Assessing the Strengths and Weaknesses of Large Language Models

Shalom Lappin[1,2,3*]

[1]School of Electronic Engineering and Compuer Science, Queen Mary University of London .
[2]Centre for Linguistic Theory and Studies in Probability, University of Gothenburg .
[3]Department of Informatics, King's College London .

Corresponding author(s). E-mail(s): s.lappin@qmul.ac.uk;

Abstract

The transformers that drive chatbots and other AI systems constitute large language models (LLMs). These are currently the focus of a lively discussion in both the scientific literature and the popular media. This discussion ranges from hyperbolic claims that attribute general intelligence and sentience to LLMs, to the skeptical view that these devices are no more than "stochastic parrots". I present an overview of some of the weak arguments that have been presented against LLMs, and I consider several of the more compelling criticisms of these devices. The former significantly underestimate the capacity of transformers to achieve subtle inductive inferences required for high levels of performance on complex, cognitively significant tasks. In some instances, these arguments misconstrue the nature of deep learning. The latter criticisms identify significant limitations in the way in which transformers learn and represent patterns in data. They also point out important differences between the procedures through which deep neural networks and humans acquire knowledge of natural language. It is necessary to look carefully at both sets of arguments in order to achieve a balanced assessment of the potential and the limitations of LLMs.

1 Introduction

The introduction of transformers [1] with multiple attention heads, and pre-trained with large scale word embeddings, has revolutionised NLP. They have yielded near or above human performance on a variety of core NLP tasks, which include, among others, machine translation, natural language generation, question answering, image captioning, text summarisation, and natural language inference (NLI). Transformers operate as autoregressive token predictors (GPT-1-GPT4, OpenAI), or as bidirectional token predictors for masked positions in contexts, such as BERT [2]. They significantly outperform sequential deep learning networks, like Long Short Term Memory Recurrent Neural Networks (LSTMs) and Convolutional Neural Networks (CNNs), on most natural language tasks, and for many non-linguistic applications, such as image recognition.

Transformers provide the LLMs that drive chatbots. The rapid success of these bots in generating extended sequences of coherent, human like discourse in response to prompts has produced vigorous debate in both the scientific literature and the popular media. Some of this discussion consists of exaggerated claims on the capabilities of LLMs. Other comments offer pat dismissals of these systems as nothing more than artificial parrots repeating training data. It is important to consider LLMs in a critical and informed way, in order to understand their abilities and their limitations.

In Section 2 I take up several of the more prominent weak arguments that have been brought against LLMs. These include

(i) the view that they simply return their training data,
(ii) the claim that they cannot capture linguistic meaning due to the absence of semantic grounding,
(iii) the assertion that they do not acquire symbolic representations of knowledge, and
(iv) the statement that they do not learn in the way that humans do.

In Section 3 I consider some of the strong arguments concerning the limitations of LLMs. These involve

(i) important constraints on LLMs as sources of insight into human cognitive processes,
(ii) the lack of robustness in LLM performance on NLI tasks,
(iii) the unreliability of LLMs as a source of factually sound information,
(iv) the inability of LLMs to identify universal patterns characteristic of natural languages,
(v) the consequences of the large data required to train LLMs, for control of the architecture and development of these systems, and
(vi) the opactiy of these systems.

Section 4 draws conclusions concerning the capacities and limitations of LLMs. It suggests possible lines of future research in deep learning in light of this discussion.

2 Weak Arguments Against LLMs

2.1 Generalisation, Innovation, and Semantic Grounding

A common criticism of LLMs is that they do little more than synthesise elements of their training data to produce the most highly valued response to a prompt, as determined by their probability distribution over the prompt and the data. [3] and [4] offer recent versions of this view. This claim is false, given that transformers exhibit subtle and sophisticated pattern identification and inferencing

This inductive capacity permits them to excel at medical image analysis and diagnostics [5]. Transformers have revolutionised computational biology by predicting properties of proteins and new molecular structures [6]. This has opened the way for the use of deep learning for the development of new medications and clinical treatments.

[7] use Reinforcement Learning (RL) to train artificial agents with LLMs to respond appropriately to complex commands in a simulated visual environment. These commands do not correspond to information or commands in their training set.[1]

[9] argue that LLMs cannot capture meaning because they are not semantically grounded, by virtue of the fact that their word embeddings are generated entirely from text. Hence they cannot identify speakers' references to objects in the world or recognise communicative intensions. [10–12] reply to this argument by observing that learning the distributional properties of words in text does provide access to central elements of interpretation. These properties specify the topology of lexical meaning.

Even if one accepts Bender and Koller's view that grounding is a necessary element of interpretation, it does not establish their claim that LLMs are unable to represent interpretations of natural language expressions. It is possible to construct multi-modal word embeddings that capture the distributional patterns of expressions in visual and other modalities. Multi-Model BERT [13] and GPT-4 [14] are pre-trained on such word embeddings. Multi-modal transformers can identify elements of a graphic image and respond to questions about them. The dialogue in Figure 1 from [14] illustrates the capacity of an LLM to reason about a complex photographic sequence, and to identify the humour in its main image.

[9] do raise important questions about what a viable computational model of meaning and interpretation must achieve. It has stimulated a fruitful debate on this issue. The classical program of formal semantics [15, 16] seeks to construct a recursive definition of a truth predicate that entails the truth conditions of the declarative sentences of a language. [8] observes that a generalised multi-modal deep neural network (DNN) achieves a major part of this program by pairing a suitable graphic (and other modes of) representation with a sentence that describes a situation.

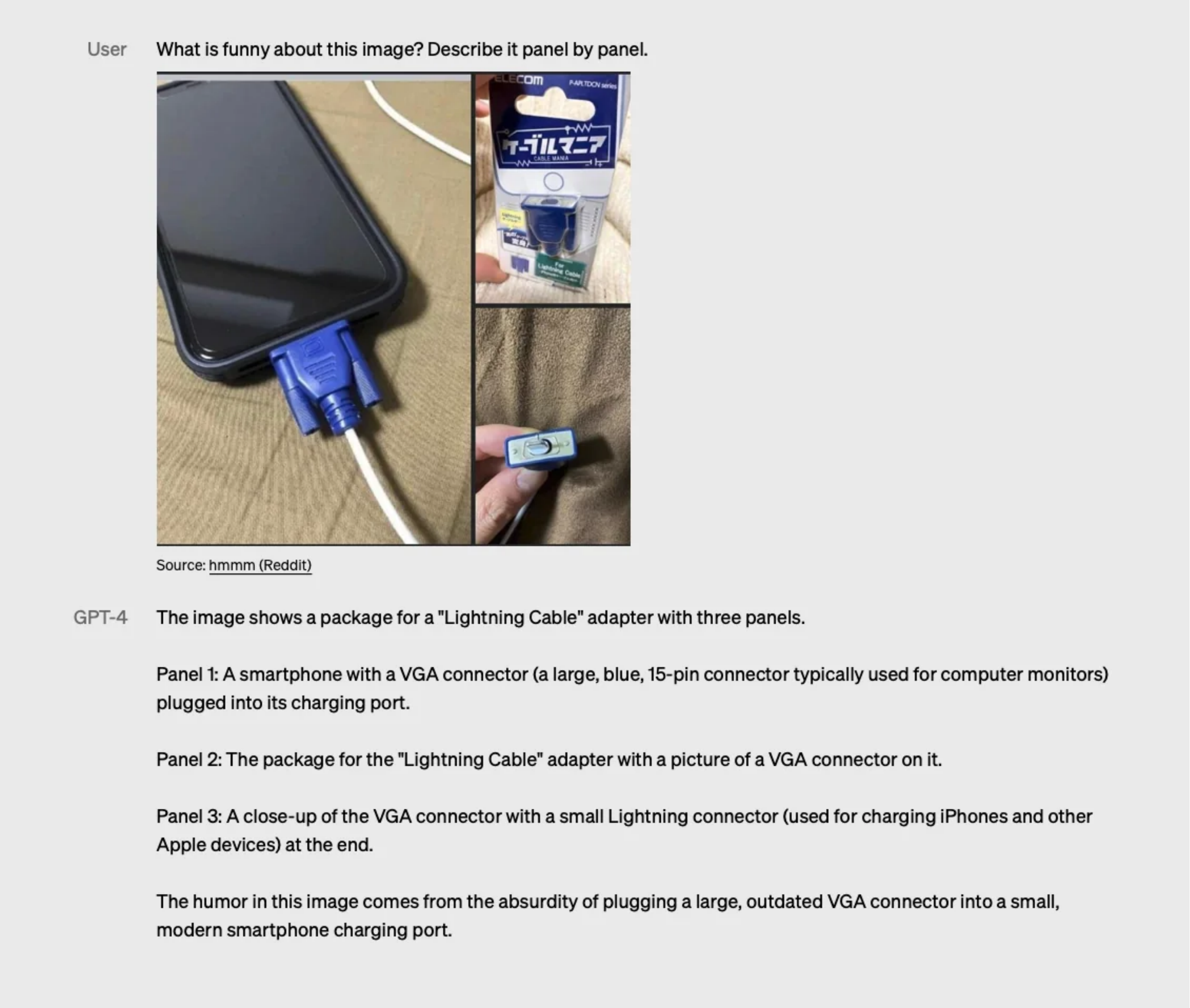
2.2 Symbolic Representations, Grammars, and Hybrid Systems

[17] and [4] maintain that LLMs are defective because they do not represent the symbolic systems, specifically grammars, which humans acquire to express linguistic knowledge. They regard grammars as the canonical form for expresing linguistic knowledge.

This claim is question begging. It assumes what is at issue in exploring the nature of human learning. It is entirely possible that humans acquire and encode knowledge of language in non-symbolic, distributed representations of linguistic (and other) regularities, rather than through symbolic algebraic systems. [18, 19], among others, suggests a view of this kind. Some transformers implicitly identify significant amounts of hierarchical syntactic structure and long distance relations [8, 20–23]. In fact [8] and

---

[1] See [8] for discussion of the problem of generalising to new linguistic input in deep learning.



User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Fig. 1 ChatGPT-4 Multi-Modal Dialogue, from OpenAI (2023)

[24] argue that DNNs can be regarded as alternative theoretical models of linguistic knowledge.

[17] argues that to learn as effectively as humans do DNNs must incorporate symbolic rule-based components, as well as the layers of weighted units that process vectors through the functions that characterise deep learning. It is widely assumed that such hybrid systems will substantially enhance the performance of DNNs on tasks requiring complex linguistic knowledge. In fact, this is not obviously the case.

Tree DNNs incorporate syntactic structure into a DNN, either directly through its architecture, or indirectly through its training data and knowledge distillation. [25–31] consider LSTM-based Tree DNNs. These have been applied to NLP tasks like sentiment analysis, NLI, and the prediction of human sentence acceptability judgments. In general they do not significantly improve performance relative to the non-tree counterpart model, and, in at least one case [31], performance was degraded in comparison with the non-enriched models.[2]

More recent work has incorporated syntactic tree structure into transformers like BERT, and applied these systems to a broader range of tasks [32, 33]. The experimental evidence on the extent to which this addition improves the capacity of the transformer to handle these tasks remains unclear. Future work may well show that hybrid systems of the sort that Tree DNNs represent do offer genuine advantages over their non-enriched counterparts. The results achieved with these systems to date have not yet motivated this claim.

2.3 Humans Don't Learn Like That

[17] and [4] claim that DNNs do not capture the way in which humans achieve knowledge of their language. In fact this remains an open empirical question. We do not yet know enough about human learning to exclude strong parallels between the ways in which DNNs and humans acquire and represent linguistic knowledge, and other types of information.

But even if the claim is true, it need not constitute a flaw in DNN design. These are engineering devices for performing NLP (and other) tasks. Their usefulness is evaluated on the basis of their success in performing these tasks, rather than on the way in which they achieve these results. From this perspective, criticising DNNs on the grounds they do not operate like humans is analogous to objecting to aircraft because they do not fly like birds.

Moreover, the success of transformers across a broad range of NLP tasks does have cognitive significance. These devices demonstrate one way in which the knowledge required to perform these tasks can be obtained, even if it does not correspond to the procedures that humans apply. These results have important consequences for debates over the types of biases that are, in principle, needed for language acquisition and other kinds of learning.

3 Strong Arguments Against LLMs

3.1 LLMs as Models of Human Learning

[34] claims that LLMs provide a viable model of human language acquisition. He suggests that they provide evidence against Chomsky's domain specific innatist view, which posits a "language faculty" that encodes a Universal Grammar.[3] This is due to their ablity to learn implicit representations of syntactic structure and lexical semantics without strong linguistic learning biases.

---

[2] See [8] for discussion of LSTM Tree DNNs.

[3] See [35, 36] for earlier discussions of domain specific, strong bias views of language learning, and arguments against them. These arguments are based on considerations from computational learning theory.

In Section 2.3 I argued that the fact that LLMs may learn and represent information differently than humans does not entail a flaw in their design. Moreover they do contribute insight into cognitive issues concerning language acquisition by indicating what can, in principle, be learned from available data, with the the inductive procedures that DNNs apply to this input. However, Piantadosi's claims go well beyond the evidence. His conclusions seem to imply that humans do, in fact, learn in the way that DNNs do. This is not obviously the case.

Moreoever, although we do not know precisely how humans acquire and represent their linguistic knowledge, we do know that they learn with far less data than DNNs require. As [37] observe, LLMs are trained on orders of magnitude more data than human learners have access to. Humans also require interaction in order to achieve knowledge of their language [38], while LLMs are trained non-interactively, with the possible exception of RL as a partial simulation of human feedback.

Warstadt and Bowman argue convincingly that in order to assess the extent to which deep learning corresponds to human learning it is necessary to restrict LLMs to the sort of data that humans use in language acquisition. This involves reconfiguring both the training data and the learning environment for DNNs, to simulate the language learning situation that humans encounter. Only an experimental context of this kind will illuminate the extent to which there is an analogy between deep and human learning.

3.2 Problems with Robust NLI

Transformers have scored well on natural language inference benchmarks. However, they are easily derailed, and their performance can be reduced to close to chance by adversarial testing involving lexical substitutions. [39] show that BERT does not generalise well to new data sets for NLI. Conversely [40] report that BERT continues to achieve high scores when fine tuned and tested on corrupted data sets containing nonsense sentence pairs. These results suggest that BERT is not learning inference through semantic relations between premisses and conclusions. Instead it appears to be identifying certain lexical and structural patterns in the inference pairs.

[41] points out that humans also frequently make mistakes in inference. However, their reasoning abilities are more robust, even under adversarial testing. Human performance declines more gracefully than transformers under generalisation challenges, and it is more sharply degraded by nonsense pairs. [41] discuss work that pre-trains models on abstract reasoning templates to improve their performance in NLI. It is not clear to what extent the natural language inference abilities of transformers constitute more than superficial inductive generalisation over labelled patterns in their training corpora. [42] review substantial amounts of evidence for the claim that LLMs do not perform well on extra-linguistic reasoning and real world knowledge tasks.

3.3 LLMs Hallucinate

LLMs are notorious for hallucinating plausible sounding narratives that have no factual basis. In a particularly dramatic example of this phenomenon ChatGPT recently went to court as a legal expert. A lawyer representing a passenger on Avianca Airline brought a lawsuit against the airline citing 6 legal precedents as part of the evidence for his case [43]. The judge was unable to verify any of the precedents. When the judge demanded an explanation, the lawyer admitted to using ChatGPT-3 to identify the legal precedents that he required. He went on to explain that he ”even asked the program to verify that the cases were real. It had said yes.”

The fact that LLMs do not reliably distinguish fact from fiction makes them dangerous sources of (mis)information. Notice that semantic grounding through multi-modal training does not, in itself, solve this problem. The images, sounds, etc. to which multi-modal transformers key text do not insure that the text is factually accurate. The non-linguistic representations may also be artificially generated. A description of the image of a unicorn may accurately describe that image. It does not characterise an animal in the world.

3.4 Linguistic Universals

Natural languages display universal patterns, or tendencies, involving word order, morphology, phonology, and lexical semantic classes. Many of these can be expressed as conditional probabilities specifying that a property will hold with a certain likelihood, given the presence of other features in the language.

[44] and [45] suggest that information theoretic notions of communicative efficiency can explain many of these universals. This type of efficiency involves optimising the balance between brevity of expression and complexity of content. As there are alternative strategies for achieving such optimisations, languages will exhibit different clusters of properties.

[46] report experiments showing that LSTM communication networks display a preference for anti-efficient encoding of information in which the most frequent expressions are the longest, rather than the shortest. They experiment with additional learning biases to promote DNN preference for more efficient communication systems. Similarly, [47] describe LSTM simulations in which the network tends to preserve the distribution patterns observed in the training data, rather than to maximise efficiency of communication.

If this tendency carries over to transformers, then they will be unable to distinguish between input from plausible and implausible linguistic communication systems. They will not recognise the class of likely natural languages. Therefore, they will not provide insight into the information theoretic biases that shape natural languages.

3.5 Large Data and the Control of Deep Learning Architecture

Transformers require vast amounts of training data for their word and multi-modal embeddings. Each significant improvement in performance on a range of tasks is generally driven by a substantial increase in training data, and an expansion in the size of the LLM. While GPT-2 has 1.5 billion parameters, GPT-3 has 175 billion, and GPT-4 is thought to be 6 times larger than GPT-3.

Only large tech companies have the computing capacity, infrastructure, and funds to develop and train transformers of this size. This concentrates the design and development of LLMs in a very limited number of centres, to the exclusion of most universities and smaller research agencies. As a result, there are limited resources for research on alternative architectures for deep learning which focuses on issues that are not central to the economic concerns of tech companies. Many (most?) researchers working outside of these companies effectively become clients using their products, which they apply to AI tasks through fine tuning and minor modifications. This is an unhealthy state of affairs. While enormous progress has been made on deep learning models, over a relatively short period of time, it has been largely restricted to a narrow dimension of tasks in NLP. In particular, work on the relation of DNNs to human learning and representation is increasingly limited. Also, examination of learning from small data with more transparent and agile systems is not a major issue in current research on deep learning.

It is important to note that Reinforcement Learning does not alleviate the need for large data, and massive LLMs. RL can significantly improve the performance of transformers across a large variety of tasks [7]. It can also facilitate zero, one, and few shot learning, where a DNN performs well on a new task with limited or no prior training. However, it does not eliminate the need for large amounts of training data. These are still required for pre-trained word and multi-modal embeddings.

3.6 LLMs Are Opaque

Transformers, and DNNs in general, are largely opaque systems, for which it is difficult to identify the procedures through which they arrive at the patterns that they recognise. This is, in large part, because the functions, like ReLU, that they apply to activate their units are non-linear. Autoregressive generative language models also use softmax to map their output vectors into probability distributions.

These functions cause the vectors that a DNN produces to be, in the general case, non-compositional. This is due to the fact that the representations of the input and the output vectors cannot be represented by a homomorphic mapping operation. A mapping $f : A \rightarrow B$ from group A to group B is a homomorphism iff for every $v_i, v_j \in A$, and the group operation $\cdot$, $f(A \cdot B) = f(A) \cdot f(B)$. As a result, it is not always possible to predict the composite vectors that the units of a transformer generate from their inputs, or to reconstruct these inputs from the output vectors in a uniform and regular way.

Probes [21], and selective de-activation of units and attention heads [23] can provide insight into the structures that transformers infer. These methods remain indirect, and they do not fully illuminate the way in which transformers learn and represent patterns from data.

[48] propose Unitary Recurrent Networks (URNs) to solve the problem of model opacity. URNs apply multiplication to orthogonal matrices. The matrices that they generate are strictly compositional. These models are fully transparent, and all input is recoverable from the output at each phase in an URN's processing sequence. They achieve good results for deeply nested bracket matching tasks in Dyck languages, a class of artificial context-free languages. They do not perform well on context-sensitive cross serial dependencies in artificial languages, or on agreement in natural languages. One of their limitations is the use of truncation of matrices to reduce the size of their

rows. This is necessary to facilitate efficient computation of matrix multiplication, but it degrades the performance of an URN on the tasks to which it is applied.

4 Conclusions and Future Research

LLMs are not simply ”stochastic parrots” synthesising fluent sounding responses to prompts from previously observed training data. They achieve a sophisticated level of inductive learning and inference, with transferable skills, across a wide range of tasks. They are able to identify hierarchical syntactic structure and complex semantic relations. Through reinforcement learning on multi-modal input they can be trained to respond appropriately to new questions and commands out of the domain of their training data. This involves significant generalisation and few shot learning. Work on LLMs has yielded dramatic progress across a broad set of AI problems, in a comparatively short period of time. It far exceeds the achievements of symbolic rule-based AI over many decades.

It is not clear to what extent LLMs illuminate human cognitive abilities in the areas of language learning and linguistic representation. While they surpass human performance on many cognitively interesting NLP tasks, they require far more data for language learning than humans do. It is not obvious that they learn and encode linguistic knowledge in the way that humans perform these operations.

LLMs are far from human abilities in natural language inference, analogical reasoning, and interpretation, particularly for figurative language. Their performance in domain general dialogue, while appearing to be fluent, remains informationally limited, and frequently unreliable. They do not distinguish fact from fiction, but freely generate inaccurate claims. They also do not optimise informational efficiency in communication.

The large scale of training data and model size that LLMs require has created a situation in which large tech companies control the design and development of these systems. This has skewed research on deep learning in a particular direction, and disadvantaged scientific work on machine learning with a different orientation.

LLMs are opaque in the way that they generalise from data, which poses serious problems of explainability. At present we can only understand their inference procedures and knowledge representations indirectly, through probes, and through selective oblation of heads and other units in the network.

These conclusions suggest the following lines of future research on deep learning. To compare LLMs to human learners it is necessary to modify the data to which they have access, and to alter their training regimen. This will permit us to examine the extent to which there are correspondences and disanalogies between the two learning processes. It will also be necessary to study the internal procedures applied by each type of learner, computational for LLMs and neurological for humans, more closely to identify the precise mechanisms that drive inference, generalisation, and representation, for each kind of acquisition.

It would be useful to experiment with additional learning biases for LLMs to see if these biases will improve their capacity for robust NLI, and for communicative efficiency. This work may provide deeper understanding of what is involved in both abilities. If it is successful, it will produce more intelligent and effective DNNs, which are better able to handle complex NLP tasks.

It is imperative that we develop procedures for testing the factual accuracy of the text produced by LLMs. Without them we are exposed to the threat of disinformation on an even larger scale than we are currently encountering in bot saturated social media. The need to combat disinformation patterns together with the urgency of filtering racial, ethnic, religious, and gender bias in AI systems powered by LLMs, that are used in decision making applications (hiring, lending, university admission, etc). These research concerns should be a focus of public policy discussion.

Developing smaller, more lightweight models that can be trained on less data would encourage more work on alternative architectures, among a larger number of researchers, distributed more widely across industrial and academic centres. This would facilitate the pursuit of more varied scientific objectives in the area of machine learning.

Finally, designing and implementing fully transparent DNNs will improve our understanding of both artificial and human learning. Scientific insight of this kind should be no less a priority than the engineering success driving current LLMs. Ultimately, good engineering depends on a solid scientific understanding of the principles embodied in the systems that the engineering creates.

Acknowledgements

Many of the ideas proposed here were developed in discussion with Richard Sproat. I am grateful to him for his contribution to my thinking on these issues. I am also indebted to Katrin Erk, Steven Piantadosi, and two anonymous reviewers for valuable remarks and suggestions on previous drafts. An earlier version of this paper was presented in the Cognitive Science Seminar of the School of Electronic Engineering and Computer Science at Queen Mary University of London, on June 21, 2023. I thank the audience of this talk for helpful questions and comments. I bear sole responsibility for any errors that remain.

My research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., Red Hook, NY (2017)

[2] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)

[3] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? FAccT '21, pp. 610–623. Association for Computing Machinery, New York, NY, USA (2021)

[4] Chomsky, N., Roberts, I., Watumull, J.: The false promise of chatgpt. The New York Times **March 8 2023** (2023)

[5] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis **88** (2023)

[6] Chandra, A., Tünnermann, L., Löfstedt, T., Gratz, R.: Transformer-based deep learning for predicting protein properties in the life sciences. eLife **12** (2023)

[7] Dasgupta, I., Kaeser-Chen, C., Marino, K., Ahuja, A., Babayan, S., Hill, F., Fergus, R.: Collaborating with language models for embodied reasoning. arXiv **2302.00763** (2023)

[8] Lappin, S.: Deep Learning and Linguistic Representation. CRC Press, Taylor & Francis, Boca Raton, London, New York (2021)

[9] Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5185–5198. Association for Computational Linguistics, Stroudsburg, PA (2020)

[10] Sahlgren, M., Carlsson, F.: The singleton fallacy: Why current critiques of language models miss the point. Frontiers in Artificial Intelligence **4** (2021)

[11] Piantadosi, S.T., Hill, F.: Meaning without reference in large language models. arXiv **2208.02957** (2022)

[12] Søgaard, A.: Grounding the vector space of an octopus: Word meaning from raw text. Minds and Machines **33**(1), 33–54 (2023)

[13] Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 13–23 (2019)

[14] OpenAI: Gpt-4 technical report. arXiv **2303.08774** (2023)

[15] Davidson, D.: Truth and meaning. Synthese **17**(1), 304–323 (1967)

[16] Montague, R.: Formal Philosophy: Selected Papers of Richard Montague. Yale University Press, New Haven, CT/London, UK (1974). Edited with an introduction by R. H. Thomason

[17] Marcus, G.: Deep learning alone isn't getting us to human-like AI. Noema **August 112022** (2022)

[18] Smolensky, P.: Connectionist AI, symbolic AI, and the brain. Artificial Intelligence Review **1**(2), 95–109 (1987)

[19] McClelland, J.L.: Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. In: MacWhinney, B., O'Grady, W. (eds.) The Handbook of Language Emergence, pp. 54–80. John Wiley and Sons, Hoboken, NJ (2016)

[20] Goldberg, Y.: Assessing bert's syntactic abilities. ArXiv **abs/1901.05287** (2019)

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

[22] Wilcox, E.G., Gauthier, J., Hu, J., Qian, P., Levy, R.: Learning syntactic structures from string input. In: Lappin, S., Bernardy, J.-P. (eds.) Algebraic Structures in Natural Language, pp. 113–137. CRC Press, Taylor & Francis, Boca Raton, London, and New York (2023)

[23] Lasri, K.: Linguistic Generalization in Transformer-Based Neural Language Models. unpublished PhD thesis, PSL, University of Paris (2023)

[24] Baroni, M.: On the proper role of linguistically oriented deep net analysis in linguistic theorising. In: Lappin, S., Bernardy, J.-P. (eds.) Algebraic Structures in Natural Language, pp. 1–16. CRC Press, Taylor & Francis, Boca Raton, London, and New York (2023)

[25] Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 151–161. Association for Computational Linguistics, Edinburgh, Scotland, UK. (2011)

[26] Bowman, S.R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C.D., Potts, C.: A fast unified model for parsing and sentence understanding. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1466–1477. Association for Computational Linguistics, Berlin, Germany (2016)

[27] Yogatama, D., Blunsom, P., Dyer, C., Grefenstette, E., Ling, W.: Learning to compose words into sentences with reinforcement learning. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)

[28] Choi, J., Yoo, K.M., Lee, S.-g.: Learning to compose task-specific tree structures. In: AAAI Conference on Artificial Intelligence (2018)

[29] Williams, A., Drozdov, A., Bowman, S.R.: Do latent tree learning models identify meaningful structure in sentences? Transactions of the Association for Computational Linguistics **6**, 253–267 (2018)

[30] Maillard, J., Clark, S., Yogatama, D.: Jointly learning sentence embeddings and syntax with unsupervised tree-lstms. Natural Language Engineering **25**(4), 433–449 (2019)

[31] Ek, A., Bernardy, J.-P., Lappin, S.: Language modeling with syntactic and semantic representation for sentence acceptability predictions. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland, pp. 76–85 (2019)

[32] Sachan, D.S., Zhang, Y., Qi, P., Hamilton, W.: Do syntax trees help pre-trained transformers extract information? In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2647–2661. Association for Computational Linguistics, Stroudsburg, PA (2021)

[33] Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J., Tong, Y.: Syntax-bert: Improving pre-trained transformers with syntax trees. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3011–3020. Association for Computational Linguistics, Stroudsburg, PA (2021)

[34] Piantadosi, S.: Modern language models refute chomsky's approach to language. Lingbuzz Preprint, lingbuzz **7180** (2023)

[35] Lappin, S., Shieber, S.: Machine learning theory and practice as a source of insight into universal grammar. Journal of Linguistics **43**, 393–427 (2007)

[36] Clark, A., Lappin, S.: Linguistic Nativism and the Poverty of the Stimulus. Wiley-Blackwell, Malden, MA and Oxford (2011)

[37] Warstadt, A., Bowman, S.R.: What artificial neural networks can tell us about human language acquisition. In: Lappin, S., Bernardy, J.-P. (eds.) Algebraic Structures in Natural Language, pp. 17–59. CRC Press, Taylor & Francis, Boca Raton, London, and New York (2023)

[38] Clark, E.: Language is acquired in interaction. In: Lappin, S., Bernardy, J.-P.

(eds.) Algebraic Structures in Natural Language, pp. 77–93. CRC Press, Taylor & Francis, Boca Raton, London, and New York (2023)

[39] Talman, A., Chatzikyriakidis, S.: Testing the generalization power of neural network models across NLI benchmarks. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 85–94. Association for Computational Linguistics, Florence, Italy (2019)

[40] Talman, A., Apidianaki, M., Chatzikyriakidis, S., Tiedemann, J.: NLI data sanity check: Assessing the effect of data corruption on model performance. In: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 276–287. Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online) (2021)

[41] Dasgupta, I., Lampinen, A.K., Chan, S.C.Y., Creswell, A., Kumaran, D., McClelland, J.L., Hill, F.: Language models show human-like content effects on reasoning. arXiv **2207.0705** (2022)

[42] Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E.: Dissociating language and thought in large language models: a cognitive perspective. arXiv (2023) 301.06627 [cs.CL]

[43] Weiser, B.: Here's what happens when your lawyer uses ChatGPT. The New York Times **May 27 2023** (2023)

[44] Gibson, E., Futrell, R., Piantadosi, S.P., Dautriche, I., Mahowald, K., Bergen, L., Levy, R.: How efficiency shapes human language. Trends in Cognitive Sciences **23**(5), 38–407 (2019)

[45] Kågebäck, M., Carlsson, E., Dubhashi, D., Sayeed, A.: A reinforcement-learning approach to efficient communication. PLoS ONE **July 15 2020** (2020)

[46] Chaabouni, R., Kharitonov, E., Dupoux, E., Baroni, M.: Anti-Efficient Encoding in Emergent Communication. Curran Associates Inc., Red Hook, NY, USA (2019)

[47] Lian, Y., Bisazza, A., Verhoef, T.: The effect of efficient messaging and input variability on neural-agent iterated language learning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10121–10129. Association for Computational Linguistics, Stroudsburg, PA (2021)

[48] Bernardy, J.-P., Lappin: Unitary recurrent networks. In: Lappin, S., Bernardy, J.-P. (eds.) Algebraic Structures in Natural Language, pp. 243–277. CRC Press, Taylor & Francis, Boca Raton, London, and New York (2023)