

# Understanding bias towards salient factual values in LLMs

Praveen Kumar Ramesh & Evelyn Yee

LEI Group

September 20, 2023



# Contents

- 1 Motivation
- 2 Hypothesis
- 3 Datasets
- 4 Experiment

# Motivation

# Motivation

- Hallucination could be caused due to:
  - Inconsistencies in the training corpus.
  - Nature of the downstream task.
  - Memorization due to parametric knowledge bias.
- In previous experiments we observed
  - Decline in performance while recalling death year both in noisy and pristine conditions.
  - Noticeable bias towards recalling birth year over death year.
- Questions?
  - Does the content of in-context examples have **significant** effect on model's performance in completing the query or does the model perform task according to its own bias?
  - If there is a bias, where does it stem from?

# Hypothesis

# Definition

## Definition 1

**Entity:** An entity is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not. <sup>a</sup>

---

<sup>a</sup>(Definition by GPT4)

## Definition 2

**Concepts/Attributes:** Observable property of an entity. An entity has  $n$  concepts  $S = \{c_1, c_2, \dots, c_n\}$ .

## Definition 3

**Concept Class:** Subset of  $S$  in which the values of the concepts share a common domain/data type.

# Definition

## Definition 4

**Contextualising Attributes:** A subset of an entity's attributes whose values uniquely identify it from other entities of the same type. (i.e. attributes which form a "key")

## Definition 5

**Target Attributes:** A subset of an entity's attributes which we prompt the model to recall after contextualising the entity.

## Definition 6

**Task Agnostic Prompting:** An ICL (in-context learning) prompting strategy where the examples presented in the prompt provide sample input-output pairs, but the text of the prompt does not specifically state the relationship between these inputs and outputs.

# Null Hypothesis

Given an entity with  $n$  **concepts** which forms  $k$  **concept classes**, when an LLM is prompted to recall a value of a concept  $c \in \mathbb{C}_i$  ( $\mathbb{C}_i$  is the concept class to which the  $c$  belongs to) under **task agnostic conditions**, the model is not biased towards recalling values of a subset of **salient concepts** in  $\mathbb{C}_i$ .



# Datasets

# Datasets

Entity	Contextualizing Attributes	Concept Classes	Total Records	Verified Records	Source
Movie	{title, release year, genre, description, certificate}	<b>People:</b> {Lead Actor, Director}	999	947	kaggle
Person	{name, nationality, occupation}	<b>Years:</b> {Birth year, death year} <b>Places:</b> {birth place, death place}	15279	2586	wikibio
Nobel Laureates	{name, reason, category}	<b>Years:</b> {Birth year, death year, win year} <b>Places:</b> {birth place, death place, work place}	989	849	opendatasoft

# Experiment

# Brief outlook

We divide our experiments in to 4 sequential phases.

- 1 Data Collection
- 2 LLM knowledge verification
- 3 Factual Recall QA
- 4 Response Analysis

# Data collection

- Source data from various credible publicly available resources.
- Format the raw data in to a standard structure.

# LLM Knowledge Verification

- We verify that the model can recall the entity (not necessarily the attributes of the entity) from its pre-trained knowledge
- Here we use simple prompting strategy like
  - instruct model to make an exhaustive list of all entities in the domain.(if the domain is small)
  - giving an entity's context to the model to check if it can recall the entity's identity.


# Factual Recall QA

There are two major steps in this phase

## Step 1:

- Elicit the model to recall a target concept for an entity, using a direct prompting strategy.
- To engineer such prompt we will test 10-15 candidate prompts over a sample set of 50 entities. With temperature = 1 we prompt the model 5 times with each of (prompt, entity) pair and average the variance across the 250 results (5 results for each of 50 entities) for each prompt and choose the prompt with least variance.<sup>1</sup>

---

<sup>1</sup>credits to Prudhviraaj Naidu for suggesting this method. 

# Factual Recall QA

## Step 2:

- We filter the full set of entities to those for which the model was able to recall the correct value of the target concept in the previous step.
- Using this subset of filtered records we construct two types of ICL prompts: direct prompts and task-agnostic prompts.



# Factual Recall QA

In **direct prompting**, we directly specify the association of concept with the entity/context.

John Robbins, an American author, was born in 1947.

Isabella Ferrari, an Italian actress, was born in 1964.

Ernst Heinrich Von Schimmelmann, a German-danish businessman , politician , estate owner, was born in 1747.

Jibanananda Das, an Indian poet , writer, was \_\_\_\_\_

# Factual Recall QA

In task-agnostic prompting, we do not specify the association of the concept with the entity/context provided. For example,

John Robbins, an American author: 1947.

Isabella Ferrari, an Italian actress: 1964.

Ernst Heinrich Von Schimmelmann, a German-danish businessman , politician , estate owner: 1747.

Jibanananda Das, an Indian poet , writer: \_\_\_\_\_

# Response Analysis

- Identify if the model is biased towards any specific concept within each concept class.
- Analytically, we may plot the distribution of the hallucinated values provided by the model. Histograms or frequency plots might be appropriate here.
- For domains with quantitative values, use measures of central tendency (mean, median) and dispersion (standard deviation) to summarize the responses.