Rethinking Reasoning Evaluation with Theories of Intelligence

David Heineman
Georgia Institute of Technology
dheineman3@gatech.edu

Abstract

Due to their linguistic and analytical performance, LLMs have attracted attention outside the NLP community and in recent months many real-world uses of language models have been implemented, despite our incomplete understanding of their intent, capabilities and inherent limitations. I explore work across qualitative studies of LLMs proposed by cognitive psychologists to empirical NLP experiments of reasoning to explain limitations with current benchmarks' ability to measure intelligence. I argue reasoning evaluation must separate generation ability from logical ability by drawing a parallel between widely accepted theories in NLP and cognitive psychology, including separation of form vs. meaning (Bender and Koller, 2020), formal vs. functional competence (Mahowald et al., 2023) and fluid vs. crystallized intelligence (Cattell, 1963). Informed by these theories, I propose a set of recommendations for the measuring rigidly defined theories of intelligence in LLMs which still allow valuable quantitative system comparisons.

1 Introduction

Catastrophic risks in accountability, fairness and bias exist in both overestimating (Bender et al., 2021) and underestimating (Bowman, 2022) the capability of large language models (LLMs). One such capability which has become central to claims of intelligence is analogical reasoning – mapping previous experiences to solve a novel problem. Claims of analogical reasoning are difficult to make without a robust definition, evaluation scheme and analysis to build upon (Mitchell and Krakauer, 2023). While orthogonal work in commonsense reasoning is a well explored topic in NLP (§4), this line of work has no grounding to support theories of general intelligence proposed in the most recent iteration of LLMs (Bubeck et al., 2023). To tie claims of intelligence to work in cognitive psychology, we can first look towards the foundational Turing (1950), a valuable thought experiment for the prerequisites (and debating the possibility) of machine intelligence, yet this work falls short of allowing a machine that 'passes' the Turing test to claim human-like intelligence (Moor, 1976). In fact, a separate line of arguments (e.g., Searle, 1980) have attempted to disprove human-like intelligence can be replicated through showing a human-like machine presents a logical contradiction. While this debate argues whether such 'thinking' machine is *theoretically possible*, recent attention to advances in LLMs highlight a growing need define rigid measurements of intelligence beyond thought experiments. Therefore, I explore how work understanding human thought supports claims about modeling language. I show how existing theories of conceptual representation can help shape our evaluation of LLMs, and highlight the inadequacy of the dominant NLP benchmarks for claims of high-level reasoning or expertise. I argue these theories of intelligence offer a robust framework for designing reasoning evaluation, and can re-shape experimentation in the age of LLMs.[1]

In this work, I begin by evaluating the state of analogical reasoning evaluation in NLP, highlighting a mismatch between current evaluation and the definition of reasoning in cognitive science. Then, I make an argument that reasoning evaluation *must* be independent from the capacity to generate fluent language by exploring three well-accepted perspectives in NLP and cognitive science. Finally, I build upon my findings to propose recommendations for future analogical reasoning evaluation.

[1]In this work, I focus on the GPT family of 'LLMs' (Radford et al., 2018), which are based on the decoder half of the Transformer architecture (Vaswani et al., 2017). At their core, LLMs are trained using a self-supervised objective on billions of tokens (typically web text, e.g. C4) (Raffel et al., 2020) to model the next word in a sequence given some set of previous context words. LLMs are quite successful on this simple objective, and recent work has shown scaling a GPT model leads to syntactically coherent and semantically meaningful outputs (Brown et al., 2020).

2 Current Evaluation Techniques

Early work in analogical problem solving argued reasoning can be interpreted as a heuristic search through some problem space, with large, less structured spaces representing increasingly complex problems (McCorduck and Cfe, 2004). Early artificial intelligence accepted that computing an entire search space is either intractable, or too cumbersome to estimate (Newell and Simon, 1975), and instead developed informed search techniques to guide the space exploration. Early tasks such as chess (Chase and Simon, 1973) or Go (Silver et al., 2016) operate over *bounded* search spaces, and these bounded search spaces have shown to be a helpful litmus tests for arguing a working memory in LLMs (e.g., Noever et al., 2020 for chess). However, reasoning *in the wild* requires developing an underlying representation of logic, and the capability to apply logic to novel, infinite, ill-structured and partially observable search spaces. In this section, we begin by exploring how current benchmarks in NLP capture this notion of analogical reasoning, highlighting the inadequacy of the current evaluation apparatus. Then, we discuss a recent article directly applying cognitive psychology benchmarks to GPT-3 and explain limitations of current work exploring analogical reasoning.

2.1 Reasoning Benchmarks

Reasoning itself is an entire sub-field in NLP, and as models become increasingly more fluent, the goalposts for reasoning have shifted quickly. SQuAD (Rajpurkar et al., 2016) sparked natural language 'understanding' as a task, introducing a dataset of 100K crowd-sourced questions about Wikipedia articles. Since then, reasoning has splintered into a vast number of highly specialized tasks, broadly covering natural language inference, question answering, commonsense reasoning and logical reasoning (Yu et al., 2023). While inference typically involves evaluating entailment and commonsense / QA rely on incorporating world knowledge, logical reasoning is the closest parallel to analogical reasoning. However, most logical datasets are artificial, such as the 200k LogicInference (Ontanon et al., 2022) and are typically either easy or entirely trivial tasks for humans (see examples in Table 1), making them unsuitable benchmarks for human-like intelligence. A separate line of work adapts questions from standardized tests such as the LSAT (Wang et al., 2022) or the Chinese Civil Servant Exam (Liu et al., 2020), with the idea that little domain knowledge is required, yet these are challenging benchmarks even to humans. However, such examples are nowhere near as controlled as those which do not heabily rely on language such as Raven's matrices (see Table 2 in §2.2) and are often synthetically generated. The latter is not an issue when system performance is easy to disambiguate, but recent work has shown reasoning benchmarks may produce different orderings of system quality as models approach human performance (Li et al., 2022). Additionally, as reasoning itself is a broad term in NLP evaluation, attempts to create multi-task benchmarks to argue for reasoning have surged in popularity, such as the GLUE (Wang et al., 2018) and Beyond the Imitation Game (BIG) (Srivastava et al., 2022) benchmarks. In BIG-BENCH, logical reasoning is the second most common task with 58 sub-tasks including identifying logical fallacies, proof verification and even parsing Pig Latin. To make an argument towards reasoning, papers introducing LLMs simply report a combined performance on logical reasoning BIG-BENCH tasks (Chowdhery et al., 2022; Rae et al., 2021). Current claims of reasoning intelligence in GPT-4 side step the question of logical ability entirely. While OpenAI (2023) reports performance on common NLP benchmarks, they also admit to (and report) data contamination, making the stability of their results questionable. In fact, Bubeck et al. (2023) lacks any discussion of analogical or commonsense reasoning, opting instead to benchmark intelligence through GPT-4's ability to answer questions, write code or solve math problems. Judging current evaluation by the criteria in §2.3, current NLP benchmarks fail to meet the prerequisites needed for a proper claim towards logical ability, leading to patchwork claims based on likely over-fit data or tasks which conflate linguistic and functional competence.

2.2 Raven's Progressive Matrices

A study of analogical reasoning, Webb et al. (2022) apply well-known cognitive psychology tasks — chiefly Raven's progressive matrices (Raven and Court, 1998) — to GPT-3, the first such application of cognitive psychology benchmarks to an LLM. They adapt 5 tasks from cognitive science literature: (i) a text-based version of matrix reasoning, (ii) letter-string analogies based on Mitchell and Hofstadter (1990), (iii) four-term verbal analogies taken

Dataset	Family	Task	Example
Entailment (2006)	Language Inference	Textual Entailment	*Sentence 1*: Musk decided to offer up his personal Tesla roadster. *Sentence 2*: Musk decided to offer up his personal car. *Answers*: **Entailment**, Neutral, Contradiction
COPA (2011)	Language Inference	Cause-and-Effect Reasoning	*Question*: Which event caused the other? *Answers*: **(A) It started raining.** (B) The driver turned the wipers on.
SQuAD (2016)	QA	First-order Question Answering	*Passage*: ... In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity ... *Question*: What causes precipitation to fall? *Answer:* **Gravity**
ROCStories (2016)	Commonsense Reasoning	Temporal Reasoning	*Passage:* Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating. *Ending:* **Karen became good friends with her roommate.**
HotpotQA (2018)	QA	Second-order Question Answering	*Passage A*: The 2015 Diamond Head Classic was a college basketball tournament ... Buddy Hield was named the tournament's MVP. *Passage B*: Chavano Rainier "Buddy" Hield is a Bahamian professional basketball player for the Sacramento Kings of the NBA ... *Question*: Which team does the player named 2015 Diamond Head Classic's MVP play for?
ARC (2018a)	Language Inference	Reasoning w/ Domain Knowledge	*Question*: Which property of a mineral can be determined just by looking at it? *Answers*: **(A) luster** (B) mass (C) weight (D) hardness
ART (2019)	Language Inference	Abductive Reasoning	*Observation 1*: Jane was a professor teaching piano to students. *Observation 2*: Jane spent the morning sipping coffee and reading a book. (A) Two of Jane's students were early for their lessons. **(B) None of Jane's students had a lesson that day.**
HellaSwag (2019)	Commonsense Reasoning	Temporal Reasoning	*Prompt*: A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She ... (A) rinses the bucket off with soap and blow dry the dog's head. (B) uses a hose to keep it from getting soapy. **(C) gets the dog wet, then it runs away again.** (D) gets into a bath tub with the dog.
FOLIO (2022)	Language Inference	Deductive Reasoning	*Prompt*: On a shelf, there are five books: a red book, a green book, a blue book, an orange book, and a yellow book. The green book is to the left of the yellow book. The yellow book is the third from the left. The red book is the second from the left. The blue book is the rightmost. (A) The red book is the third from the left. (B) The green book is the third from the left. (C) The blue book is the third from the left. (D) The orange book is the third from the left. **(E) The yellow book is the third from the left.**

Table 1: Notable NLP reasoning tasks, with **answers** when applicable, highlighting the increasing complexity and diversity of reasoning within benchmarks, yet an existing gap between current evaluation and complex tasks such as that demonstrated in Table 2. See Yu et al. (2023) for an exhaustive review.

from the UCLA Verbal Analogy Test and Sternberg and Nigro (1980), (iv) story analogies taken from Gentner and Markman (1997) (testing both near analogies, where entities and domain are shared, and far analogies, where only relations between entites are shared) and (v) analogical problem solving (via the famous Ducker's radiation problem). Examples of each task are included in Table 1. Tasks i, ii and iii can be thought of as structured analogical problems. For example, letter-string analogies (task ii) are simply a *re-representation* task, which rely on synthetic relations between terms. In fact, verbal analogies (task iii) have long been a high-performant task for NLP models built on latent representations (see Drozd et al., 2016), and may play into the strength of the self-attention mechanism hard-coding word relationships. While these experiments are more controlled and are useful in arguing linguistic competence, tasks iv and v are more useful in arguing general reasoning ability. These ill-structured problems require LLMs to both parse language into logic and communicate the solution to replicate human performance. In fact, they found GPT-3 replicated the finding of Gick and Holyoak (1980): it could only solve the radiation problem after being presented with the castle & invasion problem first.

Unlike traditional NLP experiments, their largest test set size is 60 examples, orders of magnitude smaller than even early reasoning benchmarks like ARC (Clark et al., 2018b) with 7K examples. Additionally, the human baselines were presented similar to cognitive psychology experiments, with careful selection and control of participants, which led to much more stable results than could be achieved by crowd-sourcing (Karpinska et al., 2021). This paradigm of smaller, expensive and highly-controlled testing is a helpful blueprint for testing complex abilities. While staple abilities in NLP like multi-hop QA or translation are directly observable, reasoning is difficult to capture (or even separate from the ability to generate language, see §3) and future LLM evaluation can benefit by approaching LLMs similar to human subjects. This work has shown a high quality experimental setup makes a more grounded claim towards reasoning than a large, but synthetically generated or crowd-sourced dataset, a departure from the primary evaluation paradigm in NLP.

2.3 Limitations of Current Benchmarks

Considering the current state of evaluation, we identify three overarching limitations to constructing a rigorous claim of reasoning ability in LLMs:

Loosely defined reasoning. Current work does not attempt to separate the mechanisms used in analogical reasoning. Bommasani et al. (2021) organizes this reasoning ability into three processes: (1) *universality*, a latent, domain-independent reasoning ability, (2) *grounding*, the ability to convert a novel problem into a set of universal logical symbols, (such as those discussed in Larkin and Simon, 1987) and (3) *generativity*, the ability to convert symbolic representations back into language. With this interpretation, a model of analogical reasoning requires grounding to convert a problem to its underlying logical structure, universality to process the logical problem and generativity to map the solution back to the original problem space. Current studies have yet to isolate these abilities, and such a study could highlight a specific design limitation.

Data contamination and task complexity. While extensive research has explored complex, intentional human reasoning (Miller et al., 1960), higher-level problem solving in LLMs has yet to be thoroughly understood (e.g., proving theorems, building complex software). Such an experiment would be incredibly costly to explore, as it would require building a unique symbol system alien to the training data of an LLM (e.g., a novel dataset of mathematical theorems). Existing tests for complex reasoning can be used, but as researchers have no way of searching LLM training data, no clear methodology exists to ensure they have not trained on reasoning tests. Additionally, analogical ability is thought to be a unique by-product of scaling model size, so training a custom, smaller model is an infeasible solution. Current work simply admits some data contamination exists (including Webb et al., 2022), but either restricted training data or cleverly engineered test data is needed to prevent contamination.

Unregulated language exposure. Following extensive work arguing syntactic generalization in LLMs, reasoning evaluation can benefit from a controlled training setup. For example, to propose a fair comparison between LLMs and humans, Yedetore et al. (2023) rely on the *Poverty of the Stimulus Argument* (Chomsky et al., 2011), which highlights that children do not receive enough linguistic information to learn *every* grammar rule, yet they demonstrate syntactic generalizations, and thus implicitly learn grammar through mere exposure to language. In contrast, syntactic ability in LLMs may be a bi-product of the sheer amount of different parse trees encountered in training, rather than robust human-like syntactic generalization. In their work, Yedetore et al. (2023) trained a small language model on a similar number of tokens and distribution of topics as a child would likely be exposed to in different stages of development. If a model design could learn syntactic generalizations similar to a child, then it would demonstrate similar performance on these tasks. Their evaluation setup created a test set of familiar parse trees with semantically unlikely words and performed basic linguistic tests on subject-verb agreement, filler-gap dependencies, and anaphora resolution. The LLM with child-like language data either outperformed or matched human baselines with a similar language exposure. While the study only claims LLM designs are capable of syntactic generalization, they demonstrate arguments for human-like ability can be made by modeling human-like language acquisition. As claims of human ability rely on demonstrating a model can generalize, tests of intelligence must be careful about placing strict constrains on the training setup.

	Task	Example
i	Text-based Raven's Progressive Matrices	[3] [5] [7] [1] [3] [5] [5] [7] **[1]** [_ 7] [_ 7 4 _] [4 _] [9 7] [9 7 4 8] [4 8] [9 _] [9 _ _ 8] **[_ 8]**
ii	Letter-string Analogy	accept : approve :: comfortable : ? unhappy, upset, **pleasant**, disappointed touch : robust :: colossal : ? minimum, diminutive, petite, **gargantuan**
iii	Four-term Verbal Analogy	a b c → a b c cool cool warm → **cool cool warm** b c d e → a c d e a d c b e → **a b c d e** a b c d → a b c e i i j j k k l l → **i i j j k k m m**
iv	Story Analogy	*Source story presented with near / far analogies*
v	Analogical Problem Solving	*Ducker's Radiation Problem, presented with relevant or distractor stories*

Table 2: Examples of tasks used in Webb et al. (2022).

3 Evaluating Reasoning, Not Generation

In this section, we discuss three widely accepted theories about the separation between linguistic and analytical ability, and argue evaluation must distinguish between these abilities.

Form and Meaning. In their seminal work, Bender and Koller (2020) argue *form*, the realization of language, is independent from *meaning*, the relation of form to anything external to language. Using this framework, they show meaning is grounded by *communicative intent*, the real-world goal inhabited by both speakers. While form is governed by syntactic rules and shows whether one utterance is more likely than another, communicative intent and meaning give context to an utterance and allow speakers to relate it to conceptual representations. Under this interpretation, the participation of the listener is crucial to assigning meaning to language, and as LLMs are only given training data with form, this is not a rich enough signal to learn meaning. They draw parallels between their arguments and the Chinese Room Though Experiment (Searle, 1980), pointing out that a speaker translating Chinese cannot learn the meaning of Chinese words by looking at a dictionary alone. Their meaning is connected to the physical, social and mental models represented by the language. While some have debated their assumption that real-world references are required for meaning (such as Piantasodi and Hill (2022), which argues meaning is captured by 'the way concepts relate to each other'), their framework is useful in pointing out that in-depth analyses of LMs often conflate competence in form with competence in meaning. While the two are often correlated, robust evaluation of reasoning must accept no causality exists between better linguistic and reasoning capabilities.

Formal and Functional Competence. Mahowald et al. (2023) recently proposed a separation of analyzing GPT-3.5's ability into *formal competence*, knowledge of syntactic rules, and *functional competence*, knowledge of language use, with both abilities being independent of each other. Their argument draws inspiration from fMRI brain scans taken during reasoning tasks, which show separate activation areas for language, memory, reasoning and social skills (Fedorenko and Varley, 2016). As language draws on the frontal and temporal lobes, this implies human comprehension of language and production of thought are two separate mechanisms. This is further supported by experiments of individuals with aphasia, particularly global aphasia, which impacts the comprehension and production of language. Despite lacking all linguistic ability, these individuals can solve logic puzzles, play chess and perform well on cause-and-effect reasoning tasks (Lecours and Joanette, 1980; Klessinger et al., 2007). The authors then show the hierarchical structure of human language is modeled in LLMs, as evidenced by mastery of non-local features in English. In particular, Futrell et al. (2019) treat an LSTM model similar to a human subject in a psycholinguistic study and demonstrate internal representations exist for a diverse set of complex syntactic structures and Hewitt and Manning (2019) use a probing strategy to show the distance between individual word representations in BERT reflects hierarchical sentence structure. Although this work shows human-like syntactic generalizations may be encoded in LLMs, evidence for human-like reasoning behavior is still disputed (Rogers et al., 2021) and experimental setups similar to syntactic probing have yet to be designed for reasoning. However, Mahowald et al. (2023) highlights that just as we use tools in linguistics to evaluate formal competence, we can use tools in cognitive psychology to evaluate functional competence.

Fluid and Crystallized Intelligence. Unlike the previous two dichotomies, Cattel's theory has been a foundational building block of cognitive science: crystallized intelligence is semantic knowledge from past experiences, and fluid intelligence is the ability to navigate novel situations (Cattell, 1963). This was later incorporated into Baddeley's model of working memory (Baddeley, 1992, 2000), where language and visual processing are crystallized capabilities and attention, processing (such as the phonological loop) and temporary storage are fluid capabilities. Under this interpretation, long-term semantic knowledge is an entirely separate system from logical reasoning and are supported by Baddely's experiments (e.g., Baddeley et al., 1975, 1988) Using Baddely's model of working memory as a blueprint, McClelland et al. (2020) argue modular design is necessary for fluid intelligence in LLMs (echoing real-world multi-modal model designs like Radford et al., 2021). Although modular design may seem beneficial to evaluation — evaluating reasoning would thus entail isolating the part of the model designed after working memory — model designs which separate logical processing from language or visual processing remain unstable to train, and have subpar ability in practice (Elsner and Shain, 2017). In fact, the strength of the self-attention architecture lies in that it did not make assumptions about the linearity of language previously made by the LSTM and RNN architectures (Vaswani et al., 2017). However, this does not rule out the possibility that a similar mechanism to Baddely's working memory is being implicitly learned, and such a dichotomy is useful for designing a well controlled experimental setup.

I have discussed the separation of form and meaning as a means of placing an upper bound on referenceless language learning, the separation of formal and functional competence as an evaluation tool and the separation of fluid and crystallized intelligence as a cognitive theory of intelligence, as well as their implications for evaluating reasoning. The evidence for these theories is diverse: One is supported by logical argument, one by studies of brain imaging and the last by empirical studies of human behavior. Despite their separate goals, these theories establish a common thread: the capacity to *generate* language is decoupled from that required to *reason with* language. In the following section, I show how these theories can build better reasoning evaluation.

4 Building Stronger Reasoning Evaluation

Learning from NLP reasoning benchmarks, analogical thinking in cognitive science and arguments of separation between language and reasoning, I propose a set of recommendations for reasoning evaluation in LLMs:

Clearly scale complexity. Although the dominant paradigm in NLP is to produce a single test set for an ability and interpret a model's performance across all examples equally, cognitive science sets clear boundaries on the difficulty of experiments. For example, Raven and Court (1938) uses varying difficulty levels for the types of matrices they produce, and can show a relationship between task difficulty and ability. Such a relationship would further improve the interpretability and stability of a benchmark, and would allow iterations on the same test set. Additionally, task complexity can scale to far more difficult domains, modeling high-level expert decision making like that studied by Ericsson (2009); Chi et al. (2014).

Test the *same* task across modalities. As Webb et al. (2022) demonstrate Raven's matrices can be re-formulated as a text-only problem, many analogical reasoning studies are free from the context of a specific modality. If the same underlying logical task is produced in many modalities (text and vision are the obvious choices, but arithmetic, sound and spatial reasoning are reasonable candidates as well), perhaps this can isolate the performance of an underlying reasoning mechanism (or argue against its existence). Regardless of whether model design becomes modular, multi-modal setups of the same task can isolate the performance of a learned logical mechanism, and can be used to argue for the utility of different modalities' training data on teaching reasoning ability.

Careful use of multi-task benchmarks. As discussed in §2.1, current multi-task benchmarks are reported under the umbrella of 'logical reasoning' to make claims without a grounded definition of reasoning. In fact, many of the tasks in these benchmarks can easily be gamed with a system demonstrating linguistic competence or world knowledge, rather than one which has a robust reasoning ability. While multi-task benchmarks are critical in organizing NLP datasets and allowing research to compare systems across a vast number of benchmarks simultaneously, they are not (nor claim to be) a stand-in for exhaustive analysis. The current misalignment between experimental designs and definitions of reasoning show current multi-task benchmarks cannot be used to make claims towards the kind of analogical reasoning as it is broadly understood in cognitive science, but as future work will quickly develop a suite of complex reasoning tasks, such a multi-task benchmark is still an opportunity to combine a vast number of different reasoning tests into one measure.

Balance generation & classification. While Webb et al. (2022) demonstrate a strong analysis with primarily classification tasks, a much deeper analysis can be made by testing the ability of open-ended generation. This may take the form of testing spatial reasoning – such as asking which direction a gear will spin in a line of 10 gears (Schwartz and Black, 1996) – or temporal reasoning – such as providing multiple video segments and asking the model what may happen next (Zellers et al., 2022). Classification clearly offers more stable results, yet generation could provide researchers with a richer insight into model decisions. Being careful to avoid anthropomorphizing model outputs, evaluating via generation could create richer benchmarks, recognizing a trade-off exists between stability and insight.

5 Conclusion

As I have shown, reasoning is not a monolithic goal, but an amorphous and multi-faced ability far more complex than is captured in its current state in NLP. By exploring the limitations of current work, as well as the richer body of knowledge about reasoning in cognitive science, I propose recommendations for the design and evaluation of reasoning. I posit that theories about separating syntax and semantics may translate to separating reasoning from language ability and argue this may be an effective assumption to shape evaluation work. As emphasis grows on logical capability, and LLM authors continue to make stronger claims of human-like intelligence, the NLP community has received a unique responsibility to contextualize these claims in the broader context of human cognition.

References

Alan Baddeley. 1992. Working memory. *Science*, 255(5044):556–559.

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423.

Alan Baddeley, Costanza Papagno, and Giuseppe Vallar. 1988. When long-term learning depends on short-term storage. *Journal of memory and language*, 27(5):586–595.

Alan D Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589.

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*, pages 610–623.

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.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Raymond B Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1.

William G Chase and Herbert A Simon. 1973. Perception in chess. *Cognitive psychology*, 4(1):55–81.

Michelene TH Chi, Robert Glaser, and Marshall J Farr. 2014. *The nature of expertise*. Psychology Press.

N Chomsky, RC Berwick, P Pietroski, and B Yankama. 2011. Poverty of the stimulus revisited. *Cognitive Science: A Multidisciplinary Journal*, 35(7):1207–1242.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018a. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018b. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First*

PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers, pages 177–190. Springer.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Micha Elsner and Cory Shain. 2017. Speech segmentation with a neural encoder model of working memory. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1080, Copenhagen, Denmark. Association for Computational Linguistics.

K Anders Ericsson. 2009. *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments*. Cambridge University Press.

Evelina Fedorenko and Rosemary Varley. 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1):132–153.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology*, 12(3):306–355.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicolai Klessinger, Marcin Szczerbinski, and Rosemary Varley. 2007. Algebra in a man with severe aphasia. *Neuropsychologia*, 45(8):1642–1648.

Jill H Larkin and Herbert A Simon. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100.

AndréRoch Lecours and Yves Joanette. 1980. Linguistic and other psychological aspects of paroxysmal aphasia. *Brain and Language*, 10(1):1–23.

Yitian Li, Jidong Tian, Wenqing Chen, Caoyun Fan, Hao He, and Yaohui Jin. 2022. To what extent do natural language understanding datasets correlate to logical reasoning? a method for diagnosing logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1708–1717, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

James L. McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974.

Pamela McCorduck and Cli Cfe. 2004. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

G.A. Miller, E. Galanter, and K.H. Pribram. 1960. *Plans and the Structure of Behavior*. Martino Fine Books.

Melanie Mitchell and Douglas R Hofstadter. 1990. The emergence of understanding in a computer model of concepts and analogy-making. *Physica D: Nonlinear Phenomena*, 42(1-3):322–334.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

James H Moor. 1976. An analysis of the turing test. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 30(4):249–257.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Allen Newell and Herbert A Simon. 1975. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, page 1975. ACM.

David Noever, Matt Ciolino, and Josh Kalin. 2020. The chess transformer: Mastering play using generative language models. *arXiv preprint arXiv:2008.04057*.

Santiago Ontanon, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. 2022. Logicinference: A new datasaet for teaching logical inference to seq2seq models. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

John C Raven and JH Court. 1938. *Raven's progressive matrices*. Western Psychological Services Los Angeles.

John C Raven and John Hugh Court. 1998. *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press Oxford.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Daniel L Schwartz and John B Black. 1996. Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, 30(2):154–219.

John R Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Robert J Sternberg and Georgia Nigro. 1980. Developmental patterns in the solution of verbal analogies. *Child Development*, pages 27–38.

Alan Mathison Turing. 1950. Mind. *Mind*, 59(236):433–460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2201–2216.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2022. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. *arXiv preprint arXiv:2301.11462*.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.