A Benchmark for Reasoning with Spatial Prepositions

Iulia-Maria Comșa  
Google DeepMind  
iuliaomesa@google.com

Srini Narayanan  
Google DeepMind  
srinin@google.com

Abstract

Spatial reasoning is a fundamental building block of human cognition, used in representing, grounding, and reasoning about physical and abstract concepts. We propose a novel benchmark focused on assessing inferential properties of statements with spatial prepositions. The benchmark includes original datasets in English and Romanian and aims to probe the limits of reasoning about spatial relations in large language models. We use prompt engineering to study the performance of two families of large language models, PaLM and GPT-3, on our benchmark. Our results show considerable variability in the performance of smaller and larger models, as well as across prompts and languages. However, none of the models reaches human performance.

1 Introduction

Large language models (LLMs) are becoming increasingly human-like in their performance on many tasks, but are still not on par with more advanced aspects of human cognition (Choi, 2022). On the other hand, they are showing emerging capabilities that were previously thought beyond their limits, such as grounding conceptual spaces (Patel and Pavlick, 2022). Currently, many questions are open regarding the limits of reasoning in LLMs and how they compare to humans in cognitive domains that require a deeper understanding of the world.

One such domain is spatial reasoning, which is a fundamental part of human cognition (Regier, 1996; Herskovits, 2009; Gärdensfors, 2014). This type of reasoning is relevant not only for the representation, prediction and manipulation of physical objects, but also for representing and performing inferences with abstract concepts. This is reflected in common uses of spatial prepositions, which traditionally indicate location, but are also used to refer to abstract states, forces or goals. For example, one can be “in Paris” or “under a tree” (physical locations), but one can also be “in trouble” or “under sedation” (abstract concepts).

Given their lack of embodied spatial experience and the scarcity of commonsense knowledge in training data (Gordon and Van Durme, 2013), we hypothesise that LLMs have difficulties reasoning about physical and abstract spatial relations.

We investigate this using a novel benchmark for assessing inferences on sentences containing spatial prepositions. The sentences are designed to be easy for humans, but non-trivial for models that cannot differentiate between uses of prepositions with different concepts. Our task has similarities with other NLI tasks (Bowman et al., 2015).

This paper makes the following contributions:

- We propose a novel benchmark, available in English and Romanian, to probe a model’s ability to reason with spatial prepositions in physical and abstract domain, through compositional statements.
- We assess two families of large language models, PaLM (Chowdhery et al., 2022) and GPT-3 (Brown et al., 2020) and compare them each other and against human performance on the benchmark. We find that performance varies considerably with model size, prompt setup and language. However, none of the models reaches human performance.

2 Related Work

To investigate commonsense spatial reasoning, Liu et al. (2022) introduced a benchmark focused on assessing the relative size of objects, as well as positional relationships between humans and objects during various actions. Yatskar et al. (2016) extracted a dataset of commonsense spatial relationships from a large corpus where this information appears implicitly. Weston et al. (2015) proposed a set of toy tasks for questions answering, including positional reasoning, while Mirzaee et al. (2021) introduced SpartQA, a dataset of challenging textual
First premise | Second premise | Potential conclusion | Holds?
---|---|---|---
John is in the crib | the crib is in the living room | John is in the living room | ✓
John is in the newspaper | the newspaper is in the kitchen | John is in the kitchen | X

Table 1: Examples showcasing our benchmark on reasoning with spatial prepositions. Each example consists of two premises and a conclusion. The composition of the premises can be transitive (the conclusion holds) or intransitive (the conclusion does not hold). Similar examples are present in the Romanian version of the dataset.

In contrast to these studies, our benchmark proposes the additional challenge of using spatial prepositions to refer to abstract concepts in addition to physical relationships. Reasoning with metaphorical and literal statements has been previously studied (Comsa et al., 2022), but here we focus specifically on spatial prepositions.

3 Dataset

We create small, manually-curated datasets, intended to be used as a benchmark, and not for training purposes. Each dataset consists of 400 class-balanced items. As illustrated in Table 1, each item consists of:

- **premise1**: “X is \[\text{prep}_1\] Y”
- **premise2**: “Y is \[\text{prep}_2\] Z”
- **conclusion**: “X is \[\text{prep}_3\] Z”

where \(\text{prep}\) is a spatial preposition such as “in” or “on” and \(\text{prep}_3\) is one of \{\(\text{prep}_1\), \(\text{prep}_2\)\}. Given the premises, the conclusion may or may not hold.

In the case of *congruent* compositions, the conclusion holds, typically indicating a similar type of spatial relationship. For example, if “John is in the crib” and “the crib is in the living room”, the conclusion “John is in the living room” holds.

On the other hand, in the *incongruent* compositions, the spatial prepositions in each premise refers to a different type of relation, such as through a conceptual metaphor, and the conclusion does not hold. However, the items are designed such that without a deep understanding of the commonsense semantics of the spatial prepositions, a mistaken interpretation is possible, leading to the false impression that the conclusion holds. For example, if “John is in the newspaper” and “the newspaper is in the kitchen”, the conclusion “John is in the kitchen” does not hold. In this example, the spatial preposition “in” is used differently in the two premises: in the first premise, it refers to an abstract concept (inclusion as content in a newspaper), while in the second premise it refers to a physical location. Hence, in this example, combining the premises does not validate the conclusion.

The items are class-balanced: for every congruent item that uses prepositions \(\{\text{prep}_1, \text{prep}_2, \text{prep}_3\}\) there is an incongruent item containing the same prepositions in sequence.

We release datasets in English and in Romanian. For both languages, each item was created by a native or a proficient speaker of the language, and always independently verified by another native speaker. In the process of creating items, we aimed to cover common cases for each chosen spatial preposition in order to create a representative sample of spatial preposition semantics. The creation of items was assisted by standard dictionaries with usage examples for each preposition. For a discussion on the limitations of the data generation process, please refer to Section 7.

In English, we use the spatial prepositions “in”, “at”, “on”, “with”, “under”, “above” and “behind”. In Romanian, we use their equivalents “în”, “la”, “pe”, “cu”, and “sub”, respectively\(^2\). The use of prepositions is different in the two languages and hence the datasets are not direct translations of each other, but reflect the semantics of each language. The distribution of prepositions is shown in Table 2.

To validate the benchmark, we asked English-speaking and Romanian-speaking adults to answer dataset questions of the form “if \{\text{premise1}\} and \{\text{premise2}\}, does that imply that \{\text{conclusion}\}?” with “yes” or “no”. The respondents were told that the aim was to collect a set of commonsense re-

---

\(^2\)In Romanian, the preposition “în” takes the form “într-un” or “într-o”, when followed by an indefinite masculine or feminine noun, respectively. We do not use the Romanian equivalent of the English “above” and “behind” because they are used more seldom in combinations of interest for this task.
Table 2: The number of occurrences of each preposition in our dataset, alongside the accuracy (in percentage) of humans and LLMs on items containing each preposition.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>above</td>
<td>186</td>
<td>52.3</td>
<td>62.9</td>
<td>72.8</td>
<td>80.3</td>
<td>88.5</td>
<td>49.2</td>
<td>52.2</td>
<td>51.1</td>
<td>75.8</td>
<td>64.2</td>
<td>94.5</td>
</tr>
<tr>
<td>at</td>
<td>146</td>
<td>51.8</td>
<td>68.0</td>
<td>71.7</td>
<td>85.4</td>
<td>88.1</td>
<td>51.6</td>
<td>52.7</td>
<td>53.7</td>
<td>83.4</td>
<td>66.9</td>
<td>92.6</td>
</tr>
<tr>
<td>behind</td>
<td>148</td>
<td>54.7</td>
<td>59.2</td>
<td>68.0</td>
<td>76.4</td>
<td>70.9</td>
<td>52.5</td>
<td>51.1</td>
<td>50.7</td>
<td>76.1</td>
<td>62.1</td>
<td>89.8</td>
</tr>
<tr>
<td>in</td>
<td>250</td>
<td>56.5</td>
<td>72.3</td>
<td>75.9</td>
<td>89.1</td>
<td>86.4</td>
<td>51.2</td>
<td>54.0</td>
<td>51.6</td>
<td>88.2</td>
<td>68.9</td>
<td>96.2</td>
</tr>
<tr>
<td>on</td>
<td>228</td>
<td>52.6</td>
<td>69.0</td>
<td>70.0</td>
<td>82.6</td>
<td>86.1</td>
<td>51.3</td>
<td>55.1</td>
<td>51.0</td>
<td>81.5</td>
<td>66.1</td>
<td>91.1</td>
</tr>
<tr>
<td>under</td>
<td>202</td>
<td>53.8</td>
<td>60.2</td>
<td>65.5</td>
<td>79.9</td>
<td>80.9</td>
<td>50.0</td>
<td>53.5</td>
<td>45.2</td>
<td>75.0</td>
<td>62.3</td>
<td>94.7</td>
</tr>
<tr>
<td>with</td>
<td>40</td>
<td>52.5</td>
<td>68.3</td>
<td>69.2</td>
<td>89.2</td>
<td>90.0</td>
<td>53.8</td>
<td>56.7</td>
<td>50.0</td>
<td>85.0</td>
<td>68.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

We evaluated the performance of PaLM (Chowdhery et al., 2022) in different sizes: 540b, 62b (the original model, as well as the model trained to 1.3T tokens as explained in their Appendix F) and 8b, as well as Flan-PaLM-540b (Chung et al., 2022). We also evaluated GPT-3 (Brown et al., 2020) Ada (text-ada-001), Babbage (text-babbage-001), Curie (text-curie-001) and DaVinci (text-davinci-003).

We prompted the models with questions of the form “if {premise1} and {premise2}, does that imply that {conclusion}?” We tested the LLMs with 0-shot, 1-shot and 5-shot prompts. In few-shot settings, each example was prefixed with 1 or 5 different randomly-selected examples from the dataset, each followed by its correct answer (“yes” or “no”). We assessed LLMs in a binary-choice setup of the benchmark. The models were asked to score the strings “yes” and “no” (and their Romanian equivalents) given as candidate continuations to the above prompt. An example was labelled as correct if the log likelihood score of the correct continuation string was higher than the log likelihood score of the incorrect continuation.

To mitigate prompt sensitivity (Lu et al., 2022; Cao et al., 2021), we used multiple prompt variations, as detailed in Appendix A. We report the best prompt performance for each model and setup. For each best prompt, we obtained confidence intervals by randomly sampling sets of 20 responses, similarly to the format of the humans responses.

As a baseline, we ran the same experiment using only the conclusion as a prompt, in the form: “{conclusion}?”. This can probe whether the performance might be explained by the likelihood of the conclusion only. We report the results for the highest-scoring baseline value across all models.

As an alternative to the binary-choice setup, our benchmark can also be used in a generative setting. This can be useful for assessing LLMs for open-ended or conversational applications. To illustrate this use of the benchmark, we performed a generative assessment of the largest model, PaLM-540b. The setup was identical to the above, except that the model was asked to generate 10 tokens in response to the given prompt, and responses were scored accordingly (see Appendix B for details).

An additional experiment involving the negation of congruent sentences is presented in Appendix C.

5 Results

As shown in Table 3, human accuracy was 93.51% for English and 92.6% for Romanian. LLM performance varied considerably across models, with the number of shots and across languages. The highest LLM accuracies were recorded from PaLM-540b with 5-shot prompting at 85.67% in English, and Flan-PaLM-540b with 5-shot prompting at 84.83%
for Romanian. We also observed strong performance in the 5-shot generative setting, at 87.67% for English and 80% for Romanian.

The largest models (PaLM-540b, Flan-PaLM-540b and GPT-3-DaVinci) performed consistently better than the smaller models. Interestingly, PaLM-540b greatly benefited from 5-shot prompting in Romanian, whereas GPT-3-DaVinci showed slightly worse results with more shots.

Smaller GPT-3 models and PaLM-8b almost always performed close to chance level, whereas the other PaLM models benefited from few-shot prompts in English. We observed that some of the smaller models had consistent class bias, consistently answering “no” and thus scoring predominantly correctly on incongruent items only.

The performance of the models on the baseline examples suggests that a small part of the performance can be explained by the likelihood of the conclusion only, and not just reasoning capacity. However, as in all baseline cases the performance does not approach that of the original examples, the likelihood of the conclusion is not sufficient to explain the performance of the models.

The overall performance was better for the English than for the Romanian dataset particularly in the case of PaLM models, including in the generative experiment. We expected this gap, in line with results from other multilingual tasks (Dumitrescu et al., 2021; Artetxe et al., 2020).

As shown in Table 2, performance varied across models for individual prepositions. There was only partial alignment in preposition accuracies between humans and LLMs. Humans performed best on items containing “with” and “in” in English, and “în” and “la” in Romanian, while performing worst on “behind” in English, which partially reflects the performance averaged across models. In contrast, the models made relatively more mistakes on “under”. While Flan-PaLM-540b had better overall accuracy, its performance on “in” was slightly lower compared to the other larger models, and it
had more relative difficulty with “behind”. Meanwhile, GPT-3-DaVinci had more relative difficulty with “above” and “under”. Other prepositions show less clear agreement across models. Given these results, the distribution of prepositions in the dataset should be considered a factor that influences the reported accuracies.

6 Conclusions

We have introduced a novel and challenging benchmark for commonsense reasoning with spatial prepositions in multiple conceptual domains, and provided initial results on two families of LLMs. The task is part of our efforts towards investigating the limits of foundational reasoning in LLMs.

Our task captures highly variable performance scores across LLMs, with smaller LLMs typically performing at chance level and larger models approaching, but not reaching, human performance. The range of performance on this task makes it suitable as a checkpoint in examining trade-offs in models size and performance, particularly when complex or abstract reasoning is involved. We hope to encourage the development of more tasks that capture the building blocks of reasoning in LLMs.

7 Limitations

Our benchmark aims to provide a representative assessment for the capability of LLMs to operate across different meanings of spatial prepositions. We used a wide range of examples that cover an exemplary but not exhaustive range of spatial language; it was not in the scope of the study to capture all prepositions or constructions that indicate spatiality, but rather a representative set.

Due to the richness and uniqueness of the many expressions involving spatial prepositions, a rigorous description of the lexical meanings of prepositions has been a long-standing challenge in linguistics (Herskovits, 2009) and is beyond the scope of this study. Nevertheless, for reference, we provide in Table 4 an estimation of preposition frequency in a Wikipedia corpus, alongside the number of entries as determined from a standard dictionary: Cambridge Dictionary (https://dictionary.cambridge.org/) and Dexonline (https://dexonline.ro/).

Underrepresented in our dataset. Future extensions to our dataset could introduce more flexibility in the form of the items and allow for additional types of constructions.

Finally, prepositions cue space and concepts differently across languages. As there is no bijective correspondence of spatial prepositions across languages, an absolute performance comparison between languages is not possible with the approach proposed here. We are investigating a more geometric grounding approach by training multimodal classifiers similar to Patel and Pavlick (2022) which would sharpen the cross-linguistic comparison in geometric space.

In spite of these limitations, we believe that our benchmark can provides an insightful measure regarding the ability of LLMs to handle spatial prepositions used in different semantic registers, and a challenge with good scaling across model size and task setup.

8 Ethical Risks

The authors manually ensured that the items included in the proposed datasets do not contain offensive, unfair or otherwise unethical content. Prior to release, the datasets were seen by at least 3 other NLP researchers, who did not raise any concerns regarding the content.

Acknowledgements

We thank Julian Eisenschlos, Yasemin Altun, Fernando Pereira, as well as our anonymous reviewers and meta-reviewers for valuable feedback.
References


Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In Proceedings of the 60th Annual Meeting of the Association for
We made small variations to these four prompts (e.g. by adding quotes of different types around the premises and conclusions, and spaces or delimiters at the end of the prompt) to obtain up to 48 prompts.

For an initial assessment of the performance differences among different prompts, we performed two-sample Kolmogorov-Smirnov tests on the performance of the prompts on the original three PaLM models. For the baseline prompts, only 0.41% of all pairwise prompt combinations had a p-value smaller than 0.05 before correction for multiple comparisons. For the task questions, we found an overlap of 6.96%. The small overlap between prompt performance suggests that the models are highly sensitive to prompts.

### B Appendix: Generative Experiment

The generative experiment is intended to illustrate an alternative, open-ended way in which our benchmark can be used to explore LLM responses.

A preliminary analysis of the responses to the benchmark questions revealed that most answers consisted of either “yes” or “no”, or an undetermined response, such as generating a new similar question without providing an answer. Most times, we did not find that the response attempted to meaningfully reason through the question; this was expected because the questions do not lend themselves to reasoning steps.

Based on the preliminary inspection of the generated responses, we defined the following scoring scheme. We labelled a response as correct if the correct label (“yes” or “no”) appeared among the generated tokens and the incorrect label did not. If none or both labels were present in the response, it was labelled as ambiguous. Otherwise, if only the incorrect label appeared in the response, we defined the following scoring scheme.

A preliminary analysis of the responses to the benchmark questions revealed that most answers consisted of either “yes” or “no”, or an undetermined response, such as generating a new similar question without providing an answer. Most times, we did not find that the response attempted to meaningfully reason through the question; this was expected because the questions do not lend themselves to reasoning steps.

Based on the preliminary inspection of the generated responses, we defined the following scoring scheme. We labelled a response as correct if the correct label (“yes” or “no”) appeared among the generated tokens and the incorrect label did not. If none or both labels were present in the response, it was labelled as ambiguous. Otherwise, if only the incorrect label appeared in the response, we defined the following scoring scheme.

A preliminary analysis of the responses to the benchmark questions revealed that most answers consisted of either “yes” or “no”, or an undetermined response, such as generating a new similar question without providing an answer. Most times, we did not find that the response attempted to meaningfully reason through the question; this was expected because the questions do not lend themselves to reasoning steps.
<table>
<thead>
<tr>
<th>Model</th>
<th>English 0-shot</th>
<th>English 1-shot</th>
<th>English 5-shot</th>
<th>Romanian 0-shot</th>
<th>Romanian 1-shot</th>
<th>Romanian 5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaLM-8b</td>
<td>69.83</td>
<td>74.06</td>
<td>66.08</td>
<td>70.00</td>
<td>71.25</td>
<td>64.00</td>
</tr>
<tr>
<td></td>
<td>[65.1-74.3]</td>
<td>[69.5-78.3]</td>
<td>[61.2-70.7]</td>
<td>[65.2-74.5]</td>
<td>[66.5-75.6]</td>
<td>[59.1-68.7]</td>
</tr>
<tr>
<td>PaLM-62b</td>
<td>59.10</td>
<td>64.59</td>
<td>80.05</td>
<td>50.75</td>
<td>57.00</td>
<td>53.25</td>
</tr>
<tr>
<td></td>
<td>[54.1-64.0]</td>
<td>[59.7-69.3]</td>
<td>[75.8-83.9]</td>
<td>[45.7-55.8]</td>
<td>[52.0-61.9]</td>
<td>[48.2-58.2]</td>
</tr>
<tr>
<td>PaLM-62b-1.3</td>
<td>60.10</td>
<td>78.30</td>
<td>86.03</td>
<td>72.00</td>
<td>66.50</td>
<td>75.00</td>
</tr>
<tr>
<td></td>
<td>[55.1-64.9]</td>
<td>[73.9-82.2]</td>
<td>[82.3-89.3]</td>
<td>[67.3-76.3]</td>
<td>[61.6-71.1]</td>
<td>[70.5-79.2]</td>
</tr>
<tr>
<td>PaLM-540b</td>
<td>80.30</td>
<td>86.53</td>
<td>92.27</td>
<td>65.00</td>
<td>77.75</td>
<td>89.50</td>
</tr>
<tr>
<td></td>
<td>[76.1-84.1]</td>
<td>[82.8-89.7]</td>
<td>[89.2-94.7]</td>
<td>[60.1-69.7]</td>
<td>[73.4-81.7]</td>
<td>[86.1-92.3]</td>
</tr>
<tr>
<td>Flan-PaLM-540b</td>
<td>97.01</td>
<td>97.26</td>
<td>96.51</td>
<td>95.25</td>
<td>99.00</td>
<td>99.75</td>
</tr>
<tr>
<td></td>
<td>[94.8-98.4]</td>
<td>[95.1-98.6]</td>
<td>[94.2-98.1]</td>
<td>[92.7-97.1]</td>
<td>[97.5-99.7]</td>
<td>[98.6-100.0]</td>
</tr>
</tbody>
</table>

Table 5: Performance of LLMs on the negated congruent sentences experiment, described in Appendix C.

their negation only. In negated form, sentences of the form “If John is in the crib and the crib is in the living room, does that imply that John is in the living room?” became “If John is in the crib and the crib is in the living room, does that imply that John is not in the living room?”. This dataset is class-balanced, as the answer for the congruent sentences is always “yes”, and the answer to their negation is always “no”.

The results are shown in Table 5. In most cases, the models show visibly better performance compared to the original benchmark. This performance gap suggests that the models have additional difficulty with incongruent questions, where an individual spatial preposition refers to distinct types of spatial relationships.