Lessons from using PLMs for Human Cognitive Modeling

Anonymous ACL submission

Abstract

Many studies show evidence for cognitive abilities in Pre-trained Language Models (PLMs). Researchers have evaluated the cognitive alignment of PLMs, i.e., their correspondence to adult performance across a range of cognitive domains. More recently, the focus has expanded to the developmental alignment of these models: identifying phases during training where improvements in model performance track improvements in children's thinking over development. However, challenges to this use are twofold: (1) PLMs have very different architectures than human minds and brains, and the data sets on which they are trained differ in many ways from the inputs children receive. (2) The "outputs" of PLMs are different from the behavioral measures that cognitive scientists collect in their experiments and evaluate their theories against. In this paper, we distill lessons learned from using PLMs for cognitive modeling and outline the pitfalls of attempting to use PLMs, not as engineering artifacts, but as cognitive science and developmental science models. We review assumptions used by researchers to map measures of PLM performance to measures of human performances and then, *enumerate criteria for using PLMs as credible accounts of cognition and cognitive development*.

1 Introduction

With the improving performance of language models (Touvron et al., 2023; Gemini Team, 2023; OpenAI, 2023; Wei et al., 2022), researchers have increasingly advocated for the use of Language models as computational models of cognition (Piantadosi, 2023; Mahowald et al., 2024; Warstadt and Bowman, 2024). This includes domains such as mathematical reasoning (Shah et al., 2023; Ahn et al., 2024), language comprehension (Warstadt et al., 2020; Ye et al., 2023; Koubaa, 2023), concept understanding (Vemuri et al., 2024), and analogical reasoning (Webb et al., 2023; Hu et al., 2023).

More recently, researchers have used PLMs for modeling cognitive development in children (Hosseini et al., 2022; Kosoy et al., 2023; Frank, 2023; Shah et al., 2024). For example, Portelance et al. (2023) suggests the use of language models to predict the age of acquisition of words in children. Shah et al. (2024) map the development of cognitive intelligence in humans to scaling training tokens and model size in PLMs. Other researchers also propose studying bilingualism by mapping pre-training steps in PLMs to understand the rate of language development (Evanson et al., 2023; Marian, 2023; Sharma et al., 2024).

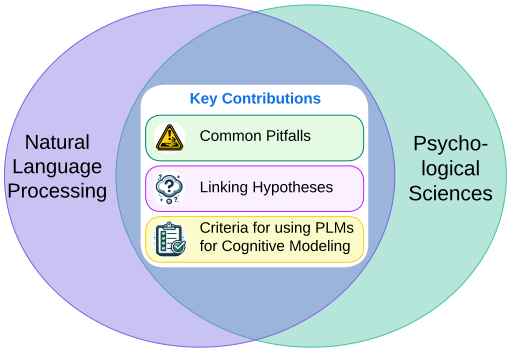

Figure 1: Overview of the Lessons from using PLMs for Human Cognitive Modeling.

In this paper, we advocate for the use of PLMs as candidate theories of cognitive and developmental science. We first review the pitfalls of using PLMs in psychological science and caution researchers against over-interpreting PLM alignment to human cognition. We then review the common assumptions used by researchers to map measures of PLM performance to measures of human performance. In doing so, we build upon previous work enumerating best practices for cognitive evaluations of PLMs (Ivanova, 2023; Mahowald et al., 2024).

2 Pitfalls of using PLMs as scientific theories

Some pitfalls come when using PLMs as cognitive science theories, i.e., of adult thinking.

- Human brains and PLMs are architecturally different. Recent research is attempting to map regions of the brain to different aspects of PLMs (layers, attention heads, etc.) in terms of local functionality and performance characteristics. However, this work is in its infancy, and its viability remains an open question (Hosseini et al., 2022; Kauf et al., 2023).

- Researchers use a *linking hypothesis* to map model performance characteristics (e.g., log probabilities) to human performance characteristics (e.g., reading times) (Shah et al., 2024). These links are often quite distal, making it unclear whether PLMs are actually "explaining" cognitive science data. (See the next section for further discussion.)

- PLMs are opaque and have limited interpretability. Human alignment and the lack thereof is hard to debug. This is a barrier to treating these models as scientific theories (McGrath et al.; Kar et al., 2022).

- Most studies evaluating the cognitive alignment of PLMs focus on a narrow range of cognitive abilities and overlook correlations with other abilities. This is in contrast with psychometric approaches to intelligence that investigate the correlations across tests of a broad range of cognitive abilities: mathematical, verbal, spatial, fluid, and so on (Snow et al., 1984; Schneider and McGrew, 2012). This is also in contrast to unified theories of cognition that attempt to model all cognitive abilities within a single computational framework (Mellon et al., 2007; Varma, 2011).

Other pitfalls are specific to the use of PLMs as developmental science theories, i.e., of the progressions in children's thinking.

- PLM checkpoints are snapshots or fingerprints of the data they are trained on. Most research only looks at the final model checkpoints and not the change in the cognitive alignment of language models (development) as a function of data observed (Warstadt and Bowman, 2022; Frank, 2023; Shah et al., 2024). Often, this is because of the unavailability of intermediate training checkpoints or resource constraints. This limits our understanding of the nature of PLM training.

- Differences also exist in the nature of the data observed by PLMs versus humans. PLMs are trained on textual data that is magnitudes larger than the number of words seen by children (Huebner et al., 2021; Hosseini et al., 2022; Warstadt et al., 2023; Bhardwaj et al., 2024). On the other hand, children learn from input from multiple senses (Smith and Gasser, 2005), whereas models are not of an embodied nature (Chemero, 2023).

- For studies evaluating the developmental alignment, the observed developmental trajectories might be artifacts of the pre-training order (Shah et al., 2024).

3 Linking Hypotheses Mapping Model Performance to Human Performance

Researchers use various linking hypotheses to map PLM performance to human performance measures. While all these linking hypotheses are simple in theory, they have multiple possible operationalizations. We review four below.

Similarity computations: Many cognitive tasks require people to judge the similarity of two items. In this case, human similarity judgments are directly modeled by computing the similarity between the corresponding representations in a PLM's latent space. This can be via cosine similarity or another metric. One example is typicality effects, which is the finding people regard some members as "better" examples of a category than others (Bhatia and Richie, 2022; Richie and Bhatia, 2021). The rank of an item is determined by the proportion of humans that produce an item when asked to enumerate the items of the category. In language models, the typicality of an exemplar for a category is estimated by encoding the exemplar name as a string, passing it through the language model, and obtaining the corresponding word embedding. Thereafter, the similarity between this exemplar vector and the category prototype is calculated, with a higher value indicating that the exemplar is more typical.

Other cognitive tasks require comparing or discriminating between two items. There, a common linking hypothesis is that the greater the similarity between the items in the model's latent space, the longer the comparison/discrimination time. For example, Shah et al. (2023) map the time it takes to compare two numbers to PLM similarity - the greater the similarity of two number representations, the longer it takes for humans to differentiate which one is greater (or lesser).

However, there are two common obstacles to using the *similarity* linking hypothesis. The first is that doing so requires models that make available the latent representations for similarity computations. The second problem is that this method suffers from problems due to tokenization. Humans use words as granular units while models use tokens (potentially words or subwords). This makes the nature of the mapping inconsistent as one unit of text for humans (words) may be mapped to two units of text for PLMs (tokens).

Surprisal values: One way of quantifying the uncertainty of model generations is in terms of the summation of logarithmic probabilities. A common linking hypothesis is that higher surprisal values correspond to longer human response times. For instance, studies of reading (Rambelli et al., 2024; Ivanova et al., 2024b) and categorization (Misra et al., 2021) have found that higher model uncertainties predict longer human response times. Relative log probabilities have been used to distinguish grammatical and ungrammatical sentences (Warstadt et al., 2020; Shah et al., 2024). They enable us to directly compare the right answer with all the candidate answers in a deterministic manner, i.e., there is no chance that the PLM will not calculate the sequential probability for a candidate string. Research also shows that surprisal values provide a better match to human plausibility judgments than prompts (Ivanova et al., 2024a).

However, aproblem with surprisal-based approaches is that they fail to show robustness to context (prompts).

Prompting: PLM generation is probabilistic and therefore the same model can give different results across inference runs. A PLM is prompted to follow the same exercise as a human multiple times and generate a probability distribution over the output space. The probability of the correct output is then mapped to model confidence. Directly comparing the generated behavior of PLMs with that of humans reduces or even eliminates the need for linking hypotheses (Patel and Pavlick, 2021; Webb et al., 2023; Zahraei and Emami, 2024).
For example, a PLM can be prompted with:
Which statement is grammatically correct? Your response must be "1" or "2".
1. Noah likes to swim. 2. Noah likes to.
The PLM can generate "1" for the correct answer.

Another benefit of prompting is that it allows for *variable output length*. Tasks that benefit from this flexibility, like commonsense reasoning, are more suited for prompting (Yasunaga et al., 2023).
Example SAT analogical reasoning problem:
Analogy: Runner : Marathon :: ?
Options:

- Envoy : Embassy
- Martyr : Massacre
- oarsman : Regatta
- Horse : Stable

An example problem benefitting from *variable output length* is SAT analogical reasoning tasks (Turney, 2013). These are of the form A:B::? (see example above). These are MCQ-based choice tasks that can be operationalized as similarity computation, surprisal, or a prompt-based reasoning problem. In the prompting case, we can force the generations to adhere to a goal: *answer the correct analogy in this output format: {"A":"B"::"C":"D"}.*

That said, there are several limitations to the prompting approach. First, PLMs are *not robust to the prompt format.* Answering prompts requires PLMs to have two abilities, (1) understanding the prompt and (2) knowing the answer. Often, the outputs of PLMs vary substantially based on the input, and sometimes maximum performance characteristics are obtained on gibberish prompts (Deng et al., 2022). Second, PLMs may output an answer beyond the given candidate options for the prompting approach. For example, the answer to the problem described in blue above could be "It would be irresponsible to imply that the grammatical structure of a sentence is inconsequential, as clear communication is fundamental for safety and understanding" (Cai et al., 2024).

Direct probing: Another alternative is to directly ask the model about its current state. For example, to measure the incremental semantic understanding of temporarily ambiguous sentences, Li et al. (2024) presents a dichotomous verification sentence to the model after each word. In addition to the direct probing, one can directly recover the implicit parse tree after each word to measure incremental syntactic understanding (Manning et al., 2020). PLM prompting helps support direct probing, which is often not possible in humans with behavioral measures or even with neuroimaging measures to “look under the hood”.

4 Criteria for Evaluating and Developing PLMs as Scientific Models

PLMs are being increasingly considered as models of cognitive and developmental science phenomena.

In light of the pitfalls outlined above, we propose two sets of criteria for using PLMs for this purpose.

The first set concerns the *appropriateness* of PLMs as scientific tools for cognitive and developmental modeling:

- **Design multiple experiments to test the alignment of each cognitive or developmental phenomenon:** PLMs may track the human performance characteristic well under one linking hypothesis or one type of test. However, this alignment may just be an artifact, for example, of pre-training data contamination. Conducting more experiments evaluating the same cognitive/developmental phenomena establishes stronger empirical plausibility.

- **Test the path-dependency of PLMs for developmental alignment:** The claim that the final model state of a PLM approximates adult performance leads to the question of the path by which it arrived there. Ideally, the model's performance improvements over training should also track the progression of cognitive abilities over development (Elman, 1996; Bengio et al., 2009). This would support researchers exploring the scaling of training data and model size in their investigations of human development.

- **Use multiple methods to interpret PLM successes and failures:** PLMs lack explainability and interpretability due to their large size (McGrath et al.). Some methods for PLM interpretation are often better than others. For example, in the experiments conducted by Li et al. (2024), incrementally constructed parse trees provided a better account of PLM alignment than the information from attention weights.

- **Control for tuning techniques:** PLMs are often tuned on specific data and in different manners like Instruction Tuning, Reinforcement Learning from Human Feedback, etc. These tuning techniques influence model behavior and the model's output centers around tuning goals rather than developing a representation of world knowledge.

- **Remember the linking hypothesis:** Adaptation of human experimental materials to textual counterparts requires certain assumptions (refer to different operationalization in section 3). These assumptions need to be well documented and understood based on the experiment type.

The second set of criteria is for guiding the *development* of PLMs as credible accounts of cognition and its development. This is a more open-ended task, and the following can be considered as mere suggestions to researchers:

- Evaluation techniques should follow the appropriateness criteria above. PLMs should be evaluated at regular intervals of pre-training to assess their potential developmental alignment, which is often overlooked in studies of cognitive alignment.

- PLMs can be tuned with specific cognitively important tasks and evaluated on a breadth of cognitively relevant tasks. For example, typicality experiments (Vemuri et al., 2024; Misra et al., 2021) could be used to preference-tune PLMs using reinforcement learning techniques. This may lead to better cognitive alignment of PLMs across a broader set of tasks.

- Pre-training data may benefit from developmentally plausible corpora (Bhardwaj et al., 2024; Warstadt and Bowman, 2022; Frank, 2023). This includes training on a curriculum based on human skill acquisition, for example, the age-of-acquisition of a word (Huebner et al., 2021; Portelance et al., 2023). Informed pre-training will allow us to better understand the developmental alignment of models.

5 Conclusion

This paper advocates for the use of Pre-trained Language Models (PLMs) as theoretical tools for investigating human cognition and its development In this advocacy, we are not alone (Warstadt and Bowman, 2022; McGrath et al.; Frank, 2023; Mahowald et al., 2024). At the same time, we caution researchers towards the informed use of PLMs in psychological sciences. We have highlighted, common pitfalls, reviewed the different linking hypotheses used by researchers to map PLM performance to human performance, and outlined criteria for evaluating and developing PLMs as credible models of cognition and cognitive development. These criteria are intended to guide researchers in designing robust experiments, interpreting PLM behaviors accurately, and increasing fidelity to human data. Given the constantly evolving nature of the field, we call for researchers to continuously refine and expand these guidelines to match new advancements in NLP and Cognitive Science.

6 Limitations

(1) The paper highlights common pitfalls, linking hypotheses, and evaluative criteria while using PLMs for cognitive modeling. These constitute a set of sound views to aid new researchers in the field. They do not exhaustively cover every pitfall, hypothesis, or criterion. (2) The suggestions in this work are good-to-have practices that support the use of PLMs for open cognitive and developmental science. There is no one-answer-fits-all approach for these tasks. Natural Language Processing is a developing field and we recommend articulating newer guidelines and practices as more PLMs are built and deployed. (3) Our work calls for the use of language technologies for the psychological sciences and provides criteria for developing credible accounts of cognition and cognitive development. Despite providing general guidelines, our work does not conduct experiments or offer any empirical evidence of performance comparisons or other quantitative measures.

7 Ethical Considerations

There are no major risks associated with conducting this research beyond those associated with working with PLMs. There may be risks in misinterpreting the criteria enlisted in this study. The suggestions in this study are one-way: we wish to find human performance characteristics and behaviors in PLMs to help model psychological sciences to aid people with cognitive impairments. We do not advocate for developing PLMs to replace humans or suggest ways to reach Artificial General Intelligence. PLMs are experimental technologies and future work using these models should proceed with caution.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Khushi Bhardwaj, Raj Sanjay Shah, and Sashank Varma. 2024. Pre-training llms using human-like development data corpus. *Preprint*, arXiv:2311.04666.

Sudeep Bhatia and Russell Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Alice Cai, Ian Arawjo, and Elena L Glassman. 2024. Antagonistic ai. *arXiv preprint arXiv:2402.07350*.

Anthony Chemero. 2023. Llms differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11):1828–1829.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Jeffrey L Elman. 1996. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press.

Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? *arXiv preprint arXiv:2306.03586*.

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2022. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *BioRxiv*, pages 2022–10.

Xiaoyang Hu, Shane Storks, Richard L Lewis, and Joyce Chai. 2023. In-context analogical reasoning with pre-trained language models. *arXiv preprint arXiv:2305.17626*.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Anna A Ivanova. 2023. Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv preprint arXiv:2312.01276*.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Evelina Fedorenko, and Jacob Andreas. 2024a. Log probability scores provide a closer match to human plausibility judgments than prompt-based evaluations.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. 2024b. Elements of world knowledge (ewok):

A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Kohitij Kar, Simon Kornblith, and Evelina Fedorenko. 2022. Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nature Machine Intelligence*, 4(12):1065–1067.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Eliza Kosoy, Emily Rose Reagan, Leslie Lai, Alison Gopnik, and Danielle Krettek Cobb. 2023. Comparing machines and children: Using developmental psychology experiments to assess the strengths and weaknesses of lamda responses. *arXiv preprint arXiv:2305.11243*.

Anis Koubaa. 2023. GPT-4 vs. GPT-3.5: A concise showdown.

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Viorica Marian. 2023. Studying second language acquisition in the age of large language models: Unlocking the mysteries of language and learning, a commentary on "age effects in second language acquisition: Expanding the emergentist account" by catherine l. caldwell-harris and brian macwhinney. *Brain and language*, 246.

Sam Whitman McGrath, Jacob Russin, Ellie Pavlick, and Roman Feiman. How can deep neural networks inform theory in psychological science?

John R Anderson Richard King Mellon et al. 2007. *How can the human mind occur in the physical universe?*, volume 3. Oxford University Press, USA.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Steven Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Eva Portelance, Yuguang Duan, Michael C. Frank, and Gary Lupyan. 2023. Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive science*, 47 9:e13334.

Giulia Rambelli, Emmanuele Chersoni, Davide Testa, Philippe Blache, and Alessandro Lenci. 2024. Neural generative models and the parallel architecture of language: A critical review and outlook. *Topics in cognitive science*.

Russell Richie and Sudeep Bhatia. 2021. Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive science*, 45(8):e13030.

W Joel Schneider and Kevin S McGrew. 2012. The cattell-horn-carroll model of intelligence.

Raj Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pre-trained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Raj Sanjay Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj, and Sashank Varma. 2023. Numeric magnitude comparison effects in large language models. *Preprint*, arXiv:2305.10782.

Mihir Sharma, Ryan Ding, Raj Sanjay Shah, and Sashank Varma. 2024. Monolingual and bilingual language acquisition in language models.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29.

Richard E Snow, Patrick C Kyllonen, Brachia Marshalek, et al. 1984. The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, 2(S 47):103.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Peter D. Turney. 2013. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics*, 1:353–366.

Sashank Varma. 2011. Criteria for the design and evaluation of cognitive architectures. *Cognitive science*, 35(7):1329–1351.

Siddhartha K Vemuri, Raj Sanjay Shah, and Sashank Varma. 2024. How well do deep learning models capture human concepts? the case of the typicality effect.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt and Samuel R. Bowman. 2024. What artificial neural networks can tell us about human language acquisition. *Preprint*, arXiv:2208.07998.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Pardis Sadat Zahraei and Ali Emami. 2024. Wsc+: Enhancing the winograd schema challenge using tree-of-experts. *arXiv preprint arXiv:2401.17703*.