Using Large Language Models as Beneficial Tools in Education

Anonymous ACL submission

Abstract

Public fame and easy open access to the ChatGPT, and the following wide use, or what could be considered misuse and abuse, of the model by some in the education and research communities, caused initially sharp negative reaction in the education and academic institutions and publishing services, aimed at detection and ban of the LLM (Large Language Models) generated texts, under efforts to combat plagiarism and chatting. Later, upon realising that such a blanket prohibition is technically problematic with the desired degree of reliability and confidence, as well as that LLMs can be legitimately used as tools for increasing productivity by taking on mundane writing tasks, the communities' attitude relaxed. The most remarkable changes in the public discourse are related to rethinking the very aims of the education system: "If some of the areas of the intellectual labour could be automated and become obsoleted by LLM, maybe it is time for education to concentrate on teaching students to think and behave not like LLMs"? Such a Constructivist view on education, considered unrealistic a century ago, now may become the only sound way forward.

1 Introduction

Large Language Models (LLM) are posed to replace a significant part of so-called intellectual labour. Students, being taught by the current education system primarily to memorise, or at least to obtain pre-packaged "knowledge", will risk being outcompeted by the more efficient LLMs on routine and trivial tasks, which require extensive information search and mundane text generation. Therefore, new education adapted to the LLMs' presence needs to find intellectual labour niches in which humans are superior to LLMs, and needs to teach students to be not like LLMs to maintain competitiveness in the new market. Hence, significant changes are needed in education, preliminaries to which, and changes themselves, we discuss in this position paper.

The contribution is organized in the following manner: Section 2 gives a brief overview of the attitude development on the LLM emergence; Section 3 discusses LLM flaws ; Section 4 outlines potential education changes to incorporate into the teaching process; and Section 5 concludes the discussion.

2 Large Language Models - a Friend or a Foe?

An explosive debut in public of the ChatGPT (Bib, 2023a) and the following similar Large Language Models (LLM) (Bib, 2023c; Chowdhery et al., 2022; Bib, 2023b; Touvron et al., 2023) also initiated a debate on LLMs' effects on education. An obvious first reaction was concern about abusing the LLMs' ability to generate human-like texts for cheating and plagiarism (Orenstrakh et al., 2023) in such examinations and tests that evaluate students in such faculties as memorisation, summarisation, reviewing, and basic analysis. Various methods of detection and prevention of using LLMs in education and academia were proposed (Tang et al., 2023; Khalil and Er, 2023; Rodriguez et al., 2022; Savelka et al., 2023).

However, the next wave of publications on the place of LLMs in education started to contemplate the thought that even if education shut the doors before LLMs, the industry would not, such as putting graduates who are not accustomed to the use of LLMs at a disadvantage. The publications started coming to the conclusion that education itself should change, not pursuing obsolete goals and not executing obsolete practices (Anders, 2023; Rudolph et al., 2023), but instead concentrating more on the areas where human-lead education (even armed with LLMs as tools) has advantages over mere LLMs in themselves (Fuchs, 2023; Cope and Kalantzis, 2019).

From the literary text analysis perspective, the generated by LLMs, though usually syntactically correct, are effete, emotionless washed-up texts, lacking linguistic variability and distinctness, and pragmatic intercity and originality (Gao et al., 2022; Chaves and Gerosa, 2021; Wilkenfeld et al., 2022; Mitrović et al., 2023). On the dynamic debating or deliberation text generation, LLMs also perform far from ideal. For example, on detecting discourse move, ChatGPT performed even worse than simple BERT models (Wang et al., 2023). Debates with ChatGPT, as everybody can see using the OpenAI interface, suffer from circular arguments, self-contradiction, and evasiveness - tendencies to please human preferences in Reinforcement Learning (RL) (Ramamurthy et al., 2022; Carta et al., 2023) - exactly those practices that nobody wants to foster in students. When used to detect manipulative discussion tactics of cyberattacks, ChatGPT also scored significantly worse than simple BERT models (Fayyazi and Yang, 2023).

General LLMs' problems with functional domains such as mathematics, reasoning, and logic (Frieder et al., 2023), emotional expressivity, wit, humour and ethics (Borji, 2023; Arkoudas, 2023), factual data, privacy, and false, bias and discrimination (Basta et al., 2019; Kurita et al., 2019; Sheng et al., 2019; Gehman et al., 2020; Bib, 2022; Bianchi et al., 2022; Weidinger et al., 2021; Tang et al.; Goldstein et al., 2023) are well documented. Machine Learning (ML) specific problems of LLMs add such issues as lack of interpretability and understanding, (Bender and Koller, 2020; Lake and Murphy, 2020; Marcus et al., 2022; Ouyang et al., 2022; Leivada et al., 2022; Ruis et al., 2022), and catastrophic ageing and forgetting by LLMs (Lazaridou et al., 2021; Hombaiah et al., 2021; Dhingra et al., 2022; McCloskey and Cohen, 1989; Parisi et al., 2019; Ratcliff, 1990; Kirkpatrick et al., 2017). When using LLMs in education, their shortcomings may not only be accounted for in the real-life application but also can be used as a foundation of fresh approaches to education to foster those qualities and skills of students that will not be made obsolete by the use of LLMs, and on the opposite, give students a competitive edge.

3 Fundamental Foundations of the LLMs' Flaws

Although implementation details of the latest models are kept proprietary, previously published research shows that LLM models are built and trained using three main principles. Traditional Natural Language Processing (NLP) tokenizing techniques include the preprocessing stage, on which "stop-words" are removed, remaining words are stemmed and lemmatized (converted to canonical dictionary form), and the Bag of Words (BoW) algorithm is used to map lemmatized words into a linear vector space, spanned on the most frequent and important words dictionary basis. The whole sentence or a bigger text is represented as a linear sum of all token vectors (or also so-called "embeddings") (Zhang et al., 2010). Such an approach is very resource usage effective but does not count in the sentence or larger text structure. For example, such sentences as: "A dog bites a man", "A man bites a dog", and "Dogs bite men" would be represented by the same embedding.

To introduce implicit elements of the linguistic structures, modern NLP models frequently use context tokenizers (Taylor, 1953) of the BERT-like family (Devlin et al., 2018). A simple illustration of the BoW and BERT embedding differences would be the former creating "DOG", "BITE", "MAN", and the latter - "nullDOGbite", "dogBITEman", "biteMANnull", "nullMANbite", "manBITEdog", "biteDOGnull". That solves the BoW's structure blindness problem but greatly increases the dimensionality of the embedding space, which is the starting point of LLMs' high computational demands and size.

The second foundation technology the LLMs use is based on the statistical n-gram approach (Brown et al., 1992). The supervised training of the Machine Learning (ML) models has a bottleneck in the manual labelling of the training data sets. To process high amounts of text and other media, LLM uses a self-supervised approach based on the Masked Language Model (MLM) (Salazar et al., 2019; Besag, 1975). In such a paradigm, part of the words are kept hidden from the ML model in training, and the purpose of the training is to find words with the highest probability of being in the hidden positions. Again, such an approach does not directly model linguistic structures but implicitly stochastically takes them into account.

To keep with the human reader's attention span and produce a coherent flow of text, LLMs have to use long context windows for MLM training of thousands of words. The brute force use of the whole continuous windows is computationally problematic; therefore, another technique of extracting the most valuable and influential context words on the predicted word gave birth to computationally tractable but still huge LLMs - Attention mechanism (Bahdanau et al., 2014; Luong et al., 2015; Gehring et al., 2016) and its Transformer implementation (Vaswani et al., 2017). In such an approach of "attention", learnable matrices are used to compute cosine or Euclidean distances between the word relevance to the projected prediction over the context window sliding, and the most consistent contributor over time is kept and used, in such a way, reducing computational demand.

The stochastic nature of the LLMs in modelling structured natural languages has been a point of fierce debate since the LLMs introduction (Bender et al., 2021; Schick and Schütze, 2020; Marcus, 2018; Blodgett and Madaio, 2021; Bommasani et al., 2021).

Another obvious problem of LLMs is the naivety of their language representation from the theoretical linguistics perspective that operates with categories of syntactic and semantic structures. The former are various kinds or relations in the mathematical sense (Combe et al., 2022; Marcolli et al., 2023), specific to particular languages, which endow non-ordered multi-sets of the morphing lexemes and are continuously mapped to the universal semantic structures (of meaning or of thought) (Chomsky, 2023) (or, possibly, to universal grammar) (Watumull and Chomsky, 2020).

Noam Chomsky especially emphasises the non-locality of such synthetic units. For example, in inflectional languages such as Balto-Slavic, or agglutinating such as Japanese, the non-locality is obvious because of their free word order, but even for the significantly sequential analytic English, Chomsky referees at the semantic attachment of an adverb to a correct verb regardless of their position and order, for example in "Intuitively, birds that fly swim" (Berwick and Chomsky, 2016).

Building models of such complex relations in LLMs, capable of discovering and retrieving such linguistic structures and, in such a way, achieving explainability and interoperability of LLMs, is a drastically undeveloped area of research (Delétang et al., 2022), frequently limited to naive methods of asking LLMs about their internals (Jiang et al., 2020).

These mechanisms introduce implicit naive syntax emulation elements by projecting hierarchical tree structures on flat sequences but with the loss of complexity. For example, in Chomsky's example, "Intuitevely" can become the sequential neighbour of "swim" by dropping "fly".

Even more complicated question of whether LLMs can model thought and intelligence, although receiving some optimistic answers (Kosinski, 2023; Bubeck et al., 2023), predominately answered negatively (Ullman, 2023; Sap et al., 2022).

From the linguistics view on natural human languages, universal semantic roles and relations between parts of a sentence, for example "Elmer threw a porcupine to Hortense", such as Actor (Elmer), Patient (porcupine), and Beneficiary (Hortense) could be mapped to syntactic roles and relations, specific to particular languages (Marantz, 1981). In English, syntactic relations between Subject, Direct and Indirect Objects are marked by the order and prepositions (to); in languages such as Balto-Slavic - by the case (nominative, accusative, dative) suffixes; in Japanese - by particles (を, に).

However, the question of what is the language of semantics/meaning, or the "language of thought", and how it is externalised into syntactic structures, is difficult even for linguistics and neuroscience of the natural human languages (Gallistel, 2011).

Surprisingly, in the last years, the voices of the critics of the limitations of the traditional narrow ML (and LLMs as part of it), such as Noam Chomsky and Garry Marcus, were joined by such big names of the narrow ML as Joshua Bengio (Lex Clips, 2023), Yann LeCun (Bib, 2023d), and even Geoffrey Hinton whose students built ChatGPT (Metz, 2023).

4 Education Ameliorating Horizons In the Context of LLMs

Although LLMs lack agency, structural representation of the language, and real-world picture (Browning, 2022; Floridi, 2023), they, under human teacher supervision, could still be used to help foster those abilities in students. Such non-commodified abilities to behave not like LLM (LLMs behaviour is described by Ben Goertzel as "competent mediocrity" (Charrington, 2023)), will remain in high value and demand.

Educational methodologies founded on initiative, curiosity, and active actionable students' construction of knowledge (Vygotsky, 2012; Beilin, 1992; Shchedrovitsky, 1995), and therefore demanding high educator involvement, hence prohibitively costly, with the routine and trivial tasks delegated to LLMs may become practically sound. We want students to be "competent", for which goal LLMs may be useful tools and examples, but also not "mediocre", for which LLMs may be used as counter-example tools. It's been observed that LLM-generated scientific paper abstracts are easily identified by humans based on Goertzel's "competent mediocrity" style, though such estimates have a noticeable false positive error - people also write papers in such a style (Gao et al., 2022).

Sporadic research in applying LLMs to education change in the active direction is visible in publications. For example, one of the routine tasks a competent educator may be released from, but a general eye on, is the trace of the students' discourse flow (Wang et al., 2023), or teamwork feedback (Katz et al., 2023). Constant feedback, personalized and adaptive learning (Annuš, 2023), student initiative and psychometrics (Katz et al., 2023), collaborative, transparent and diverse intelligence (Cope et al., 2021). LLMs and other AI models are inherently student-driven, and it's up to the education system, particularly up to its change, to view and experience that drive as a threat or benefit (Dai et al., 2023; Haensch et al., 2023).

We propose systematic research on the use of LLMs and other AI methods in practical implementation methods of education of constructing knowledge and understanding, such as (but not limited to):

- Fostering a big picture view, understanding, and based on them, first-hand actionable application, experimentation and implementation of the knowledge.
- Continuous, recursive (i.e. changing assignments) feedback (aizuchi - a rare Japanese loan into English linguistic jargon (Kita and Ide, 2007)).
- Pursuit of student questions and interests. Interactive (i.e. self-assigning) and co-acting (together with pedagogue) learning.
- Non-disciplinary or non-didactic learning, self-involved assessment.
- Dynamic knowledge acquisition, with each step in it being a challenge for the student, seemingly impossible, but with guidance and work achievable, building confidence in own abilities.
- Collaborative, social learning - learning through teaching other students.
- Emotion and sentiment expression aware and competent learning and teaching.

5 Discussion and Conclusions

5.1 Limitations

The presented review is in no way comprehensive and exhaustive - a number of publications on various aspects of LLM creation and use are published at an astonishing rate, and the very LLM landscape is changing quickly, outpacing academic publishing cycles. The research results are frequently contradicting, not merely because some of them are not rigorous - the research field is so vast that available results are fragmented and patchy, depending on the initial conditions that hardly can cover comprehensively all possible aspects of the LLM use. Inevitably, this opinion piece is incomplete in its foundations and subjective in proposals.

5.2 Risks

A significant change in the education system, especially if it is related to a significant cost increase, and hence, applied to limited society strata, can lead to further societal disparity. However, the risks of keeping the outdated education system that produces an incompetent and unneeded workforce can be even greater.

5.3 Conclusions

Under the likely perspective of LLMs taking on a significant share of the previously thought of "intellectual" labour, education needs to shift its goals and methods to fostering students' abilities and habits that differentiate them from LLMs. That requires gaining a better understanding of what LLMs can not successfully do, not only from the empirical perspective but also from the first principles laying in the foundations of LLM. Building the education system from human strengths, such as agency, individual initiative and interest, social collaboration, emotional involvement, and structural view of the language and world picture, would likely require significant and expensive education system change, the core of which would likely align with the Constructivist view on it Of Vygotsky and Piaget.

References

2022. Doctor GPT-3: hype or reality? - Nabla, https://www.nabla.com/blog/gpt-3. [Online; accessed 5. Sep. 2022].

2023a. GPT-4, https://openai.com/research/gpt-4. [Online; accessed 16. Mar. 2023].

2023b. Introducing LLaMA: A foundational, 65-billion-parameter language model, https://ai.facebook.com/blog/large-language-model-llama-meta-ai. [Online; accessed 17. Mar. 2023].

2023c. Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance, https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html. [Online; accessed 17. Mar. 2023].

2023d. Post | LinkedIn. [Online; accessed 11. Aug. 2023].

Brent A Anders. 2023. Is using chatgpt cheating, plagiarism, both, neither, or forward thinking? *Patterns*, 4(3).

Norbert Annuš. 2023. Chatbots in education: The impact of artificial intelligence based chatgpt on teachers and students. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(4):366–370.

Konstantine Arkoudas. 2023. ChatGPT is no stochastic parrot. But it also claims 1 > 1. | Medium. *Medium*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings.

Harry Beilin. 1992. Piaget's enduring contribution to developmental psychology. *Developmental psychology*, 28(2):191.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Robert C Berwick and Noam Chomsky. 2016. *Why only us: Language and evolution*. MIT press.

Julian Besag. 1975. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale.

Su Lin Blodgett and Michael Madaio. 2021. Risks of AI foundation models in education. *CoRR*, abs/2110.10024.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

Jacob Browning. 2022. AI And The Limits Of Language. *NOEMA*.

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

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. *arXiv preprint arXiv:2302.02662*.

The Twiml Ai Podcast with Sam Charrington. 2023. Are Large Language Models a Path to AGI? with Ben Goertzel - 625. [Online; accessed 9. Aug. 2023].

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. The impact of chatbot linguistic register on user perceptions: a replication study. In *International Workshop on Chatbot Research and Design*, pages 143–159. Springer.

Noam Chomsky. 2023. Genuine explanation and the strong minimalist thesis. *Cognitive Semantics*, 8(3):347–365.

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

Noemie Combe, Yuri I Manin, and Matilde Marcolli. 2022. Geometry of information: Classical and quantum aspects. *Theoretical Computer Science*, 908:2–27.

Bill Cope and Mary Kalantzis. 2019. Education 2.0: Artificial intelligence and the end of the test. *Beijing International Review of Education*, 1(2-3):528–543.

Bill Cope, Mary Kalantzis, and Duane Searsmith. 2021. Artificial intelligence for education: Knowledge and its assessment in ai-enabled learning ecologies. *Educational Philosophy and Theory*, 53(12):1229–1245.

Yun Dai, Ang Liu, and Cher Ping Lim. 2023. Reconceptualizing chatgpt and generative ai as a student-driven innovation in higher education.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, et al. 2022. Neural networks and the chomsky hierarchy. *arXiv preprint arXiv:2207.02098*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Reza Fayyazi and Shanchieh Jay Yang. 2023. On the uses of large language models to interpret ambiguous cyberattack descriptions. *arXiv preprint arXiv:2306.14062*.

Luciano Floridi. 2023. Ai as agency without intelligence: on chatgpt, large language models, and other generative models. *Philosophy & Technology*, 36(1):15.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Kevin Fuchs. 2023. Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse? In *Frontiers in Education*, volume 8, page 1166682. Frontiers.

Charles Randy Gallistel. 2011. Prelinguistic thought. *Language learning and development*, 7(4):253–262.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv*, pages 2022–12.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations.

Anna-Carolina Haensch, Sarah Ball, Markus Herklotz, and Frauke Kreuter. 2023. Seeing chatgpt through students' eyes: An analysis of tiktok data. *arXiv preprint arXiv:2303.05349*.

Spurthi Amba Hombaiah, Tao Chen, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Dynamic language models for continuously evolving content. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp Data Mining*. ACM.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Andrew Katz, Siqing Wei, Gaurav Nanda, Christopher Brinton, and Matthew Ohland. 2023. Exploring the efficacy of chatgpt in analyzing student teamwork feedback with an existing taxonomy. *arXiv preprint arXiv:2305.11882*.

Mohammad Khalil and Erkan Er. 2023. Will chatgpt get you caught? rethinking of plagiarism detection. *arXiv preprint arXiv:2302.04335*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks.

Proceedings of the national academy of sciences, 114(13):3521–3526.

Sotaro Kita and Sachiko Ide. 2007. Nodding, aizuchi, and final particles in japanese conversation: How conversation reflects the ideology of communication and social relationships. *Journal of Pragmatics*, 39(7):1242–1254.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations.

Brenden M. Lake and Gregory L. Murphy. 2020. Word meaning in minds and machines.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, C d M d'Autume, Sebastian Ruder, Dani Yogatama, et al. 2021. Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*.

Evelina Leivada, Elliot Murphy, and Gary Marcus. 2022. Dall-e 2 fails to reliably capture common syntactic processes.

Lex Clips. 2023. Yoshua Bengio: From System 1 Deep Learning to System 2 Deep Learning (NeurIPS 2019). [Online; accessed 11. Aug. 2023].

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Alec Marantz. 1981. *On the nature of grammatical relations*. Ph.D. thesis, Massachusetts Institute of Technology.

Matilde Marcolli, Noam Chomsky, and Robert Berwick. 2023. Mathematical structure of syntactic merge. *arXiv preprint arXiv:2305.18278*.

Gary Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of dall-e 2.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Cade Metz. 2023. 'The Godfather of AI' Quits Google and Warns of Danger Ahead. *N.Y. Times*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Michael Sheinman Orenstrakh, Oscar Karnalim, Carlos Anibal Suarez, and Michael Liut. 2023. Detecting llm-generated text in computing education: A comparative study for chatgpt cases. *arXiv preprint arXiv:2307.07411*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-domain detection of gpt-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233.

Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1).

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Jaromir Savelka, Arav Agarwal, Christopher Bogart, Yifan Song, and Majd Sakr. 2023. Can generative pre-trained transformers (gpt) pass assessments in higher education programming courses? *arXiv preprint arXiv:2303.09325*.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118.

G P Shchedrovitsky. 1995. *Selected works*. Shola Kulturnoi Politiki.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining?

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Lev S Vygotsky. 2012. *Thought and language*. MIT press.

Deliang Wang, Dapeng Shan, Yaqian Zheng, Kai Guo, Gaowei Chen, and Yu Lu. 2023. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, page 515–519. International Educational Data Mining Society.

Jeffrey Watumull and Noam Chomsky. 2020. Rethinking universality. *Syntactic architecture and its consequences II*, page 3.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

J Nan Wilkenfeld, Bei Yan, Jujun Huang, Guirong Luo, and Kristina Algas. 2022. "ai love you": Linguistic convergence in human-chatbot relationship development. In *Academy of Management Proceedings*, volume 2022, page 17063. Academy of Management Briarcliff Manor, NY 10510.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52.