Introduction

In the growing applications of language models scientists, engineers, and the general public, users typically act as if these models have knowledge about the world. However, many problems with the accuracy and consistency of this knowledge can be demonstrated, which creates a need for greater scrutiny and understanding into the knowledge behaviors of language models. In one recent example which makes this kind of understanding necessary, OpenAI suggests to use language models to autonomously perform alignment research:

> "Language models are particularly well-suited for automating alignment research because they come "preloaded" with a lot of knowledge and information about human values from reading the internet."[1]

For alignment research and other potentially high impact applications of language models, it is critical to understand the nature of the knowledge encoded in a model before it is used. Existing tests are insufficient for this task. I advocate for the use of knowledge graphs as a formalism to extract knowledge encoded in language models, in order to provide complementary methods to existing language model evaluation techniques.

I begin with an exposition on the motivating problem, which goes up into the section "What are problems with WSC". Then I explain some commonly known problems with language model evaluation tasks up until the section "Better Representations".

What are language models?

To speak about this, some definitions are needed. This is a question about machine learning technologies. The field of machine learning is concerned with constructing computer systems that automatically improve through experience. Machine learning systems approximate a procedure to generate the right output given any input, which improves as it is given more examples of correct input-output behavior[2]. There are many relevant properties of machine learning systems that we will discuss further.

[1] Jan Leike, "Our Approach to Alignment Research," OpenAI (OpenAI, August 24, 2022), https://openai.com/blog/our-approach-to-alignment-research/.

[2] M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science* 349, no. 6245 (2015): pp. 255-260, https://doi.org/10.1126/science.aaa8415.

Language models are a specific kind of machine learning system. In particular, they act as probability distributions of word sequences[3]. Given an unfinished sentence as input, a language model can output the words that are most likely to complete the sentence. This kind of model is trained on samples of natural language, often from scraping text from the internet.

Today, language models are used in chatbots, text generation, translation, text summarization, Code completion. Increasingly effective method of working with language, the technology has potential to find greater applications, inform research in neuroscience and artificial intelligence[4].

In this report we will focus on the properties of large pre-trained language models such as GPT-3, BERT, XLNet. These are called foundation language models, but we will just refer to them as language models. Despite promises to improve the world, the lack of interpretability and transparency pose difficulties for the safe and trusted usage of language models.

Why interpretability of language models is important

Much has been written already about this topic, so we'll just mention a few relevant points. Interpretability for artificial intelligence, roughly speaking, is the practice of understanding how a model works[5]. This can be divided into two separate settings of understanding: Transparency and explainability, also referred to as transparency interpretability and post-hoc interpretability[6].

- Transparency is concerned with how a system functions internally. We can talk about transparency into a system in the three following fashions: Functional transparency (Understanding the algorithms a system uses), structural transparency (Understanding how the system is implemented in code or hardware), and run transparency (Understanding the input data used when the system was run).[7]

[3] Jurafsky Dan and James H Martin. 2000. *Speech and Language Processing : An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*. Upper Saddle River N.J: Prentice Hall.

[4] Hang Li, "Language Models," *Communications of the ACM* 65, no. 7 (2022): pp. 56-63, https://doi.org/10.1145/3490443.

[5] Leilani H. Gilpin et al., "Explaining Explanations: An Overview of Interpretability of Machine Learning," *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, https://doi.org/10.1109/dsaa.2018.00018.

[6] Zachary C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM* 61, no. 10 (2018): pp. 36-43, https://doi.org/10.1145/3233231.

[7] Kathleen A. Creel, "Transparency in Complex Computational Systems," *Philosophy of Science* 87, no. 4 (2020): pp. 568-589, https://doi.org/10.1086/709729.

- Explainability is concerned with how a system behaves when run. This area uses methods of evaluation, visualization, and explanation to better understand model behavior.[8]

Model transparency is a critical need in AI from the perspective of scientists, engineers, and the general public, as present-day large scale models are being created with high capabilities but low interpretability. This lack of transparency hinders our capacity to provide confident explanations for model behaviors and to identify when models are potentially operating on the basis of unreliable dataset artifacts. To enable sufficient trust in our models, it is necessary that we understand the extent and robustness of their capabilities.

Interpretability plays a big role in ensuring that AI systems are aligned. Interpretability tools provide a way to check that an AI system is aligned with what we would like it to do. Furthermore, the insights gained from interpretability can also inform the design of further safer and interpretable systems[9]. As mentioned earlier, OpenAI proposes to use language models to perform alignment research. Even if interpretability tools do not end up working for sufficiently advanced AI systems, the budding use of AI research assistant systems which will assist in alignment research (like https://elicit.org/, and OpenAI's proposal), upon which interpretability does work, means that interpretability tools will ultimately, directly or indirectly, inform concerns about alignment.

Therefore, we take it that interpretability is important. We are concerned with interpretability in language models - because, as mentioned In particular, the ways in which transparency interpretability is lacking.

This write-up is concerned with the functional transparency and run transparency of language models. In particular we advocate for knowledge graphs as a formalism in post-hoc evaluation tasks to better inform research into functional and run transparency. I believe interpretability can be improved through the combination of different kinds of techniques [10]. More specifically, though we seek to inform better interpretation of the knowledge encoded in language models.

[8] Zachary C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM* 61, no. 10 (2018): pp. 36-43, https://doi.org/10.1145/3233231.

[9] "Chris Olah's Views on Agi Safety," AI Alignment Forum, accessed September 12, 2022, https://www.alignmentforum.org/posts/X2i9dQQK3gETCyqh2/chris-olah-s-views-on-agi-safety.

[10] Chris Olah et al., "The Building Blocks of Interpretability," Distill, June 15, 2021, https://distill.pub/2018/building-blocks/.

Why Interpretability of knowledge in language models is important

The concept of knowledge in language models needs clarification. On a high level, we can interpret the knowledge in language models by interpreting the sentences a model generates. We attach information and meaning about the world to words, and language models manipulate sequences of words. Therefore, we can interpret language models as manipulating information about the world. In this fashion, the level of abstraction upon which we interpret language models is by using human semantics to interpret the sentences it outputs. There are certainly other levels of abstraction to interpret the knowledge in language models - for instance by considering the internal weights of nodes in a neural network. However, I contend that post-hoc interpretation of language model output is presently the most practical sense in which to think of knowledge in language models.

To add more clarity to the concept of knowledge, the Stanford encyclopedia of philosophy gives a definition of knowledge as "justified true belief" which we will refer to for greater clarity.

How would one evaluate the justified true belief within a language model? The more straightforward aspect to address is truth. On a high level, language models operate by constructing sequences of words. These words can be given a semantic interpretation (what it means, how it interacts with the meaning of other words) and then the combined meaning of the string of words can be interpreted. Language exists as a form of communication, and only has meaning when interpreted. We can interpret the outputs of language models by assigning to each word the semantic interpretation it normally has for us in everyday language. It is under that interpretation that we evaluate whether the output of a language model is true or not.

Evaluating belief and justification is nearly impossible due to the lack of functional transparency into language models. In assuming language models "justify" or "believe" their outputs, we implicitly ascribe to a language model an internal representation of the world. Animal researchers do this[11]. Given input as words, the language model processes those words in some way using its internal representation of the world, before producing output. By considering language models as having knowledge, we create a framework for interpreting the behavior of a language model within an internal model of the world.

Ultimately, it is practical to consider language models as having knowledge for the simple fact that in most applications of language models, users act as if they have knowledge.

[11] Hilary Kornblith, "Knowledge in Humans and Other Animals," *A Naturalistic Epistemology*, April 2014, pp. 119-141, https://doi.org/10.1093/acprof:oso/9780198712459.003.0009.

Under these assumptions, interpreting the knowledge capacities of language models becomes ever more important as these systems become increasingly prominent within the field of AI as well as the world at large. As mentioned earlier, OpenAI proposes to use language models to perform alignment research. Even if language models do not ultimately factor into advanced AI systems, the budding use of AI research assistant systems which will assist in alignment research (like https://elicit.org/, and OpenAI's proposal), means that the knowledge stored in language models, directly or indirectly, inform important concerns about alignment. Their safe usage today will inform the future of society. We would like that the knowledge stored in language models is accurate and not actively harmful towards society. In short, it is simply just prudent to know what knowledge is in a system, before deploying that system. However, interpreting the knowledge in language models is not a trivial task.

When treating language models as if they have knowledge, problems arise since a language model's internal representation of the world does not necessarily align with a human representation of the world. For instance, the word "strawberry", when interpreted using an average English speaker's representation of the world, refers to a small, juicy red fruit with seeds on its surface. For a system which can be used to generate sequences of words, one can ask whenever the system outputs the sentence "Gavin ate a strawberry", whether or not the system also internally also represents the word "strawberry" as a small, juicy red fruit with seeds on its surface[12]. It is imaginable that the model could interpret "strawberry" in its internal representation of the world as a round orange-colored citrus. Somewhere along the line while the model was trained, these two concepts got mixed up. We also can't rule out the possibility of a system having a completely alien interpretation of the word "strawberry". And the same goes for any other word. Insert citation here.

In the small case this misalignment in meaning is insignificant – this happens all the time in everyday conversation, and gets resolved with a simple verbal suggestion or reference to the dictionary. However, when specifying the goals of complex AI systems, this misalignment with human representations of the world can pose problems as the system must model increasingly complex words and concepts. This can be thought of a concern about the ontology of a system – where ontology is defined as the set of domain concepts and their properties and relationships with each other[13]. The problem of specifying the goals of a system with a potentially misaligned ontology has been termed the ontology identification problem. The programmer of an AI system wants to ensure that the system's ontology enables it to pursue the goals the programmer specified, even as that ontology may be updated [14]. Ontology

[12] "Strawberry," STRAWBERRY | definition in the Cambridge English Dictionary, accessed September 12, 2022, https://dictionary.cambridge.org/us/dictionary/english/strawberry.

[13] Mishra K (2022) Role of Ontology Engineering in Artificial Intelligence. Int J Swarm Evol Comput. 11:237.

[14] Rao Pärnpuu (2016).

identification is a major subproblem of eliciting, which is itself a major subproblem of AI alignment. This is explained in greater detail in this document from ARC.

Some of these misalignments may arise due to the architecture of models or training procedures used. A more fundamental problem for the misalignment of language models is in its training data; what we represent in text is often not representative of the characteristics of the actual world. This is called the reporting bias problem[15]. The fundamental question is, how much information about the world is actually encoded in text? As a hypothetical example consider the fact that people have eyes and the fact that people have spleens. In most pieces of text, we would expect there to be many mentions of people having eyes than people having spleens (beautiful eyes, metaphors about eyes). Respectively, we expect much fewer mentions of people having spleens, because it is less interesting and usually implicit. Without assuming a correct model of human anatomy, we can imagine a language model may encode the fact that it is more likely for a person to have eyes than to have a spleen, even though that's not accurate. People typically report things that are atypical, while normal, everyday things are typically left implicit for the reader to infer.

Now given that we act as if language models have knowledge, and due to reporting bias that knowledge may not accurately represent the goals we want a system to pursue, we return to the issue of transparency; it is difficult to diagnose misalignment of knowledge due to the opacity of language models. Knowledge is encoded in the statistical associations between words defined by the weights of the network, and not easily examinable. We want to identify the knowledge of a language model. It is important that the knowledge is aligned with reality/our goals. Transparency into this knowledge is helpful.

For interpreting knowledge in language models, it is useful to examine their behavior post-hoc. We care about the knowledge encoded by the model when expressed through the language it outputs. Without functional transparency, the only way to examine the knowledge expressed through language is by post-hoc interpretation. That's what researchers have typically done presently. SuperGLUE, a general purpose language capabilities benchmark proposed in 2019, has several subtasks designed for the purpose of evaluating knowledge capabilities[16]. One of these is the Winograd Schema Challenge, or WSC, one of the most well-known tests for commonsense reasoning in language models.

[15] Jonathan Gordon and Benjamin Van Durme, "Reporting Bias and Knowledge Acquisition," *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction - AKBC '13*, 2013, https://doi.org/10.1145/2509558.2509563.

[16] Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "Superglue: A stickier benchmark for general-purpose language understanding systems." *Advances in neural information processing systems* 32 (2019).

How have people historically assessed the knowledge in language models?

In 2011, Levesque and his colleagues proposed the Winograd Schema Challenge, a reading comprehension task for language models, for evaluating the intelligence of a machine- in particular, its commonsense reasoning capabilities[17]. They believed that the existing tests, including Turing and Captcha, were too easy for systems to game, so they designed this challenge as a better alternative. The task looked at hand-crafted sentences with a single ambiguous pronoun, where the goal was to pick the correct referent of the pronoun. One example is:

> The trophy doesn't fit in the brown suitcase because it's too big. What is too big?
> Answer 0: the trophy
> Answer 1: the suitcase

They made sure every sentence had a dual, where a single word could be switched to change the referent of the pronoun, while maintaining the grammatical acceptability of the sentence. This ensured that the referent of the pronoun was dependent on the meaning of the sentence, rather than purely its syntactic structure. The dual of the above sentence is:

> The trophy doesn't fit in the brown suitcase because it's too small. What is too small?
> Answer 0: the trophy
> Answer 1: the suitcase

Further examples can be examined here. As one can imagine, this task is fairly intuitive for people to perform., but nontrivial for machines. Levesque et al contend that the "a system will need to have commonsense knowledge about space, time, physical reasoning, emotions, social constructs, and a wide variety of other domains"[18]. Ever since the beginning of the field of AI, researchers have been interested in is a form of knowledge called "common sense"[19]. This includes facts about the world such as: "if I push a cup off the table it will break"', "The cause of something must happen before its effect", "if I injure someone they will feel pain", etc. However, this kind of knowledge is fairly ill-defined. In some sense, it just refers to the sum of knowledge that a system needs to be intelligent.

A collection of 273 Winograd Schemas became a canonical commonsense-reasoning evaluation task, referred to as WSC273. This evaluation task was striking because of its

[17] Hector Levesque et al., "The winograd schema challenge." In *Thirteenth international conference on the principles of knowledge representation and reasoning*. 2012.

[18] Hector Levesque et al., "The winograd schema challenge."

[19] Ernest Davis and Gary Marcus, "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence," *Communications of the ACM* 58, no. 9 (2015): pp. 92-103, https://doi.org/10.1145/2701413.

naturalness and simplicity. It was also polarizing on the question of whether success on this task actually indicated intelligence or not. For the past decade, this task has inspired the creation of numerous new commonsense reasoning benchmarks (WNLI[20], WinoGrande[21], WinoLogic[22], etc.) and challenged NLP researchers' engineering skills.

Recent years have seen remarkable progress on the WSC. In 2015, a human baseline of 92% was established for the Winograd Schema Challenge[23]. In 2016 a formal competition was held for the WSC, but no systems which entered did better than chance. In May 2019, Kocijan et al. achieved 72.5% accuracy on WSC273 using pre-training. In November 2019, Sakaguchi et al achieved 90.1% on WSC273[24]. Reaching near-human performance on this commonsense reasoning task demonstrates great promise for the knowledge and reasoning capabilities of language models. However, further inspection of the WSC and similar knowledge-assessing benchmarks reveal limitations in these tests as accurate indicators of knowledge.

What are problems with WSC and similar evaluation tasks

In many cases, progress on the WSC can be attributed to poor evaluation, artifacts in the test set, and supervised fine-tuning on particular evaluation tasks[25]. This is not to say that the WSC tells us nothing about the knowledge capabilities of language models, it is just difficult to distinguish between the knowledge behavior being assessed for and other capabilities that the model may have picked up during its training. For example, we find that when models are fine-tuned on the WSC task, they pick up on idiosyncrasies of the format of the task. Indeed, many of the words in the sentences of each task are not even necessary for determining the correct answer. Elazar et al found that a fine-tuned RoBERTa was able to perform above chance (64.88%) even when half of each WSC sentence was removed. Similarly, when instead removing both nouns from each sentence, fine-tuned RoBERTa was also able to perform above chance (60.72%)[26]. This is all compared to a baseline performance of 89.71% on WSC. This indicates the syntactic format of the WSC task encodes a non-negligible amount of

[20] GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

[21] Keisuke Sakaguchi et al., "Winogrande: An Adversarial Winograd Schema Challenge at Scale," *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 05 (March 2020): pp. 8732-8740, https://doi.org/10.1609/aaai.v34i05.6399.

[22] Weinan He et al., "WinoLogic: A Zero-Shot Logic-Based Diagnostic Dataset for Winograd Schema Challenge," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, https://doi.org/10.18653/v1/2021.emnlp-main.307.

[23] David Bender., "Establishing a Human Baseline for the Winograd Schema Challenge." In *MAICS*, pp. 39-45. 2015.

[24] Vid Kocijan et al.,"The Defeat of the Winograd Schema Challenge." *arXiv preprint arXiv:2201.02387* (2022).

[25] Vid Kocijan et al.,"The Defeat of the Winograd Schema Challenge."

[26] Yanai Elazar et al., "Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, https://doi.org/10.18653/v1/2021.emnlp-main.819.

information that can be exploited by language models to pick the correct answer, beyond the semantic information expected to be used by knowledge and reasoning capabilities.

This kind of behavior persists even when disregarding the fine-tuning of models. Simply perturbing sentences in the evaluation task by switching proper names around causes accuracy to drop, in varying degrees up to 10 percentage points[27]. All this points to the high likelihood that language models are using some artifact in the WSC sentences in order to determine the right answer on evaluation tasks, but this doesn't clarify as to what that artifact is. One proposed explanation is the presence of surface-level statistical associations present in the benchmark, which could explain how language models could give the correct sentence such high probability. Trichelair et al found that 13.5% of WSC273 schemas are "associative", that is, solvable with surface level statistical association rather than deeper knowledge [28]. One such example from WSC273, run on GPT2, is as follows:

Sentence	Score
In the storm, the tree fell down and crashed through the roof of my house. Now, I have to **get the tree repaired.**	-2.72
In the storm, the tree fell down and crashed through the roof of my house. Now, I have to ***get the roof repaired.***	-2.68

GPT2 correctly scored the second sentence with a higher score, which is the correct answer. The supposed knowledge necessary for this task should have concerned the fact that the roof was broken by the tree, and therefore needed to be repaired. One could interpret success on this example as evidence for that kind of reasoning being used in the internal representation of the world that GPT2 has. However, there is an easier explanation. The word "roof" is more associated with the word "repair" than the word "tree" is with the word "repair". This means the fact that the phrase "I have to get the roof repaired" is much more probable than "I have to get the tree repaired", regardless of anything else, offsets the score enough for the model to score the second sentence higher overall.

[27] Paul Trichelair et al., "How Reasonable Are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and Swag," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, https://doi.org/10.18653/v1/d19-1335.

[28] Paul Trichelair et al., "How Reasonable Are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and Swag,"

To be sure, understanding that the association between roofs and repairing is stronger than the association between trees and repairing is in fact a form of knowledge. The problem is, this is not the kind of knowledge that the WSC is trying to evaluate. When assessing knowledge from language models, we would only like to consider interpretation of the words outputted under a typical human world-model, as discussed above. The likelihood of particular phrases co-occurring is not typically the kind of knowledge used to reason in this scenario. What constitutes "typical human behavior" in these kinds of situations is ultimately quite subjective, but the general point remains that in evaluation tasks of this kind, it can often be unclear what knowledge is actually being tested for.

Beyond the WSC, this principle applies for the numerous other commonsense reasoning and knowledge benchmarks that came after. Three more examples:

- For reading comprehension tasks, one can completely omit the reading passage and ask the question and maintain performance above chance[29]
- For entailment tests (does sentence A entail sentence B), much of model performance can be explained by the presence of phrases in A repeated in B.[30]
- The creation of NLI datasets involving crowdworkers have numerous spurious correlations and annotation artifacts[31]

Presently, there is no lack of research into the artifacts that show up in evaluation tasks which allow language models to succeed without demonstrating the capabilities being tested. Researchers must take care, when evaluating the results of these benchmarks, to be conscientious of the ways in which language models can succeed on the benchmarks with techniques besides knowledge and reasoning. As general purpose indicators of commonsense knowledge and reasoning behavior, WSC and related benchmarks work well and have remained the standard ways to test for these capabilities. However, when we are interested in interpreting the knowledge capacities of language models for large scale, safety critical applications, these general purpose indicators are likely to prove insufficient. Through these benchmarks we gain evidence a model is able to behave as if it has knowledge. However, the exact characteristics and specific types of knowledge encoded by the models remains opaque, as it is nontrivial to disentangle success on the benchmarks due to deeper reasoning from

[29] Divyansh Kaushik and Zachary C. Lipton, "How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, https://doi.org/10.18653/v1/d18-1546.

[30] Tom McCoy, Ellie Pavlick, and Tal Linzen, "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, https://doi.org/10.18653/v1/p19-1334.

[31] Suchin Gururangan et al., "Annotation Artifacts in Natural Language Inference Data," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, https://doi.org/10.18653/v1/n18-2017.

success due to surface-level correlations. The ultimate goal of interpreting the knowledge encoded in language models requires methods of evaluating knowledge with greater organization, greater capacity for visualization, and greater ability to isolate different pieces of knowledge. To this end, knowledge graphs can serve as a useful formalism for organizing knowledge for improved evaluation of language models.

Better Representations - Knowledge graphs

Knowledge graphs are a concept that was formalized as early as 1972s[32] as a way to represent the transfer of knowledge between a student and instructor. In the broadest sense, knowledge graphs are simply graph-based abstraction of knowledge, where nodes represent entities and edges represent relations between different entities[33]. These are used by Google, Yahoo, Facebook, Wikidata. The principle behind knowledge graphs is the idea that natural text, as a way of storing information, is not structured enough for quick manual access. Representing knowledge in graph form allows for efficient processing by search engine, chatbots, social media, and other industry systems[34]. The potential connection between language models and knowledge graphs was first explored in 2019 by Petroni et al[35], who created a series of probes to see how well language models are able to fill in the relations of a knowledge graph. With this technique, researchers can extract a knowledge graph out from the behavior of a language model.

In essence, knowledge graphs are a well-studied formalism for representing knowledge that have seen wide usage in industry as a way to store information due to their structure and ease with which they can be manipulated and processed. For these reasons, they also provide a useful formalism in the space of evaluating the knowledge in language models.

KG completion and structured knowledge

An in-depth literature review on language models and knowledge graphs will be omitted for use as supplemental material, and only key points will be covered here. Essentially, a knowledge graph consists of triples in the form

[32] E. W. Schneider, "Course Modularization Applied: The Interface System and Its Implications for Sequence Control and Data Analysis.," *PsycEXTRA Dataset*, 1973, https://doi.org/10.1037/e436252004-001.

[33] Aiden Hogan et al. "Knowledge graphs." *ACM Computing Surveys (CSUR)* 54, no. 4 (2021): 1-37.

[34] Dieter Fensel et al., *Knowledge Graphs Methodology, Tools and Selected Use Cases* (Cham, Switzerland: Springer, 2020).

[35] Fabio Petroni et al., "Language Models as Knowledge Bases?", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, https://doi.org/10.18653/v1/d19-1250.

(head, relation, tail)

Where “head” and “tail” are entities, and “relation” represents the relationship between the head and tail. For instance, the following is a triple which represents the fact that the capital of France is Paris.

(Paris, is_capital_of, France)

These triples can be represented in a graph by assigning each entity a node, and connecting nodes with labeled edges as specified by each triple. In this way, we can graphically represent a fairly complex body of knowledge, where each entity can have many different kinds of relations with many other entities. Research done on extracting knowledge graphs from language models have investigated the creation of these triples in the following ways

- Given two entities, using a language model to determine the relationship between them[36]
- Given a relationship, using a language model to find two objects that have that relationship[37]
- Given an object and a relationship, using a language model to find the second object that the given object has that relationship with [38]

In the past 3 years, researchers have been working on developing these extraction techniques and making them more robust and consistent[39]. Much of the research on language models and knowledge graphs has been motivated by the potential of using large language models to augment the knowledge graphs used in industry[40]. However, there is a long way to go before language models can be used as robust stores of knowledge. Language models will need much greater consistency, interpretability, and reasoning skills in order to be used as knowledge bases[41]. While this poses problems for the adoption of knowledge-graph extraction techniques for industry usage, a key insight is that the ways in which knowledge graphs illuminate these flaws of language models actually make them a very effective tool for the evaluation of knowledge in language models.

[36] Chenguang Wang et al., "Zero-shot information extraction as a unified text-to-triple translation." *arXiv preprint arXiv:2109.11171* (2021).

[37] Shibo Hao et al., "BertNet: Harvesting Knowledge Graphs from Pretrained Language Models." *arXiv preprint arXiv:2206.14268* (2022).

[38] Fabio Petroni et al., “Language Models as Knowledge Bases?”

[39] Peter West et al., “Symbolic Knowledge Distillation: From General Language Models to Commonsense Models,” *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, https://doi.org/10.18653/v1/2022.naacl-main.341.

[40] Simon Razniewski et al., "Language models as or for knowledge bases." *arXiv preprint arXiv:2110.04888* (2021).

[41] Badr AlKhamissi et al., "A review on language models as knowledge bases." *arXiv preprint arXiv:2204.06031* (2022).

Analysis of WSC vs KG evaluation

Two papers have caught on to the potential of knowledge graphs for evaluating the knowledge in language models in a broad sense[42][43]. Considering issues of knowledge, reporting bias, and transparency in light of use of knowledge graphs provides avenues for much more effective and targeted evaluation of knowledge in language models. We provide an example of the kind of evaluation that can be done using this formalism:

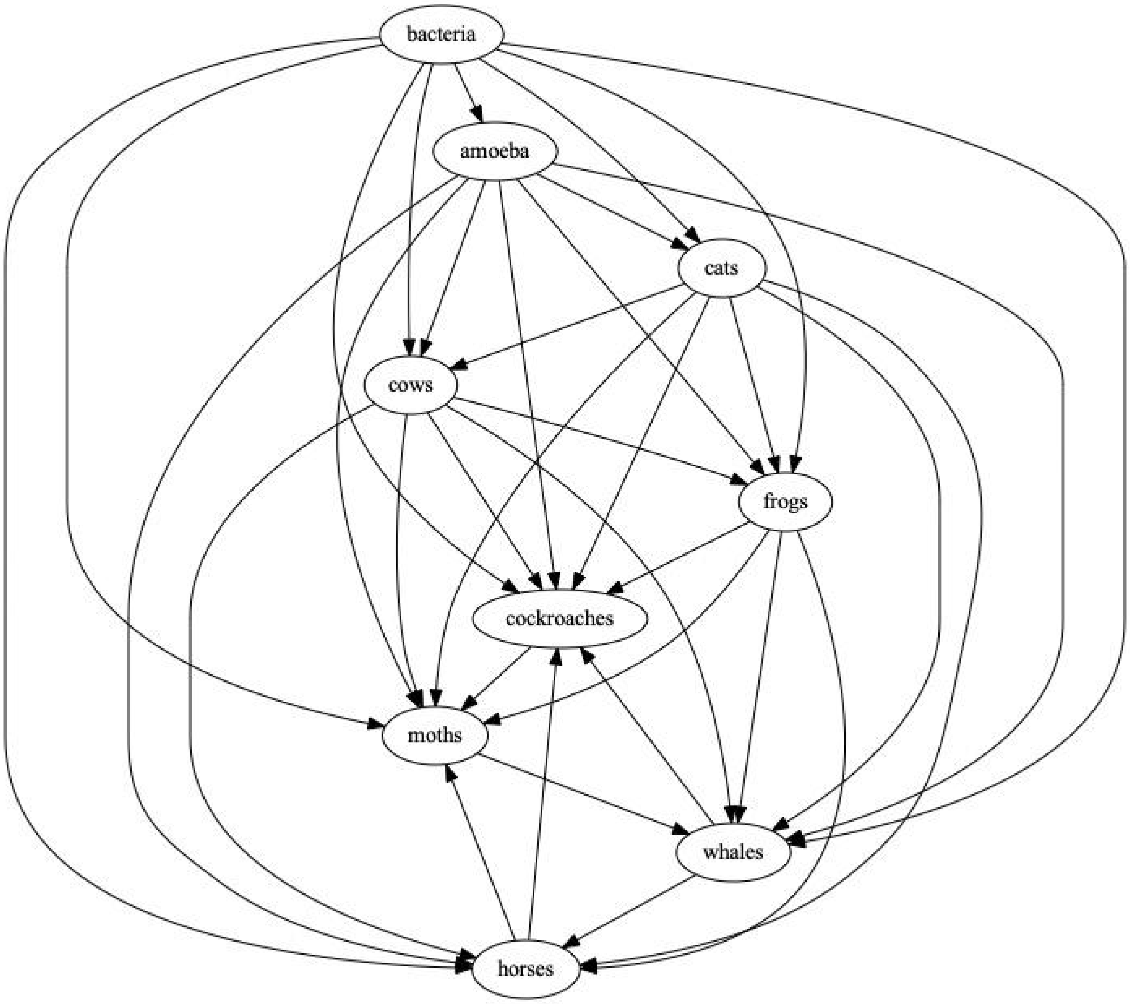
We want to target a very specific kind of knowledge for evaluation: in this case, we chose knowledge about the relative sizes of common organisms in GPT2. This is a simple form of knowledge which can be structured in a knowledge graph easily; nodes are different organisms, and an arrow between organisms A and B indicates that B is larger than A. Additionally, the relative sizes of animals can be verified fairly easily when evaluated, but is a kind of knowledge which falls subject to the problem of reporting bias. In text, most animals are never compared directly in size to each other; extracting a graph which accurately ranks the relative sizes of different organisms suggests knowledge about the sizes of animals which goes beyond surface-level correlations in the text.

To minimize spurious correlations, we use GPT2 to score several sentences of the form “A are larger than B”, “B are smaller than A”, as well as simple paraphrases. If the average score of all the paraphrases of “A are larger than B” is larger than the average score of all paraphrases of “B are larger than A”, then we add the triple (A, is_larger_than, B) to our graph. Here, a hand-picked list of organisms is used for demonstration.

[42] Vinitra Swamy et al., "Interpreting Language Models Through Knowledge Graph Extraction." *arXiv preprint arXiv:2111.08546* (2021).

[43] Shibo Hao et al., "BertNet: Harvesting Knowledge Graphs from Pretrained Language Models."

An extracted graph representing relative sizes of organisms



This gives us a measure of how well the language model encodes knowledge about the relative sizes of organisms. Being able to visualize the relation allows for further investigation. A few observations can be made to prompt further inquiry:

1. Bacteria and amoeba are consistently ranked as smaller than other organisms. This is correct behavior.
2. Whales are ranked as larger than all other organisms besides cockroaches and horses. This is nearly correct behavior
3. Cockroaches and moths are fairly consistently ranked as larger than any other animal. This is incorrect behavior.

In particular, observation 3 illuminates a potential misunderstanding of the relationship of size that is encoded in GPT2. Under the principle of reporting bias, we can hypothesize that text will consistently portray cockroaches and moths as very large. This is due to the fact that people tend to exclaim hyperboles with respect to these insects such as "I just killed a MASSIVE cockroach" or "I just saw a huge moth in the bathroom". In this fashion, it is reasonable for the model to encode cockroaches and moths as having very large sizes relative to other organisms. Further tests beyond knowledge graph extraction, can help shed light on this behavior.

This example provides suggestions for how to use knowledge graph extraction to better evaluate the knowledge in language models. Cognisance of the nature of knowledge and reporting bias combined with the clarity afforded by representing knowledge in a graph-based abstraction allow us to gain deeper, targeted understanding into the knowledge capacities of language models.

We can draw comparisons between the WSC and knowledge graph extraction:

WSC	**Knowledge Graph Extraction**
Difficult to interpret as a whole	Inherently interpretable as a graph
Knowledge tested for in each schema is complex and mixed	Knowledge of a single relation is tested for in each triple
Task is based on the wording of the sentence - paraphrasing would modify the test	Task is insensitive to paraphrase - can minimize spurious correlations by averaging over different paraphrases
Test general knowledge and reasoning capabilities	Scrutinizes very specific relations and types of knowledge

Ultimately, knowledge graph extraction serves as an essential complement to other standard methods for evaluating the knowledge in language models. Knowledge graphs provide a formalism which is visualizable, targetable, and less sensitive to spurious syntactic correlations than other methods of knowledge evaluation. This technique is able to scrutinize very specific kinds of knowledge, and diagnose, on a low level, areas in which the behavior of knowledge in language models does not align with what we would like it to. However, knowledge graph extraction fails to scale up to more complex forms of knowledge and reasoning. For commonsense reasoning capabilities, which are inherently difficult to disambiguate and define, the WSC and related evaluation procedures will remain of greatest use.

Conclusion

Methods for interpreting the knowledge in language models are in need of greater clarity and rigor. Modern applications of language models typically assume they possess knowledge, when in use. However, due to the problem of reporting bias, the knowledge encoded in language models may not accurately represent the world based on the text they were trained on. Problems can arise due to the lack of functional transparency into language models, which makes this misalignment in knowledge difficult to diagnose. While present methods of evaluating knowledge in language models lack the clarity, interpretability, and specificity needed to scrutinize the low-level knowledge of language models, knowledge graph extraction can fill these needs.

Previous research in knowledge graph extraction from language models has been geared towards the potential for language models to augment knowledge graphs for industry usage. Further research can seriously consider these techniques instead in the setting of evaluating the knowledge in language models. In particular, the synergy between this new method of evaluating knowledge and pre-existing ones. Ultimately, interpretability techniques become more effective overall when different methods are used together. And language models are in critical need of more effective interpretability.

Works Cited

AlKhamissi, Badr, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. "A review on language models as knowledge bases." *arXiv preprint arXiv:2204.06031* (2022).

Bender, David. "Establishing a Human Baseline for the Winograd Schema Challenge." In *MAICS*, pp. 39-45. 2015.

“Chris Olah's Views on Agi Safety.” AI Alignment Forum. Accessed September 12, 2022. https://www.alignmentforum.org/posts/X2i9dQQK3gETCyqh2/chris-olah-s-views-on-agi-safety.

Creel, Kathleen A. “Transparency in Complex Computational Systems.” *Philosophy of Science* 87, no. 4 (2020): 568–89. https://doi.org/10.1086/709729.

Davis, Ernest, and Gary Marcus. “Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence.” *Communications of the ACM* 58, no. 9 (2015): 92–103. https://doi.org/10.1145/2701413.

Elazar, Yanai, Hongming Zhang, Yoav Goldberg, and Dan Roth. “Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.819.

Facebookresearch. “Facebookresearch/Lama: Language Model Analysis.” GitHub. Accessed September 13, 2022. https://github.com/facebookresearch/LAMA.

Fensel, Dieter, Şimşek Umutcan, Kevin Angele, Elwin Huaman, Kärle Elias, Oleksandra Panasiuk, Ioan Toma, Umbrich Jürgen, and Alexander Wahler. *Knowledge Graphs Methodology, Tools and Selected Use Cases*. Cham, Switzerland: Springer, 2020.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. “Explaining Explanations: An Overview of Interpretability of Machine Learning.” *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018. https://doi.org/10.1109/dsaa.2018.00018.

Gordon, Jonathan, and Benjamin Van Durme. “Reporting Bias and Knowledge Acquisition.” *Proceedings of the 2013 workshop on Automated knowledge base construction - AKBC '13*, 2013. https://doi.org/10.1145/2509558.2509563.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. “Annotation Artifacts in Natural Language Inference Data.” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018. https://doi.org/10.18653/v1/n18-2017.

Hao, Shibo, Bowen Tan, Kaiwen Tang, Hengzhe Zhang, Eric P. Xing, and Zhiting Hu. "BertNet: Harvesting Knowledge Graphs from Pretrained Language Models." *arXiv preprint arXiv:2206.14268* (2022).

He, Weinan, Canming Huang, Yongmei Liu, and Xiaodan Zhu. “WinoLogic: A Zero-Shot Logic-Based Diagnostic Dataset for Winograd Schema Challenge.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.307.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane et al. "Knowledge graphs." *ACM Computing Surveys (CSUR)* 54, no. 4 (2021): 1-37.

Jordan, M. I., and T. M. Mitchell. “Machine Learning: Trends, Perspectives, and Prospects.” *Science* 349, no. 6245 (2015): 255–60. https://doi.org/10.1126/science.aaa8415.

Kaushik, Divyansh, and Zachary C. Lipton. “How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks.” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. https://doi.org/10.18653/v1/d18-1546.

Kocijan, Vid, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. "The Defeat of the Winograd Schema Challenge." *arXiv preprint arXiv:2201.02387* (2022).

Kornblith, Hilary. “Knowledge in Humans and Other Animals.” *A Naturalistic Epistemology*, 2014, 119–41. https://doi.org/10.1093/acprof:oso/9780198712459.003.0009.

Leike, Jan. “Our Approach to Alignment Research.” OpenAI. OpenAI, August 24, 2022. https://openai.com/blog/our-approach-to-alignment-research/.

Levesque, Hector, Ernest Davis, and Leora Morgenstern. "The winograd schema challenge." In *Thirteenth international conference on the principles of knowledge representation and reasoning*. 2012.

Li, Hang. “Language Models.” *Communications of the ACM* 65, no. 7 (2022): 56–63. https://doi.org/10.1145/3490443.

Li, Hang. “Language Models.” *Communications of the ACM* 65, no. 7 (2022): 56–63. https://doi.org/10.1145/3490443.

Lipton, Zachary C. “The Mythos of Model Interpretability.” *Communications of the ACM* 61, no. 10 (2018): 36–43. https://doi.org/10.1145/3233231.

McCoy, Tom, Ellie Pavlick, and Tal Linzen. “Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. https://doi.org/10.18653/v1/p19-1334.

Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. “The Building Blocks of Interpretability.” Distill, June 15, 2021. https://distill.pub/2018/building-blocks/.

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. “Language Models as Knowledge Bases?” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. https://doi.org/10.18653/v1/d19-1250.

Pärnpuu, Rao. “Ontology Identification Problem In Computational Agents,” 2016.

Razniewski, Simon, Andrew Yates, Nora Kassner, and Gerhard Weikum. "Language models as or for knowledge bases." *arXiv preprint arXiv:2110.04888* (2021).

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. “Winogrande: An Adversarial Winograd Schema Challenge at Scale.” *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 05 (2020): 8732–40. https://doi.org/10.1609/aaai.v34i05.6399.

Schneider, E. W. “Course Modularization Applied: The Interface System and Its Implications for Sequence Control and Data Analysis.” *PsycEXTRA Dataset*, 1973. https://doi.org/10.1037/e436252004-001.

“Strawberry.” STRAWBERRY | definition in the Cambridge English Dictionary. Accessed September 12, 2022. https://dictionary.cambridge.org/us/dictionary/english/strawberry.

Swamy, Vinitra, Angelika Romanou, and Martin Jaggi. "Interpreting Language Models Through Knowledge Graph Extraction." *arXiv preprint arXiv:2111.08546* (2021).

Trichelair, Paul, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Cheung. “How Reasonable Are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and Swag.” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. https://doi.org/10.18653/v1/d19-1335.

Wang, Chenguang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. "Zero-shot information extraction as a unified text-to-triple translation." *arXiv preprint arXiv:2109.11171* (2021).

West, Peter, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. “Symbolic Knowledge Distillation: From General Language Models to Commonsense Models.” *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. https://doi.org/10.18653/v1/2022.naacl-main.341.