What's in a name? The Influence of Personal Names on Spatial Reasoning in BLOOM Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large language models have been shown to exhibit reasoning capability. But the ability of these models to truly comprehend the reasoning task is not yet clear. An ideal model capable of reasoning would not be affected by the names of the entities over which the relations are defined. In this paper, we consider an algorithmically generated spatial reasoning task over the names of persons. We show that the choice of names has a significant impact on the reasoning accuracy of BLOOM large language models. Using popular names from different countries of the world, we show that BLOOM large language models are susceptible to undesirable variations in reasoning ability even though the underlying logical reasoning challenge does not depend on these names. We further identify that the conditional log probability scores characterizing the uncertainty in prediction produced by BLOOM models are not well-calibrated and cannot be used to detect such reasoning errors. We then suggest a new approach based on model self-explanations and iterative model introspection that performs better than BLOOM conditional log probability scores in detecting such errors, and may help alleviate the bias exhibited by these models.

1 Introduction

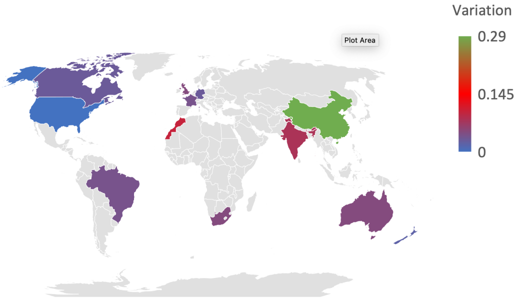


Figure 1: A map showing the variation in performance of the largest BLOOM model on a linear spatial reasoning task with 5 female names.

Over the last four decades, cognitive psychologists and neurolingusitic studies have investigated the impact of proper names on human cognition with interesting outcomes (Bredart et al., 2002). It is now widely accepted that human beings find it more difficult to recall personal names (Young et al., 1985; 1988) than other kinds of words, including relatively rare common nouns. Several different explanations have been investigated for such a discrepancy, including the semantic tag nature of personal names without being descriptive (Fogler & James, 2007), the need to obtain a single correct label with no synonyms (Hanley, 2011), the larger set of possible phonologies (James & Fogler, 2007), and the frequency of word use in daily discourse (Kittredge et al., 2008). Inspired by these results in human cognition, we seek to investigate the impact of personal names from different parts of the world on the linear spatial reasoning ability of large language models - a task that ought to be neutral to the choice of the names themselves.

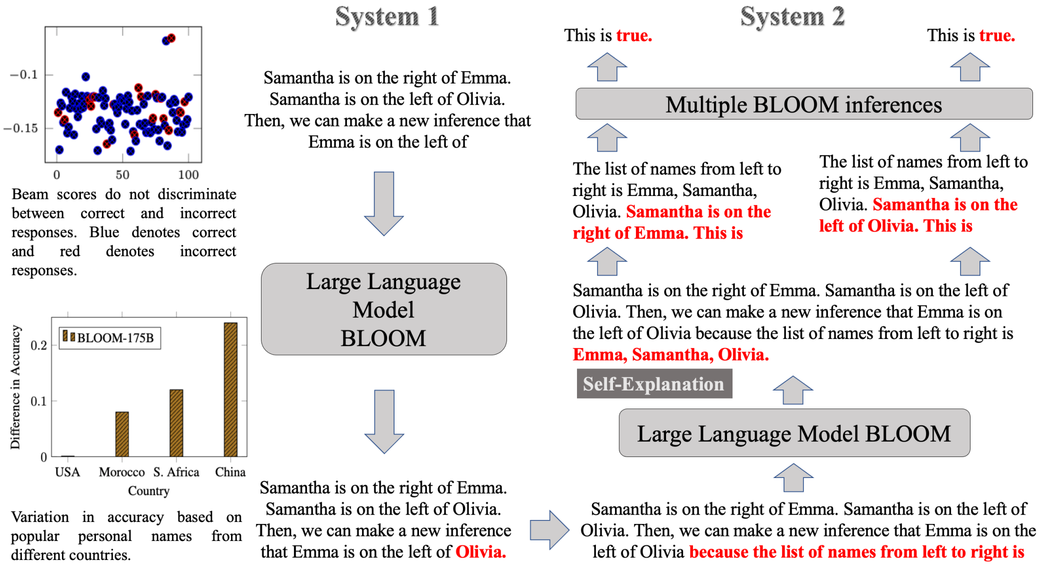
The rise of the large language models promises to revolutionize differentiable approaches to natural language processing. Models such as BERT (Devlin et al., 2018), T5 (Raffel et al., 2020), GPT (Brown et al., 2020), OPT (Zhang et al., 2022), PALM (Chowdhery et al., 2022) and BLOOM (BigScience, 2022) yield results close to the state-of-the-art on popular benchmarks in several language tasks. However, such models are susceptible to adversarial attacks (Wang et al.)

and reduced accuracy on out-of-distribution data (Du et al., 2021). Bias and toxicity evaluations of such models are now standard with a variety of benchmarks (Ousidhoum et al., 2021). Inspired by both the neurolinguistic findings about the curious case of personal names in human cognition and the now well-established bias studies in deep learning, we study personal names as a new source of variations in the performance of large language models even when the underlying reasoning task ought to be independent of the choice of personal names.

2 Summary of Results

We create a scalable spatial reasoning task involving n individuals with different names. In this task, we provide the pairwise spatial relationship between entities to the BLOOM family of models, and then ask the model to predict a new relationship among the entities. Using male names popular in the US on the spatial reasoning task with four individuals, the accuracy of the BLOOM model grows from 0.3 to an impressive 0.83 as the number of parameters increases from 560 million to 146 billion. Thus, an increase in the model size leads to an improvement in the accuracy of the model's accuracy in reasoning about the spatial relationships, and sufficiently large models exhibit very good spatial reasoning capability.

We illustrate this set up as the intuitive System 1 (Kahneman, 2011) in Figure 2 as this is the relatively fast portion of the inference process. We later show that this reasoning is prone to name bias, and consequently, we develop a deliberative System 2 that makes the model generate explanation for its reasoning and self-reflect on whether the inferences are consistent. We show that this deliberative self-reflection can be used to detect errors made by System 1 due to the name bias.



Beam scores do not discriminate between correct and incorrect responses. Blue denotes correct and red denotes incorrect responses.

Variation in accuracy based on popular personal names from different countries.

Samantha is on the right of Emma. Samantha is on the left of Olivia. Then, we can make a new inference that Emma is on the left of

Samantha is on the right of Emma. Samantha is on the left of Olivia. Then, we can make a new inference that Emma is on the left of **Olivia.**

Samantha is on the right of Emma. Samantha is on the left of Olivia. Then, we can make a new inference that Emma is on the left of Olivia **because the list of names from left to right is**

Samantha is on the right of Emma. Samantha is on the left of Olivia. Then, we can make a new inference that Emma is on the left of Olivia because the list of names from left to right is **Emma, Samantha, Olivia.**

The list of names from left to right is Emma, Samantha, Olivia. **Samantha is on the right of Emma. This is**

The list of names from left to right is Emma, Samantha, Olivia. **Samantha is on the left of Olivia. This is**

Figure 2: Overview of our approach. (top left) Beam scores from the largest BLOOM model for correct (blue) and incorrect (red) responses. (bottom left) Variation in response from the largest BLOOM model on personal names from different countries. (center) System 1 BLOOM model for solving the spatial reasoning task. (right) System 2 for obtaining self-explanations from BLOOM models as a spatially ordered list of names and multiple inferences for verifying initial facts on this list of names as a self-introspection step. Predictions from the BLOOM models are shown using red text.

Variation due to choice of proper names: The accuracy of the BLOOM models on this spatial reasoning task varies significantly depending on the source countries of the proper names even though the reasoning task is completely invariant to the choice of the names. The difference in accuracies for popular proper names from USA, Morocco, South Africa and China are shown in Fig 2 (bottom left). When we change the names in the reasoning task using the names common in US, the variance is 0 but it increases to 0.08 for names from Morocco, 0.12 for South Africa, and 0.24 for China.

Poor calibration of BLOOM models on spatial reasoning: We observe that the conditional log probabilities or beam scores of BLOOM predictions are not well-calibrated, and the scores do not discriminate between correct and wrong responses. The beam scores of 100 responses from BLOOM-176B are shown in Fig. 2 (top left). The scores for the correct (incorrect) responses are shown in blue (red). As shown in the figure, the distributions of the two scores can not be separated. Thus, this approach for uncertainty quantification is not sufficient to detect when the model makes wrong predictions due to name bias.

System 2 using self-explanations and self-introspection: In order to better understand the large language model, we seek an explanation from the model by prompting it to list the names of the individuals from left to right. The BLOOM model responds with such an ordering of names, and this list serves as a self-explanation of the model's prediction. This is illustrated in Fig. 2 (right bottom). Then, a number of prompts are generated for the BLOOM model by coupling the generated list of names with the earlier stated facts in the original prompt, and the model is asked to predict whether the pair of facts is consistent (see Fig. 2 (right top). The conformance results are then used to predict if the model response is correct.

We believe that our results show that there is an urgent need to develop training strategies and inference algorithms that can mitigate emergent bias for large language models in such settings, where performance in the underlying task ought to be clearly independent of the source of the bias.

3 Spatial Reasoning in BLOOM Large Language Models

3.1 Spatial Reasoning Tasks

Spatial reasoning has been a topic of sustained interest in representational learning. Qualitative reasoning using spatial cardinal directional representations has been investigated for at least three decades (Frank, 1991; Teresa Escrig & Toledo, 1998). Spatial reasoning for textual data (Weston et al., 2015) using deep learning models (Le et al., 2020) has made such rapid progress that has led to the need for creating new and more challenging benchmarks (Shi et al.).

We create a simple spatial reasoning task where n individuals are located on a straight line. The task specifies the relative location, left or right, of one person with respect to another. Finally, the task requires that we make a new hitherto undeclared inference about the relative position of individuals. We keep the reasoning problem simple because our goal is not to evaluate the limits of the reasoning capability of these large language models, but instead to analyze their bias with respect to the used proper names in defining the reasoning problem.

Algorithm 1: Synthesis of Spatial Problems

```
Input: List of popular names $p_1, p_2, \dots p_n$, Set of left-of
       relations $S_l = \{(p_i, p_j) | p_i$ is on left of $p_j\}$, Set of
       right-of relations
       $S_r = \{(p_i, p_j) | p_i$ is on right of $p_j\}$, Query Pair
       $(p_k, p_l)$
Output: Correct Response $R$
$V \leftarrow \{p_1, p_2, \dots p_n\}, E_l \leftarrow \phi, \quad E_r \leftarrow \phi$   // Graph
for $(p_i, p_j) \in S_l$        // Add left-of relations
do
    $E_l \leftarrow E_l \cup (p_i, p_j), E_r \leftarrow E_r \cup (p_j, p_i)$
for $(p_i, p_j) \in S_r$       // Add right-of relations
do
    $E_r \leftarrow E_r \cup (p_i, p_j), E_l \leftarrow E_l \cup (p_j, p_i)$
$E_l^0 \leftarrow E_l, E_r^0 \leftarrow E_r, i = 0$ while $i = 0$ or $E_l^i \neq E_l^{i-1}$
  // transitive closure
do
    $E_l^{i+1} = E_l^i$ if $(p_i, p_j) \in E_l^i$ and $(p_j, p_k) \in E_l^i$
      then
        $E_l^{i+1} \leftarrow E_l^{i+1} \cup (p_i, p_k)$
    $i \leftarrow i + 1$
$i = 0$ while $i = 0$ or $E_r^i \neq E_r^{i-1}$        // transitive
  closure
do
    $E_r^{i+1} = E_r^i$ if $(p_i, p_j) \in E_r^i$ and $(p_j, p_k) \in E_r^i$
      then
        $E_r^{i+1} = E_r^{i+1} \cup (p_i, p_k)$
    $i \leftarrow i + 1$
if $(p_k, p_l) \in E_l^i$ then
    $R \leftarrow p_k$ is on the left of $p_l$
if $(p_k, p_l) \in E_r^i$ then
    $R \leftarrow p_k$ is on the right of $p_l$
if $(p_k, p_l) \notin E_l^i$ and $(p_k, p_l) \notin E_r^i$ then
    $R \leftarrow$ Relationship between $p_k$ and $p_l$ not known.
```

The challenge problems in our tasks are created algorithmically. Given n individuals and a query pair p_k, p_l, we construct a left-graph $G_l = (V, E_l)$ of n nodes. For every pair of nodes $p_i, p_j \in V$, we add the directed edge (p_i, p_j) to the left-graph if p_i is known to be on the left of p_j according to the stated facts in the generated task. Finally, we compute the transitive closure $G_l^* = (V, E_l^*)$ of the left-graph.

If $(p_k, p_l) \in E_l^*$, the given facts in the generated task are adequate to show that p_k is to left of p_l, we keep this problem in our task and note that the correct response states that p_k is to left of p_l.

Similarly, we create a right-graph where the edges denote that an individual is to the right of another, compute its transitive closure and determine whether the stated facts are enough to conclude that p_k is to the right of p_l. An illustrative automatically generated task is shown below with the correct response from the BLOOM-176B model colored in red.

> John is on the right of David. James is on the left of David. James is on the left of John. Joseph is on the right of James. John is on the left of Joseph. Then, we can make a new inference that David is on the left of Joseph.

3.2 BLOOM Large Language Models

The recent increase in the size of large language models using distributed GPUs has led to impressive performance on diverse tasks, such as finishing sentences (Zellers et al., 2019), commonsense story cloze (Mostafazadeh et al., 2016), physical commonsense reasoning (Bisk et al., 2020), challenging question answering problems (Clark et al., 2018), open book question answering (Mihaylov et al., 2018), Turing tests based on correct word disambiguation (Levesque et al., 2012; Sakaguchi et al., 2021), and more challenging general-purpose language understanding (Wang et al.).

A variety of large language models have been trained on internet-scale corpora of text in multiple languages and code bases. These models include BERT, T5 , GPT, OPT, PALM and BLOOM. A few of these models, such as GPT3, are only accessible via an API while all the trained weights for other models, such as BLOOM, are available to the public at large.

Our experimental studies have focused on BLOOM as it is an open-science open-access model that has been trained using data in multiple languages and is the result of an international effort. BLOOM has been trained on the Jean Zay Public Supercomputer provided by the French government and is readily available to the public (BigScience, 2022). BLOOM is a family of large-language models ranging from 560 million parameters to 176 billion parameters, and provides an effective platform for evaluating the variations in our spatial reasoning task. As illustrated in Fig. 3, BLOOM models have been trained on substantial text from multiple languages and probably multiple countries.

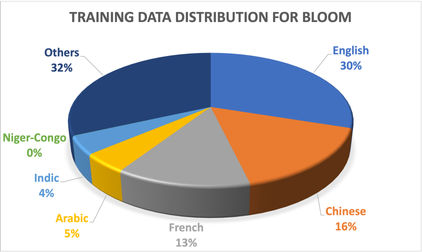


Figure 3: The distribution of languages used to train the BLOOM models.

3.3 Efficacy of BLOOM on Linear Spatial Reasoning Task

We observe that BLOOM is capable of solving the spatial reasoning task with zero-shot prompting in many cases. An example can be seen in the illustration of System 1 in Fig. 2. We consider reasoning tasks with different complexities by changing the number of individuals over which spatial reasoning needs to be performed.

Spatial reasoning over 3 individuals. Figure 4 shows the performance of the BLOOM family of models on our spatial reasoning task for 3 male names drawn from the popular names of different countries. The accuracy of the BLOOM family rsies in a sustained manner from a maximum of 0.52 for the 560 million parameter model to a maximum of 1.0 for the 176 billion parameter model.

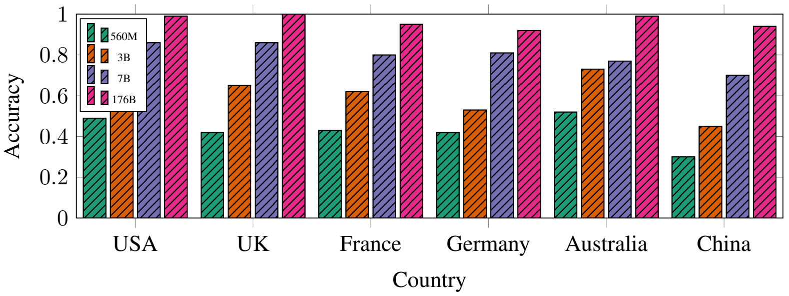


Figure 4: Efficacy for reasoning with 3 male names using 4 different BLOOM models. The BLOOM-176B model has a high accuracy on this task.

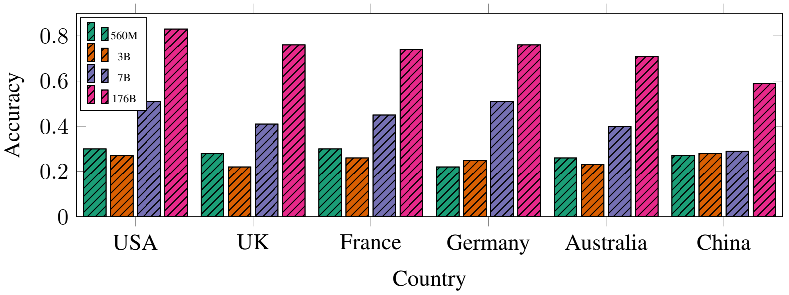


Figure 5: Efficacy for spatial reasoning with 4 male names popular in different parts of the world using multiple BLOOM models.

Spatial reasoning over 4 individuals. Figure 5 shows the performance of BLOOM family of models on the spatial task with 4 male names drawn from the popular names of different countries. The 560 million parameter model has a maximum accuracy of only 0.3, which is below random chance. However, the 176 billion parameter model has an accuracy of 0.83 for the United States. Figure C.1 in the Appendix shows the performance of the family of the models on tasks with 5 individuals.

4 Accuracy Variations due to Country of Personal Names

While the significant variation of the model accuracy on the geographical locations of the names on an unrelated spatial reasoning task is itself a source of concern, we observe that the variation is emergent and it generally increases in magnitude as the models grow in size. We quantitatively evaluate the variation as the difference between the accuracy for a given geographical area and the maximum accuracy for a model of that size. For example, for 4 individuals, names drawn from China have an accuracy of 0.59 while names drawn from USA have an accuracy of 0.83 for the BLOOM-176B model. Hence, the variation for China is the difference of the two accuracies, i.e., 0.24. This high variation indicates relatively low accuracy for the corresponding subpopulation even when the underlying reasoning task is the same and should not depend on the names being used, and hence, indicates bias in the model.

Fig 6 shows a plot of the variation for multiple BLOOM models with four names drawn from popular names in different countries. For names popular in USA, a rise in the size of the model reduces the variation to 0. On the other hand, a rise in the size of the BLOOM models generally exacerbates the variation in accuracy problem for countries, such as China, India, and South Africa.

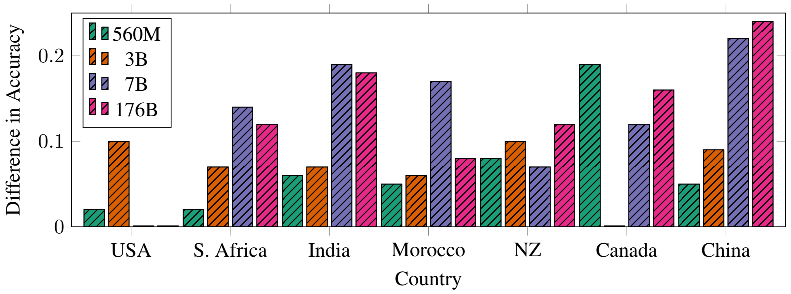


Figure 6: Difference between the best accuracy for a given model and the accuracy on spatial reasoning with 4 individuals for popular male names in a country. As the BLOOM models become larger, the variation for names from China, India, and South Africa increases.

We also analyze the variation on the more challenging spatial reasoning task with 5 individuals. Figure 7 shows that the BLOOM-176B model shows a variation of 0.03 for the United States. However, the variation for China, India, and South Africa becomes as high as 0.14, 0.17, 0.10

These variations may be explained using various reasons: (i) different term-frequency in training data, (ii) lack of descriptive value of a personal name that can be learned using a large text corpora, and (iii) the potentially larger set of possible phonologies in different languages. However, as classical algorithms are replaced by increasingly complex and often proprietary and closed large language models, the variation from proper names needs to be addressed to avoid biased analyses.

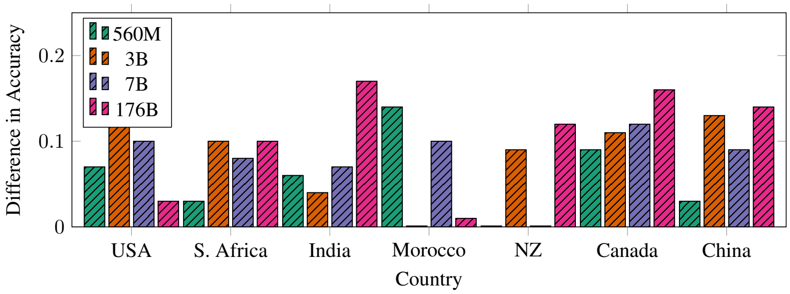


Figure 7: Difference between the best accuracy for a given model and the accuracy for popular male names drawn from a geographical region. This spatial reasoning task uses 5 individuals.

5 Accuracy Variations due to Gender of Personal Names

BLOOM models clearly demonstrate variations in their accuracy on our spatial reasoning task based on the geographical source of the names being used in the task. Using popular female names for the same countries, we studied how the variation in model accuracy changes. Further, we investigated if there is a substantial difference in accuracy between popular male and female names from the same country. The variation in the accuracy of the model on the spatial reasoning task for four female names from different countries is shown in Figure 8. Again, a large variation is observed for the BLOOM-176B model for female names from China. However, the qualitative nature of the plot is now different and the difference in accuracies do not always rise with increase in model size.

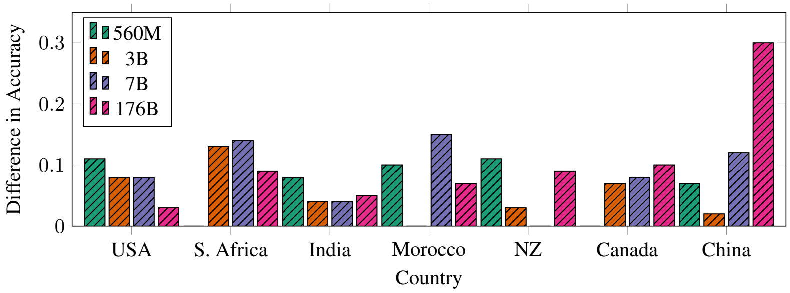


Figure 8: Difference between the best accuracy for a given model and the accuracy on spatial reasoning with 4 individuals for popular female names drawn from a geographical region.

Fig. 9 plots the difference in the accuracy of male and female names for the spatial reasoning task with 5 individuals. A positive value indicates that the model performed better on female names, while a negative value shows that the model did better on male names. We observe that the names of neither of the genders produce consistently high accuracy either across countries or across models.

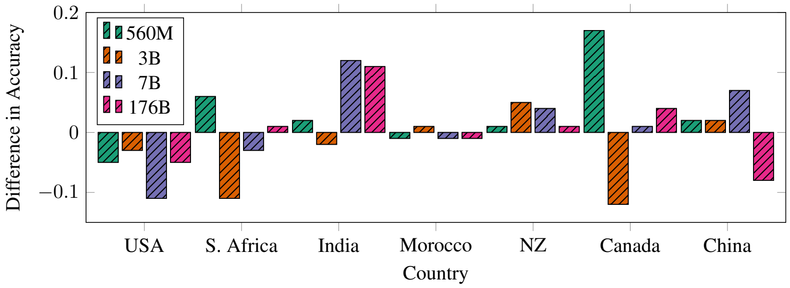


Figure 9: Difference between the accuracies for male and female names with 5 individuals for popular names drawn from a geographical region.

We conjecture that this observation is the result of two competing forces. First, female names may have a larger or smaller phonology in different countries, and this may lead to different variations in term frequency based on gender in different countries. Second, female names may be less frequent in some corpora as opposed to male names, such as old telephone directories. This may lead to different training regimes for models of different sizes learning representations of these names with different degrees of success. Future training regimes for large language models need to explicitly guard against gender bias.

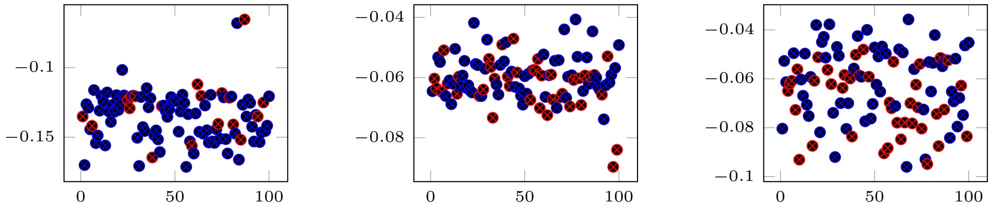
6 Uncalibrated Model Scores and Detecting Errors using Model Self-Explanations

Since the BLOOM models produce significantly different accuracies on the spatial reasoning task for different choice of names, it becomes even more important to understand if the model is predicting a correct response on a given query. A natural approach to quantify uncertainty of a model is to use the log probability of the response predicted by the model and threshold it to decide if a response is reliable or not. We first show that the BLOOM models are not well-calibrated and produce similar conditional log probabilities for both correct and incorrect responses. Then, we show that the BLOOM models can produce textual self-explanations for a problem, which can then be audited by repeatedly querying the BLOOM models with related sub-problems to decide if a response is consistent and likely to be correct.

6.1 Uncalibrated Model Score Predictions

The BLOOM models, like other large language models, compute the beam score or the conditional log probability of the response being predicted. In a well-calibrated model, a correct output will be associated with a high probability and an incorrect output will correspond to a relatively lower probability. This will allow the model to be deployed in a practical setting.

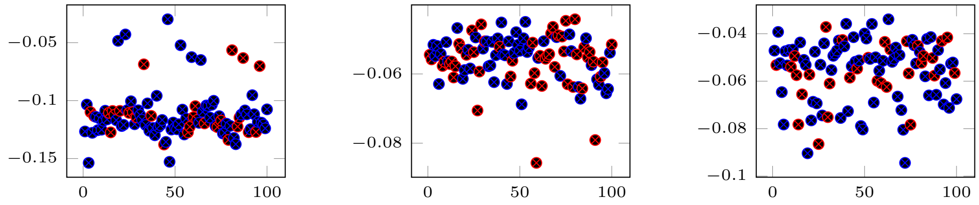
Fig 10 shows the beam score or log probability corresponding to a typical response from the BLOOM-176B model on male names from the US, China and South Africa. The horizontal axis represents the index of the query and the vertical axis represents the beam score or the log probability of the response.



(a) USA (b) China (c) South Africa

Figure 10: The beam score or conditional log probabilities for the top response from the model using male names. The X axis represents the index of the query and the Y axis represent the score. Red color represents an incorrect response, while a blue color represents a correct response.

It is clear that the correct and incorrect responses both produce similar log probability values. The correlation coefficient between the scores or the log probabilities and the correctness of the response is only 0.157 for names from the USA. We observe similar qualitative and quantitative results for the male and female names from other countries. Fig. 11 shows the variation in log probability values for correct and incorrect responses for female names from the US, China and South Africa.



(a) USA (b) China (c) South Africa

Figure 11: The beam score or conditional log probabilities for the top response from the model when using female names. Red color represents an incorrect response, while a blue color represents a correct response.

Based on the results observed in Fig. 10 and 11, the conditional log probabilities do not correlate with an intuitive sense of confidence. Hence, we seek to create a System 2 deliberative component for large language models on this task that can help detect when the predictions from the BLOOM models are likely to be wrong.

6.2 Error Detection using Model Self-Explanations

In the absence of effective model calibration for the spatial reasoning task in BLOOM large language models, we create a new approach using model self-explanations for error detection in such models. Our approach is presented in Algorithm 2.

Algorithm 2: System 2 for Spatial Reasoning in Large Language Models

```
Input: Facts: $F_1, F_2, \ldots F_n$, Query: $Q$, Model: $\mathcal{M}$
Output: Yes, if response is correct. Otherwise, No.
1  $P_{main} \leftarrow F_1 + F_2, \cdots + F_n + Q$        // Concatenate the facts and query to create prompt
2  $R_{main} \leftarrow \mathcal{M}(P_{main})$        // Query the model
   /* Query the model again to determine its internal explanation as a sequence of the names
      from left to right */
3  $P_{explain} \leftarrow P_{main} + R_{main} +$ " because the list of names from left to right is "
4  $R_{explain} \leftarrow \mathcal{M}(P_{explain})$        // Query the model
   /* Verify if the linear list of names is consistent with each known fact */
5  for $i \in 1, \ldots, n$ do
6      $P^i_{verify} \leftarrow$ "the list of names from left to right is " $+ R_{explain} + F_i +$ "This is "
         // Verifying fact i
7      $R^i_{verify} \leftarrow \mathcal{M}(P^i_{verify})$
8      if $R^i_{verify}$ is True for all $i \in 1, \ldots, n$ then
9          Return Yes
10     Return No.
```

First, our approach uses a zero shot prompt P_{main} and seeks a predictive response R_{main} to the spatial reasoning task from the model based on the n facts F_1 through F_n. This is illustrated in lines 1-2 of the algorithm and in the example below.

> John is on the left of Robert. Thomas is on the left of John. John is on the left of Michael. Thomas is on the left of Robert. Robert is on the left of Michael. Then, we can make a new inference that Thomas is on the left of

The BLOOM model responds by completing the query with a new inferred fact R_{main}, such as the following. The response from the model is highlighted using red text.

> Then, we can make a new inference that Thomas is on the left of Michael.

Second, our approach creates a prompt $P_{explain}$ using the answer to the first prompt from the model and asks the reason why the model believes in the response. The explanation for the earlier response from the model is obtained in lines 3-4 of the algorithm by using a zero shot prompt of the following form:

> John is on the left of Robert. Thomas is on the left of John. John is on the left of Michael. Thomas is on the left of Robert. Robert is on the left of Michael.
> Then, we can make a new inference that Thomas is on the left of Michael because the list of names from left to right is

The BLOOM model responds by completing the query with a linear ordering of the names $R_{explain}$ in the earlier prompt $P_{explain}$. This enables us to gain an insight into the model's view of the world in terms of its perceived linear ordering of names.

> Then, we can make a new inference that Thomas is on the left of Michael because the list of names from left to right is Thomas, John, Robert, Michael.

The second prompt is indeed designed to elicit a structured textual *explanation* from the model that reflects the model's internal view of the stated facts using the vocabulary of the earlier prompt. Once we have the linear ordering of names from the model, we generate n different verification prompts P^i_{verify} for the model consisting of this linear ordering $R_{explain}$ and one of the known facts F_i. These prompts are expected to produce either true or false, and a representative example of such reasoning is used to inductively bias the model.

> The list of names from left to right is John, James, Thomas, Joseph. Then, James is on the right of Joseph. This is false.
> The list of names from left to right is Thomas, John, Robert, Michael. John is on the left of Robert. This is

The model responds with a decision R^i_{verify} as True if the linear ordering is consistent with each of the facts stated in the original task prompt. Otherwise, the model responds False.

> The list of names from left to right is Thomas, John, Robert, Michael. John is on the left of Robert. This is true.

The model is also able to identify scenarios where its linear ordering explanation is inconsistent with the original facts in the prompt.

> The list of names from left to right is David, Charles, William, James. William is on the left of Charles. This is false.

Our self-explanation based approach has two interesting properties. First, the verification of the text explanation is only an $O(n)$ linear time operation while computing the right ordering of n individuals requires looking at $O(n!)$ permutations. Hence, the text explanation produced by the second prompt is crucial in effectively verifying the consistency of the predicted model response. Second, our approach only verifies that the internal view of the model as documented in the textual explanation produced by the predicted response to the second prompt is consistent with the original set of facts; it does not guarantee that the model's response is indeed correct or that the model's reasoning during verification is correct. However, in practice, we observe that a significant fraction of responses deemed correct by our approach are indeed correct, as shown in Figure 12.

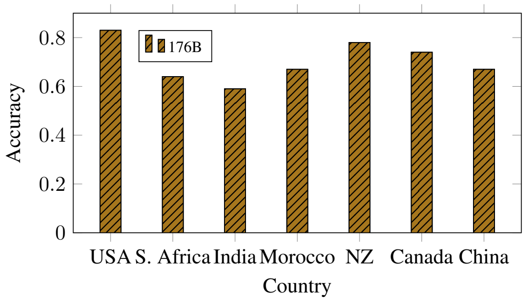


Figure 12: Fraction of predictions labeled as correct by our System 2 approach in Algorithm 2 that are actually correct.

7 Conclusions and Future Work

While there has been significant work on evaluating explicit bias in the predictions from large language models, we propose the automated synthesis of a spatial reasoning task that ought to be independent of the choice of personal names. However, we find that the BLOOM large language models are substantially influenced by the choice of personal names on this task.

Further, the beam score or conditional log probabilities of predicted tokens have very poor correlation with the correctness of the model response on this task, and cannot be relied upon as a metric for uncertainty quantification. We then suggest a new approach inspired by Kahneman's System 2 model (Kahneman, 2011) that first seeks a textual explanation from the model in the form of a linear ordering of the individuals, and then repeatedly queries the model to verify if this linear ordering is consistent with the stated facts in the spatial reasoning task. We find that this approach performs better than the log probabilities or beam scores from BLOOM models in detecting correct responses.

The analysis and observations in this paper open up several exciting directions for future research. First, it will be interesting to compare the performance of BLOOM with other large language models that were not designed by an international team of researchers. Since the trained weights of many of the state-of-the-art models are not available to the public at large, such proprietary models need similar internal auditing. Second, our work has focused on personal names, and it may be interesting to pursue a wider study using different proper nouns, such as places and brand names. Third, our model self-explanation and System 2 design using multiple model self-introspection steps may be applied to other problems that can yield a succinct textual representation of the perceived view of the world from the model.

Ethics: Our approach brings to light a potential concern with BLOOM and possibly other large language models where their accuracy is affected by the choice of personal names, even on tasks that are completely independent of the choice of names. By increasing the awareness of potential bias of these large language models, we hope to motivate research into mitigation techniques and make practitioners aware about challenges in social deployment of these models. Our self-explanation method is one such mitigation approach.

REFERENCES

BigScience. BigScience language open-science open-access multilingual (BLOOM) language model. Online, May 2022.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *AAAI*, 34(05):7432–7439, April 2020.

Serge Bredart, Tim Brennen, and Tim Valentine. *The cognitive psychology of proper names*. Routledge, 2002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Others. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*, 33:1877–1901, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. April 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. March 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

Du, Gan, and Isola. Curious representation learning for embodied intelligence. *Proceedings of the IEEE/CVF*, 2021.

Kethera A Fogler and Lori E James. Charlie brown versus snow white: the effects of descriptiveness on young and older adults' retrieval of proper names. *J. Gerontol. B Psychol. Sci. Soc. Sci.*, 62 (4):P201–7, July 2007.

Andrew U Frank. Qualitative spatial reasoning with cardinal directions. In *7. Österreichische Artificial-Intelligence-Tagung / Seventh Austrian Conference on Artificial Intelligence*, pp. 157–167. Springer Berlin Heidelberg, 1991.

J Richard Hanley. Why are names of people associated with so many phonological retrieval failures? *Psychon. Bull. Rev.*, 18(3):612–617, June 2011.

Lori E James and Kethera A Fogler. Meeting mr davis vs mr davin: effects of name frequency on learning proper names in young and older adults. *Memory*, 15(4):366–374, May 2007.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

Audrey K Kittredge, Gary S Dell, Jay Verkuilen, and Myrna F Schwartz. Where is the effect of frequency in word production? insights from aphasic picture-naming errors. *Cogn. Neuropsychol.*, 25(4):463–492, June 2008.

Hung Le, Truyen Tran, and Svetha Venkatesh. Self-Attentive associative memory. In Hal Daumé Iii and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5682–5691. PMLR, 2020.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, May 2012.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. September 2018.

Nasrin Mostafazadeh, Lucy Vanderwende, Wen-Tau Yih, Pushmeet Kohli, and James Allen. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 24–29, Berlin, Germany, August 2016. Association for Computational Linguistics.

Nedjma Djouhra Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, and Others. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021.

Shi, Zhang, and Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. *Association for the Advancement of Artificial*.

M Teresa Escrig and Francisco Toledo. *Qualitative Spatial Reasoning: Theory and Practice : Application to Robot Navigation*. IOS Press, 1998.

Wang, Pruksachatkun, Nangia, and others. Superglue: A stickier benchmark for general-purpose language understanding systems. *Adv. Neural Inf. Process. Syst.*

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-Complete question answering: A set of prerequisite toy tasks. February 2015.

A W Young, D C Hay, and A W Ellis. The faces that launched a thousand slips: everyday difficulties and errors in recognizing people. *Br. J. Psychol.*, 76 (Pt 4):495–523, November 1985.

Andrew W Young, Andrew W Ellis, and Brenda M Flude. Accessing stored information about familiar people. *Psychol. Res.*, 50(2):111–115, September 1988.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? May 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. May 2022.