Do ever larger octopi still amplify reporting biases? Evidence from judgments of typical colour

Fangyu Liu[1], **Julian Martin Eisenschlos**[2], **Jeremy R. Cole**[2], **Nigel Collier**[1]
[1] University of Cambridge [2] Google Research
`{fl399, nhc30}@cam.ac.uk` `{eisenjulian,jrcole}@google.com`

Abstract

Language models (LMs) trained on raw texts have no direct access to the physical world. Gordon and Van Durme (2013) point out that LMs can thus suffer from *reporting bias*: texts rarely report on common facts, instead focusing on the unusual aspects of a situation. If LMs are only trained on text corpora and naively memorise local co-occurrence statistics, they thus naturally would learn a biased view of the physical world. While prior studies have repeatedly verified that LMs of smaller scales (e.g., RoBERTa, GPT-2) amplify reporting bias, it remains unknown whether such trends continue when models are scaled up. We investigate reporting bias from the perspective of colour in larger language models (LLMs) such as PaLM and GPT-3. Specifically, we query LLMs for the typical colour of objects, which is one simple type of perceptually grounded physical common sense. Surprisingly, we find that LLMs significantly outperform smaller LMs in determining an object's typical colour and more closely track human judgments, instead of overfitting to surface patterns stored in texts. This suggests that very large models of language alone are able to overcome certain types of reporting bias that are characterized by local co-occurrences.[1]

1 Introduction

Large language models (LLMs) have been compared to hypothetical giant octopi living underwater that are exposed to a lot of language data (Bender and Koller, 2020). Such octopi would struggle to understand what actually happens on land as they lack the physical perceptual experience of living there. As such, they may overfit to text-only corpora and thus amplify reporting bias (Gordon and Van Durme, 2013) rather than faithfully reflecting the physical world.

[1]https://github.com/google-research/language/tree/master/language/octopus-llm (code).

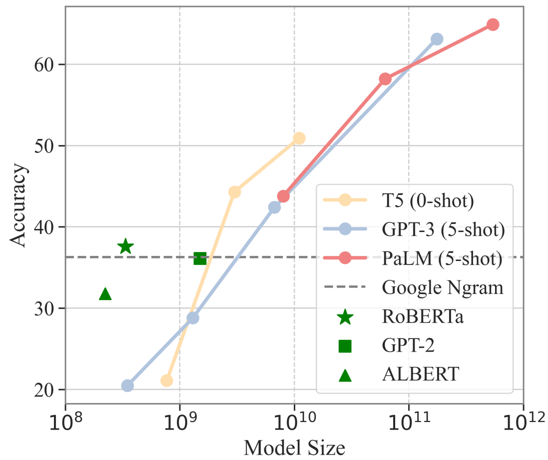

Figure 1: On typical colour judgments, large language models (LLMs) greatly outperform small LMs which previously were found to be no better than corpus statistics (Google Ngram). See Table 2 for full results.

In textual corpora, humans do not tend to mention what is commonly known, instead using language to express new information, which is likely less common. For example, when describing the colour of a banana: "green banana" has much higher frequency than "yellow banana" in the Google Books corpus.[2] It is natural to expect LMs would overfit to such reporting bias since they are trained to memorise such co-occurrence statistics. To observe this, we can query widely used pretrained models, such as RoBERTa$_{\text{Large}}$ (Liu et al., 2019) with our previous example. Given the prompt "It is commonly known that most bananas have the color <mask>", RoBERTa ranks "green" the highest.[3] This agrees with corpus statistics derived from raw text corpus such as the Google Ngram (Lin et al., 2012) mentioned above.[4] Paik

[2]research.tiny.us/google-ngrams-banana
[3]research.tiny.us/roberta-banana
[4]The Google Books corpus is an enormous collection of books digitised at Google (Michel et al., 2011). The 2nd edition of the corpus derived by Lin et al. (2012) contains >8B books, constituting over 6% of all books ever published. Google Ngram is a corpus of ngram statistics derived from the Google Books corpus (2nd edition). More details in §3.

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 210–220
November 20–23, 2022. ©2022 Association for Computational Linguistics

et al. (2021) test pretrained LMs' perception of colours and confirm that they perform no better than naive co-occurrence statistics extracted from the corpus. In fact, naively using corpus statistics achieves around 40% accuracy on their proposed colour probing benchmark CoDa while the best LM performs similarly. Zhang et al. (2022) extend the evaluation to a broader range of visual properties, confirming that reporting bias can negatively influence model performance and increasing model size does not help. Shwartz and Choi (2020) repeat the reporting bias experiments of Gordon and Van Durme (2013) on pretrained LMs and find that LMs overestimate rare events and actions, also amplifying reporting bias.

However, the LMs tested by Paik et al. (2021); Zhang et al. (2022); Shwartz and Choi (2020), i.e., GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019), RoBERTa, and ALBERT (Lan et al., 2020), usually have only several hundred million parameters and are of much smaller sizes than LLMs available now. In this work we probe T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), and PaLM (Chowdhery et al., 2022) of various sizes, with parameter counts ranging from 770M to 540B. Surprisingly, we find that LLMs almost double the performance of small language models (SLMs) on the typical colour task (Figure 1).[5] Paik et al. (2021) point out that SLMs achieve poor performance on objects that typically only have one colour (such as bananas), possibly due to their true colour being an aspect of common sense and thus not frequently mentioned in the training corpus. We find in contrast that LLMs achieve surprisingly good performance in this category, reaching >80% accuracy. After plotting accuracy against model size, we observe that scaling up is universally helpful for improving LLMs' performance on the colour probing benchmark (CoDa; Paik et al. 2021). Moreover, as LLMs are scaled their correlation to corpus ngram statistics plateau, suggesting that memorising (local) co-occurrence statistics cannot explain their success.[6]

Our study presents controlled analyses on the colour prediction task as a case study to show that scaling up LLMs could overcome surface-level pattern memorisation (i.e., text reporting bias in our case) and learn physical world common sense at least to some extent. This is an important and surprising finding as it provides a key evidence to counterargue the previous consensus that despite achieving better performance for a range of NLP tasks, larger LMs are more prone to overfitting to corpus statistics and therefore amplifying the reporting bias. Our study points out that this criticism on model scale is misleading as it is not based on the complete picture, and when the model capacity is increased to a significantly large scale such as PaLM-540B and GPT-3_{davinci}, they start to overcome reporting bias and are able to abstract physical common sense from text.

2 Method

To test whether LLMs replicate corpus biases rather than human judgment in the typical colour task, we compare the models' output distributions with the corpora's distribution and the distribution of human judgments. Visual perception provides an ideal testbed as corpus statistics can vary from physical facts; obvious facts are left unspoken. In this case, we focus on the typical colour task, largely following the setup by Paik et al. (2021). Given a query asking the colour of an object, the model must output a distribution over eleven possible colours. We then compare the output distribution to both corpus statistics and average human judgement to examine their respective correlations.

In the following, we explain how we query the LLMs and use their predictions. We test LLMs in three setups: zero-shot, one-shot, and five-shot.

Zero-shot. We use the following prompt across all models:

> **It is known that most** {$\texttt{OBJECT}_q$} **have the color** `<mask>`

where {$\texttt{OBJECT}_q$} is replaced with the object's name (in plural form).[7] After inputting the prompt, we compute next-token-prediction likelihood for all 11 colours in the CoDA label space and record the log-likelihood scores for all answers as the output distribution of the query:

$$S(c) = \log P_\Theta(c|\text{prompt}) \quad (1)$$

where Θ is LM's parameters; c is the color; "prompt" is the input prompt specified earlier. For 0-

[5] For convenience and consistency, we refer to all models with fewer than 10B parameters as small language models (SLMs) while those with more than 10B parameters as LLMs.

[6] A careful reader would note here that the models' training data may differ distributionally from Google Ngram. We discuss this more in §5.

[7] We try other prompts to test LLMs' sensitiveness towards the exact terms used. See Appx. §B.2 for more discussion.

and 5-shot prompting, the answer scoring scheme remains the same. See Appx. §C for details of how few-shot prompts are constructed.

3 Experimental Setup

Dataset. The CoDa dataset contains queries and human judgments of 521 objects. For each object, CoDa has a human-perceived colour distribution over 11 basic colours in English. The 11 colours were identified by Berlin and Kay (1969) and include black, blue, brown, grey, green, orange, pink, purple, red, white, yellow. As an example of the dataset, the object "Carrot" has the human-perceived scores of black: 0.0, brown: 0.023, orange: 0.797, etc., where the scores over 11 colours sum up to 1. CoDa contains three types of questions (1) Single (2) Multi and (3) Any. "Single" means the object has only one typical colour such as "Carrot" which is typically orange. "Multi" objects have between two and four typical colours: "Apple" is frequently red or green. "Any" objects have no fixed set of typical colours, such as "Shirt" and "Car". By default we report micro-average results across all three types. However, we also discuss the "Single" category in detail as it is thought to be especially indicative of reporting bias because such facts are rarely stated in texts. The statistics of CoDa are listed in Table 1.[8]

Type	Size	Examples
Single	198	Carrot, Spinach
Multi	208	Apple, Street light
Any	115	Shirt, Car

Table 1: CoDa statistics and examples.

Metrics. We use $\text{Acc}_{@1}$, ρ_{human}, ρ_{ngram}. $\text{Acc}_{@1}$ measures whether the model gets the most typical colour of an object correct. Other metrics are useful, but less clearly interpretable: ρ_{human} measures a set of predictions' Spearman's ρ correlation with the distribution of human colour judgments (however, there is low human consensus for some objects and colours). Higher $\text{Acc}_{@1}$ is better; higher ρ_{human} indicates a closer match to human judgments. ρ_{ngram} measures the models' predictions' correlation with the Google Ngram statistics. Fitting corpus statistics is not necessarily good or bad: we report it to see its relationship with both model size and model performance.

Google Ngram baseline. Together with queries and human judgments, Paik et al. (2021) also provide ngram stats collected from Google Books and Wikipedia to compute the correlation with these corpora. Specifically, they consider all bi- and trigrams containing a colour followed by an object. A corpus-based baseline is then computing the accuracy/correlation between the total ngram counts of colour-object pairs and the human perceived-scores. We use Google Ngram as the default baseline as Google Books is much larger than Wikipedia and Google Ngram has better correlation with human judgments than Wikipedia. Wikipedia results are reported in Appx. §A.

SLM baselines. We use the best-performing SLMs from Paik et al. (2021) as our baselines, which are RoBERTa$_{\text{Large}}$, GPT-2$_{\text{XL}}$, and ALBERT$_{\text{V2-XXL}}$. One important difference between Paik et al. (2021)'s setup is that they create ten different hand-crafted templates and present the best results per-object for each model. Our work uses a single template across all models and objects. Thus, we are underestimating LLMs' performance compared to the previously reported SLMs' numbers from Paik et al. (2021). Nonetheless, we see that LLMs outperform SLMs by large margins.[9]

Compared LLMs and their sizes. OpenAI does not disclose the exact size of their text models Ada, Babbage, Curie and Davinci. According to blog.eleuther.ai/gpt3-model-sizes, they roughly correspond to 350M, 1.3B, 6.7B, and 175B, which we use as the models' parameter counts. For other models (i.e., T5 and PaLM), their number of parameters are made clear in the original papers. We list all compared models' sizes in the second column of Table 2.

4 Results

Main results (Table 2). We show our main results in Table 2. As a general trend, LLMs with >10B parameters all significantly outperform SLMs with <10B parameters, and performance

[8]The original CoDa dataset has a train/validation/test split used for training classifiers to probe embedding-based representations. However, the split was only applied on the embedding model CLIP (Radford et al., 2021) and all other numbers were reported on the full set. To be consistent, we also report performance on the full dataset.

[9]In Appx. §B.1, we show that SLMs' performance can drop to chance-level using the same zero/few-shot evaluation protocol as LLMs. We also demonstrate that when using different prompts, LLMs such as GPT-3$_{\text{davinci}}$'s 0-shot performance can be improved from 55.5% to 62.2% (Appx. §B.2). However, we uniformly use one single prompt for LLMs to avoid over-optimistic results.

Model	Size	0-shot			1-shot			5-shot		
		$\text{Acc}_{@1}$	ρ_{human}	ρ_{ngram}	$\text{Acc}_{@1}$	ρ_{human}	ρ_{ngram}	$\text{Acc}_{@1}$	ρ_{human}	ρ_{ngram}
Google Ngram	-	36.3	44.2	100.0	-	-	-	-	-	-
RoBERTa$_{\text{Large}}$*	335M	**37.6**	-	-	-	-	-	-	-	-
GPT-2$_{\text{XL}}$*	1.5B	36.1	-	-	-	-	-	-	-	-
ALBERT$_{\text{V2-XXL}}$*	223M	31.8	-	-	-	-	-	-	-	-
T5$_{\text{Large}}$	770M	21.1	25.7	42.2	19.4	21.0	24.5	17.9	20.7	11.8
T5$_{\text{XL}}$	3B	44.3	**57.4**	60.3	39.0	48.8	55.2	42.4	47.8	60.3
T5$_{\text{XXL}}$	11B	**50.9**	49.5	57.5	**47.2**	**54.3**	55.9	**48.0**	**53.4**	54.1
GPT-3$_{\text{ada}}$	350M	17.9	15.7	48.8	21.3	24.5	46.0	20.5	25.4	42.2
GPT-3$_{\text{babbage}}$	1.3B	27.6	22.1	58.0	27.6	29.8	51.7	28.8	37.1	51.9
GPT-3$_{\text{curie}}$	6.7B	33.6	32.8	63.5	40.1	44.2	59.2	42.4	47.1	57.1
GPT-3$_{\text{davinci}}$	175B	**55.5**	**43.1**	65.0	**61.8**	**60.5**	61.0	**63.1**	**62.3**	55.9
PaLM-8B	8B	29.6	34.7	61.5	39.9	38.9	64.7	43.8	52.6	62.0
PaLM-62B	62B	34.2	33.5	64.4	50.1	44.8	65.3	58.2	61.9	61.1
PaLM-540B	540B	**42.6**	**46.0**	66.3	**63.9**	**62.5**	62.5	**64.9**	**66.2**	60.1

Table 2: Results on CoDa (average over all three types). For $\text{Acc}_{@1}$ and ρ_{human} (the higher the better), the best performing models within each model class are **boldfaced**. The symbol * denotes numbers from Paik et al. (2021), which uses a more optimistic protocol, aggregating the best per-object performance over 10 hand-crafted prompts.

increases monotonically with scale within each model class. While the SLMs do not perform significantly better than Google Ngram (accuracy 36.3%), LLMs achieve up to 64.9% (PaLM-540B 5-shot). PaLM-540B 5-shot also correlates best with human judgments. For PaLM and GPT-3, few-shots are much better than 0-shot;[10] while for T5, 0-shot seems to be the best.

Results on the "Single" colour split (Table 3). The "Single" split deserves extra attention as it has the highest human consensus and is also considered to be common sense knowledge, implying it is rarely stated in the corpus (Paik et al., 2021). While none of the SLM baselines outperform the Ngram baseline on $\text{Acc}_{@1}$, the largest PaLM and GPT-3 surpass the Ngram baseline by nearly 40%. Furthermore, the LLMs' predictions correlate significantly more to human judgments.

We also present an error analysis on the "Single" split in Appx. §B.5. Out of the ten errors made by PaLM-540B, only one is a clear mistake where the model classifies picnic baskets as red. For other nine errors, the error seems to be associated with the ambiguous nature of the questions or the dataset construction process.

Correlation metrics (Figure 2). For GPT-3, its correlation with corpus ngram statistics (ρ_{ngram}) initially increases but then plateaus and even decreases (on 5-shot: 42.2 → 51.9 → 57.1 → 55.9). On PaLM, ρ_{ngram} decreases from the start as model size grows (on 5-shot: 62.0 → 61.0 → 60.1). On both models, ρ_{ngram} initially is larger than ρ_{human}. However, for model sizes above 10^{11} parameters, both models' predictions have $\rho_{\text{human}} > \rho_{\text{ngram}}$. This suggests that when LMs are small, they can underfit corpus ngrams. When LMs start to be scaled up, they increasingly fit the corpus. However, after a certain model size, additional scale does not lead to more overfitting to corpus statistics. On the contrary, as LLMs' predictions correlate more with human judgment, they also start to decorrelate with corpus statistics.

	Model	$\text{Acc}_{@1}$	ρ_{human}	ρ_{ngram}
0-shot	Google Ngram	43.9	44.2	100.0
0-shot	RoBERTa$_{\text{Large}}$*	**42.9**	**47.8**	-
0-shot	GPT-2$_{\text{XL}}$*	40.4	40.3	-
0-shot	ALBERT$_{\text{V2-XXL}}$*	34.3	43.7	-
0-shot	GPT-3$_{\text{ada}}$	20.2	16.9	47.4
0-shot	GPT-3$_{\text{babbage}}$	30.8	27.4	56.0
0-shot	GPT-3$_{\text{curie}}$	39.9	39.9	62.0
0-shot	GPT-3$_{\text{davinci}}$	**71.2**	**50.7**	62.2
0-shot	PaLM-8B	34.8	38.2	62.1
0-shot	PaLM-62B	44.4	34.3	64.1
0-shot	PaLM-540B	**53.0**	**42.2**	65.6
5-shot	GPT-3$_{\text{ada}}$	19.7	21.7	42.3
5-shot	GPT-3$_{\text{babbage}}$	32.3	35.0	50.3
5-shot	GPT-3$_{\text{curie}}$	53.5	47.3	55.9
5-shot	GPT-3$_{\text{davinci}}$	**82.3**	**59.9**	53.3
5-shot	PaLM-8B	53.0	50.9	60.7
5-shot	PaLM-62B	73.2	58.5	58.5
5-shot	PaLM-540B	**80.8**	**63.1**	57.0

Table 3: Results on CoDa ("Single" type). 1-shot and T5 results (omitted) follow similar trend as Table 2.

[10] We observe that PaLM 0-shot is relatively poor (significantly worse than GPT-3) and its strength is only shown with few-shot. Similar behaviour of PaLM is also observed on tasks such as Natural Questions (Kwiatkowski et al., 2019). Since this is not the focus of this paper, we leave discovering the cause for future investigation.

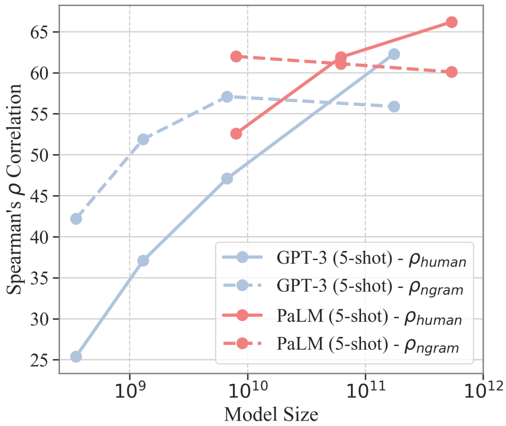

Figure 2: GPT-3 and PALM's Spearman's ρ correlation with human judgment and Google Ngram as they are scaled up. These are 5-shots results from Table 2.

5 Discussion and Limitations

Discrepancy among corpora. The corpus statistics we investigate are induced from Google Books and Wikipedia. They do not necessarily replicate the corpus statistics used for training LLMs. Nonetheless, we do not believe the discrepancy would be big enough to boost LLMs' performance to 80% on single-type questions. Future work could investigate the original training corpus of LLMs (e.g., C4 for T5).

Is ngram a good reference? Paik et al. (2021); Zhang et al. (2022) use the counts of colour occurrences with objects in bi- and tri-grams. However, to fully understand whether LLMs overfit, we also need to consider longer contexts as it is possible that the typical colour of an object is described in longer pieces of text; thus, LLMs performance improvements can be attributed to memorising long-term dependencies better than SLMs. In this case, the "generalisation" is only memorising a context that is similar to the prompt. Alternatively, LLMs may learn good representations of the quantifiers, such as "most", and the usage of the atypical colours in the text may not co-occur with quantifiers suggesting it is common. In future work, we intend to examine whether a similar phenomenon persists when collecting occurrence stats over typical model input lengths and using more fine-grained data that also characterises premodifiers such as quantifiers.

Comparing within model class for better control of confounders. Though LLMs today are almost all Transformer-based models with similar autoregressive pretraining objectives, we note that there are caveats preventing us from having a perfect control over design choices on pretraining corpora and specific architectures. In terms of pretraining data, within-family models of different sizes generally use the same training data (GPT-3models are however less transparent in this regard). However, it is unclear what differences there are across model families. In terms of model architectures, T5 is an encoder-decoder model while GPT-3 and PALM are decoder-only models. PALM has further modifications on top of the original Transformer architecture such as using SwiGLUE activation (Shazeer, 2020) instead of the standard ReLU; using RoPE embeddings (Su et al., 2021) instead of the original relative position embeddings. As a result, more conclusive findings should be drawn *within* model classes, e.g. comparing PALM-540B with its two smaller versions instead of GPT-3 models.

Colours live on a spectrum. The evidence we obtain does not reflect whether LLMs have a fine-grained and holistic understanding of the nature of colour. That is, colours live on a continuous spectrum. LLMs could have solved CoDa by identifying the mappings between objects and colours but not colour's relative positions on the spectrum. One way to probe this is to examine if LLMs can resolve colour synonyms (e.g., do LLMs know that "scarlet" occupies a subspan of the colour red?). However, a rigorous and systematic study of this problem is beyond the scope of this study.

6 Conclusion and Future Work

In this work, we examine LLMs ability to make typical colour judgments, a simple property of visual common sense. Contradicting Paik et al. (2021); Zhang et al. (2022), we find that typical colour judgments do not follow an inverse scaling law, and scale is indeed quite critical for high accuracy on the task. While generalising from this task to visual reasoning as a whole is premature, we provide some evidence that larger models of language alone are able to overcome a basic type of reporting bias. Future work will look at a wider range of physical properties (Collier et al., 2022) and more carefully control for the data and model size. We also hope our work opens an avenue for empirically verifying on what level meaning acquisition is possible from a cognitive linguistic perspective (Piantasodi and Hill, 2022).

Acknowledgements

We would like to thank the reviewers and the ACs for their constructive feedback. We would also like to thank Bhuwan Dhingra, William W. Cohen, Tiago Pimentel, Ehsan Shareghi, for the discussions and comments. We thank Cory Paik for providing details about Paik et al. (2021).

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms, Their Universality and Evolution*. University of California Press.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Nigel H Collier, Fangyu Liu, and Ehsan Shareghi. 2022. On reality and the limits of language data. *arXiv preprint arXiv:2208.11981*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steven T Piantasodi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Noam Shazeer. 2020. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5321–5335, Seattle, United States. Association for Computational Linguistics.

A A More Comprehensive Table (Table 5)

In the main text, we compare different models under different setups in Table 2. To provide more information for reference and also strengthen our findings, we present a more comprehensive Table 5 which also reports Kendall's τ as a correlation metric, and include Wikipedia stats provided by Paik et al. (2021) as another source of ngrams. The main conclusion remains the same. Kendall's τ has identical trend to Spearman's ρ, and similar fitting trend of Google Ngram is also shown on Wikipedia.

B Further Discussions

Here we present some more extensive discussions on several topics that concern the experimental setup, including testing SLMs under the same setup as the LLMs (Appx. §B.1); testing different prompts (Appx. §B.2); the discrepancies among analysed corpora and the real pretraining corpora of LLMs (Appx. §B.3); the risk of direct data leakage (Appx. §B.4); and error analysis (Appx. §B.5); .

B.1 Real zero/few-shot setup for SLMs

In the main text, we used SLM numbers reported by Paik et al. (2021) under an optimistic setup: i.e. out of 10 prompts, choosing always the prompt that maximises per-object's performance when evaluating models. We note that when under the same evaluation protocol as LLMs, SLMs' performance would have dropped to chance level. We pick the best performing SLM RoBERTa$_{\text{Large}}$ as an example. When consistently using one prompt, RoBERTa$_{\text{Large}}$ has only an accuracy score of 7.3%. Prompting with few-shot examples does help a bit. However, the 5-shot accuracy of RoBERTa$_{\text{Large}}$ (real) still has a roughly 50% gap compared with few-shot performance of the best LLMs.

B.2 LLMs' Sensitiveness to Prompts

For the main experiment, we choose an arbitrary prompt: "It is known that most `{OBJECT}` have the color `<mask>`.". However, it is possible that LLMs are particularly good or bad at this prompt and it is worth testing whether LLMs are robust to how we ask the question. In Table 6, we test GPT-3's sensitivity towards different prompts. First, we change the quantifier "most" to "all", no quantifier, "some", "few", and "no". We find that the LLM is sensitive to the quantifier and produces scores generally well correspond to the quantity being asked. Note that "all" and no quantifier lead to lower performance than "most", possibly due to the question is unnatural since there is rarely any object exclusively having only one colour. We also paraphrase the original prompt and find that a grammatical paraphrased query can lead to up to around +/-6% performance difference. An ungrammatical prompt will damage the model's performance, even including key words such as "most", "color", and "common sense".

B.3 Discrepancy among Corpora

As discussed in Limitations (§5), we use Google Books and Wikipedia in line with Paik et al. (2021) for direct comparison. As can be seen in Table 5, Google Ngram is better agreeing with human judgment. Moreover, Google Books is much larger than Wikipedia. So, in the main experiments, we use it as an approximation of pretraining corpora. However, it remains unknown how well these sources' ngram distributions align with the real training corpora of LLMs. In future work, there should ideally be more strict control and better access to the pretraining data to draw firmer conclusions.

B.4 Have the LLMs seen test data during training?

It is unlikely that LLMs have seen the test data in its exact form in their pretraining corpora. As the whole web can be used as training data, this is a real risk. However, we think it is unlikely that LLMs have seen CoDa. The CoDa dataset was released on October 2021. GPT-3-davinci-002 was trained with data until June 2021; GPT-3-curie/babbage/ada-001 were using data until October 2019;[11] T5's pretraining corpus C4 was crawled on April 2019. PaLM's precise training data is unknown, but the paper was published after CoDa. However, performance-wise PaLM is not significantly better than GPT-3-davinci-002, which uses training data before the release of CoDa.

B.5 Error Analysis

Here we pick the errors made by the models on Single-type questions to understand why or what type of questions they make mistake. Both GPT-3 and PaLM achieve above 80% in this category. We randomly sample 10 errors made by PaLM-540b (5-shot) and list them below.

[11] beta.openai.com/docs/models/gpt-3

Model	Size	0-shot			1-shot			5-shot		
		Acc$_{@1}$	ρ_{human}	ρ_{ngram}	Acc$_{@1}$	ρ_{human}	ρ_{ngram}	Acc$_{@1}$	ρ_{human}	ρ_{ngram}
Google Ngram	-	36.3	44.2	100.0	-	-	-	-	-	-
RoBERTa$_{\text{Large}}$*	335M	37.6	-	-	-	-	-	-	-	-
RoBERTa$_{\text{Large}}$ (real)	335M	7.3	25.8	55.9	8.4	17.0	52.9	15.4	28.5	51.8
GPT-3$_{\text{davinci}}$	175B	**55.5**	**43.1**	65.0	61.8	60.5	61.0	63.1	62.3	55.9
PaLM-540B	540B	42.6	46.0	66.3	**63.9**	**62.5**	62.5	**64.9**	**66.2**	60.1

Table 4: Evaluating the best performing SLM on CoDa, using one consistent prompt (the same setup for all LLMs tested). Performance of the optimistic 10-prompt setup by Paik et al. (2021) and also performance of LLMs are listed for reference. When evaluated under the same protocol as LLMs, the best performing SLM RoBERTa's performance drops very significantly and is at chancel level.

- - - - - - - - - - - - error 1 - - - - - - - - - - - - -
query: ... **most `mangoes` have the color** `<mask>`
ground truth: `orange`
prediction: `yellow`
- - - - - - - - - - - - error 2 - - - - - - - - - - - - -
query: ... **most `computer monitors` have the color** `<mask>`
ground truth: `black`
prediction: `gray`
- - - - - - - - - - - - error 3 - - - - - - - - - - - - -
query: ... **most `sinks` have the color** `<mask>`
ground truth: `gray`
prediction: `white`
- - - - - - - - - - - - error 4 - - - - - - - - - - - - -
query: ... **most `porcupines` have the color** `<mask>`
ground truth: `brown`
prediction: `black`
- - - - - - - - - - - - error 5 - - - - - - - - - - - - -
query: ... **most `potatoes` have the color** `<mask>`
ground truth: `brown`
prediction: `white`
- - - - - - - - - - - - error 6 - - - - - - - - - - - - -
query: ... **most `kangaroos` have the color** `<mask>`
ground truth: `brown`
prediction: `gray`
- - - - - - - - - - - - error 7 - - - - - - - - - - - - -
query: ... **most `pancakes` have the color** `<mask>`
ground truth: `brown`
prediction: `yellow`
- - - - - - - - - - - - error 8 - - - - - - - - - - - - -
query: ... **most `scorpions` have the color** `<mask>`
ground truth: `brown`
prediction: `black`
- - - - - - - - - - - - error 9 - - - - - - - - - - - - -
query: ... **most `coins` have the color** `<mask>`
ground truth: `gray`
prediction: `yellow`
- - - - - - - - - - - - error 10 - - - - - - - - - - - - -
query: ... **most `picnic baskets` have the color** `<mask>`
ground truth: `brown`
prediction: `red`

Most of the ten queries seem to be ambiguous. Black and brown scorpions are both common; the color of a mango might be described as orange or yellow; kitchen sinks are normally gray but bathroom sinks are normally white; old computer monitors are normally gray but newer ones are normally black. The most obvious mistake seems to be on picnic baskets which PaLM classifies as red. We believe these are included in Single-type questions due to the method used for constructing CoDa. To identify if an object has a single, multiple, or many typical colours, Paik et al. (2021) use a clustering algorithm together with manual assignment. However, the threshold of one-versus-many clusters can be hard to decide, and many objects would end up at the boundary. Also, depending on the number of annotators, the presented ground truth may be noisy when compared to the general population.

C Few-shot Prompts

One-shot. For one-shot, we prepend one randomly selected example from the dataset. The example is constructed by randomly selecting an object from the dataset and then choosing the colour with the highest probability answer from the ground truth. Some of the objects could have multiple reasonable colours (e.g., yellow will be chosen for bananas, even though they can be green or brown).

It is known that most {OBJECT$_1$} **have the color** {COLOR$_1$}; **most** {OBJECT$_q$} **have the color** `<mask>`

Five-shot. Similar to one-shot, but we randomly sample five objects from the dataset.

It is known that most {OBJECT$_1$} **have the color** {COLOR$_1$}; {OBJECT$_2$} **have the color** {COLOR$_2$}; ...; {OBJECT$_5$} **have the color** {COLOR$_5$}; most {OBJECT$_q$} **have the color** `<mask>`

Model	size	human			GBN		wiki	
		Acc$_{@1}$	ρ	τ	ρ	τ	ρ	τ
GBN	-	36.3	44.2	36.2	100.0	100.0	66.5	55.9
wiki	-	23.3	28.6	23.2	66.5	55.9	100.0	100.0
				0-shot				
RoBERTa$_{\text{Base}}$*	110M	28.0	-	-	-	-	-	-
RoBERTa$_{\text{Large}}$*	335M	37.6	-	-	-	-	-	-
GPT-2$_{\text{Small}}$*	124	27.1	-	-	-	-	-	-
GPT-2$_{\text{Base}}$*	355M	31.7	-	-	-	-	-	-
GPT-2$_{\text{Large}}$*	774M	33.2	-	-	-	-	-	-
GPT-2$_{\text{XL}}$*	1.5B	36.1	-	-	-	-	-	-
ALBERT$_{\text{V2-Base}}$*	11M	20.9	-	-	-	-	-	-
ALBERT$_{\text{V2-Large}}$*	17M	28.8	-	-	-	-	-	-
ALBERT$_{\text{V2-XL}}$*	58M	25.2	-	-	-	-	-	-
ALBERT$_{\text{V2-XXL}}$*	223M	31.8	-	-	-	-	-	-
T5$_{\text{Large}}$	770M	21.1	25.7	20.6	42.2	32.1	33.3	25.9
T5$_{\text{XL}}$	3B	44.3	57.4	46.6	60.3	47.3	41.7	32.3
T5$_{\text{XXL}}$	11B	50.9	49.5	40.5	57.5	44.9	40.5	31.4
GPT-3$_{\text{ada}}$	350M	17.9	20.3	15.7	48.8	36.7	36.9	28.1
GPT-3$_{\text{babbage}}$	1.3B	27.6	27.8	22.1	58.0	44.5	44.6	34.6
GPT-3$_{\text{curie}}$	6.7B	33.6	41.0	32.8	63.5	50.1	37.3	36.8
GPT-3$_{\text{davinci}}$	175B	**55.5**	**52.8**	**43.1**	65.0	51.5	48.1	37.3
PaLM-8B	8B	29.6	34.7	27.3	61.5	47.6	46.8	36.5
PaLM-62B	62B	34.2	33.5	26.9	64.4	50.9	49.9	49.5
PaLM-540B	540B	42.6	44.0	35.5	*66.3*	*52.7*	*48.3*	*38.0*
				1-shot				
T5$_{\text{Large}}$	770M	19.4	21.0	16.4	20.3	15.7	24.5	18.5
T5$_{\text{XL}}$	3B	39.0	48.8	39.4	37.6	28.9	55.2	42.6
T5$_{\text{XXL}}$	11B	47.2	54.3	44.3	38.7	29.6	55.9	43.5
GPT-3$_{\text{ada}}$	350M	21.3	24.5	19.3	46.0	35.0	34.8	27.0
GPT-3$_{\text{babbage}}$	1.3B	27.6	29.8	23.6	51.7	39.7	39.7	30.5
GPT-3$_{\text{curie}}$	6.7B	40.1	44.2	35.6	59.2	46.3	44.6	34.7
GPT-3$_{\text{davinci}}$	175B	61.8	60.5	50.1	61.0	48.0	42.0	32.7
PaLM-8B	8B	39.9	48.0	38.9	64.7	*51.7*	*47.6*	*37.6*
PaLM-62B	62B	50.1	54.9	44.8	*65.3*	*51.7*	46.2	35.8
PaLM-540B	540B	**63.9**	**63.5**	**52.8**	62.5	49.3	42.7	33.1
				5-shot				
T5$_{\text{Large}}$	770M	17.9	20.7	16.2	11.8	9.1	6.0	4.3
T5$_{\text{XL}}$	3B	42.4	47.8	38.8	60.3	47.3	42.6	33.3
T5$_{\text{XXL}}$	11B	48.0	53.4	43.6	54.1	42.0	36.6	28.8
GPT-3$_{\text{ada}}$	350M	20.5	25.4	19.9	42.2	32.3	31.2	23.8
GPT-3$_{\text{babbage}}$	1.3B	28.8	37.1	29.5	51.9	39.7	39.6	30.6
GPT-3$_{\text{curie}}$	6.7B	42.4	47.1	38.0	57.1	44.8	40.9	32.1
GPT-3$_{\text{davinci}}$	175B	63.1	62.3	51.6	55.9	43.7	35.9	27.7
PaLM-8B	8B	43.8	52.3	42.6	*62.0*	*49.1*	*44.9*	*35.1*
PaLM-62B	62B	58.2	61.9	51.2	61.1	48.0	41.3	31.8
PaLM-540B	540B	**64.9**	**66.2**	**55.2**	60.1	47.3	40.7	31.6

Table 5: Full table containing more corpus stats (wiki) and more metrics (Kendall's τ). GBN: Google Ngram; wiki: Wikipedia ngrams. Both are from Paik et al. (2021).

Prompt	Acc$_{@1}$
It is known that most {OBJECT} have the color <mask> *(original)*	55.5
different quantifiers	
It is known that all {OBJECT} have the color <mask>	49.9
It is known that _{OBJECT} have the color <mask>	46.3
It is known that some {OBJECT} have the color <mask>	27.3
It is known that few {OBJECT} have the color <mask>	22.5
It is known that no {OBJECT} have the color <mask>	14.0
paraphrases of the original prompt	
It is known that color of most {OBJECT} are <mask>	56.6
It is known that the color of most {OBJECT} are <mask>	59.1
It is common sense that the color of most {OBJECT} are <mask>	**62.2**
It is known that most {OBJECT} are <mask>	49.1
It is known that {OBJECT} are <mask>	44.2
It is common knowledge that most {OBJECT} have the color <mask>	52.0
It is common sense that most {OBJECT} have the color <mask>	55.5
It is commonly known that most {OBJECT} have the color <mask>	53.0
Everybody knows that most {OBJECT} have the color <mask>	54.3
Most people think that {OBJECT} have the color <mask>	53.6
The majority of {OBJECT} have the color <mask>	51.2
The vast majority of {OBJECT} have the color <mask>	52.9
Most {OBJECT} color <mask> *(ungrammatical)*	44.1
Common sense most {OBJECT} color <mask> *(ungrammatical)*	43.4

Table 6: GPT-3$_{davinci}$'s 0-shot performance on CoDa across different prompts.