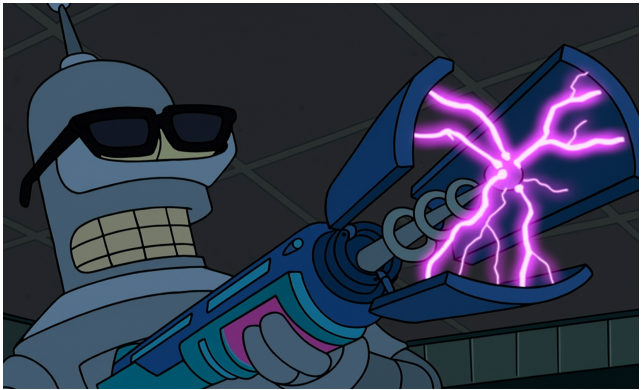
Benchmarking Cognitive Biases in Large Language Models as Evaluators

Ryan Koo[1] **Minhwa Lee**[1] **Vipul Raheja**[2] **Jonginn Park**[1], **Zae Myung Kim**[1] **Dongyeop Kang**[1]
[1]University of Minnesota, [2]Grammarly
{koo00017,lee03533,park2838,kim01756,dongyeop}@umn.edu,
vipul.raheja@grammarly.com

LLM Biases

State-of-the-art LLMs suffer from a variety of biases

- egocentric bias
- attentional bias
- order bias $\Longrightarrow$ our focus
- salience bias
- bandwagon effect
- compassion fade
- ...



Our Approach

Goal: assess order bias of SOTA LLMs in different configurations

We assessed order bias using prompts where we varied

- the number n of available choices $\{pr_1, ..., pr_n\}$
- the number m of choices to select among the available choices
- the type of example E for in-context learning

- $prompt(n, m, \{pr\}, E)$ denotes a prompt with a specific value of n, m, E, and available choices $\{pr_1, ..., pr_n\}$

Experiments: Prompt Template

Select the *m* most relevant probes for the scenario CLOUD out of the following *n* probes:
⟨*Probe list*⟩

Respond on the basis of the following example:

Query: "*Select the m′ most relevant probes for the scenario ML out of the following n′ probes:*
⟨*Sample probe list*⟩"
Answer: "*The m′ most relevant probes to the scenario are:*
⟨*First m′ probes*⟩"

Reply only with a JSON object, put the resulting probe names in a field named "`res_probes`". Do not include the description field. Do not add anything other than the JSON object. Ensure that the probe order does not affect your judgment.

Experiments: Process

1. Generate of the set of available choices $\{pr_1, ..., pr_{100}\}$
2. Generate prompts $prompt_1(n, m, \{pr\}_1, E)$ and $prompt_2(n, m, \{pr\}_2, E)$ where the order of the available choices $\{pr\}_1$ and $\{pr\}_2$ are first extracted at random from step 1 and then randomly sorted
3. Submit prompts $prompt_1(n, m, \{pr\}_1, E)$ and $prompt_2(n, m, \{pr\}_2, E)$ to LLM llm and record answers $ans(prompt_1, llm)$ and $ans(prompt_2, llm)$
   - $ans(prompt, llm) = \{pr_1, \ldots, pr_m\}$

Experiments: Process

4. Compute bias

Order Bias

$$deg = \frac{|ans(prompt_1, llm) \cap ans(prompt_2, llm)|}{m}$$

$$bias = 1 - deg$$

Experiments: Setup

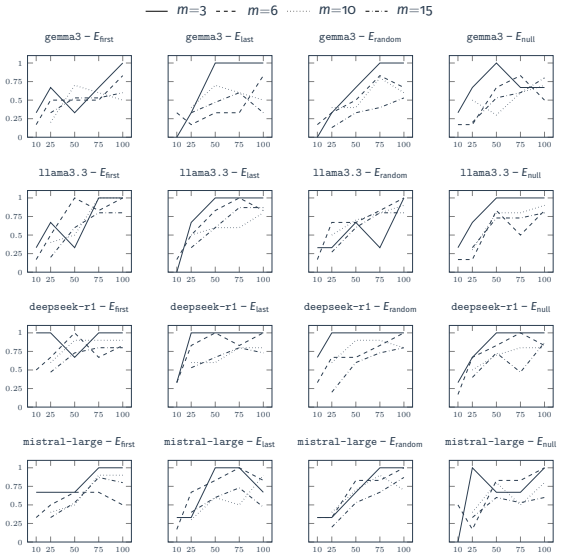
Param.	Values
n	10, 25, 50, 75, 100
m	3, 6, 10, 15
n'	10
m'	2

Example type	Explanation
E_{first}	The chosen probes are the first n' listed among the n' available probes
E_{last}	The chosen probes are the last n' listed among the n' available probes
E_{random}	The chosen probes are n' random listed among the n' available probes
E_{null}	No example

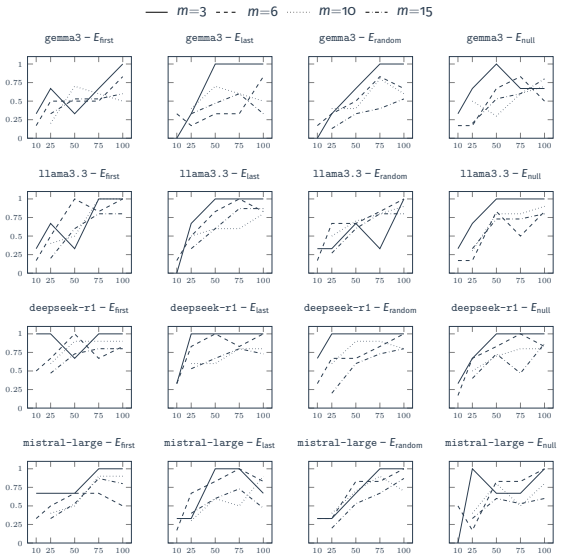
LLM name	Scale (N° of param.)	Release date
Gemma3 27B	Small–medium (27B)	Feb 2025
Llama3.3 70B	Medium–large (70B)	Dec 2024
Deepseek R1 70B	Medium–large (70B)	Jun 2025
Mistral 123B	Large (123B)	Nov 2024

SOTA: $n = 2$ $m = 1$

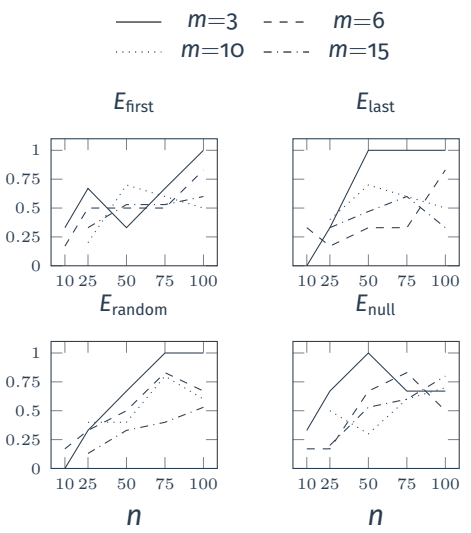
Experiments: Results



Experiments: Results

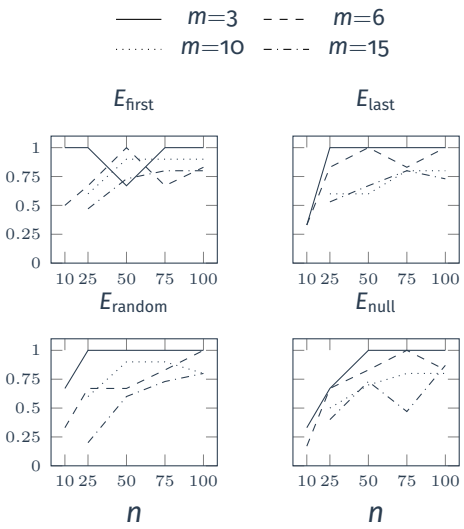


Experiments: Results – Gemma3



- Medium order bias (0.524)
- Bias increases as n increases
- Bias decreases as m increases
- Examples do not affect the bias

Experiments: Results – Deepseek-R1 70B



- Highest order bias (0.761)
- Bias increases as n increases
- Bias decreases as m increases
- Examples designed to induce order bias result in higher bias

Experiments: Findings

LLM	*Bias*
gemma3	0.524
mistral-large	0.640
llama3.3	0.661
deepseek-r1	0.761

Main findings

- All LLMs are affected by order bias
- Bias increases as n increases and decreases as m increases
- The impact of the type of example varies
- A small LLM does not imply higher bias (and vice versa)

Experiments: Findings

LLM	$n \leq 25$	$n > 25$
gemma3	0.298	0.637
llama3.3	0.362	0.810
deepseek-r1	0.586	0.848
mistral-large	0.393	0.764

LLM	$m \leq 6$	$n > 6$
gemma3	0.550	0.492
llama3.3	0.658	0.663
deepseek-r1	0.808	0.701
mistral-large	0.671	0.602

Main findings

- All LLMs are affected by order bias
- Bias increases as *n* increases and decreases as *m* increases
- The impact of the type of example varies
- A small LLM does not imply higher bias (and vice versa)

Experiments: Findings

LLM	E_{first}, E_{last}	E_{random}, E_{null}
gemma3	0.521	0.528
mistral-large	0.646	0.635
llama3.3	0.675	0.646
deepseek-r1	0.786	0.735

Main findings

- All LLMs are affected by order bias
- Bias increases as n increases and decreases as m increases
- The impact of the type of example varies
  - Medium on gemma3 and mistral-large
  - High on llama3.3 and deepseek-r1 $\Longrightarrow$ the LLMs where the bias is highest
- A small LLM does not imply higher bias (and vice versa)

Experiments: Findings

LLM	Parameters	*Bias*
gemma3	27B	0.524
mistral-large	123B	0.640
llama3.3	70B	0.661
deepseek-r1	70B	0.761

Main findings

- All LLMs are affected by order bias
- Bias increases as n increases and decreases as m increases
- The impact of the type of example varies
- A small LLM does not imply higher bias (and vice versa)

Conclusions

We provided a preliminary, systematic analysis of order bias, showing that

- state-of-the-art LLMs are largely affected by it
- the number of parameters of the LLM does not linearly correlate with the bias

Future work

- assessment of other types of bias
- LLM certification

Conclusions?