# Interpreting docstrings without using common sense: the private science of very large language models*

Darren Abramson†, Ali Emami‡

August 2021

## 1 Introduction

Codex is a machine learning model of natural and programming languages to which OpenAI provides limited third-party access.[1] Github Copilot is a commercial product that is built on Codex.[2] In this paper, we describe some scientific concerns with Codex/Copilot that dovetail with its widely discussed ethical and legal problems. Our focus is on the scientific problems that attend Codex, with consequent weaknesses for the Copilot commercial service. In our view, ethical and scientific weaknesses are closely tied, and we describe this with a few instances.

The argument of the paper is as follows: GPT-3, the natural language model on which Codex is built, and that services such as Copilot ultimately depend on, suffers from scientific deficiencies. First we present critical remarks on Copilot's structure and underlying language model. We then present paths forward for these, identifying specific architectural features that prevent GPT-3 from competing with recent advances due to freely distributed and licensed research software.

## 2 Scientific concerns with Codex/GPT-3

In this section we summarize the unstable scientific state of affairs into which Codex and its parent language model, GPT-3 [1] falls. We argue that these technologies suffer from the usual deficiencies of closed science. In the next section



†Associate Professor, Department of Philosophy, Dalhousie University. da@dal.ca

‡Assistant Professor, Department of Computer Science, Brock University. aemami@brocku.ca

[1]https://openai.com/blog/openai-codex/

[2]https://copilot.github.com/

we argue that, conversely, those parts of computational linguistics and artificial intelligence that continue to hew closer to the values of the Free Software Foundation are succeeding at improving the speed and quality of the science of natural language models.

## 2.1 Black-boxes & paired programming

As is the case with many deep learning models, Codex (based on its parent model, GPT-3) can be described as a *black-box* model, to wit, a model that "takes a sequence of query inputs, and returns corresponding outputs, while keeping internal states such as model architecture hidden" [2]. Despite its impressive performance on a number of natural language processing tasks, ranging from news article generation, arithmetic, and story generation, there exists to date no reliable means to understand or interpret the rationale behind its prediction decisions.

Advances in deep learning have led to the widespread belief that performance comes at the cost of interpretability. On the other hand, numerous results in a variety of different machine learning applications have been emerging that demonstrate the very contrary. For example, the emergence of non-linear models with interpretability constraints (i.e., *glass-box* or *white-box* models) [3, 4, 5, 6] have been shown to perform just as well as unconstrained models, which may mask a multitude of possible serious mistakes [6]. In the criminal justice system, it has been repeatedly demonstrated that black box models for predicting future crime are not any more accurate than simple and interpretable predictive models based on age and criminal history [7, 8]. This result has also been shown to hold in computer vision where deep neural networks constrained for interpretability lead to more transparent computations without doing so at the expense of accuracy (e.g., [9, 10, 11]). The unveiling of such results has led to the proposal of a landmark competition in AI, called the Explainable Machine Learning Challenge[3].

For domains that are potentially lower stake (e.g., machine translation, topic classification, question answering), the prospect of a simple model explaining its prediction may understandably be overshadowed by a stronger, black-box model that boasts superior prediction accuracy. However, towards high-stakes machine learning application domains where interpretability should be regarded as an inseparable component of the output (e.g., healthcare, financial systems, and criminal justice systems) this compromise simply cannot be afforded. Similarly, in the case of "paired programming" to which GitHub openly imputes Copilot, the importance of interpretability is self-evident.

Paired programming is an activity in which, typically, two human programmers cooperate in solving a programming task. The purpose of a competent paired programmer is to help their partner programmer identify hidden problems, question their assumptions, and inform them about alternate and possibly more efficient solutions. Copilot, for now, fails to contribute in this way and,

[3] https://community.fico.com/s/explainable-machine-learning-challenge

quite the contrary, abides blindly by the assumptions of the programmer and directs all its resources towards producing continuations (and may do so by regurgitating chunks from the training data [12]) based on the immediate context of what has been so-far typed by the programmer.

The coupling of these two realities – the first, of the countless results in the literature challenging the belief that accuracy should not be compromised for interpretability, as well as the second, that many domains, such as that of paired programming, require that the system be as interpretable as it is competent, implies that the purport of Copilot's success it not just inaccurate, but potentially harmful to the greater good. It has facilitated the marketing and selling of proprietary, complex, and too-large-to-recreate black box models for high-stakes decisions when otherwise simple, tractable and interpretable models exist for the same tasks. As such, it allows the model creators to profit without considering harmful consequences to the affected individuals. One may even wonder if it isn't the very complicated architecture and massive size of these models that helps suppress the possibility for criticism in the first place.

## 2.2 A lack of common sense

Machine learning models of speech, language and translation have piqued the interest of both researchers and industries alike through their success on a variety of benchmarks (e.g., [13, 14, 15]). However, natural language models exhibit a number of deficiencies, including brittleness, lack of generalizability, and the inability to model compositionality; these have have long plagued neural networks for other domains also [16, 17, 18]. Compositionality is a feature of representations that supports behaviors that are systematically related to one another. For example, consider commutativity for addition and subtraction. A machine learning model of arithmetic that was able to correctly compute $5 + 3$ but not $3 + 5$ would appear to lack compositional representations.

Of particular concern lately is that such models seem to lack common sense, which, by virtue of being shared and "common", is rarely stated explicitly in the training corpora and therefore poses a distinct challenge for data-hungry approaches. Without common sense, systems' output, for example, in language modelling, can seem glaringly *un*intelligent at deployment [19].

The problem of common sense is no less pervasive in the task of code completion for which Codex claims expertise. For example, in a recent critical discussion[4] on Copilot, generated code from the model, given a docstring input asking for the *optional* compression of a file, in fact *always* compressed the file. This inability to capture the effects of modifiers on the meaning of a sentence corroborates recent findings [20]. Additionally, in the original paper for Codex, authors concede that the model has a tendency to recommend syntactically incorrect or undefined code, and struggles to parse through increasingly long and higher-level or system-level specifications [21]. The question, therefore, is why, despite these concerning demonstrations, Co-pilot's strengths seem to

[4] https://www.fast.ai/2021/07/19/copilot/

preponderate its flaws.

The very fact that Codex can generate reasonably appearing code for a given problem with a non-zero chance of passing a unit test is nonetheless impressive. At the same time, the manner in which the positive results are reported warrants deeper investigation, if not for ethical reasons, then purely on a scientific basis. The *pass@k* metric employed by Codex's authors corresponds to the probability that a single sample of code among $k$ pools generated would pass a single unit test.

After having fine-tuned the sampling parameter of temperature, *pass@1* of Codex is reported at roughly 28%. Whether or not success with respect to a single, arbitrary unit test, and under the regime of parameter tuning should be interpreted beyond this specific scope is surely a matter of discretion. In fact, many results showcased by the developers have been admitted to requiring careful "priming" [21]. This limitation is only compounded by the fact that the model is black-box, making the grasp of why certain priming works while other approaches fail as challenging as interpreting the output in the first place.

Codex is both architecturally similar to the natural language model GPT-3 and trained on a pre-existing GPT-3 model already trained on very large crawls of the open web and other text databases. The authors explain that "Surprisingly, we did not observe improvements when starting from a pre-trained language model, possibly because the fine-tuning dataset is so large. Nevertheless, models fine-tuned from GPT converge more quickly, so we apply this strategy for all subsequent experiments."[21]

## 2.3 Methods for assessing common sense

It is important to acknowledge that the work being done at OpenAI is truly innovative. Even 10 years before the first GPT model was presented the scientific community struggled to build systems that leveraged the vast quantities of unlabeled text available freely on the web to produce better machine representations for use in natural language tasks [22].

The first GPT paper is focused on fine-tuning: after the model is trained on improving its ability to predict next tokens in natural language corpora, it is given a task-specific output layer and trained on some 'downstream task' (e.g., a benchmark) [23]. Despite demonstrating significant progress in fine-tuned performance on these tasks, in the 'Analysis' section, the authors explain the importance of measuring 'zero-shot behaviours': accuracy of the language model on tasks that it was not trained to solve, but can solve as a result of its predictive optimization on next tokens.

They say "We'd like to better understand why language model pre-training of transformers is effective. A hypothesis is that the underlying generative model learns to perform many of the tasks we evaluate on in order to improve its language modeling capability and that the more structured attentional memory of the transformer assists in transfer compared to LSTMs" [23]. In Section 3 we compare results with an approach that differs significantly in both architecture and methodology and show improvement with less training and smaller model

sizes. In particular, we question whether fine-tuning is an effective method for measuring common sense knowledge in a language model, and suggest an alternative.

In the previously mentioned critique of the science behind Copilot, coming from a pioneer of the modern successes in machine learning for image recognition[5], the author helped to originate the major step forward for natural language processing (MLP) of fine-tuning a pre-trained language model. That author argues that Copilot is a "buggy mess", and that progress will emerge from its competitors. Part of our task here is to argue that the best alternatives will likely emerge from free software.

Fine-tuning as a metric for language model performance has come under scrutiny. Shuffled corpora are collections of natural language texts that have been randomized for preserving word frequencies but not ordering relations between short sequences of words. [24] show that language models pre-trained on shuffled corpora exhibit similar fine-tuned performance to language models pre-trained on normal texts. They conclude that fine-tuned probes are too weak; instead, we believe fine-tuned probes should be replaced by zero-shot measurement. Zero-shot measurement is a method by which a language model is tested on some task without any additional training beyond its language modeling objective. In contrast, fine-tuning typically involves adjusting the weights of a language model through training on a sample of data from the task the model is being evaluated on, often with the addition of an output layer corresponding to the task's objective.

A recent industry white paper makes it clear why even few-shot measurement of language models ought to be avoided: " Our decision to use zero-shot learning was driven by its simplicity and deterministic behavior, which does not depend on the selection of examples shown during few-shot learning."[6]

Few-shot learning involves choosing some example or examples to pair test examples with when measuring the model, a decision with considerable combinatorial latitude. Fine-tuning involves not only using hundreds or thousands of training examples, but also hyperparameters that can be manipulated for as many GPU hours as are available. For this reason, some authors of benchmarks for common sense post warnings on research ethics along with their datasets. Here is an example:

> 4. Research ethics
>
> In providing both the development and test sets, we are relying on competitors to exercise ethical research practices.
>
> Researchers should not study the 500 questions in the COPA test set, and avoid any temptation to alter their systems toward the content of this particular set of questions. Researchers should evaluate the performance of their systems on the COPA test set only once, after they have concluded all of their efforts to improve perfomance on the

[5] https://www.fast.ai/2021/07/19/copilot/

[6] White paper linked from https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1

COPA development set. Findings of unethical research practices can ruin your career, so obey the rules.[7]

Therefore we believe that because GPT-3 is demonstrably inferior to much smaller language models for common sense as expressed through zero-shot measurement, pair coding AI may not bode well for it in its current state.

# 3 The paths forward

In this section we advocate for our preferred methods for AI-supported coding and for language model assessment in support of building systems that possess natural language understanding.

## 3.1 Retrieval-based approaches

Shortly after the innovation of deep learning-based approaches such as Long Short-Term Memory, the attention mechanism, and the Transformer architecture, significant improvements on a variety of natural language processing tasks were accomplished. On the other hand, as discussed previously, these approaches were accompanied by a variety of issues, interpretability being the major concern in our case. In light of this, a number of dense retrieval-based methods for common sense, that either augment deep learning models or replace them altogether, have been recently proposed and appear more modular as well as interpretable. Consider a pre-transformer system that automatically retrieves snippets from search engines resembling a sample test instance and uses it to reason on-the-fly about the solution to common sense problems (The Knowledge Hunter; [25]).

More recently, language model pre-training was augmented with a latent knowledge retriever, allowing the model to retrieve and attend over documents from a large corpus such as Wikipedia and was demonstrated to be both superior in performance to conventional black-box language models as well providing qualitative benefits such as interpretability and modularity [26]. Based on a very similar idea, an open-source chatbot, dubbed Blender Bot 2.0, combines an information retrieval component with a deep learning based generator, and seeks relevant information both in its long-term memory and from documents it finds by searching the internet to generate high quality, as well as up-to-date responses [27].

These retrieval-based methods suggest an important research direction for Copilot, that promises both improvements in performance, but more importantly, in terms of the long-term benefit and edification of the user. Specifically, just as is the natural inclination for a user, prior to Copilot, is to *search* up a solution related to their problem and retrieve both a sample solution, as well as a context and explanation, retrieval-augmented approaches could potentially offer the same benefits. Although not as "convenient" as simply relying on a

[7] https://people.ict.usc.edu/ gordon/copa.html

quick code continuation, this latter mechanism would commit more closely to the purpose of pair programming; it likely results in the user learning far more about the problem and the possible space of solutions. Incidentally, Microsoft has created a related but lesser-known product called "API Usage Examples"[8] that looks for examples online of people using the API or library that the user is working with, and provides examples of real code and its use-case, along with links to the source of the example. Its relation to Copilot shadows that between the mentioned retrieval based-methods and their purely black-box counterparts, towards which we are more enthusiastic.

### 3.2 Zero-shot language model assessment

A technique recently developed twice [28], [29] leverages a phenomenon that may be familiar to an educator trying to help a student while maintaining fairness in an exam setting. Suppose there is some arbitrary instruction: 'Describe Fodor and Pylyshyn's argument that connectionist systems cannot explain human natural language ability.' A student asks 'Could you help me? I don't understand this question'.

Suppose you repeat the question to the student, but emphasizing a different word each time: "*Describe* Fodor and Pylyshyn's argument..."; "Describe *Fodor* and Pylyshyn's argument..."; "Describe Fodor *and* Pylyshyn's argument..." and so on. Arguably, you haven't given the student any information not available to the other students in the classroom who might not be able to hear the conversation. But might the repeated reflection help the student?

In the [28] context, transformer architectures were applied to the problem of scoring sequences according to hypotheses in the context of speech recognition, neural machine translation, and linguistic acceptability. Strangely, despite the approach being a natural fit for common sense tasks, the authors do not apply their pseudo-log likelihood approach to that problem.

That paper has a nice diagram of the concept of the PLL; rather than reproduce it let us first explain via example from a widely influential common sense database. 'Winograd schemas' are pairs of sentences that, by differing in a single word (a semantic change) change the reference of a word in that sentence.[9] Suppose a masked language model is considering the Winograd sentences F1: 'Frank was upset with Tom because the toaster Frank had bought from him didn't work' and F2: 'Frank was upset with Tom because the toaster *Tom* had bought from him didn't work.' Since masked language models are built to provide the likelihood of a word given a context – optimizing for this is a standard objective function for training them – there is an obvious way to do this.

To score the pair for BERT, simply choose the sentence for which the likelihood of the differing word, when masked, is higher: compare the likelihood of the tokens 'Frank' and 'Tom' for the masked language models' forward pass on

[8] https://marketplace.visualstudio.com/items?itemName=VisualStudioExptTeam.vscodeintellicode-insiders

[9] Terry Winograd first presents this idea in section 1.6.5 of [30]. An influential collection of similar examples is presented by [31].

the sentence 'Frank was upset with Tom because the toaster `MASK` had bought from him didn't work'.

Pseudo-log likelihoods, instead, are the result of a forward pass for *every* word in the sentence. To compare F1 and F2, we iteratively mask each word in the sentences, take the log of the likelihood of the masked word given the sentence, and then sum. The log has the nice property of magnifying differences between scores. For F1 this would involve summing the log likelihood of 'Frank' in '`MASK` was upset with Tom because...' with the log likelihood of 'was' in 'Frank `MASK` upset with Tom because...', and so on, for every word in the sentence. Note the similarity to helping the student by emphasizing each word, in turn.

Masked language models provide a likelihood of word given context, and for two sentences that differ in a single word, difference in likelihood seems to be *amplified* when we measure all possible words given context, not just the word that *differs* given context. As we show in the next section, the dramatic magnifying effect of PLLs suggest that not only the regime of fine-tuning, but also non-bidirectional language models, are being eclipsed in the domain of common sense.

The authors of [28] produce dramatic improvements over existing approaches with their methods, and point out that the method strongly favours bidirectional language models over unidirectional ('causal') models (see their 'Winston Churchill' example, and discussion in the conclusion). It is important to note that Codex/Copilot is built on the early technological decision to train hundreds of thousands of hours of GPU time on a unidirectional model, and OpenAI/Microsoft have every right to try to capitalize on their investment. However, if common sense is a pre-requisite to thoughtful, verbal interaction with human beings, then Codex/Copilot is a sunk-cost fallacy reified into a product.

### 3.3 Recent results on common sense

In this section we summarize some recent and forthcoming results building first on the work of [28] and also [29]. The authors of [28] license their code under an Apache license and state in their abstract that they intend for their contribution to "enable plug-and-play use of the growing number of pretrained MLMs". Given the naturalness of their techniques for common sense language tasks, we have applied them to a number of public datasets using a variety of public language models. The superiority of `albert-xxlarge-v2`[10] became apparent quickly in our experimentation.

In [32] we compare the performance of RoBERTa [33] and AlBERT [34] language models using PLLs on the Winograd data set [31] and a much larger, crowd-sourced data set with similar form: the Winogrande data set, train-xl split [35]. The BERT variant PLLs for Winogrande score strictly better on the Winogrande data split than reported zero-shot GPT-3 results [1]. GPT-

[10] https://huggingface.co/albert-xxlarge-v2

3 reports higher zero-shot values than our otherwise state-of-the-art zero-shot results for Winograd, but explicitly state that there is cross-contamination between the Winograd examples and their very large, web-crawled pre-training corpus. Our best results using the `albert-xxlarge-v2` with PLLs are 81.05% for the Winograd data set and 76.71 for the Winogrande data set.

We also investigate the performance of PLLs with bidirectional languge models in [32] on adversarially 'perturbed' variations of the Winograd data set [36]. We find that PLLs improve on other scoring methods both for accuracy averaged across all perturbed data sets, but also 'accuracy delta', $\mathbf{\Delta_{Acc.}}$: the average change in accuracy when Winograd schemas are systematically perturbed. `albert-xxlarge-v2` did not score the lowest $\mathbf{\Delta_{Acc.}}$ (-4.54) but did show the highest average accuracy across all perturbed Winograd data sets: 79.64%

In [37] we show that across every model we measured, natural language model performance using PLLs was better on Winograd than on Winogrande. We suspect that this might be because of the poorer quality of crowdsourced data for common sense; work in progress attempts to answer this question with human subjects. There is uncertainty in both the methodologies used by the research community to assess language model performance on common sense and also the data sets to which we apply those methods. It is difficult to imagine progress on the associated open research questions without free access to models, unavailable for both Copilot and Codex.

It came to our attention recently that a second group of researchers independently discovered the utility of pseudo-log likelihoods; in this case, they were applied explicitly to zero-shot benchmarking on a number of popular common sense results [29]. That paper's authors have independently published a freely available repository for computing PLLs.[11] By taking the trivial step of modifying their code to use `albert-xxlarge-v2` we produced results that were better on most (but not all) measures than their approach.

Practical considerations are significant when building on the work of others. The advantage of the first paper/repository we became aware of for PLLs ([28]) is that the code scales automatically to multiple GPUs if available, dramatically improving the speed to compute them. A recent publication cites the computational cost of PLLs as a reason to limit their use only to 'base' sized models [38]. Our own ability to report results using `albert-xxlarge-v2` is due to the availability of national, public supercomputing resources available to researchers in the country we work in. The [29] formulation has the advantage of being much more readable and compact code, but does not automatically scale to multiple GPUs.

Finally, consider the 'Timedial' dataset, which we first heard about via blog post at Google AI.[12] With [37] we provide a script which, using mlm-scoring [28], produces a .csv of sentence scoring for substitutions. The top-2 accuracy is over 75% with masked PLLS from `albert-xxlarge-v2`, exceeding the best fine-tuned accuracy score reported in [39]. Free software can, we hope, produce

[11] https://github.com/XuhuiZhou/CATS

[12] https://ai.googleblog.com/2021/08/two-new-datasets-for-conversational-nlp.html

language models that understand enough human common sense to be useful partners for programming and other tasks.

# 4 Conclusion

The last decade has seen considerable advances in natural language processing using machine learning. By hiding the GPT-3 model behind an API wall, and building a commercial service on top of it that hides its underlying functioning, OpenAI has demonstrably chosen a path for innovation that is slower and technologically poorer than free and open alternatives. We target the reader who erroneously believes that innovation in natural language understanding can only come from ever-larger language models pre-trained using quantities of data and compute unavailable to individual researchers. Instead, free software has produced advances through smaller language models and methods that can be scrutinized and improved by anyone with a stake in their success. Well-known ethical advantages of free software systems and the scrutiny they permit produce scientific value also, in distributed innovation representing the diverse values of AI practitioners.

# References


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

[2] S. J. Oh, B. Schiele, and M. Fritz, "Towards reverse-engineering black-box neural networks," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 121–144, Springer, 2019.

[3] R. Caruana, S. Lundberg, M. T. Ribeiro, H. Nori, and S. Jenkins, "Intelligible and explainable machine learning: Best practices and practical challenges," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3511–3512, 2020.

[4] J. Kalin, M. Ciolino, D. Noever, and G. Dozier, "Black box to white box: Discover model characteristics based on strategic probing," in *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 60–63, IEEE, 2020.

[5] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2015.

[6] C. Rudin and B. Ustun, "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice," *Interfaces*, vol. 48, no. 5, pp. 449–466, 2018.

[7] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, "Learning certifiably optimal rule lists for categorical data," 2018.

[8] N. Tollenaar and P. G. M. van der Heijden, "Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 176, no. 2, pp. 565–584, 2013.

[9] Y. Ming, P. Xu, H. Qu, and L. Ren, "Interpretable and steerable sequence learning via prototypes," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), p. 903–913, Association for Computing Machinery, 2019.

[10] C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that: deep learning for interpretable image recognition," *CoRR*, vol. abs/1806.10574, 2018.

[11] Y. Li, m. Murias, s. Major, g. Dawson, K. Dzirasa, L. Carin, and D. E. Carlson, "Targeting eeg/lfp synchrony with neural nets," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[12] N. Dehouche, "Plagiarism in the age of massive generative pre-trained transformers (gpt-3)," *Ethics in Science and Environmental Politics*, vol. 21, pp. 17–23, 2021.

[13] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[14] X. Liu, K. Duh, L. Liu, and J. Gao, "Very Deep Transformers for Neural Machine Translation," *arXiv e-prints*, p. arXiv:2008.07772, Aug. 2020.

[15] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," in *Proceedings of the 8th international conference on signal processing systems*, pp. 142–147, 2016.

[16] E. Davis and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Communications of the ACM*, vol. 58, no. 9, pp. 92–103, 2015.

[17] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

[18] G. Marcus, F. Rossi, and M. Veloso, "Beyond the turing test.," *AI Magazine*, vol. 37, no. 1, 2016.

[19] P. Trichelair, A. Emami, J. C. K. Cheung, A. Trischler, K. Suleman, and F. Diaz, "On the evaluation of common-sense reasoning in natural language understanding," *arXiv preprint arXiv:1811.01778*, 2018.

[20] A. Emami, I. Porada, A. Olteanu, K. Suleman, A. Trischler, and J. C. K. Cheung, "ADEPT: An adjective-dependent plausibility task," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 7117–7128, Association for Computational Linguistics, Aug. 2021.

[21] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba, "Evaluating large language models trained on code," 2021.

[22] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[23] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training,"

[24] K. Sinha, R. Jia, D. Hupkes, J. Pineau, A. Williams, and D. Kiela, "Masked language modeling and the distributional hypothesis: Order word matters pre-training for little," *arXiv preprint arXiv:2104.06644*, 2021.

[25] A. Emami, N. De La Cruz, A. Trischler, K. Suleman, and J. C. K. Cheung, "A knowledge hunting framework for common sense reasoning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 1949–1958, Association for Computational Linguistics, Oct.-Nov. 2018.

[26] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *arXiv preprint arXiv:2002.08909*, 2020.

[27] M. Komeili, K. Shuster, and J. Weston, "Internet-augmented dialogue generation," *arXiv preprint arXiv:2107.07566*, 2021.

[28] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, 2020.

[29] X. Zhou, Y. Zhang, L. Cui, and D. Huang, "Evaluating commonsense in pre-trained language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9733–9740, 2020.

[30] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[31] H. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge," in *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

[32] D. Abramson, "AI's Winograd moment; or, how should we teach machines common sense? Guidance from cognitive science," in *Artificial Intelligence and Human Enhancement: Affirmative and Critical Approaches in the Humanities* (H. Nagl-Docekal and W. Zacharasiewicz, eds.), De Gruyter; expected April., 2022.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[35] K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8732–8740, 2020.

[36] M. Abdou, V. Ravishankar, M. Barrett, Y. Belinkov, D. Elliott, and A. Søgaard, "The sensitivity of language models and humans to winograd schema perturbations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7590–7604, 2020.

[37] D. Abramson and A. Emami, "An application of pseudo-log-likelihoods to natural language scoring," *arXiv preprint arXiv:2201.09377*, 2022.

[38] P. Laban, L. Dai, L. Bandarkar, and M. A. Hearst, "Can transformer models measure coherence in text? re-thinking the shuffle test," 2021.

[39] L. Qin, A. Gupta, S. Upadhyay, L. He, Y. Choi, and M. Faruqui, "Timedial: Temporal commonsense reasoning in dialog," *arXiv preprint arXiv:2106.04571*, 2021.