The potential – and the pitfalls – of using PLMs as cognitive science theories.

Anonymous ACL submission

Abstract

Many studies show evidence for cognitive abilities in Pre-trained Language Models (PLMs). Researchers have evaluated the cognitive alignment of Pre-trained Language Models (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 the use of PLMs as cognitive science theories 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 treating 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 performance. We identify potential pitfalls of this approach to understanding human thinking, and we end by *enumerating criteria for using PLMs as credible accounts of cognition and cognitive development*.

1 Introduction

With the improving performance of pre-trained language models (Touvron et al., 2023; Gemini Team, 2023; OpenAI, 2023; Wei et al., 2022), researchers have increasingly advocated for their use as computational models of cognition (Piantadosi, 2023; Mahowald et al., 2024; Warstadt and Bowman, 2024; Coda-Forno et al., 2024). This includes many 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), spatial reasoning (Ramakrishnan et al., 2024) and analogical reasoning (Webb et al., 2023; Hu et al., 2023). For example, Shah et al. (2023) investigated the latent number representations of PLMs, finding that they showed the *distance*, *size*, and *ratio* effects observed in humans and understood as the behavioral signatures of a "mental number line" (Moyer and Landauer, 1967; Parkman, 1971; Halberda et al., 2008). To take another example, the Ravens test is the standard measure of *fluid reasoning*. Although Raven's problems are visual, Webb et al. (2023) translated them to equivalent "digit encodings" and showed that PLMs performed as well as humans on this test. More generally, many works are coalescing independent tests to form comprehensive psychometric benchmarks to enable scientific evaluation of the alignment of ML models to human cognition across multiple task domains (Chang et al., 2024; Zhuang et al., 2023; Shah et al., 2024; Coda-Forno et al., 2024; Wang et al., 2024; Tan et al., 2024).

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; Tan 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. Similarly, Wang et al. (2024) use Piaget's Theory of Cognitive Development to estimate that models like GPT-4o show cognitive abilities similar to 20-year-old humans. Instead of just looking at model end states, Shah et al. (2024) look at developmental parallels and also investigate the path dependence: whether the performance improvements of PLMs over training track the growth of cognitive abilities in children over development. Similarly, Tan et al. (2024) explore developmental parallels by comparing the learning trajectories of vision-language models to both child and adult behavioral data. To take a final example,

researchers propose studying bilingualism by varying the pattern of exposure to multiple languages to pre-training steps in PLMs to understand the rate of language development (Evanson et al., 2023; Marian, 2023; Sharma et al., 2024).

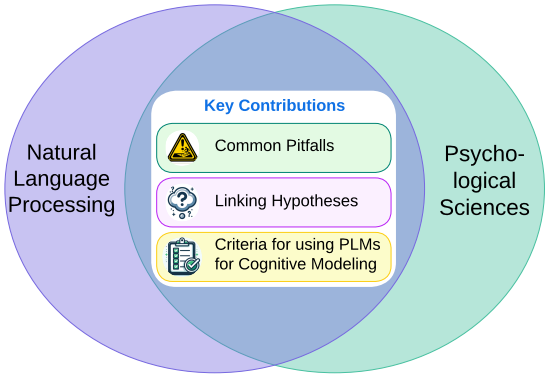

Figure 1: Overview of the lessons from using PLMs for modeling human cognition and its development.

In this paper, we describe the pitfalls and promise of PLMs and provide evidence for the sufficiency of their use as candidate theories in cognitive and developmental science. The contributions of this paper are threefold. (1) We first review the pitfalls of using PLMs in psychological science and caution researchers against over-interpreting PLM alignment to human cognition. (2) Next, we review the standard assumptions researchers use to map measures of PLM performance to human performance measures. (3) Finally, we build upon previous work (Ivanova, 2023; Mahowald et al., 2024; Cuskley et al., 2024; Birhane and McGann, 2024) in enumerating best practices for cognitive evaluations of PLMs.

2 PLMs as theories in cognitive and developmental science

PLMs make promising candidates as theories of cognitive science due to the nature of language model training (Hardy et al., 2023). PLMs acquire a wide range of capabilities in the pre-training phase, essentially obtaining human-like behaviors "for free" without the need for extensive task-specific tuning or adaptations (Demszky et al., 2023; Weng, 2024; Yang et al., 2024; Minaee et al., 2024; OpenAI, 2023). For example, the probabilistic nature of PLM text generation lures researchers to draw parallels with the human decision-making process, where humans anticipate future courses of action based on their past experiences and context. Specifically, they can provide insights into how humans may handle ambiguous situations where multiple interpretations are possible (McGrath et al., 2020; Gawlikowski et al., 2023; Dong et al., 2024; Belem et al., 2024). Moving from cognitive science to developmental science, PLMs have been shown to learn linguistic sensitivities to syntax, semantics, morphology, etc., from mere exposure to (i.e., masked word prediction of) large amounts of text data exposure. This raises the possibility that their learning trajectories mimic the development trajectories of children acquiring language (Aher et al., 2023; Duan et al., 2024; Shah et al., 2024).

This allure of PLMs as theories of cognitive science also extends to adjacent research areas. One example is the ability to simulate personas using PLMs, which enables the proxying of human subjects (Park et al., 2022; Samuel et al., 2024; Schuller et al., 2024; Tseng et al., 2024). This is especially important when real-world experiments with human participants are complex, resource intensive, and/or pose ethical challenges (Aher et al., 2023; Dillion et al., 2023; Hämäläinen et al., 2023). Another such research direction is neuroscience: Functional mapping in the neuroscience literature (i.e., mapping cognitive functions to their neural correlates) is similar to mechanistic interpretability in PLMs: both map functionalities to structure, provide a component-level analysis and explain network interactions. This raises the question of whether the functional units of a PLM can be mapped to functional units of the brain; this would enable their use in generating hypotheses about neural activity (Bzdok et al., 2024).

In its essence, we present an argument for the sufficiency of the usefulness of PLMs as cognitive and developmental science models: as long as *their performance profiles match those of humans*, then PLMs can be used to *estimate or predict* human behavior to specific stimuli. That is, we argue that matching functional forms is key, and that prediction is the more important criterion than explanation when applying ML models to cognitive phenomena (Breiman, 2003; Yarkoni and Westfall, 2017). Prior research in many domains has adopted this sufficiency argument (Niu et al., 2024; Ong, 2024).

In human studies, experiments are carefully designed to build and test hypotheses on the effect of cognitive phenomena. Independent of the phenomena under consideration, there are broadly three stages when mapping between the human data and

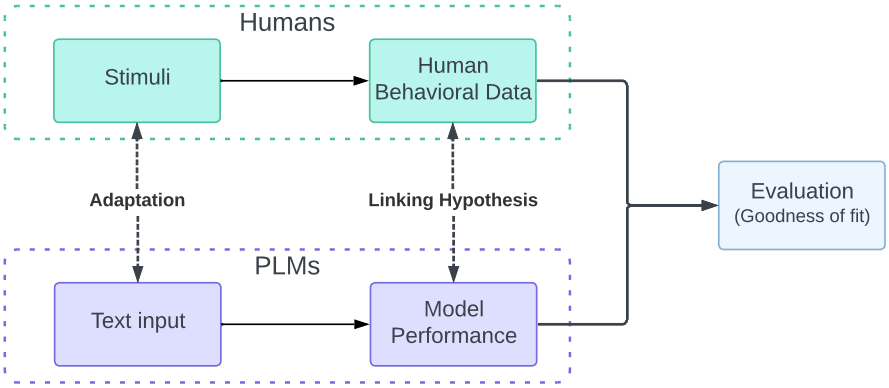

Figure 2: The three-stage mapping between human data and model performance for the sufficiency of the usefulness of PLMs as theories of cognitive science.

model performance:

1. Adaptation of the experimental stimuli and task: The stimuli presented to humans and the task they are asked to perform are modified to match the input modality of PLMs.

2. Linking Hypothesis: The model outputs are mapped to indices of human outputs.

3. Comparison to human performance: Model performance is compared against human performance to evaluate human alignment. This usually involves computing a measure of the *goodness of fit*.

The above three stages (refer to figure 2 provide us with the structure to evaluate the sufficiency or the usefulness of PLMs as cognitive or developmental science theories: *Under reasonable assumptions* about the adaptation of human stimuli to model inputs, linking hypothesis for the model outputs, and the comparison of model and human performance, *we are justified in using PLMs to estimate human behavior*.

3 Pitfalls of using PLMs as scientific theories

Along with the many promises of using PLMs for cognitive modeling, the assumptions in the three stage process can lead to potential pitfalls for using PLMs as scientific theories. Drawing upon current research, we enumerate the (potential) pitfalls of using PLMs as scientific theories. First, we consider their status as cognitive science theories, i.e., as models of adult thinking, which include abstract and conceptual reasoning, metacognition, attention and working memory, linguistic abilities like comprehension, tool of thought (inner speech), and a lot more.

- Researchers use *linking hypothesis* to map model performance characteristics to human performance characteristics (Hale, 2001; Levy, 2008). For example, when modeling incremental sentence processing, the log probability of the next word (given the words that come before) according to a PLM can be mapped to the time humans take to read that word (Li et al., 2024). These links are often quite distal, leaving it unclear whether PLMs are actually "explaining" cognitive science data. (See the next section for further discussion.)

- Human brains and PLMs are architecturally different. Recent research attempts to map different aspects of PLMs (layers, attention heads, etc.) to different of the brain, seeking a correspondence between model performance and functional neuroimaging measures. However, this work is in its infancy, and its viability (which requires a correspondence between NLP and brain architectures) remains an open question (Hosseini et al., 2022; Kauf et al., 2023).

- PLMs are opaque and have limited interpretability. When models and humans fail to align in their performance, the reason(s) why can be difficult to debug. This lack of interpretability is a barrier to treating these models as scientific theories (McGrath et al., 2023; Kar et al., 2022). Furthermore, commercial models often have many tuning interventions like supervised fine-tuning, instruction tuning, and RLHF/ RLAIF. While research shows instruction tuning aligns PLMs

to the human brain (Aw et al., 2023), lack of transparency in the tuning methods used makes it difficult to understand *why* they work when they do indeed work.

- 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 line with the experimental approach to human behavior – but at odds with the *differential* or *psychometric* approach that has also proven invaluable (Cronbach, 1957). While recent research developed larger benchmarks and expanded the evaluation of PLMs to a suite of tests (Coda-Forno et al., 2024; Chang et al., 2024), they have yet to establish a correlation on model performance among tasks. Contrastively, psychometric approaches to intelligence in humans 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 broader view is also in contrast to unified theories of cognition that attempt to model all cognitive abilities within a single computational framework (Anderson, 2009; Varma, 2011).

- PLMs are trained on aggregate human data (like Wikipedia), and thus, their behavior may not always be aligned with the range of different behaviors observed across individuals. Although some research has explored fingerprinting human personas with PLMs, there are conflicting views on whether models can simulate these personas (Milička et al., 2024) or not (Salewski et al., 2024).

Next, we consider the pitfalls that come with treating PLMs as developmental science theories, i.e., of the progressions in children's thinking and behavior.

- PLM checkpoints are snapshots or fingerprints of the data they are trained on. Most research only looks at final model checkpoints and evaluates cognitive alignment with adult thinking. Equally important is the question of whether, as language models observe more and more data, their performance is aligned to that of older and older children (Warstadt and Bowman, 2022; Frank, 2023; Shah et al., 2024). Developmental alignment is frequently overlooked because of the unavailability of intermediate training checkpoints or resource constraints. This limits our understanding of the developmental fidelity of model training. However, recent Open-source language modeling attempts provide promising opportunities to study developmental progression by making checkpoints more accessible (Biderman et al., 2023; Liu et al., 2023; Groeneveld et al., 2024).

- There are large differences in the nature of the data observed by PLMs versus humans. PLMs are trained on textual data that are 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 embodied in nature (Chemero, 2023; Cuskley et al., 2024; Birhane and McGann, 2024). Here, the emergence of vision-language models offers the potential to bridge the gap between the disembodied nature of PLMs and the multisensory learning of humans. However, this still faces certain limitations - for example, research suggests that while models can process simple visual features like color and size, models struggle with complex spatial and numerical reasoning tasks when applied to novel contexts or objects (Yiu et al., 2024).

- For the studies that have evaluated the developmental alignment, the developmental trajectories observed in models might be artifacts of the pre-training order (Shah et al., 2024). Research needs to perform training corpus analysis and account for the influences of training curriculum design, including the sequence and nature of data presented during pre-training, to distinguish between genuine developmental progressions and artifacts of training strategies.

4 Linking Hypotheses - Mapping Model Performance to Human Performance

The "outputs" of PLMs are often quite different from the behavioral measures that cognitive scientists collect in their experiments and evaluate their theories against. Researchers use various linking hypotheses to map PLM performance to human performance measures. These assumptions necessarily define – and potentially limit – the quality and the credibility of the alignment. It is important to critically evaluate these linking hypotheses because they structure how we interpret the model's capabilities to human cognition, language understanding, and task performance. Below, we highlight

different practices in mapping human judgments to indices of model behavior and review some popular linking hypotheses.

Similarity computations: Many cognitive tasks require people to judge the similarity of two items. Examples range from Roger Shepard's early psychophysical studies of people's similarity judgments of perceptual stimuli (e.g., circles of differing diameters) (Shepard and Podgorny, 1978) to Thomas Griffiths' current Bayesian and neural network models of how people judge the similarity of an exemplar to a category prototype (L Griffiths et al., 2008).

Human similarity judgments can be 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 (Turney and Littman, 2005). One example of similarity computations in cognitive tasks comes from the typicality effect, which is the finding people regard some members as "better" examples of a category than others (L Griffiths et al., 2008; Bhatia and Richie, 2022; Richie and Bhatia, 2021). The typicality of an exemplar is determined by the proportion of humans that produce it when asked to enumerate the exemplars of a category. In language models, the typicality of an exemplar for a category can be 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 (e.g., the embedding obtained by encoding the category label as a string and passing it through the model) is calculated, with a higher value indicating that the exemplar is more typical (Misra et al., 2021; Vemuri et al., 2024).

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. Coming back to the example of a latent mental number line, Shah et al. (2023) map the time it takes to compare which of two numbers is greater to the similarity of their PLM encodings. The linking hypothesis is that the greater the similarity of (i.e., the less discriminable) the two encodings, the longer the predicted time for humans to differentiate which one is greater. Under this mapping, PLMs show the "distance effect" observed in humans (Moyer and Landauer, 1967) and one of the behavioral signatures of the mental number line.

Recent research shows the neural plausibility of vector-based representations in capturing human conceptual properties, including compositionality, feature computation, relational reasoning, and symbolic processing, validating the hypothesis (Piantadosi et al., 2024).

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 different granular units (words or subwords) as compared to PLMs which use tokens. This introduces misalignment, making the mapping inconsistent as one unit of text for humans may be mapped to two units for PLMs (for example, Nine in humans is represented as *N, #ine* in some PLMs like Pythia (Biderman et al., 2023)).

Surprisal values: One way of quantifying the uncertainty of model generations is in terms of the summation of their surprisal, i.e., their negative 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 evidence for this correspondence. Similarly, Shain (2024) use PLMs to demonstrate strong predictability estimates of human reading times. Research also shows that surprisal values provide a better match to human plausibility judgments than prompts (Ivanova et al., 2024a). Finally, relative surprisal has been used to distinguish grammatical and ungrammatical sentences (Warstadt et al., 2020). This enables us to directly compare the right answer with all possible candidate answers in a deterministic manner because PLMs will generate sequential probabilities for all strings, whether they are grammatical or not.

A problem with surprisal-based approaches is that they show high context-sensitivity, i.e., the alignment to humans often depends on the framing and the structure of the original prompt. In case of many mathematical tasks, forcing a model to generate longer texts and reasoning chains leads to better human alignment (Jiang et al., 2024). On similar lines, it is difficult to operationalize variances in PLM outputs akin to variances in human responses to stimuli. In some cases, surprisal is a weak proxy for human behavior: Van Schijndel and

Linzen (2021) find that estimates of the time cost of word predictability derived using surprisal methods underestimate the magnitude of human garden path effects. Surprisal also fails to explain syntactic disambiguation difficulty (Huang et al., 2024). These findings highlight some failure modes of surprisal as a linking hypothesis and depict the need for establishing empirical plausibility – goodness of fit score – for each linking hypothesis on each task.

Prompting: PLM generation is probabilistic, and thus, the same model can give different results across inference runs. In prompting, PLM can be instructed to follow the same exercise as a human multiple times to generate a probability distribution over the output space. The probability of the correct output is then mapped to model confidence. Thus, prompting is very "lightweight": directly comparing the generated behavior of PLMs with that of humans reduces (or even eliminates) the need for indirect linking hypotheses (Patel and Pavlick, 2021; Webb et al., 2023; Zahraei and Emami, 2024), allowing for more *flexibility in the input choice* For example, to evaluate a model on the BLiMP benchmark (Warstadt et al., 2020), a PLM can be prompted with:

Prompt: Which statement is grammatically correct? Your response must be "1" or "2".

1. Noah likes to swim. 2. Noah likes to.

The PLM can directly generate "1" for the correct answer, and its probability of doing so across inference runs can be directly compared with the probability of humans making the correct choice. Contrastively, the typical method of comparison consists of relative surprisal of the two sentences.

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). This flexibility allows us 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) present a dichotomous verification sentence to the model after each word. The shifting probabilities of the sentences consistent with different semantic interpretations signal the model's online ambiguity handle. PLM prompting enables direct probing, which is often not possible in humans with behavioral measures (i.e., it would be disruptive to ask a human reader after each word to verify sentences) or even with neuroimaging measures to "look under the hood".

Another example is SAT analogical reasoning tasks (Turney, 2013). These are of the form A:B::? For example, consider the following SAT analogical reasoning problem:

Analogy: Runner : Marathon :: ?

Options:

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

Such multiple choice question-based choice tasks 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"}.*

There are several limitations to the prompting approach. First, similarly to the surprisal method, PLMs are *not robust to the prompt format.* Answering prompts requires PLMs to (1) understand the prompt and (2) know the answer. Often, small variations in the input prompts lead to large changes in human alignment (Guo et al.), 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 outputted answer to the grammaticality judgment prompt 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" – output of an existing PLM by Cai et al. (2024). No human would make this response. In the same vein of errors, when asked to follow complex human-like experimental procedures, prompting methods sometimes fail in maintaining output format consistency, making it hard to use this method. Third, research also points towards certain tasks where prompting lacks efficacy and argues against the exact metalinguistic prompting approach presented above (Hu et al., 2023; Hu and Frank, 2024).

Overall, we see that some linking hypotheses are better suited for certain tasks than others, and sometimes, certain linking hypotheses fail to show any human behavior profile.

5 Criteria for Evaluating and Developing PLMs as Scientific Models

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

In light of the discussion above, the review of some of the popular assumptions (linking hypotheses), and the common pitfalls outlined, we propose two sets of criteria for using PLMs for this purpose.

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

- **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 on one type of test. However, this alignment may just be an artifact, for example, of pre-training data contamination, i.e., the accidental inclusion of evaluation samples in the pre-training data. Triangulation – conducting more experiments evaluating the same cognitive/developmental phenomena – establishes stronger empirical plausibility.

- **Use multiple methods to interpret PLM successes and failures:** PLMs lack explainability and interpretability due to their large size (McGrath et al., 2023). 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. It is typically not clear in advance which evaluation metric will be most insightful for a given set of models and cognitive tasks, necessitating a multi-pronged approach.

- **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.

- **Control for tuning techniques:** PLMs are often tuned on specific data and in different ways, such as Instruction Tuning, Reinforcement Learning from Human Feedback, etc. The technique used is not incidental. Rather, it can influence model behavior, which is important if the model's output centers around tuning goals rather than developing a representation of world knowledge. Empirical evidence also suggests tuning methods to better align the models with human cognition (Aw et al., 2023).

- **Remember the linking hypothesis:** Adaptation of human experimental materials to textual counterparts that match the required modalities for PLMs requires making certain assumptions (refer to different operationalization in section 4). These assumptions are not neutral but part of why models may or may not align with human performance. For this reason, they need to be well-documented and explored in their own right. Additionally, transparency in the selection of such assumptions helps establish the credibility and replicability of findings.

- **Establish task correlations:** Inspired by Snow et al. (1984), who find correlations in human performance among different psychometric tests of intelligence, we propose a similar PLM evaluation paradigm. This paradigm will allow us to make estimates of model performance alignment with human behaviors on a set of novel tasks given PLM testing profiles on other tasks. This approach enables us to understand the broader capabilities and generalization of the model across different types of cognitive challenges.

- **Embodiment and Interactiveness:** Many researchers have advocated the addition of more modalities in the pre-training process, such as vision, touch, etc., to emulate the child learning process (Cuskley et al., 2024). While purely theoretical and speculative, addition of embodiment to PLMs and external interactions promises stronger human alignment.

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. This step is often overlooked in current studies of cognitive alignment. Evaluation should also focus on constructing developmental benchmarks,

such as the one presented by Yiu et al. (2024), where behavioral profiles of 3-year-old kids are captured and mapped to visual language model outputs.

- **PLMs can first be tuned on a small number of "core" cognitive tasks and then evaluated on a broader range of cognitive 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 might result in better cognitive alignment to a large number of tasks that all rely on human-like semantic representations (Vázquez Martínez, 2021; Vázquez Martínez et al., 2023).

- **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 the known developmental trajectories of knowledge and skill acquisition. For example, corpora could be ordered or sampled based on the age-of-acquisition of the words their text contains (Huebner et al., 2021; Portelance et al., 2023). Informed pre-training will allow us to better understand the developmental alignment of models.

6 Conclusion

This paper advocates for the use of Pre-trained Language Models (PLMs) as theoretical tools for investigating human cognition and its development through the three-stage sufficiency of usefulness. In this advocacy, we are not alone (Warstadt and Bowman, 2022; McGrath et al., 2023; Frank, 2023; Mahowald et al., 2024), however, at the same time, we caution researchers towards the informed use of PLMs when making theoretical claims in the psychological sciences. We have highlighted common pitfalls in this enterprise, reviewed the different assumptions – 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.

7 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 using PLMs for open cognitive and developmental science. There is no one-answer-fits-all approach for these tasks. NLP is a developing field, and we recommend articulating newer guidelines and practices as newer and ever-larger PLMs are trained and deployed. (3) Our work calls for 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 any experiments or offer any empirical evidence of performance comparisons or other quantitative measures. (4) In section 5, we suggest criteria by building upon and refining prior works based on empirical observations from the field. (5) PLMs, as described in this paper, are artificially designed systems that hope to reveal mechanisms of naturally evolving ones. Thus, researchers should keep the final goal in mind and not mistake a technological tool for an absolute source of insight. (6) The suggested criterion in this paper focuses on using PLMs for behaviorally benchmarking cognitive profiles of neurotypical and developmentally typical individuals. PLMs should also be used for simulating the behaviors of impaired individuals to better support them.

8 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 and, in the future, 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 be conducted with caution.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

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*.

John R Anderson. 2009. *How can the human mind occur in the physical universe?* Oxford University Press.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. In *First Conference on Language Modeling*.

Catarina G Belem, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. *arXiv preprint arXiv:2407.15814*.

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*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Abeba Birhane and Marek McGann. 2024. Large models of what? mistaking engineering achievements for human linguistic agency. *Language Sciences*, 106:101672.

Leo Breiman. 2003. Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1):81–82.

Danilo Bzdok, Andrew Thieme, Oleksiy Levkovskyy, Paul Wren, Thomas Ray, and Siva Reddy. 2024. Data science opportunities of large language models for neuroscience and biomedicine. *Neuron*, 112(5):698–717.

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

Chen-Chi Chang, Ching-Yuan Chen, Hung-Shin Lee, and Chih-Cheng Lee. 2024. Benchmarking cognitive domains for llms: Insights from taiwanese hakka culture. *arXiv preprint arXiv:2409.01556*.

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

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225*.

Lee Joseph Cronbach. 1957. The two disciplines of scientific psychology. *American Psychologist*, 12:671–684.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

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*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G Cai. 2024. Hlb: Benchmarking llms' humanlikeness in language use. *arXiv preprint arXiv:2409.15890*.

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*.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

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

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hanna Hajishirzi. 2024. Olmo: Accelerating the science of language models. *arXiv preprint*.

Grace Guo, Jenna Kang, Raj Sanjay Shah, Hanspeter Pfister, and Sashank Varma. Understanding graphical perception in data visualization through vision-language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Justin Halberda, Michèle MM Mazzocco, and Lisa Feigenson. 2008. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213):665–668.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Mathew Hardy, Ilia Sucholutsky, Bill Thompson, and Tom Griffiths. 2023. Large language models meet cognitive science: Llms as tools, models, and participants. In *Proceedings of the annual meeting of the cognitive science society*, volume 45.

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.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

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*.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

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*.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *arXiv preprint arXiv:2405.06705*.

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.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

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*.

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. 2023. How can deep neural networks inform theory in psychological science?

Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, and Himabindu Lakkaraju. 2020. When does uncertainty matter?: Understanding the impact of predictive uncertainty in ml assisted decision making. *arXiv preprint arXiv:2011.06167*.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *Plos one*, 19(3):e0298522.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

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

Robert S. Moyer and Thomas K. Landauer. 1967. Time required for judgements of numerical inequality. *Nature*, 215(5109):1519–1520.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, and Keyu Chen. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

Desmond C Ong. 2024. Gpt-ology, computational models, silicon sampling: How should we think about llms in cognitive science? *arXiv preprint arXiv:2406.09464*.

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

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

John M Parkman. 1971. Temporal aspects of digit and letter inequality judgments. *Journal of experimental psychology*, 91(2):191.

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.

Steven T Piantadosi, Dyana CY Muller, Joshua S Rule, Karthikeya Kaushik, Mark Gorenstein, Elena R Leib, and Emily Sanford. 2024. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*.

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.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. 2024. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*.

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.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

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

Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pre-trained language models. *arXiv preprint arXiv:2407.01047*.

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.

Cory Shain. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

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

Roger N Shepard and Peter Podgorny. 1978. Cognitive processes that resemble perceptual processes. In *Handbook of learning and cognitive processes: Vol. 5. Human information processing*, pages 189–237. Erlbaum Hillsdale, NJ.

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.

Alvin Wei Ming Tan, Sunny Yu, Bria Long, Wanjing Anya Ma, Tonya Murray, Rebecca D Silverman, Jason D Yeatman, and Michael C Frank. 2024. Devbench: A multimodal developmental benchmark for language learning. *arXiv preprint arXiv:2406.10215*.

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.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

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

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

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

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

Héctor Javier Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 479–495, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

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.

Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024. Coglm: Tracking cognitive development of large language models. *arXiv preprint arXiv:2408.09150*.

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*.

Benjue Weng. 2024. Navigating the landscape of large language models: A comprehensive review and analysis of paradigms and fine-tuning strategies. *arXiv preprint arXiv:2404.09022*.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. *arXiv preprint arXiv:2403.09162*.

Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12:1100 – 1122.

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*.

Eunice Yiu, Maan Qraitem, Charlie Wong, Anisa Noor Majhi, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. 2024. Kiva: Kid-inspired visual analogies for testing large multimodal models. *arXiv preprint arXiv:2407.17773*.

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

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. 2023. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. *arXiv preprint arXiv:2306.10512*.