Connecting Language to the World: Development of Natural Language Environments

Jay Orten
jay.orten@gmail.com

Abstract

Generally, artificial agents can only successfully navigate a system insofar as they have some degree of understanding of the system. This understanding can be perceived as a world model. Determining whether or not language models have the ability to develop and act within a world model is significant because it determines how we react to and utilize language models. We explore ways to test the possibility of world models, and how language models may be able to establish world models via the development of natural language environments. Furthermore, we propose specific research directions for these objectives.

1 Introduction

In *Tractatus Logico-Philosophicus* (Wittgenstein, 1922), Austrian philosopher Ludwig Wittgenstein made the following statement:

> *"The limits of my language*
> *mean the limits of my world."*

It is commonly understood that, by this statement, Wittgenstein meant to imply that we can only understand and navigate the world insofar as we can express it through language. Our internal knowledge and usage of language is a corpus, and any corpus of text describes a system, or world model. The closer our world model is to reality, the more fully we can understand reality (Johnson-Laird, 2010).

Such a statement is a thought-provoking proposition in the context of language models. Currently, Large Language Models (LLMs) are trained on massive corpora derived from internet sources, which represent an incredibly complex system. Given this, it may not be surprising that language models can struggle to navigate the system entirely correctly. If, however, LLMs have the capacity to develop world models, then there may be methods to train models that can act more robustly and contain more meaning and understanding in specific environments.

In this paper, we explore the possibility of world models through the lens of natural language environments. In this context, natural language environments can be seen as specialized training data for the specific task of world model exploration and development. To develop background on this approach, we review the current perceptions of meaning and understanding in NLP. With this background, we explore techniques for developing natural language environments. Finally, we propose specific research directions for this area of work.

2 Meaning and Understanding in Natural Language Generation

In order to motivate world model development, we begin with a discussion of understanding in language models. Understanding in this context is a slippery concept largely because it is naturally vague. Additionally, understanding and meaning in artificial agents is a topic of incredible depth and respectable age. The purpose of this paper is not to be an exhaustive review of the field of understanding in artificial intelligence. Rather, we propose a few possible interpretations of understanding and meaning in language models, and invite the reader to form their own judgements.

By one argument, understanding in language is a result of form and meaning. This is the position taken by Bender et al. (2021) with their designation of language models as "stochastic parrots", suggesting that models are nothing more than complex statistical distributions that can parrot learned language in an extremely coherent-sounding but not actually meaningful way. Because models only operate on linguistic form, they lack meaning and therefore possess no actual understanding of the world. As further argued by Bender and Koller (2020), achieving meaning in language is impossible without communicative intent, which language models lack. Communicative coherency is necessarily bound in social and relational context, which connects language to the world. Because language models allegedly possess no communicative intent, the coherency of generational text is entirely determined by our own perceptions. In other words, it's a one sided relationship.

The self-created relationship between user and model is broken when models hallucinate. Hallucinations - when LLM's produce fabricated or completely illogical content - are currently a point of high research interest, and for good reason (Bubeck et al., 2023). A model that hallucinates has the capability to spread false but convincing information (Zhang et al., 2023). This and other limitations are detrimental to LLM's as tools and products. In many ways, hallucinations are the antithesis of understanding.

However, Bender and Koller (2020) also suggest that, for a model that is grounded in the world via data augmentation, "meaning can conceivably be learned to the extent that the communicative intent is represented in that data". They propose that meaning cannot be learned from form alone, but requires additional context surrounding the form.

In rebuttal to Bender and Koller (2020), Piantadosi and Hill (2022) claim that reference to the real world does not necessarily determine meaning. For instance, we can imagine abstract, impossible, or nonexistent concepts that can still hold meaning for us, such as "aphid-sized accordion" or "king of San Francisco". Meaning, rather, comes from relationships between concepts, and this can exist without referents. In this way, significant meaning can be derived from semantic structure alone, as long as that structure is adequately well-defined.

In order to center the conversation around research directions, Bisk et al. (2020) frames the problem of natural language understanding in the structure of five World Scopes. They suggest that language with meaning is derived from our ability to understand the world through all of our senses, our ability to experiment and test the world, and our ability to connect socially with others. While current NLP methods rely solely on large internet based corpora, achieving actual meaning with language generation requires grounding in the world by these additional aspects of perception, embodiment, and social connection. Based on this view, the next natural question becomes: how do we grant natural language models these aspects?

Discussions regarding understanding seem to suggest that meaning and understanding in language are gained from the possession of an internal world model, and in order for language models to function beyond simple pattern machines, they must possess such a world model. A world model is simply some kind of internal representation of the processes that produce training sequences. The idea of developing a world model via grounding is common within the domain of Reinforcement Learning (RL) (Nottingham et al., 2023), and is being explored in the concept of language modeling (Li et al., 2023). However, many approaches to world models in NLP are cast in the perspective of using language models to generate world model information (Wong et al., 2023; Hao et al., 2023; Yoneda et al., 2023), rather than focused on the investigation of the inherent existence of world models within language models themselves.

Given the current body of research, there remain a number of questions of interest:

- Is it possible for language models to contain robust world models?
- Do large language models as they currently exist possess world models?
- How could a world model be detected in a language model?
- What could be done to assist language models in their development of world models?

3 Techniques for Building World Models

In this paper, we approach all of the previously mentioned questions through the scope of natural language environments. As suggested by Bender and Koller (2020) and Bisk et al. (2020), developing meaning in language requires information beyond natural semantic structure. It seems requisite, therefore, that we train language models on additional information representing perception, embodiment, and social connection. This additional training data models a world, which we refer to here as a natural language environment. The power of such an environment is that it can ideally be engineered for any task.

In this section, we propose two primary potential methods of developing world models via natural language environments: grounding through senses, and well-defined modeling environments.

Grounding and Senses Grounding models is not a new area of study. Common in the reinforcement learning domain, grounding is a critical, and often implicit, component of RL that can be applied in a variety of ways in order to give an agent some connection to it's learning environment.

There has been much study centered around giving RL agents grounding via large language model text generation. A common approach is to use the capabilities of Large Foundation Models (LFMs) such as GPT4 to generate text representing world state (Yoneda et al., 2023), goals for the agent (Nottingham et al., 2023; Ahn et al., 2022; Song et al., 2023), and even code to execute some kind of action in the environment(Singh et al., 2022). These techniques are possible through the internal knowledge that LFM's already contain. More recently, and perhaps most impressively, Voyager combines goal and code generation to create a lifelong learning embodied agent capable of playing Minecraft (Wang et al., 2023).

Currently, there is less research on giving language models themselves further grounding. This is typically referred to as multi-modal NLP, based on the idea that language models learn better in multi-modal environments, like babies (Ororbia et al., 2019). One possible approach is to utilize Variational Auto-Encoders (VAEs) to encode images into sequential data (Ramesh et al., 2021). Such an approach was used by Ha and Schmidhuber (2018) to predict future frames for a RL agent, enabling the agent to learn within it's own simulated dream. Many of these mentioned approaches are older and utilize architectures such as Recurrent Neural Networks (RNNs). The networks used by Ha and Schmidhuber (2018) are relatively tiny compared to the size of large language models currently. A more recent application of multi-modal NLP is GPT4 (OpenAI, 2023), which has the ability to answer questions based on images.

In general, grounding for language models could be accomplished by the inclusion of additional training data representing contextual information, as Bender and Koller (2020) suggest. The ultimate goal would be to give language models additional senses, beyond just language. One fascinating implementation of this is Decision Transformer (Chen et al., 2021), which models reinforcement learning as sequential vectors representing return, state, and action. The predicted actions of the transformer model match or exceed offline RL baselines on certain tasks. Xiang et al. (2023) accomplished grounding by fine-tuning LLM's on data representing embodied experiences that were gathered from a virtual environment.

Well-defined Modeling Environments Currently, large language models are trained on massive corpora containing data with a vast variety in content and quality (OpenAI, 2023; Touvron et al., 2023). In contrast, Eldan and Li (2023) explored the minimal requirements for a language model to generate coherent, reasonable text. They found that a simplified vocabulary and specialized training data yielded a small model that would often outperform it's larger counterparts which were trained on more complicated corpora. This compelling success could be attributed to a simplified vocabulary, or world scope. Such simplification may be allowing the model to more easily build a general world model because it does not need to navigate the incredibly complex, stochastic nature of large text corpora.

Given a limited, completely deterministic environment, a language model with the capability of building a world model should be able to learn the system well enough to interpolate to correct movements within the system that did not exist within the training data. Basic algebra is a perfect example of such an environment. Numeracy to large magnitudes naturally emerges within embeddings (Wallace et al., 2019). It has been found that small transformers can learn a mapping function within constrained digit lengths, but struggle to extrapolate beyond (Lee et al., 2023). Increased performance can be gained by representing arithmetic tasks in more explicit surface representations, but length remains a challenge (Nogueira et al., 2021). There is evidence, however, that neural networks can develop generalized neuronal mechanisms for arithmetic functions (Hanna et al., 2023).

Games like chess exist in environments that are deterministic, well-defined, but still incredibly complex. It's possible to train a language model exclusively on move sequences and observe that the model has practically learned the rules on it's own (DeLeo and Guven, 2022). Because the environment is well-defined, these models have the ability to track the implicit state of objects in the environment over time (Toshniwal et al., 2022). Li et al. (2023) conducted similar tests on a model trained on move sequences from the game Othello and found evidence that the model develops an internal representation of the board state.

In order to develop robust world models, LLM's may simply require environments that are less broad and more deterministic. If language models are capable of developing world models, it should manifest itself in such environments.

4 Experimental Possibilities

Given these approaches (grounding and well-defined environments) to world model development, we propose, in no particular order, a number of possible experiments to test the existence and development of world models in language modeling.

4.1 Vocab Size vs Performance

TinyStories (Eldan and Li, 2023) achieved fluent and even superior performance with models two magnitudes smaller then their comparable counterparts. This was accomplished through simplified training data and a reduced vocabulary. The authors, however, did not explicitly allude to what exactly is responsible for such incredible performance. If we assume that the limited vocabulary is responsible, then it may be interesting to explore the trade off between performance and vocabulary size; do overly complicated vocabularies act as a bottleneck for LLM performance? There is evidence to suggest that this is so (Gowda and May, 2020), but it seems to remain an open question. If semantically simplified training data is responsible for the performance of TinyStories, then our hypothesis of the importance of well-defined modeling environments is only validated.

4.2 State Tracking

Experiments with environments such as chess suggest that language models have the capability to track states and entities of an environment, even implicitly (Toshniwal et al., 2022). Chess is a good environment for state-tracking, but it may be overly complex for the purposes of observing state tracking. Kim and Schuster (2023) tested this ability specifically on smaller examples, finding that models that have been trained on code corpora specifically posses a certain aptitude for non-trivial state tracking. However, the authors only tested three models: Flan-T5, GPT-3, and GPT3.5. It seems requisite that this ability be further tested on models that have been trained from scratch on specialized tasks, with controlled training data, in order to systematically test state tracking ability.

4.3 Arithmetic with Specialized Form

Despite already receiving notable research attention (Qian et al., 2022; Luo et al., 2023; Yuan et al., 2023; Sharma et al., 2022; Zhou et al., 2022), arithmetic abilities in language models remains an open question. Recent results such as Minerva (Lewkowycz et al., 2022) strongly suggest that capable reasoning models are not simply memorizing training data, but developing the ability to generalize in-distribution (Wang et al., 2021). Because algebra is a completely deterministic system that can be limited, it acts as an ideal environment to test world model development. While it's not expected that language models should act like calculators if they do develop internal world models, navigation within such a well-defined system should be observable. To our knowledge, no previous research has investigated arithmetic tasks in the context of natural language environments and world models.

An application of natural language environments in this context would involve further exploration on how different surface representations of simple algebra functions affect the performance of language models learning such a task (Nogueira et al., 2021; Yuan et al., 2023). Furthermore, well-engineered training data could be utilized to conduct controlled tests on model performance with data not present in training.

4.4 Grounding in Natural Language Environments

Granting language models "senses" by encoding contextual data, such as images, positions, or movement appears to be an area with untapped potential. If we approach language models as universal function approximators for sequential data, anything that we can encode sequentially represents an opportunity for developing a natural language environment.

Philosophically speaking, all senses are inherently sequential; it is the natural passage of time that brings meaning to our senses. For instance, vision could be seen as a sequence of frames, where each frame is a member of an incredibly large, nearly infinite vocabulary. Given an environment where the number of possible frames is limited, a language model could theoretically learn to see from the input of sequential frame data. With the addition of additional inputs/outputs in the training data, a model could be taught to make certain decisions based on input representing senses and directions.

For example, suppose a task existed whereby a drone needed to deliver cargo from point A to point B. In order to finish the task properly, the drone must attend to wind speed, direction, and other external factors. As such, a dataset could be constructed of input/output pairs every second, where input consists of external sensor data and instructions, and output consists of actions taken given the input.

This approach, similar to that of Chen et al. (2021), could be adapted and tested on any number of environments and tasks, either virtual or physical. For example, Xiang et al. (2023) gathered experiences of an agent acting in a virtual environment in order to fine-tune a LLM to navigate prompts relating to the environment successfully.

4.5 Winograd Tests with Specialized Data

Winograd schemas, as proposed by Terry Winograd in 1972 (Winograd, 1972), consist of a pair of sentences, such as:

The city councilmen refused the demonstrators a permit because they feared violence.

The city councilmen refused the demonstrators a permit because they advocated violence.

In this example, the test is to determine who "they" refers to in both sentences. As can be seen, the sentences are semantically ambiguous, and answering the test correctly requires a certain commonsense understanding of language and the world that only humans seem to naturally posses. As such, Levesque et al. (2012) proposed that Winograd schemas would be an excellent test for AI systems.

At this point in time, the Winograd Schema Challenge (WSC) has arguably been "solved" by large transformer-based language models (Kocijan et al., 2023). For instance, Sakaguchi et al. (2019) achieved 90% accuracy on the WSC. However, Kocijan et al. (2023); Sakaguchi et al. (2019) note a few possible reasons for this performance, and why the WSC is essentially flawed as a benchmark:

- Researchers became overly focused on conquering the challenge, leading them to narrowly focus on gaming the system rather than developing commonsense reasoning in agents.
- The task is too narrowly defined, meaning it cannot be an effective measure of commonsense reasoning.
- Artifacts and leakage from training corpora means the model is parroting, not understanding.

Effective commonsense reasoning benchmarks are not the focus of this work. However, Winograd schemas, or similar statements, may remain useful insofar as they can be indicators of successful natural language environment navigation in artificial agents. For example, suppose a language model were trained on data representing a highly constrained environment. Within this system, a limited number of entities interact with each other spatially, such as items being placed in a box. If the training data were specific and instructive enough, could the model learn enough about the interactions and properties within the system in order to respond correctly to vague statements about the system that weren't part of the training data?

5 The Problem of Data

Natural language environments rely heavily on the presence of high-quality, specialized data. As such, the experimental possibilities proposed in section 4 are ideal candidates for the application of data-centric AI (DCAI) approaches. In opposition to model-centric AI, DCAI focuses on the power that high quality data has to improve model performance and control model behaviour (Zha et al., 2023b). Rather than focusing on using the same data and changing the model, DCAI focuses on using the same model, and changing the data (Jarrahi et al., 2022). DCAI encompasses a broad range of data-related tasks such as collection, labeling, reduction, augmentation, quality assurance, and more (Zha et al., 2023a). Most importantly, DCAI considers how more capable models can be trained through less data via techniques such as data synthesis (Motamedi et al., 2021).

One possibility for data collection is the utilization of virtual environments. Virtual environments are cheap, easy to use, highly flexible, and friendly to data collection, making them ideal frameworks for testing many of these proposed approaches. For instance, Wang et al. (2023) utilized Minecraft as a virtual environment, and Ha and Schmidhuber (2018) trained models on VizDoom (Wydmuch et al., 2019). Xiang et al. (2023) used VirtualHome (Puig et al., 2018) to generate embodied experiences. Other virtual environments include VRKitchen (Gao et al., 2019) and CHALET (Yan et al., 2019).

The NetHack Learning Environment, based on the classic terminal game, was created as a RL environment that is complex, scalable, and challenging (Küttler et al., 2020). To this point in time, no agents have been able to finish the game, although progress has been made (Hambro et al., 2022). As such, the NetHack Learning Environment is currently positioned as a long term challenge for RL research.

6 Conclusion

Large language models as they currently exist contain some grasp on our reality, although it certainly isn't the entire picture. This incomplete comprehension may account for their lack of performance in certain tasks. The question is, how can we give LLM's a more complete picture of reality? Doing so may require bridging the gap between models and agents.

In this paper, we have reviewed current opinions regarding meaning and understanding within large language models, and whether language models possess the ability to develop internal world models. We have proposed specific research directions to both test the existence and development of world models through the use of natural language environments. It is hoped that such approaches will bring clarity to current research trends.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as i can, not as i say: Grounding language in robotic affordances.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling.

Michael DeLeo and Erhan Guven. 2022. Learning chess with language models and transformers. In *Data Science and Machine Learning*. Academy and Industry Research Collaboration Center (AIRCC).

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Xiaofeng Gao, Ran Gong, Tianmin Shu, Xu Xie, Shu Wang, and Song-Chun Zhu. 2019. Vrkitchen: an interactive 3d virtual environment for task-oriented learning.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

David Ha and Jürgen Schmidhuber. 2018. World models.

Eric Hambro, Sharada Mohanty, Dmitrii Babaev, Minwoo Byeon, Dipam Chakraborty, Edward Grefenstette, Minqi Jiang, Daejin Jo, Anssi Kanervisto, Jongmin Kim, Sungwoong Kim, Robert Kirk, Vitaly Kurin, Heinrich Küttler, Taehwon Kwon, Donghoon Lee, Vegard Mella, Nantas Nardelli, Ivan Nazarov, Nikita Ovsov, Jack Parker-Holder, Roberta Raileanu, Karolis Ramanauskas, Tim Rocktäschel, Danielle Rothermel, Mikayel Samvelyan, Dmitry Sorokin, Maciej Sypetkowski, and Michał Sypetkowski. 2022. Insights from the neurips 2021 nethack challenge.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model.

Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. 2022. The principles of data-centric ai (dcai).

Philip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The defeat of the winograd schema challenge.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. 2020. The nethack learning environment.

Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. 2023. Teaching arithmetic to small transformers.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct.

Mohammad Motamedi, Nikolay Sakharnykh, and Tim Kaldewey. 2021. A data-centric approach for training deep neural networks with less data.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of transformers with simple arithmetic tasks.

Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. 2023. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling.

OpenAI. 2023. Gpt-4 technical report.

Alexander G. Ororbia, Ankur Mali, Matthew A. Kelly, and David Reitter. 2019. Like a baby: Visually situated neural language acquisition.

Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2022. Limitations of language models in arithmetic and symbolic induction.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Mandar Sharma, Nikhil Muralidhar, and Naren Ramakrishnan. 2022. Overcoming barriers to skill injection in language modeling: Case study in arithmetic.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2022. Progprompt: Generating situated robot task plans using large language models.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2022. Chess as a testbed for language model state tracking.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings.

Cunxiang Wang, Boyuan Zheng, Yuchen Niu, and Yue Zhang. 2021. Exploring generalization ability of pretrained language models on arithmetic and logical reasoning.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models.

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, Inc., USA.

Ludwig Wittgenstein. 1922. Tractatus logico-philosophicus. *Filosoficky Casopis*, 52:336–341.

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought.

Marek Wydmuch, Michał Kempka, and Wojciech Jaśkowski. 2019. ViZDoom Competitions: Playing Doom from Pixels. *IEEE Transactions on Games*, 11(3):248–259. The 2022 IEEE Transactions on Games Outstanding Paper Award.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2023. Language models meet world models: Embodied experiences enhance language models.

Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. 2019. Chalet: Cornell house agent learning environment.

Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R. Walter. 2023. Statler: State-maintaining language models for embodied reasoning.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks?

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. 2023a. Data-centric ai: Perspectives and challenges.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023b. Data-centric artificial intelligence: A survey.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning.