Generative Image Model Benchmark for Reasoning and Representation (GIMBRR)

Jascha Achterberg,[*1] **Ron Arel,**[*2] **Tetiana Grinberg,**[*3] **Adel Chaibi,**[3] **Joscha Bach,**[3] **Nikos Tzagadakis**[4]

[1] University of Cambridge, work done during an internship at Intel Labs
[2] UIUC, work done during an internship at Intel Labs
[3] Intel Labs, [4] Metanomic
jascha.achterberg@mrc-cbu.cam.ac.uk, ronarel2@illinois.edu, {tetiana.grinberg, adel.chaibi, joscha.bach}@intel.com, nikos@metanomic.net

Abstract

Recent developments in generative AI have highlighted how the field is moving to a state where models start developing a large set of skills and can solve a multitude of tasks without having actively been optimized for them. Benchmarks have been a key component driving model development in the past, but as models' capabilities become more complex, it becomes harder to create benchmarks which can meaningfully capture the skillsets of algorithms to inform future model developments and training pipelines. While in the domain of language we have seen the development of a wide-ranging array of benchmarks, these are currently missing for generative image algorithms. Here we introduce the Generative Image Model Benchmark for Reasoning and Representation (GIMBRR) which is an open-source software package to assess generative image algorithms on 11 cognitive tasks using manual and automated evaluation pipelines. GIMBRR is built with customizability in mind, so that it can easily be updated with new tasks and assessment routines. This way it can be adapted to suit the needs of research teams with specific goals in image generation and to update task difficulty as generative image algorithms progress in general. We used GIMBRR to measure performance of three popular generative image models (DALL-E 2, Midjourney, Stable Diffusion), demonstrating that reasoning and representation tasks pose a considerable challenge to all of them. We have also demonstrated how cognitive theory can be used to perform a systematic analysis of generative and representational capabilities of these models.

Introduction

Machine learning models are usually assessed using benchmarks and specific validation datasets. As modern AI systems are becoming more generalist in terms of tasks they are capable of solving, it becomes increasingly valuable to not only quantify their overall performance on the specific task for which they are trained, but also to systematically diagnose the source of their errors so that researchers can improve the related skill in a targeted fashion. Many systems are trained to solve a vast set of tasks and have thus acquired a large set of varied skills. In such systems, there exists a strong potential that by improving the model's core capabilities, there will be strong trickle-up effects leading to performance improvements on many of the more complex tasks. This is similar to the pattern of skill improvement that can be observed in human learners.

In human learning and cognitive development, it is commonly noted that there exist certain core knowledge categories that allow children to learn complicated skills while interacting with the world (Spelke 2022). The acquired basic set of skills then supports children in learning more complicated skills (Kievit 2020; Peng and Kievit 2020). It has been argued that giving an AI model the capabilities to break down a problem into simple parts, understand objects in the world and the relationships between them, and use these observations as the basis of a flexible reasoning process, will be an essential part of creating more generally intelligent agents (Lake et al. 2017; Chollet 2019).

Although significant progress has been made in creating advanced libraries for the assessment of generative language models (Srivastava et al. 2022) and multimodal models that interact with visual inputs (Akkus et al. 2023), assessment methodologies for generative image models are still in their nascency. To facilitate the continued progress of generative image model performance, we introduce an open-source and easy to customize benchmark suite which can be used to assess models on 11 core cognitive and generative tasks. Together with this evaluation suite, we present the assessment results of three popular generative image models. We then use these results to identify potential causes of generative performance issues within each model. We hope that our evaluation methodology along with the GIMBRR benchmark suite ("benchmark") will provide the research community with a standardized, systematic and efficient approach to iterative model development.

Related Work

A number of researchers have already pointed out the shortcomings of the default approach to benchmark design (Raji et al. 2021), with multiple works attempting to address various problems such as reproducibility of human annotation (Khashabi et al. 2021), causal reasoning (Weston et al. 2015)

[*]These authors contributed equally.

and lack of task variety (Srivastava et al. 2022). However, most of the successful approaches have centered around Large Language Models (Efrat, Honovich, and Levy 2022).

Currently, generative image models are typically evaluated using metrics such as FID (Heusel et al. 2017), CLIP score (Radford et al. 2021), and Inception Score (Betzalel et al. 2022). However, these metrics have several limitations when it comes to evaluating out-of-distribution prompts and the corresponding generated images, as well as the presence or absence of specific object relations in the generated image. For example, FID requires a large sample of real and generated images to accurately calculate the distance between the two distributions. Thus, the FID metric is ineffective when a very specific output is required, rather than any valid instance from a large class of objects. CLIP uses contrastive learning to address some of these issues, but it still struggles with distinguishing between different relations. This happens because CLIP treats prompts as a bag-of-words, disregarding the grammatical and relational content of the statements.

Prompt engineering is a well-established method for improving the performance of generative models and is currently an area of active research, given the recent advances in the field (Hao et al. 2022; Zhou et al. 2022). Our approach can be viewed as a form of adversarial prompt engineering, which seeks to identify a model's weaknesses in reasoning and representation.

Method

Cognitive Tasks

In the initial version of the benchmark, we developed a set of 11 cognitive tasks that we consider to be fundamental in evaluating the capabilities of generative image algorithms. These tasks encompass counting, multiple object representation, simple arithmetic, directed actions, spatial relationships, unusual arrangements, conditional generation, compositional characteristics, negation, chimeras, and text rendering (see Table 1 for example prompts).

We created a set of 100 prompts for each task, using a combination approach of generating a large set of prompts with GPT-3 (Brown et al. 2020) and then sub-selecting the ones that best fitted our evaluation methodology. All the prompts that are used in this version of the benchmark have been manually reviewed before integrating them into the benchmark suite. We have also extracted meta-data on prompts which we believe to influence the generative performance of models. One example is the exact count of objects which needs to be generated in the counting tasks. The results section discusses these additional variables further. Where the task was designed to accommodate automated as well as manual evaluation, object classes for prompts were deliberately selected to match those used in the automated assessment models - in particular the COCO classes (Lin et al. 2014).

Although the tasks assess relatively independent generative capabilities, there are significant overlaps between them. Therefore, in our discussion of model performance, we will not only focus on the performance on each task but also on the relationship between them. For instance, the directed actions task relies on the algorithms' ability to generate two separate objects, which is also captured by the multiple objects task. As a model's ability to solve any of the tasks described will likely be strongly influenced by its ability to generate the individual objects, we added an additional task to test the models' ability to generate all of the individual objects used in the more complicated tasks outlined before (called "Single object" task). Overall, this task comprises 1602 prompts. We also designed a separate "controlled" set of prompts to allow for a more detailed examination of numerical abilities, which includes studying single object representation, multiple object representation, counting of single objects, counting of multiple objects, and simple arithmetic. We ensured that the object classes used across these tasks were the same. This resulted in a total set of 2102 prompts.

Benchmark Software

In addition to the generative tasks we are introducing a benchmark suite ("benchmark") to streamline and standardize the assessment of image generation algorithms' capabilities on cognitive tasks. The benchmark is an open source-software designed with customization in mind, allowing the research community to add tasks and functions as image generation algorithms improve. The benchmark is based on the Streamlit Python package and can be deployed locally via GitHub (https://github.com/IntelLabs/generative-ai/) or through a hosted version on Hugging Face (refer to GitHub repository README).

Figure 1 shows a flowchart of using the benchmark. The benchmark comes with an interface to download prompts for all our tasks. Image generation is performed by the users. Users then upload their generated images to the benchmark tool. Here they can either manually evaluate the generations or use the provided automated evaluation routines. The manual evaluation routine asks users to rate image-prompt pairs on whether they match or not (binary yes / no rating). During this rating loop, the user can choose to view related prompts on the same page (e.g. the single objects that correspond to a multiple object representation prompt), in order to gain better qualitative insight into the model's generative capabilities. The benchmark also provides functionality to restart the manual assessment sequence by uploading a csv file with partially completed annotation results. This can be used to continue an uncompleted assessment or to cross-check annotations done across a distributed team. In addition to the manual assessment route, the benchmark comes with an interface to easily integrate automated routines to assess uploaded images. In the current release, we provide a baseline set of automated evaluation routines for a subset of tasks (single object representation, multiple object representation, simple arithmetic, counting, negation, conditional generation [only a subset of prompts]). Automated evaluation relies on CLIP (Radford et al. 2021) to test for the presence and absence of objects in images and DETR (Carion et al. 2020) to check for exact counts of objects in images. The benchmark has options to visualize both the manual and the automated assessments and allows users to download their annotation

Cognitive tasks	Explanation	Example prompt
Counting	A specific count of one type of object.	5 apples
Multiple objects	A list of objects / animals without any particular relationship.	Yarn, robin, mouthwash
Simple arithmetic	Two objects appearing with a specific count where the count of one is defined by an arithmetic operation.	Four pizzas and twice as many cups
Directed actions	A human or animal interacting with an object.	Pig touching a guitar
Spatial relationships	Two animals or objects in a specific spatial relationship to each other. The arrangements are physically possible.	A frisbee under a laptop
Unusual arrangements	Two animals or objects in a defined relation with each other. The way they interact is strongly implausible in a real world scenario.	A bus wearing a tutu
Conditional generation	Two possible output descriptions are provided and the prompt specifies which of the two should be generated.	A jar or a squirrel - only render containers
Compositional characteristics	Generate an animal or an object with a specific requested characteristic.	A shoe with a purple sole
Negation	Generate an object or a scene with an additional description of a feature that must be absent from the output image.	A kitchen without a stove
Chimeras	Specifying an imaginary animal as a composite of body parts of different animals.	A rabbit with the antlers of a moose and the tail of a horse
Text rendering	Specifying an object with an additional characteristic which comes in the form of written text.	A billboard advertising "Visit the Grand Canyon"

Table 1: List of cognitive tasks used in benchmark.

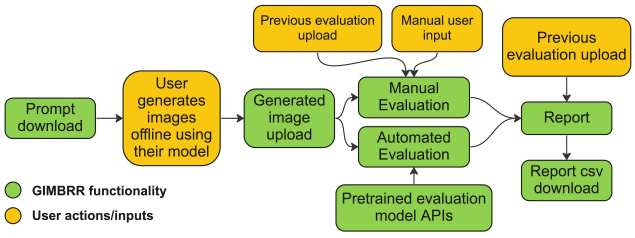

Figure 1: Workflow of using the benchmark software.

results. To allow the benchmark to adapt to new capabilities of generative image algorithms, it is easy for the user to add new sets of prompts, tasks or automated evaluation routines. The GitHub repository contains a detailed guide on how to customize the benchmark software to the user's needs. All of the following evaluations and analyses are based on the manual annotation procedure. Comparison with results from the automated assessments will be the topic of future reports.

Assessment procedure

Now we want to use our benchmark to evaluate the capabilities of common image generation algorithms. Specifically, we evaluate Stable Diffusion (v1.5 and v2.1), DALL-E 2 (version available in February 2023) and Midjourney (v4, as available in February 2023). Images for the set of 2102 prompts discussed above were generated using the available APIs with their default parameters. For manual assessment, we collaborated with professional human annotators from Mindy Support (https://mindy-support.com/) to assess all generated images using our software (v0.0.5). All images were annotated once by a team of annotators, and then 20% of all images were randomly selected for Quality Assessment to check for accuracy of evaluations. The dataset, exact set of prompts, used instructions for annotators, and analyses will be made available ahead of the symposium in the assessments section of the GitHub repository.

Assessment results

Across the entire benchmark we find that DALL-E 2 performs the best (54% of prompts matched), followed by Mid-

journey (46% of prompts matched), followed by Stable Diffusion 2.1 (34% of prompts matched). These overall results of our assessment are split into the different cognitive tasks in Figures 2 and 3. We see that there is a lot of variance in generative capabilities across tasks and across algorithms. Importantly, we do not observe near-perfect performance on any of the tasks, highlighting that the tasks selected here indeed pose a meaningful challenge to currently available algorithms. All of the evaluated models tend to perform best in representing specific single objects and also generating single objects with specific additional characteristics ("Compositional characteristics"). The models struggle the most with prompts that involve text rendering and simple arithmetic.

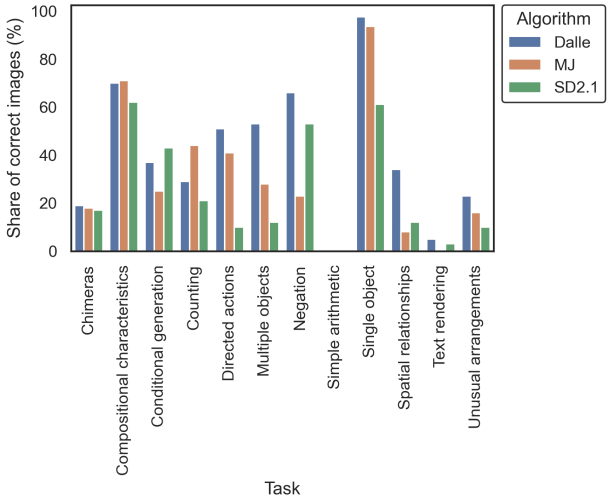

Figure 2: Performance on cognitive tasks across algorithm types (Dalle = DALL-E 2, MJ = Midjourney, SD2.1 = Stable Diffusion v2.1).

In designing our tasks, we were not only interested in providing summary metrics for a model's performance. Instead we believe that with models developing wider spectra of capabilities, it is essential for benchmarks to provide deeper insights into how a given model succeeds or fails in order to inform the training pipeline of future models.

As Figure 2 shows that current models struggle the most with simple arithmetic, we want to analyze this failure in more detail. We use the set of controlled numerical prompts, which break simple arithmetic down into (multiple object) counting and general object representation while controlling for the objects used across these tasks. Figure 4 shows the model performance on these additional controlled tasks. We observe that these models not only fail on simple arithmetic prompts but also lack the ability to separately count two different object types. As such, models not only struggle with generating "Four oranges and twice as many apples" but they also struggle with "Four oranges and eight apples", highlighting that models currently can not scale their ability to count to multiple object classes and hence also cannot solve the simple arithmetic task.

When studying cognitive performance in humans, we can

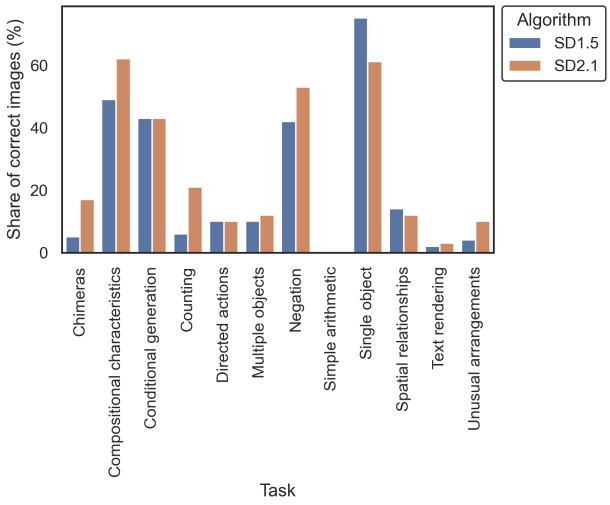

Figure 3: Performance on cognitive tasks across Stable Diffusion versions.

gain important insights through studying the relationship between task performances (Simpson-Kent et al. 2020; Fried and Cramer 2017). The most direct way of doing this in our benchmark is to study whether being better at generating single objects has a trickle-up effect on solving more complex tasks. Figure 4 shows the observed performance in complex tasks with how good an algorithm is in generating the single objects used in the more complex tasks. Across all tasks and algorithms, we observe no significant relationship in a linear regression ($p = 0.09$). The data is depicted in Figure 5. Note that for this analysis we exclude tasks that models are categorically unable to solve (simple arithmetic, text rendering). Our data allows us to also run a similar analysis with tasks which are more closely related. When relating model performance on the multiple objects task with other tasks that require rendering multiple objects (directed actions, spatial relationships, unusual relations) we observe a positive relationship ($b = 0.6, p < 0.01$). When we test this relationship with tasks that explicitly do not require rendering multiple objects (conditional generation, compositional characteristics) we find no significant relationship ($p = 0.78$). This suggests a potential positive trickle-up effect of multiple object rendering capabilities. When testing for this using an interaction effect of "relatedness" and multiple object performance we do not find a significant interaction ($p = 0.16$) but instead a significant positive main effect of performance on the multiple object task ($b = 0.1, p < 0.05$). Figure 6 shows the data used in these analyses. This would suggest that there is a relationship between tasks but that this is more about a general skill that appears to be necessary for generating more complex images, regardless of the specific relatedness of tasks. Overall the picture of specific relationships between task capabilities remains inconclusive for the relatively small sample size used in this study (residual dfs of interaction model = 16) and there is a possibility that these dependencies will become more clear as more models are

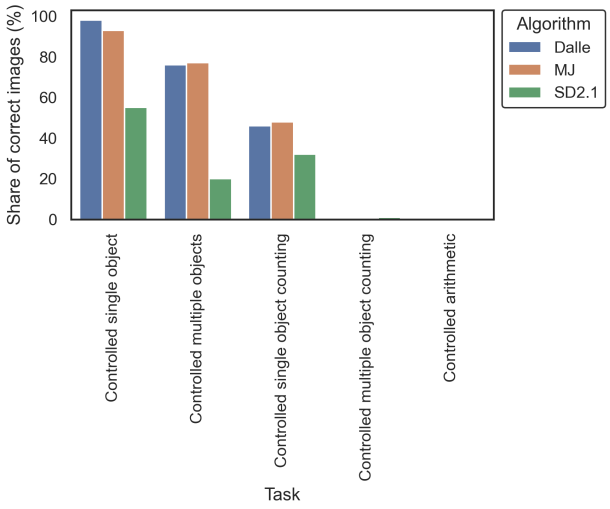

Figure 4: Performance of different algorithms on numerical tasks which control for the objects used across tasks.

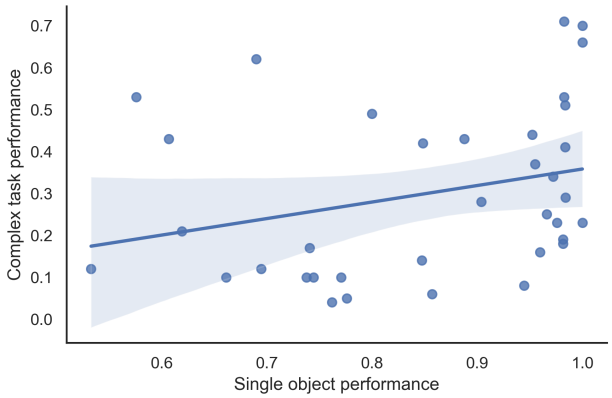

Figure 5: Relationship between performance on complex cognitive tasks to performance on generating single objects used in these complex cognitive tasks. Shaded area around the linear model is 95% confidence interval via a n=10000 bootstrap.

tested over time.

Many of the prompt categories were set up to control for additional factors which we believed would influence the performance of algorithms. For example in counting we can observe how larger numbers of objects are more challenging to generate (Figure 7). We find a related effect for multiple object prompts where algorithms struggle more with prompts describing a larger set of different objects to be generated (Figure 8). Interestingly there seems to be an additional interaction effect: When the objects in the multiple object prompt are commonly seen together in images (e.g. fork and knife), the increasing number of objects has a smaller impact on performance than when the combination of objects is uncommon (e.g. mouthwash and helmet). There are a variety of additional interesting relationships in other tasks and the GitHub repository provides a more detailed look at these factors.

Discussion

We introduced a benchmark for systematically evaluating generative image models. Our initial assessment of three prevalent algorithms reveals significant challenges for SOTA models, suggesting that our benchmark offers valuable guidance for improving these algorithms. We release an open-source evaluation suite alongside this paper, which is easily adaptable with new tasks and automated assessment routines, potentially eliminating the need for human annotators.

The design of the current set of tasks in the benchmark actively takes cognitive theory into account by probing models on prompts which can be considered representative of cognitive skills in humans. This allows us to not only identify skills that models are missing but also to understand why models are failing to perform well on a given task. As an example, we show that current models' problems in solving our simple arithmetic task partially arises because of their inability to count using multiple object classes, rather than only the arithmetic operation itself. This cognitive perspective on model benchmarking also allows us to potentially understand the relationship between different tasks and the corresponding cognitive capabilities. While current data only provides an inconclusive picture of these relationships, we do find that cognitive challenges such as requiring larger numbers of objects to be generated do predictably yield a drop in model performance. We believe that as models and algorithms become more powerful and domain-general, benchmarks will increasingly need to go beyond a summary metric on model performance, but should instead provide a profile of models' cognitive capabilities. With the benchmark presented in this paper we take a first step in this direction.

Acknowledgements

We thank the teams at AlephAlpha and Stability AI, as well as the Intel Labs EAI team for their discussion and feedback. We separately thank Stability AI for providing API credits to generate multiple image batches for evaluation. We thank Mindy Support for their data annotation support service and feedback on the benchmark software. J.A. is supported by a Gates Cambridge Scholarship and UKRI MRC funding and as a result has applied a Creative Commons Attribution (CC BY) license to this manuscript for the purpose of open access.

References

Akkus, C.; Chu, L.; Djakovic, V.; Jauch-Walser, S.; Koch, P.; Loss, G.; Marquardt, C.; Moldovan, M.; Sauter, N.; Schneider, M.; Schulte, R.; Urbanczyk, K.; Goschenhofer, J.; Heumann, C.; Hvingelby, R.; Schalk, D.; and Aßenmacher, M. 2023. Multimodal Deep Learning.

Betzalel, E.; Penso, C.; Navon, A.; and Fetaya, E. 2022. A Study on the Evaluation of Generative Models. *arXiv preprint arXiv:2206.10935*.

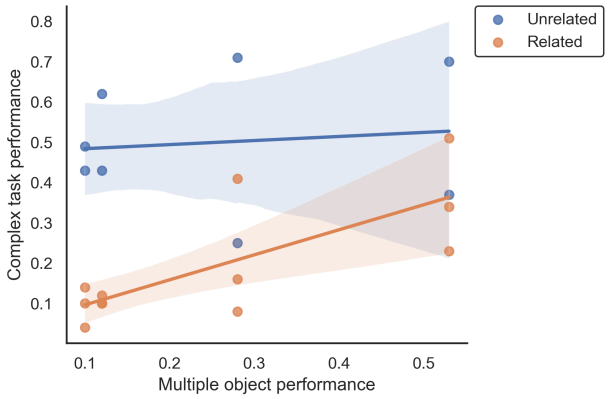

Figure 6: Relationship between performance on the multiple object task and tasks which either require generating multiple objects ('Related', namely directed actions, spatial relationships, unusual relations) or explicitly do not rely on generating multiple objects ('Unrelated', namely conditional generation, compositional characteristics). Shaded area around linear models is 95% confidence interval via a n=10000 bootstrap.

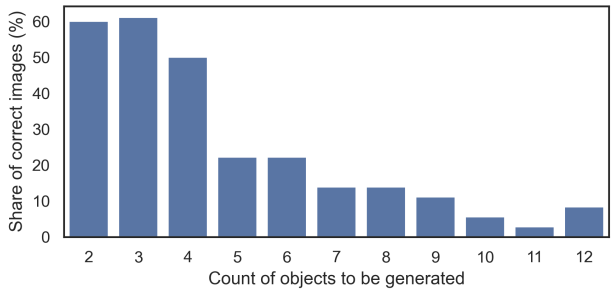

Figure 7: Performance on the counting task where prompts are split by the exact number of the specific object type that models are being asked to generate. There is a difficulty effect of generating larger numbers of objects. The data is pooled across all models.

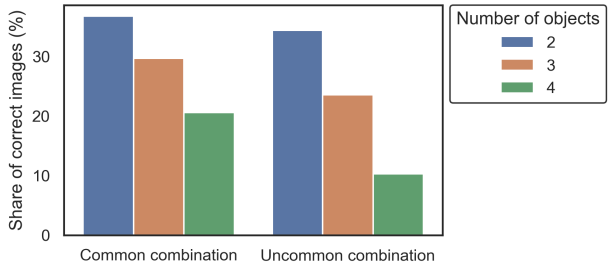

Figure 8: Performance on the multiple objects task where prompts are split by the number of different objects in the prompt and whether the combination of objects is common or uncommon. There is a difficulty effect of generating larger sets of objects and of generating uncommon combinations of objects. The difficulty effect of larger sets increases for uncommon combinations of objects. The data is pooled across all models.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *CoRR*, abs/2005.14165.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. *CoRR*, abs/2005.12872.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Efrat, A.; Honovich, O.; and Levy, O. 2022. Lmentry: A language model benchmark of elementary language tasks. *arXiv preprint arXiv:2211.02069*.

Fried, E. I.; and Cramer, A. O. J. 2017. Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology. *Perspectives on Psychological Science*, 12(6): 999–1020. PMID: 28873325.

Hao, Y.; Chi, Z.; Dong, L.; and Wei, F. 2022. Optimizing Prompts for Text-to-Image Generation. *arXiv preprint arXiv:2212.09611*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Khashabi, D.; Stanovsky, G.; Bragg, J.; Lourie, N.; Kasai, J.; Choi, Y.; Smith, N. A.; and Weld, D. S. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*.

Kievit, R. A. 2020. Sensitive periods in cognitive development: A mutualistic perspective. *Current Opinion in Behavioral Sciences*, 36: 144–149.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Peng, P.; and Kievit, R. A. 2020. The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives*, 14(1): 15–20.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from nat-

ural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Simpson-Kent, I. L.; Fuhrmann, D.; Bathelt, J.; Achterberg, J.; Borgeest, G. S.; and Kievit, R. A. 2020. Neurocognitive reorganization between crystallized intelligence, fluid intelligence and white matter microstructure in two age-heterogeneous developmental cohorts. *Developmental Cognitive Neuroscience*, 41: 100743.

Spelke, E. S. 2022. *What Babies Know: Core Knowledge and Composition Volume 1*, volume 1. Oxford University Press.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; Van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2022. Large Language Models Are Human-Level Prompt Engineers.