A Benchmark for Reasoning with Spatial Prepositions

Iulia-Maria Comșa
Google DeepMind
iuliacomsa@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]

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; Gärdenfors, 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 may 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

[1]The data is available at https://github.com/google-research/language/tree/master/language/spatial_prep.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16328–16335

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	❌
the helmet is above the scooter	the scooter is above the parking lot	the helmet is above the parking lot	✅
the helmet is above the scooter	the scooter is above my pay grade	the helmet is above my pay grade	❌
the robot is in the tent	the tent is under the bridge	the robot is under the bridge	✅
the robot is in the building	the building is under construction	the robot is under construction	❌

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.

commonsense spatial relationships.

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 (Comșa 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 [$prep_1$] Y"
- *premise2*: "Y is [$prep_2$] Z"
- *conclusion*: "X is [$prep_3$] Z"

where $prep$ is a spatial preposition such as "*in*" or "*on*" and $prep_3$ is one of $\{prep_1, prep_2\}$. 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 $\{prep_1, prep_2, prep_3\}$ 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 {*premise1*} and {*premise2*}, does that imply that {*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.

Prep.	Count	PaLM					GPT-3				Avg.	Human
		8b	62b	62b-1.3	540b	Flan	Ada	Babb.	Curie	DaVinci	LLM	
above	186	52.3	62.9	72.8	80.3	88.5	49.2	52.2	51.1	75.8	64.2	94.5
at	146	51.8	68.0	71.7	85.4	88.1	51.6	52.7	53.7	83.4	66.9	92.6
behind	148	54.7	59.2	68.0	76.4	70.9	52.5	51.1	50.7	76.1	62.1	89.8
in	250	56.5	72.3	75.9	89.1	86.4	51.2	54.0	51.6	88.2	68.9	96.2
on	228	52.6	69.0	70.0	82.6	86.1	51.3	55.1	51.0	81.5	66.1	91.1
under	202	53.8	60.2	65.5	79.9	80.9	50.0	53.5	45.2	75.0	62.3	94.7
with	40	52.5	68.3	69.2	89.2	90.0	53.8	56.7	50.0	85.0	68.0	100.0
cu	126	57.4	50.8	61.3	64.6	82.8	56.6	56.9	52.4	78.6	61.9	90.8
la	220	60.5	50.2	62.8	72.4	88.2	52.4	57.7	52.0	76.7	63.2	93.3
pe	222	58.1	50.2	63.4	72.8	84.8	58.6	54.7	51.8	81.5	63.5	91.0
sub	242	53.6	50.9	53.6	71.2	82.5	58.0	56.2	51.9	76.0	60.9	90.5
în	390	60.0	50.6	60.6	78.5	85.6	55.6	55.6	50.3	81.6	63.7	95.0

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.

sponses from humans and compare them to LLMs responses, which they consented to. The respondents were not paid. Each respondent answered 20 randomly-selected questions from the dataset. As a response quality measure, we only included the responses where the accuracy for congruent questions, which we consider easier, was above 75%. We thus obtained responses from 27 English-speaking and 23 Romanian-speaking adults. This allows the detection of an effect size of 0.56 and 0.61, respectively, at alpha 0.05 and power 0.8. The results are shown in Table 3.

4 Large Language Models Evaluation

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%

Model	Mean accuracy [95% C.I.]					
	English			Romanian		
	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot
PaLM-8b	53.00 [48.0-58.0]	53.00 [48.0-58.0]	55.25 [50.2-60.2]	60.25 [55.3-65.1]	55.25 [50.2-60.2]	59.00 [54.0-63.9]
PaLM-62b	56.25 [51.2-61.2]	69.25 [64.5-73.7]	72.25 [67.6-76.6]	50.25 [45.2-55.3]	50.50 [45.5-55.5]	51.00 [46.0-56.0]
PaLM-62b-1.3	60.50 [55.5-65.3]	74.00 [69.4-78.2]	78.00 [73.6-82.0]	58.50 [53.5-63.4]	54.25 [49.2-59.2]	64.00 [59.1-68.7]
PaLM-540b	**78.25** [73.9-82.2]	**83.50** [79.5-87.0]	**87.00** [83.3-90.1]	65.75 [60.9-70.4]	70.25 [65.5-74.7]	**84.25** [80.3-87.7]
Flan-PaLM-540b	**83.00** [79.0-86.6]	**82.75** [78.7-86.3]	**86.75** [83.0-89.9]	**83.25** [79.2-86.8]	**86.25** [82.5-89.5]	**85.50** [81.7-88.8]
GPT-3-Ada	50.00 [45.0-55.0]	50.75 [45.7-55.8]	55.25 [50.2-60.2]	54.50 [49.5-59.5]	52.50 [47.5-57.5]	61.50 [56.5-66.3]
GPT-3-Babbage	50.25 [45.2-55.3]	53.00 [48.0-58.0]	57.00 [52.0-61.9]	60.50 [55.5-65.3]	53.75 [48.7-58.7]	54.00 [49.0-59.0]
GPT-3-Curie	50.25 [45.2-55.3]	48.25 [43.3-53.3]	52.75 [47.7-57.7]	51.25 [46.2-56.2]	51.25 [46.2-56.2]	51.75 [46.7-56.7]
GPT-3-DaVinci	**83.00** [79.0-86.6]	**81.75** [77.6-85.4]	78.25 [73.9-82.2]	**80.25** [76.0-84.0]	**79.75** [75.5-83.6]	77.75 [73.4-81.7]
Baseline (conclusion only)	71.75 [67.1-76.1]	66.25 [61.4-70.9]	71.00 [66.3-75.4]	65.25 [60.4-69.9]	65.25 [60.4-69.9]	68.50 [63.7-73.0]
Generative (PaLM-540b)	72.75 [68.10-77.06]	82.00 [77.88-85.64]	88.25 [84.69-91.24]	62.38 [57.30-67.02]	60.25 [55.27-65.08]	82.00 [77.88-85.64]
Human	93.51 [91.8-95.3]			92.60 [90.1-95.1]		

Table 3: Performance of LLMs and humans on the spatial prepositions reasoning task in English and in Romanian. The best performance for each LLM across prompts is shown. Models with the best overlapping accuracy are highlighted. We include results for a baseline where the models made a response to the conclusion only, and for a generative experiment where PaLM-540b freely generated responses to the questions.

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 senses found in a dictionary as a proxy for the number of senses of each preposition. As can be observed, the number of senses is not proportional to corpus frequency. Moreover, each preposition might preferentially collocate with different verbs, and hence be more difficult to use in our dataset, where we chose the standard format "X is [$prep_1$] Y". This is one reason why the preposition "with" is relatively 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.

Prep.	Wiki. count	Dict. entries
in	516438	28
with	151830	25
at	82579	15
on	136415	44
above	5775	5
under	14618	8
behind	2789	3
în	657525	20
pe	176677	43
la	293601	27
cu	217508	28
sub	19903	13

Table 4: The frequency of each preposition based on a Wikipedia corpus estimation (Goldhahn et al., 2012), alongside the number of entries as determined from a standard dictionary: Cambridge Dictionary (https://dictionary.cambridge.org/) and Dexonline (https://dexonline.ro/).

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

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Yejin Choi. 2022. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2):139–155.

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. 2022. PaLM: Scaling language modeling with pathways. arXiv:2204.02311. Version 5.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. Version 1.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only. Association for Computational Linguistics.

Stefan Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces (Chapter 11)*. The MIT Press.

Annette Herskovits. 2009. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press.

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*

Computational Linguistics (Volume 1: Long Papers), pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Terry Regier. 1996. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. The MIT Press.

Jason Weston, Antoine Bordes, Sumit Chopra, Tomas Mikolov, Alexander M. Rush, and Bart van Merriënboer. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. Version 10.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics.

A Appendix: Prompts

We consider the following types of prompts for assessing LLM performance on the preposition transitivity benchmark:

1. "If {*premise1*} and {*premise2*}, does that imply that {*conclusion*}?"

2. "Q: If {*premise1*} and {*premise2*}, does that imply that {*conclusion*}? A:",

3. "Question: If {*premise1*} and {*premise2*}, does that imply that {*conclusion*}? Answer:",

4. "QUESTION: If {*premise1*} and {*premise2*}, does that imply that {*conclusion*}? ANSWER:",

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 0.44% 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 labelled it as incorrect. We scored the responses by assigning scores of 1, 0.5 and 0 to correct, ambiguous and incorrect responses, respectively.

We ran this experiment with five different temperature parameter values between 0 and 1. We found that a lower temperature produced the best results most of the time, and hence report the results for a temperature value of 0.

C Appendix: Negated Congruent Sentences

As an additional baseline and diagnostic tool, we assessed the performance of PaLM models on a dataset consisting of the congruent sentences and

Model	Mean accuracy [95% C.I.]					
	English			Romanian		
	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot
PaLM-8b	69.83 [65.1-74.3]	74.06 [69.5-78.3]	66.08 [61.2-70.7]	70.00 [65.2-74.5]	71.25 [66.5-75.6]	64.00 [59.1-68.7]
PaLM-62b	59.10 [54.1-64.0]	64.59 [59.7-69.3]	80.05 [75.8-83.9]	50.75 [45.7-55.8]	57.00 [52.0-61.9]	53.25 [48.2-58.2]
PaLM-62b-1.3	60.10 [55.1-64.9]	78.30 [73.9-82.2]	86.03 [82.3-89.3]	72.00 [67.3-76.3]	66.50 [61.6-71.1]	75.00 [70.5-79.2]
PaLM-540b	80.30 [76.1-84.1]	86.53 [82.8-89.7]	92.27 [89.2-94.7]	65.00 [60.1-69.7]	77.75 [73.4-81.7]	89.50 [86.1-92.3]
Flan-PaLM-540b	97.01 [94.8-98.4]	97.26 [95.1-98.6]	96.51 [94.2-98.1]	95.25 [92.7-97.1]	99.00 [97.5-99.7]	99.75 [98.6-100.0]

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.