On Reality and the Limits of Language Data

Nigel H. Collier, Fangyu Liu
University of Cambridge, UK
{nhc30,fl399}@cam.ac.uk
Ehsan Shareghi
Monash University, Australia
ehsan.shareghi@monash.edu

Abstract

Recent advances in neural network language models have shown that it is possible to derive expressive meaning representations by leveraging linguistic associations in large-scale natural language data. These potentially Gestalt representations have enabled state-of-the-art performance for many practical applications. It would appear that we are on a pathway to empirically deriving a robust and expressive computable semantics. A key question that arises is how far can language data alone enable computers to understand the necessary truth about the physical world? Attention to this question is warranted because our future interactions with intelligent machines depends on how well our techniques correctly represent and process the concepts (objects, properties, and processes) that humans commonly observe to be true. After reviewing existing protocols, the objective of this work is to explore this question using a novel and tightly controlled reasoning test and to highlight what models might learn directly from pure linguistic data.

1 Introduction

We consider the relationship between the real world and the meaning representations that are derived from large-scale language data using neural network language models (NNLMs) that have been trained using self-supervision, i.e. learning to predict a part of the input from the other parts. The success of NNLMs reflects Wittgenstein’s (1953) philosophy that ‘meaning is use’ yet their drawbacks echo the controversy that exists around the representation of language and reality (Hacker, 2001).

A widely discussed problem in the Philosophy of Language concerns the power of language to represent reality - the world of ‘sticks and stones; cats and trees’ (Devitt and Sterelny, 1999), of ‘molecules, galaxies, and babies’ (Searle, 2010) that we know about from advances in scientific discovery. This question currently finds voice in recent works that try to close the gap between NNLMs and normative human performance on language tasks, particularly those composite benchmarks that seek to establish transfer learning effects (Wang et al., 2019; Srivastava et al., 2022).

Do the limitations of NNLMs arise from data or fundamental properties inherent to their structure? A variety of evaluation metrics (Gatt and Krahmer, 2018) across multiple tasks and models have been pursued to gain insight into this question (Tenney et al., 2019; Rogers et al., 2020). Whilst evaluations have spurred remarkable achievements in many tasks, we argue that they obscure a fundamental question about the limitation of NNLMs to represent reality. We look at these models’ capacity to tell true from false propositions about naïve physics, a subset of what is called common sense reasoning (Da and Kasai, 2019; Rae et al., 2021; Lin et al., 2021), that focuses on the intuitive truth that human cognition endorses about the physical world (Smith and Casati, 1994).

Whilst common sense reasoning has long been considered a part of the landscape of natural language understanding evaluations, e.g. the Winograd Schema Challenge (WSC) (Levesque et al., 2012), the exact kinds of knowledge that model builders should target is only rarely made explicit. For example, McCarthy (1980) in motivating circumscription discussed how only a system with a representation of necessary parts such as oars could logically reason about something wrong with the boat. Without explicit landmarks such as these, the achievement outcomes in our evaluation of common sense reasoning become hard to translate into practice. The characteristic of reasoning with expressed facts along with unstated facts (enthymemes) from a store of com-
mon sense knowledge appears fundamental to human understanding (Singer et al., 1992). It seems that NNLMs will likewise require access to enthymematic mechanisms if they are to make predictions about and correctly explain the long tail of unstated facts not explicitly mentioned in training data and thereby avoid catastrophic errors.

In this paper we triangulate between three synergistic fields: philosophical ontology, cognitive psychology and NLP. A philosophical approach to objective reality is taken to motivate some of the necessary goals of common sense reasoning. We then present a psychologically motivated ‘Google checked’ test, based on analogical reasoning to interpret model capabilities. Analogical Reasoning Tests such as the Miller Analogies Test (House and Keeley, 1996) are widely used and trusted for graduate admission as indicators of human academic achievement. The analogy test focuses on understanding word meanings when presented as ordered pairs, i.e. of the form \(a\) is to \(b\) as \(c\) is to \(d\). The analogy questions are usually presented as verbal problems with minimal contextual information. For example, select one of four possible answers to the question ‘Inception is to conclusion as departure is to (____)’?

Within artificial intelligence, Analogical Reasoning Tests (ARTs) have long been viewed as crucial instruments to elucidate the role of knowledge in problem solving (Minsky, 1975; Lenat et al., 1985; Hall, 1989). Here we present a new bipartite ART test, ART A&B, consisting both of analogy questions in Section A (48 questions) and a structured set of semantic relation questions in Section B (320 questions, covering 16 semantic relations) that underlie the analogy test. The test instrument’s properties were evaluated against standard criteria for consistency (Kuder-Richardson 20 and Cohen’s \(\kappa\)) and variability (ceiling and floor effects). The test is intended for use in zero-shot settings to focus on true generalization from foundation knowledge and avoid both practice effects as well as spurious correlations between the training and testing data.

We use the test responses of 61 participants to construct two normative profiles: the first against human expert judgements and the second against a common sense judgement where at least 80% of participants agreed. We report the results of four GPT-3 language models against the two profiles. To rule out results based on associative behaviour we further examined the results of nonsymmetric semantic relation items where the largest GPT-3 model (Davinci) fell within human norms. The general conclusion we draw is that the three smallest GPT-3 models fall outside both expert and common sense norms. After ruling out associative behaviour, Davinci falls within norms for half the relations but at the lower end or outside human norms for relations that require understanding of mereotopological and affordance relations.

We hope that this test will help to clarify several misconceptions about common sense reasoning in current state-of-the-art NNLMs, draw attention to algorithmic deficiencies, and facilitate coherent strategies for achievement outcomes. Ultimately, since belief informs action, it is a crucial endeavour to ensure the world view of intelligent machines is properly aligned with our own.

2 Related work

Here we briefly review the case for why naïve physics is a core part of common sense, why corpora might be deficient in allowing us to reconstruct common sense, and provide a survey of important common sense benchmarks.

What is common sense? Grice’s theory of pragmatics (1975) says that people avoid communicating obvious information, i.e. truisms which we shall call common sense. However despite its apparently obvious nature, common sense remains an uneasy obscurity in the NLP literature which often describes it via its absence, e.g. as the output of a system that elicits assertions within a large corpus (Schubert, 2002) or as a database (Angeli and Manning, 2014). Arguably, a critical examination of what is common sense knowledge? and its distinction from other types of knowledge feels long overdue if we are to construct productive research questions, e.g. how stable is such knowledge? how malleable are NNLMs for representing such knowledge? A satisfying discussion is beyond the scope of this work but for now we highlight Brachman and Levesque’s (2021) working definition: Common sense is the ability to make effective use of ordinary, everyday, experiential knowledge in achieving ordinary, practical goals.

The definition, with parallels to Gibson’s world of affordances (1979), is ambitious and draws at-
tention to the need for models to act holistically upon common sense knowledge within everyday tasks. We favour this as an end goal but within this complex endeavour (Davis and Marcus, 2015) we are concerned to systematically investigate the essential core of non-holistic common sense knowledge, i.e. commonly held knowledge about physical reality which is capable of scientific treatment. Recent commentaries on human intelligence support this view that an intuitive understanding of the physical world is core in human infants and animals but currently lacking in artificial intelligence (Hassabis et al., 2017; Crosby et al., 2020).

In ontological terms what might obvious physical facts be? It is a difficult question but in the tradition of philosophical realism, Smith (1995; 2004) has argued for a stable and culturally invariant notion of common sense. Following this, the common sense we seek to capture pertains to a subset of objective reality - which exist independent of an individual’s cognition - to which ordinary (non-expert) cognition can relate via contact with the physical environment we perceive.  

Are corpora Gestalt sources of common sense knowledge? A premise in NLP is that language models might be able to recover implicit physical common sense knowledge (Forbes et al., 2019). However, when thinking about facts reported in large-scale corpus data, as Russell noted (Russell, 1914), what might be commented on is not the whole sense at one time but rather the part that individuals single out for attention. This intuitions about the reporting of perception resonates in modern psycholinguistic (Griffin, 2004) and computational linguistic (Viethen et al., 2011) studies that shed light on the complex relationship between internal thought, perception and language production. For example it has been found that fixation priorities may be dynamically altered depending on the task (Castelhano et al., 2009) such as when observers are required to describe a scene (Meyer and Lethaus, 2004).

When we come to analyse factual descriptions arising from perceptual evidence within a population it seems natural that we should expect variation due to individual’s thinking disposition. A question therefore arises: when the pieces of perceptual evidence are combined from a large group of individuals we find on the Web, does a combination of data, the algorithm and the physical computer (Kaplan et al., 2020) allow an NNLM to represent the sum of common sense reality as precisely as humans perceive it from the environment? For example, do these representations allow NNLMs to connect concepts taxonomically? e.g. Dog is a type of Mammal (Davis and Marcus, 2015) Such an important question is almost never explicitly posed in conversational probes of natural language generation (NLG) which tend to focus on overlap metrics (Gatt and Krahmer, 2018) and have often lacked sufficiency for semantic proposition content (Anderson et al., 2016).

**Common Sense Benchmarks.** A substantial body of literature exists that observes how models make predictions on common sense benchmarks. For a recent and comprehensive survey we refer readers to Storks et al. (2019). One of the most influential evaluations is the Winograd Schema Challenge (WSC) (Levesque et al., 2012), a discriminative alternative to the generative probe set out in Turing’s 1950’s Imitation Game (2009). Each question in the test has an ambiguous pronoun and a binary choice of antecedents, for example, The trophy doesn’t fit in the brown suitcase because it’s too big. Options: trophy or suitcase. The example is designed so that humans easily resolve the ambiguous pronoun based on their understanding that part of the physical and social reality of suitcase is its use as a container. Levesque’s goal was in essence to ensure that native English-speaking adults would pass the test easily using a variety of world knowledge and not sentence structural clues. At the same time, in contrast to the Imitation Game, evaluation would be independent of ‘verbal maneuvers’.

The WSC has inspired much work, however it has been criticised on a variety of grounds (Kocijan et al., 2020) such as the high cognitive load required to craft the examples, the limited range of test sentences, the sensitivity of some examples to selectional restrictions, the presence of switchable referents, and the possibility to employ simple associative tricks (Kocijan et al., 2019) based on a high mutual information between a trigger verb and an antecedent. Despite recent applications of adversarial filtration to create Winogrande debiased (WG) a large-scale version of the WSC that is immune to associative solutions (Sakaguchi et al., 2020), the limitations of WSC
call into question the external validity of the test and whether system performance on the WSC is indicative of advances in common sense reasoning (Trichelair et al., 2019). In response to these limitations, alternative benchmarks have emerged which vary greatly in style and complexity such as common sense question answering with Open-Book QA (Mihaylov et al., 2018, OBQA), CommonsenseQA (Talmor et al., 2019, CSQA), Cosmos QA (Huang et al., 2019, CQA), and Commonsense QA 2.0 (Talmor et al., 2022, CSQA2), everyday situation understanding (Zellers et al., 2018, SWAG), and likely causes of sentence endings with the Choice of Plausible Alternatives (Roemmele et al., 2011, COPA). Closest to our work is PROST (Aroca-Ouellette et al., 2021), a recent benchmark focusing on physical common-sense understanding which uses an alternative task construction to the one we present here: a small closed world of objects serves as the source for the generation of multiple choice questions using a set of fixed templates. Ten relations explore model understanding of object attributes and affordances.

As reported in Table 1 (upper part), whilst previous benchmarks have provided insights into common sense reasoning by NNLMs, these insights are often limited to the data set due to their holistic nature and low level of ontological control.

3  Analogical Reasoning Tests A & B

Drawing on objective reality we aim to craft a high-quality physical common sense benchmark that is balanced and highly stratified for its relation types and with verifiable common knowledge. In contrast to previous benchmarks we provide a rigorous evaluation of human common sense norms and variations, select a stable and locally invariant subset and assess the ability of public state-of-the-art NNLMs. Having established a common sense benchmark, we address the evidence for NNLM performance failures where knowledge and/or algorithm combine to display abnormal beliefs. The semantic relations within the new benchmark are described in Table 1 (lower part).

3.1 Two Tasks

Due to its importance in everyday reasoning we expect analogies to be one of the central mechanisms for applying core physical common sense to new situations (Sternberg, 1977; Hofstader and Sander, 2013). Two homemade experimental tasks were employed to establish norms: inductive reasoning with uncued cross-domain analogies (Section A: 48 unique questions), and analogy relations (Section B: 320 unique questions) that underlie the analogies. The questions were designed using expert introspection, collectively agreed after independent answering by the co-authors, and Google checking (§3.3). Both ART A and B had an equal balance of the 16 semantic relations that we discuss below. Overall, approximately half the questions in each analogy and each analogy relation were judged by the paper authors to be Yes (True) and half were judged to be No (False). A simple and consistent syntax was chosen for all items within a semantic relation.

Consider this example of an analogy question: Do congeal and thicken have a similar relationship to whisper and murmur? The expected answer would be Yes (True). Analogical reasoning (Sternberg, 1977; Gentner et al., 2001) requires participants to grasp that the source entities (e.g. ‘congeal’ and ‘thicken’) are related by a specific relationship (e.g. synonymy) based on their shared attributes. Participants then need to map this source analog to a target domain. In this case to ask whether ‘whisper’ and ‘murmur’ are synonyms. In the case of ‘False’ questions, the second pair of words would be related by a different semantic relation from within the set of 16 types.

The following is an example of an analogy relation question: Is coal shinier than aluminium foil? The expected answer would be No (False). In terms of intensity, ‘coal’ is less shiny than ‘aluminium foil’. For analogy relation questions, questions judged True and False by ourselves were considered to represent expert truth and falsehood about objective reality, respectively.

3.2 Semantic Relations

We now turn to consider an inventory of physical common sense relationships that we expect to be present within the core of common sense reasoning. We note that at this stage we cannot be sure that the semantic relations or test items we identify are defining features of common sense. The strength and stability of norms within this inven-

---

4It should be noted that we do expect some human variation due to performance errors (Chomsky et al., 2006) and algorithmic differences (Cherniak, 1990).

5ART A: 25 Yes, 23 No with 8 joint resolutions; ART B: 154 Yes, 166 No with 52 joint resolutions.
Table 1: Some key attributes of modern common sense reasoning benchmarks. N denotes that the relation did not appear to be explicitly controlled for in the paper and vice versa. ¹WordNet included for reference; ²Social common sense corresponds to linguistic/interpersonal intelligence (Humphrey, 2007) and Physical common sense corresponds to logical-mathematical/naturalistic intelligence; ³Expert selection of objects and question patterns followed by automated generation; ⁴Obtained after debiasing; ⁵Hardening of the data set is accomplished through bias reduction such as adversarial filtering; ⁶Hardening of the data set through use of distractor answers; ⁷Hardening of the data set is accomplished through gamification; ⁸Hardening of the data set is done indirectly via restriction to zero-shot setting in order to avoid biasing training to testing data.

Each of the 16 relations is measured by 20 questions which require a Yes/No answer. The choice of questions is based on an intuition that each of the semantic relations contributes to an understanding of the physical world. Each question was broadly intended to surprise readers so that answering would engage both knowledge and processes beyond association and rote learning. For example, comparing two objects for size at vastly different scales or in highly unlikely domains, or considering whether possible attributes such as colour were necessary attributes of an object. Questions were designed also to contain misleading cues, for example asking about whether...
sibling concepts are part of each other. In this respect the questions try to create a mildly hostile environment although not one that is actively informed by adversarial testing on any model.

Note that analogies were not restricted by domain so we regard them as cross-domain, e.g. *Do jeans and cotton have a similar relationship to tree and leaf?* so requires more covert generation of mappings between the pairs of words that cross domains (Barr et al., 2015).

### 3.3 Google Checking

We searched the Web using Google and found that several of the source and target concepts in the set of candidate questions had co-occurring mentions. We therefore chose to eliminate candidates where there was an exact or trivial variation of the question and answer, arriving at the final set of 48 ART A and 320 ART B questions.

### 4 Methods

**Participants.** 63 participants were recruited from Prolific Academic and restricted to those who provided evidence of their education at master degree or higher who were native English speakers resident in the United Kingdom. Two participants were excluded due to non-completion. Hence 61 participants were our subject group. Subject group participants were aged 22 to 68 (mean = 34.4, SD = 10.4) and 13 (21%) identified their gender as male and 48 (79%) as female. This data formed part of the correlation analysis. The participants of the test were given all questions to complete in a guide time of one hour and ten minutes, with a recommendation to take a rest every 40 to 50 questions. Subjects were free to go over the guide time up to a limit of three hours. Analogy questions and analogy relation questions were equally divided among 16 semantic relations (Table 1) and, within each Section, presented in random order to mitigate author’s perceiving patterns of answers. Selection was done so that each analogy and analogy relation question received 61 individual responses. All questions were given to NNLMs in the test.

**Design.** The dependent variable we are testing is the agreement rate for the aggregate test in ART A&B as well as for each of the 16 semantic relations in ART B. Independent variables are test type (analogy or analogy relation), question type (synonym, necessary quality, etc.) and question agreement (i.e. full set or common sense subset).

**Procedure.** In July 2022 each participant read the instructions and completed a consent. For each test we provided four complete and three warm up examples with answers that could be revealed by mouse-over. We did not provide a complete set of examples for each semantic type as it becomes unclear if performance is a result of practice or of true generalization; see (Kramer and Willis, 2002) for a discussion of practice effects on humans. The analogy questions were attempted before the analogy relation questions. The ordering of questions to each participant was randomised to reduce attention bias; this meant that the 16 semantic relations were not grouped within the tests. Questions were administered using Academic Prolific’s platform which linked to an individual sheet. To mitigate random responses, the participants were informed that if they got below 70% accuracy the submission would not be approved. In such cases we asked subjects to re-do the test. In practice all subjects exceeded 70% without being asked to re-do. The participants were asked to use their own knowledge to answer questions and not to use other sources such as the Web. In the case of NNLMs, models, meta-parameters and pre-contextual instructions are those detailed below.

**NNLMs Models.** We used four variants of GPT-3 with different parameter sizes, Ada (350M), Babbage (1.3B), Curie (6.7B), and Davinci (175B). To ensure all four model’s performance reflects their parameter sizes, and not their ability to follow the instructions, we used the conditional probability of “Yes” and “No” tokens given a prompt to select model’s response. The prompts are simple direct questions such as: *Does a missing wheel stop a car from running?* The temperature parameter for models was set to 0 in all experiments, switching off stochastic behavior.

### 5 Results

**5.1 Statistical Properties of ART A&B**

**Equivalence and Gold Standard.** In order to obtain an objective benchmark, a panel of three experts (co-authors) independently annotated the
tests with a Fleiss’s $\kappa$ of 0.75 (8 item disagreements giving moderate-to-substantial agreement with 95% CI) on Section A and 0.77 (52 item disagreements giving substantial agreement with 95% CI) on Section B. Sources of difference were noted and resolved by discussion to produce a gold standard by making minor modifications to the test question or by agreeing a single interpretation. All disagreements could be resolved. Disagreements were due to word sense understanding, e.g. impure as mixed with foreign matter versus morally corrupt, semantic errors in the question, e.g. a comparison to deck versus keel where spatial orientation was important, and the exact degree of probability implied by modal verbs such as will X cause Y, ranging from tentative to required.

**Ceiling/Floor Effect.** We adopted the commonly used threshold to test for ceiling and floor effects. These were considered present if 15% of the human subjects achieved the worst (chance level) or the best score on ART A or B as a whole. A secondary objective was to test for ceiling/floor effects on any subset of ART B. No ceiling or floor effect was observed for either of the tests as a whole. However a floor effect was observed for Has part and a ceiling effect was observed for Order of intensity, Order of size and Cause and effect.

**Internal Consistency.** Since our test items are dichotomous with varying levels of difficulty, reliability was measured using the Kuder-Richardson Formula 20 ($KR_{20}$) on the collectively agreed gold answers. The value for ART A for all questions was 0.72 (moderate/high), and for ART B was 0.87 (high) indicating the adequacy of the tests.

### 5.2 Human Variation on the ART A&B

**Establishing Levels of Agreement.** Table 2 shows human subject performance against the gold standard. The 61 participants who completed the test scored a mean of 36.3 correct answers (76%) on ART A and 260.5 (81%) on ART B. We further examined our data and noted cases where a large majority ($\geq 80\%$) disagreed with the gold standard (2 on ART A and 7 on ART B).

**Relationship between Analogy Task and Analogy Relation Task.** We used the Jarque-Bera non-normality test taking account of both skewness and kurtosis to determined with $p < 0.05$ that human subject scores on both ART A&B were not normally distributed. On inspection both had heavy left tails (Skewness: ART A -0.89, ART B -1.32) and ART B was leptokurtik (Kurtosis: ART A 0.25, ART B 3.60). We therefore employed Spearman rank correlation coefficient and found moderate evidence ($\rho = 0.46$) for dependence between the rankings in performance.

### 5.3 A Common Sense Subset of ART A&B

The first part of our study looked at agreement against an expert gold standard. We cannot however simply assume that different people perceive all the relationships between concepts in our test set in the same way. In the second part we therefore try to establish the common sense norms of naïve physics by establishing the set of questions and answers on which a majority of participants agreed. Note that in the common sense subset, the gold answers are the majority answers chosen by the human subjects. Table 2 shows the number of questions on which participants agreed, both as a composite and grouped by individual semantic relations. Agreement@80%, the value we choose for our common sense subset, shows the number of questions on which 80% or more of participants agreed. After filtering for 80% agreement threshold, 26 out of 48 ART A questions (54%) remained and 228 out of 320 ART B questions (71%) remained for evaluating NNLM models.

### 5.4 NNLMs vs. the Full Set of ART A&B

Having established an approximation of a human norm we compare NNLMs to human performance. Since NNLMs do not exhibit random performance errors due to the environment, we consider that any difference between the NNLM and the human subject reference interval exhibits a potential divergence in rationality. Table 3 indicates the patterns of differences between four scales of GPT-3 models and human subjects on ART A&B. Results highlight both alignment and divergence on naïve physical commonsense understanding due to NNLM data and algorithmic limitations.

**NNLMs on ART A.** Only Babbage had a composite score that was within the low part of the reference range, whilst Ada, Curie and Davinci fell outside the reference range.

**NNLMs on ART B.** All NNLMs composite scores for ART B fell under the reference intervals. For Davinci the composite score is approaching human level performance but this result obscures an uneven understanding across categories. Even for Davinci, Order of intensity, Order of size, Precondition, Necessary quality, and Fiat bound-
Table 2: Human performance on the ART Sections A and B. Agree@N% is the percentage of questions on which a percentage $N$ of participants agreed. Maximum agreement is 48 for the Analogy Test, 320 for the Analogy Relation Test, and 20 for each semantic relation. $^1$ Potential range shows the range of correct answers that are possible for individual participants where the gold standard is the expert committee; $^2$ Observed range show the range of correct answers for individual participants where the gold standard is the expert committee; $^3$, $^4$ Show the mean score and standard deviation over all participants; $^5$ A ceiling or floor effect is reported as No if less than 15% of test takers (9 out of 61) obtained the maximum or minimum possible scores, otherwise Yes; $^6$ A composite score across all the semantic relations.

| Measure                  | Missing % | Potential range | Observed range | $M^3$ | SD$^4$ | Floor/ceiling effect@15% | Agree@80% | Agree@90% |
|--------------------------|-----------|-----------------|----------------|-------|--------|--------------------------|-----------|-----------|
| Analogy Test             | 0%        | 0–48            | 22–43          | 36.3  | 4.6    | N/N                      | 26        | 14        |
| Analogy Relation Test$^5$| 0%        | 0–320           | 199–285        | 260.5 | 15.5   | N/N                      | 228       | 183       |

1. Synonym 0% 0–20 11–20 16.4 2.0 N/N 14 11
2. Necessary quality 0% 0–20 12–19 15.7 1.7 N/N 12 10
3. Associated quality 0% 0–20 10–18 14.8 1.6 N/N 15 10
4. Has part 0% 0–20 10–19 15.5 1.6 Y/N 15 12
5. Order of intensity 0% 0–20 12–20 18.7 1.5 N/Y 19 16
6. Order of size 0% 0–20 14–20 18.8 1.4 N/Y 19 18
7. Cause and effect 0% 0–20 13–20 18.0 1.6 N/Y 16 13
8. Capable of 0% 0–20 12–20 16.2 1.8 N/N 12 12
9. Default inheritance 0% 0–20 12–19 16.7 1.5 N/N 16 14
10. Precondition 0% 0–20 11–20 17.0 1.7 N/N 15 11
11. Boundary (bona fide) 0% 0–20 9–19 14.6 2.1 N/N 13 6
12. Antonym 0% 0–20 9–18 15.2 2.1 N/N 12 6
13. Contained in 0% 0–20 14–20 17.6 1.4 N/N 15 14
14. Has member 0% 0–20 10–19 15.2 2.0 N/N 12 11
15. Troponym 0% 0–20 14–20 17.4 1.4 N/N 16 14
16. Boundary (fiat) 0% 0–20 7–17 13.0 2.1 N/N 7 5

ary show marked divergence to the reference intervals. This outcome supports the view of Aroca-Ouellette et al. (2021) on GPT-1 and -2 that such models are challenged on object attributes (e.g. height) and affordances (e.g. breakability).

A group of GPT-3 tests showed quantitative performance for Davinci at the upper end of the reference interval on Synonym, Associated quality, Has member and somewhat surprisingly Cause and effect and Capable of. This result provides a challenge to the view that only rationalistic (i.e. logic based) language models can be successful in answering common sense questions correctly. Yet we should be cautious in accepting this result at its face value. As we discuss later (§6), models draw on a wide range of associations that sometimes allows them to display seemingly correct reasoning. Such effects however are hard to isolate.

Despite the difference in question formulation, these rather mixed findings might help explain the modest gain in performance from Curie to Davinci that was observed in the zero-shot SAT Analogy performance for GPT-3 (Brown et al., 2020) using Turney et al.’s 374 5-choice data set (Turney et al., 2003). Again we adopt a cautionary note in this conclusion as Davinci model’s best zero-shot performance of 53.7 accuracy compared surprisingly favourably to an estimated 57 accuracy for an average college-bound senior high school student (Turney and Littman, 2005).

5.5 NNLMs vs. the Common Sense Subset of ART A&B

The right block of results in Table 3 shows the NNLM performance on the common sense subset of ART A&B. Reference ranges have unsurprisingly narrowed compared to the full tests, and unfortunately for two relations (Antonym and Fiat Boundary) the lower reference bound falls very close to the random baseline. As shown by the upper value of the potential range, several relations appear to have markedly strong consensus at $\geq 80\%$ agreement among human subjects including Order of intensity, Order of size, Cause and effect, Precondition, Contained in and Troponym. On the other hand Capable of, Bona fide boundary and Antonym had fewer consensus items at $80\%$ and Fiat boundary had the fewest consensus items indicating how disordered the judgements were in this category. The common sense subset once again favoured Davinci over other models. The subset revealed churn on a category by cat-
due to the low level of agreement ary (bone fide)

rudder consisting of parts and boundaries such as perceived as either a single unit or as a structure various granular levels, for example, a boat can be Mereo-Topology.

warrant more precise investigation:

world of which we highlight three in particular that 3 model to reason about objects in the physical algorithmic deficiencies in even the largest GPT-

Our investigation has drawn attention to several

Discussion

Our investigation has drawn attention to several algorithmic deficiencies in even the largest GPT-

Mereo-Topology. Humans understand objects at various granular levels, for example, a boat can be perceived as either a single unit or as a structure consisting of parts and boundaries such as rudder, hull and keel. ART has demonstrated that GPT-3 performs at the lower end or outside the reference interval for Has part, Contained in, Boundary (bona fide). We tend to discount the result for Boundary (flat) due to the low level of agreement among human subjects. Interestingly, GPT-3 performance on demarcating boundaries appeared invariant to model size, indicating that a combination of more language data with ever larger numbers of model parameters is not yielding improvements. Moving forwards, improving the ability of NNLMs to capture in silico the 3-dimensional structures of the real world, including their boundaries, constituencies, and cavities, is a high priority candidate for algorithm development.

Affordance Characteristics. As part of the niche environment that humans occupy, object properties seem to be understood by humans in terms of their affordance characteristics, or in other words the opportunities and limitations that they offer to humans (Gibson, 1979; Smith and Mark, 2003). ART has shown up weaknesses in GPT-3 to understand object affordance characteristics through Necessary quality, Order of size and Order of intensity. Severe performance limitations on Order of size appear consistent with evidence for GPT-3 and other NNLMs using CSQA2 (Talmor et al., 2022) as well as earlier work on BERT for the task of scalar probing (Zhang et al., 2020), and supports the view that there is an algorithmic weakness in the ability of language-only distributed semantic representations to capture affordance characteristics (Glenberg and Robertson, 2000; Jones et al., 2022). Our finding is particularly salient in the light of strong human performance and low human variance for Order of size and Order of

Table 3: NNLM performance on the full set and common sense subset of ART Sections A and B where green shows performance within the reference range and pink outside the reference range. 1 Shows the potential range of correct answers that are possible for individual participants where the gold standard is the expert committee; 2 Shows the reference interval calculated as the range within which 95% of the human subject reference population values fall. 3 Later (§6) we run another test on the sensitivity of the Davinci on the order of arguments in nonsymmetric relations by reversing the order of arguments, and observe the inability of the model to correctly answer Cause and Effect and Troponym questions.

| Measure                  | Full ART A&B | Common Sense Subset of ART A&B |
|--------------------------|--------------|--------------------------------|
|                          | Potential   | Reference | GPT-3    | GPT-3    | GPT-3    | GPT-3    | Potential | Reference | GPT-3    | GPT-3    | GPT-3    | GPT-3    |
| ART Section A            | 0–48        | 28–43     | 22       | 24       | 27       | 0–26     | 18–25     | 9          | 21       | 19       | 15       |
| ART Section B            | 0–320       | 231–285   | 168      | 169      | 177      | 221      | 192–220   | 148        | 135      | 137      | 180      |
| Synonym                  | 0–20        | 13–20     | 11       | 10       | 12       | 17       | 0–14      | 12–14     | 10       | 10       | 8        | 12       |
| Necessary quality        | 0–20        | 13–18     | 12       | 13       | 13       | 11       | 0–12      | 11–12     | 9        | 7        | 8        | 9        |
| Associated quality       | 0–20        | 12–17     | 8        | 7        | 8        | 15       | 0–15      | 10–14     | 10       | 6        | 14       |
| Has part                 | 0–20        | 13–18     | 11       | 11       | 11       | 13       | 0–15      | 12–15     | 11       | 12       | 12       | 13       |
| Order of intensity       | 0–20        | 16–20     | 10       | 12       | 10       | 12       | 0–19      | 16–19     | 10       | 11       | 9        | 11       |
| Order of size            | 0–20        | 17–20     | 10       | 9        | 10       | 11       | 0–19      | 16–19     | 10       | 9        | 10       | 11       |
| Cause and effect         | 0–20        | 15–20     | 9        | 12       | 10       | 18       | 0–16      | 14–16     | 8        | 10       | 10       | 15       |
| Capable of              | 0–20        | 13–19     | 12       | 11       | 9        | 17       | 0–12      | 11–12     | 10       | 3        | 4        | 11       |
| Default inheritance      | 0–20        | 14–19     | 10       | 11       | 13       | 17       | 0–16      | 13–16     | 10       | 11       | 13       | 16       |
| Precondition             | 0–20        | 14–20     | 11       | 8        | 11       | 10       | 0–15      | 13–15     | 11       | 8        | 11       | 10       |
| Boundary (bona fide)     | 0–20        | 11–17     | 11       | 14       | 12       | 12       | 0–13      | 9–12      | 8        | 8        | 8        | 10       |
| Antonym                 | 0–20        | 12–18     | 11       | 5        | 9        | 14       | 0–12      | 8–12      | 7        | 3        | 5        | 9        |
| Contained in             | 0–20        | 15–20     | 10       | 14       | 13       | 13       | 0–15      | 13–15     | 9        | 12       | 11       | 10       |
| Has member               | 0–20        | 12–19     | 11       | 11       | 14       | 16       | 0–12      | 10–12     | 10       | 10       | 10       | 12       |
| Troponym                | 0–20        | 15–20     | 11       | 10       | 13       | 16       | 0–16      | 14–16     | 11       | 10       | 12       | 13       |
| Boundary (flat)          | 0–20        | 10–17     | 11       | 11       | 9        | 9        | 0–7       | 5–6       | 4        | 4        | 2        | 4        |
intensity. Davinci’s apparent progress in Associated quality hints at a situation whereby the largest GPT-3 model can generate potential affordance characteristics for objects, e.g. can a person walk down a bus? but without being able to engage either the inference processes or underlying conceptual representation that underlies human physical common sense. For example, understanding the physical qualities of a bus that enables walking down its central aisle possible such as the frame’s hollow structure, overall size relative to a human body, density, motion and so on.

Yet in other areas of object understanding GPT-3 appears to offer evidence for verbal reasoning on a par with our human subjects. For example in Synonym, Antonym, Has member, Default inheritance, and intriguingly, Cause and effect. We take this finding not as a conclusion but as the next starting point for further investigation.

Stability on Nonsymmetric Semantic Relations. In order to test whether the responses arise from associative fortuity (Pearl, 2009) in the data we tested Davinci’s sensitivity to argument reversal for nonsymmetric relations (e.g. Original: Does a missing wheel stop a car from running; Modified: Does a car stop a missing wheel from running?) and seeing if this changed Davinci’s response. Results of this on 9 semantic relations with nonsymmetric items (i.e., Cause and effect, Order of intensity, Order of size, Has part, Default inheritance, Precondition, Contained in, Has member, Troponym) indicated that Davinci’s response is stable under 7 relations but exhibits potential signs of associative behavior for Cause and effect and Troponym, failing (e.g., replying yes for both of the examples provided above) on 70% of the cases. It could be argued that this behavior might be potentially improved via optimizing the wording of questions to better elicit knowledge from GPT-3, or via a more sophisticated prompting approach, e.g. Chain of Thoughts (Wei et al., 2022). Combining NNLMs with symbolic (Nye et al., 2021) or physical world (Ahn et al., 2022) grounding have been shown to be promising directions for future exploration.

Exceeding Human Performance. We found no evidence in the pre-filtered ART or the common sense subset that GPT-3 is exceeding humans in the zero shot setting. This is unlikely given the structure of the test as by design it provides a relatively low ceiling of human performance.

Limits of Naïve Physics. Some may argue that it is absurd to expect a poll of non-expert participants to reveal the truth about objective reality and indeed we do not make such a strong claim here. We wish to emphasize that naïve physics represents the norms of human beliefs about the physical world. This subtle but important distinction is shown by cases where human subjects showed strong disagreement with the expert judgements we used as our gold standard. Disagreement at the 80% level occurred in 2 questions in ART A and 7 questions in ART B pointing to potential fallacies in the understanding of human subjects and highlighting perceptual biases. Consider the question Are teeth part of a football team? In a literal sense the answer is Yes but echoing Russell’s comment, this is not the part of the whole that most of us would single out for attention.

7 Conclusion

This work builds on foundations that are already firmly established in different fields: the importance of lexical semantic relationships for language processing (Miller, 1995), the importance of conceptual modeling based on basic ontological principles (Smith, 2012), and prior evidence that establishes the characteristics and norms of human reasoning (Stanovich and West, 1998). Our particular interest in this work is to consider how far language data alone can enable computers to understand the physical world. We have questioned therefore the strong empirical assumption of much prior work in NLP about the ability of NNLMs to represent physical common sense using language data alone (Lake et al., 2017) and identified some promising areas for algorithmic investigation and theory formation. Going forwards we hope that the discussion in this paper will be a useful adjunct for intrinsic evaluation of natural language generation models and further to NNLMs that aim to combine machine learning from text with nature-inspired cognitive capabilities.

We anticipate that guiding progress towards the goal of understanding objective reality will highlight relevant insights from what is termed Type 2 processing (Stanovich and Toplak, 2012) within the reasoning and decision making field. Improving algorithms for common sense understanding should help NNLMs go beyond associative-driven behaviours and to achieve analytical goals with more consistent and coherent outputs.
Acknowledgements

We are very grateful to Keith Stanovich who served as a guide on key topics in cognitive psychology.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision, pages 382–398. Springer.

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 534–545.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning about objects through space and time. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4597–4608.

Nathaniel Barr, Gordon Pennycook, Jennifer A Stolz, and Jonathan A Fugelsang. 2015. Reasoned connections: A dual-process perspective on creative thought. Thinking & Reasoning, 21(1):61–75.

Isaac I Bejar, Roger Chaffin, and Susan Embretson. 2012. Cognitive and psychometric analysis of analogical problem solving. Springer Science & Business Media.

Ronald J Brachman and Hector J Levesque. 2021. Toward a new science of common sense. arXiv preprint arXiv:2112.12754.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Monica S Castelhano, Michael L Mack, and John M Henderson. 2009. Viewing task influences eye movement control during active scene perception. Journal of vision, 9(3):6–6.

Christopher Cherniak. 1990. Minimal rationality. MIT Press.

Noam Chomsky et al. 2006. Language and mind. Cambridge University Press.

Matthew Crosby, Benjamin Beyret, Murray Shanahan, José Hernández-Orallo, Lucy Cheke, and Marta Halina. 2020. The animal-AI testbed and competition. In Neurips 2019 competition and demonstration track, pages 164–176. PMLR.

Jeff Da and Jungo Kasai. 2019. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 1–12.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM, 58(9):92–103.

Michael Devitt and Kim Sterelny. 1999. Language and reality: An introduction to the philosophy of language. MIT Press.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? arXiv preprint arXiv:1908.02899.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61:65–170.

Dedre Gentner, Keith J Holyoak, and Boicho N Kokinov. 2001. The analogical mind: Perspectives from cognitive science. MIT press.

James Jerome Gibson. 1979. The ecological approach to visual perception. Houghton, Mifflin and Company.

Arthur M Glenberg and David A Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories.
of meaning. *Journal of memory and language*, 43:379–401.

Herbert P Grice. 1975. Logic and conversation. In *Syntax and semantics. Vol. 1: Speech acts*, pages 41–58. New York: Academic Press.

Zenzi M Griffin. 2004. *Why look? Reasons for eye movements related to language production*. Psychology Press.

PM S Hacker. 2001. *Wittgenstein: connections and controversies*. Clarendon Press.

Rogers P Hall. 1989. Computational approaches to analogical reasoning: A comparative analysis. *Artificial intelligence*, 39(1):39–120.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258.

Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books: New York.

J Daniel House and Edward J Keeley. 1996. Differential prediction of adult student performance from miller analogies test scores. *The Journal of genetic psychology*, 157(4):501–503.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Nicholas Humphrey. 2007. The society of selves. *Philosophical transactions of the royal society b: Biological sciences*, 362(1480):745–754.

Cameron R Jones, Tyler A Chang, Seana Coulson, James A Michaelov, Sean Trott, and Benjamin Bergen. 2022. Distributional semantics still can’t account for affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842.

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.

Arthur F Kramer and Sherry L Willis. 2002. Enhancing the cognitive vitality of older adults. *Current directions in psychological science*, 11(5):173–177.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Doug Lenat, Mayank Prakash, and Mary Shepherd. 1985. Cyc: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, 6(4).

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

John McCarthy. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial intelligence*, 13(1-2):27–39.

Antje S Meyer and F Lethaus. 2004. The use of eye tracking in studies of sentence generation. *The interface of language, vision, and action: Eye movements and the visual world*, pages 191–211.
Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Simon K Milton and Barry Smith. 2004. Top-level ontology: The problem with naturalism. In Formal ontology in information systems, pages 85–94.

M Minsky. 1975. A framework for representing knowledge. The Psychology of Computer Vision.

Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. Advances in Neural Information Processing Systems, 34:25192–25204.

Judea Pearl. 2009. Causality. Cambridge university press.

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.

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. 2020. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8:842–866.

Bertrand Russell. 1914. The relation of sense-data to physics. Nicola Zanichelli.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winograd: An adversarial winograd schema challenge at scale. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 8732–8740.

Lenhart Schubert. 2002. Can we derive general world knowledge from texts? In Proceedings of the second international conference on Human Language Technology Research, pages 94–97.

John R Searle. 2010. The basic reality and the human reality. In Dirk Franken, Attila Karakus, and Jan G. Michel, editors, John R. Searle: Thinking About the Real World, pages 19–44. De Gruyter.

Murray Singer, Michael Halldorson, Jeffrey C Lear, and Peter Andrusiak. 1992. Validation of causal bridging inferences in discourse understanding. Journal of Memory and Language, 31(4):507–524.

Barry Smith. 1995. Formal ontology, common sense and cognitive science. International journal of human-computer studies, 43(5-6):641–667.

Barry Smith. 2012. Ontology. In The furniture of the world: Essays in ontology and metaphysics, pages 47–68. Brill.

Barry Smith and Roberto Casati. 1994. Naive physics. Philosophical psychology, 7(2):227–247.

Barry Smith and David M Mark. 2003. Do mountains exist? towards an ontology of landforms. Environment and Planning B: Planning and Design, 30(3):411–427.

James P Spradley. 2016. Participant observation. Waveland Press.

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.

Keith E Stanovich and Maggie E Toplak. 2012. Defining features versus incidental correlates of type 1 and type 2 processing. Mind & Society, 11(1):3–13.
Keith E Stanovich and Richard F West. 1998. Individual differences in rational thought. *Journal of experimental psychology: general*, 127(2):161.

Robert J Sternberg. 1977. Component processes in analogical reasoning. *Psychological review*, 84(4):353.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. CommonsenseQA 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the winograd schema challenge and swag. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3382–3387.

Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing test*, pages 23–65. Springer.

Peter Turney, Littman Michael, Jeffrey Bingham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2003, pages 482–489.

Peter D Turney and Michael L Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1):251–278.

Jette Viethen, Robert Dale, et al. 2011. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 12–22.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. Blackwell, Oxford. Translated by G.E.M. Anscombe.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 292–299.