Development of Cognitive Intelligence in Pre-trained Language Models

Anonymous ACL submission

Abstract

Recent studies show evidence for *emergent cognitive abilities* in Large Pre-trained Language Models (PLMs). The increasing *cognitive alignment* of these models has made them candidates for cognitive science theories. Prior research into the emergent cognitive abilities of PLMs has been *path independent* to model training, i.e. has only looked at the final model weights and not the intermediate steps. However, building plausible models of human cognition using PLMs also requires aligning their performance during training to the developmental trajectories of children's thinking. Guided by psychometric tests of human intelligence, we choose four task categories to investigate the alignment of ten popular families of PLMs and evaluate each of their available *intermediate and final training steps*: Numerical ability, Linguistic abilities, Conceptual understanding, and Fluid reasoning. We find a striking regularity: regardless of model size, the developmental trajectories of PLMs consistently exhibit a window of maximal alignment to human cognitive development. Before that window, training appears to endow "blank slate" models with the requisite structure to be poised to rapidly learn from experience. After that window, training appears to serve the engineering goal of reducing loss but not the scientific goal of increasing alignment with human cognition.

1 Introduction

Large Pre-trained Language Models (PLMs) like Google's Gemini (Team et al., 2023), Meta's LLaMA 2 (Touvron et al., 2023), and OpenAI's GPT 4 (OpenAI, 2023a) show human-level or even super-human performance on many cognitive performance tasks. This is true in 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), contributing to the hype of claims of reaching Artificial General Intelligence (AGI).

Such claims deserve to be scrutinized. Human intelligence is multi-faceted. Furthermore, there is a massive disparity between the training data scale of PLMs and humans. PLMs unintentionally acquire human performance characteristics from the corpora they are trained on, through residues of the values, beliefs, and biases of the authors of the texts (Pellert et al., 2024). We approach the human alignment of PLMs by grounding evaluation in frameworks for *psychometric intelligence*. Psychometric measures of intelligence include multiple subtests spanning a range of abilities, including mathematical thinking, language comprehension, spatial thinking, fluid reasoning, and conceptual understanding (Snow et al., 1984; Carroll, 1993; Sternberg, 2000; McGrew, 2009; Haier, 2023). In this work, we choose representative assessments of different facets of human intelligence, modified for the required textual modality, to evaluate the *cognitive alignment* of PLMs.

A second goal of our work is to move beyond cognitive alignment to also evaluate the *developmental alignment* of PLMs. The claim that the final model state of a PLM approximates adult performance leaves open the question of the path by which it arrived there. Ideally, the model's performance improvements over training also track the progression of cognitive abilities over development (Elman, 1996; Bengio et al., 2009). This potential parallelism would be stronger evidence for PLMs as cognitive science models. Researchers are increasingly addressing this question by building PLMs trained on a developmentally plausible corpus of child-directed speech, transcribed dialogue, and children's literature (Huebner et al., 2021; Warstadt et al., 2023; Bhardwaj et al., 2024).

We ask the question of developmental alignment in a theoretically important way: Is the cognitive alignment of PLMs achieved in a *path-independent*

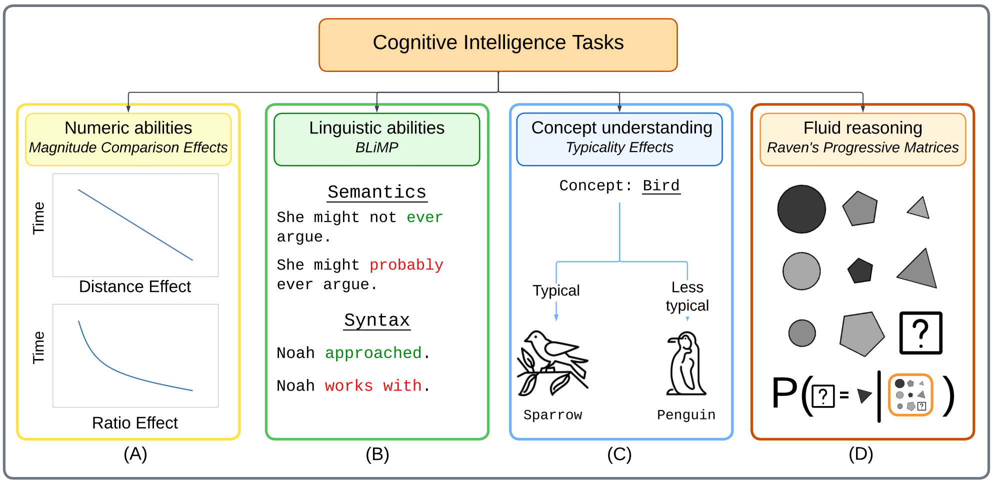

Figure 1: A list of cognitive intelligence tasks under consideration.

or *path-dependent* manner? Prior studies focusing on the cognitive alignment of PLMs have only established path independence: that models at the end of training approximate adult performance across a range of domains. Here, we also evaluate path dependence: Do the performance improvements of PLMs over training track the growth of these abilities in children over development (Holyoak et al., 1984)? We ask this question for models of different sizes and track their developmental alignment over millions and billions of training tokens. If path independence holds, this opens up applications that we detail in the conclusion.

To summarize, our key contributions are as follows:

- **Cognitive Modelling using AI:** We test the appropriateness of PLMs for cognitive modeling by evaluating whether their performance profiles match those of humans.
- **Developmental trajectories in LLM pre-training and scaling:** Previous studies have only evaluated the final training checkpoints of PLMs for their cognitive plausibility, and have neglected the question of developmental trajectories. Here, we also ask: Can PLMs be used to model human developmental trajectories despite the training data scale mismatch between PLMs and humans?
- **Representative tasks:** We choose representative tasks of human-like psychometric intelligence tests in PLMs. These tasks evaluate numeric, linguistic, conceptual, and fluid intelligence. We propose these to be a *prerequisite* to using PLMs for cognitive modeling.

2 Related work

2.1 Psychometric theories of intelligence

Previous intelligence assessments in AI have looked at singular dimensions, such as numeric abilities (Zhuang et al., 2023; Fang et al., 2024). Rather than choose cognitive abilities in a piece-meal fashion, we look to psychometric theories of intelligence for guidance (Sternberg, 2000). These theories distil performance on a large number of subtests into a small number of latent factors. Despite popular attention to "general intelligence" and the latent factor g, there is a long history of theories positing that intelligence is composed of multiple domain-specific abilities. An important, early domain-specific theory of intelligence, (Thurstone, 1938), included seven "primary abilities". The most widespread psychometric theory of intelligence today, the Cattell-Horn-Carrol (CHC) theory (Carroll, 1993; McGrew, 2009), includes among its "broad" abilities quantitative knowledge, reading and writing ability, fluid reasoning, and "comprehension" knowledge (a subcomponent of which is conceptual understanding). We evaluate the cognitive and developmental alignment of PLMs along these four abilities.

2.2 Emergent cognitive abilities in Language Models

Recently, the performance of language models has improved as they have increased in size from millions to billions of parameters, trained on larger corpora, and further tuned in novel ways (instruction tuned, RLHF). This has led to researchers increasingly advocating for the use of PLMs as cognitive models (Piantadosi, 2023; Warstadt and Bowman, 2024). Increasing the number of parameters of the

Table 1: Summary of assessments.

Cognitive Domain	Task	Source	License
Numeric Abilities	Magnitude Comparison Effects	(Shah et al., 2023)	(cc by 4.0)
Linguistic Abilities	BLiMP	(Warstadt et al., 2020)	(cc by 4.0)
Concept Understanding	Typicality Effects	(Vemuri et al., 2024; Castro et al., 2021)	(cc by 4.0)
Fluid reasoning	Raven's Progressive Matrices	(Hu et al., 2023)	(cc by 4.0)

models has given rise to *Emergent Abilities* that cannot be predicted by extrapolating from the performance of smaller models (Wei et al., 2022a). Emergent abilities have been observed in a variety of task types such as multi-task language understanding (Hendrycks et al., 2021), grounded conceptual mapping (Patel and Pavlick, 2022), and truthfulness (Lin et al., 2021). In recent works, Hoffmann et al. (2022); Biderman et al. (2023) show the benefits of training a model for more tokens on problem-solving (Wei et al., 2022b), common-sense reasoning (Sakaguchi et al., 2021), arithmetic abilities (Biderman et al., 2023), and linguistic performance (Paperno et al., 2016). Although the presence of emergent abilities extends to cognitive science domains (Wei et al., 2022b; Goertzel, 2023; Hagendorff, 2023), prior studies have been piecemeal in their approach and have failed to (1) consider multiple cognitive abilities as specified by theories of psychometric intelligence and (2) move beyond cognitive alignment to also evaluate the developmental alignment of PLMs over training.

2.3 Pre-trained language model use in developmental modeling

Recently, researchers have begun advocating for the use of PLMs for modeling cognitive development in children (Kosoy et al., 2023; Salewski et al., 2024). For example, Portelance et al. (2023) and Bhardwaj et al. (2024) suggest the use of language models to predict the age of acquisition of words in children. Researchers have also proposed studying second language acquisition and 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). We investigate the assumption that the performance of an intermediate training checkpoint of PLMs maps to the age of child development by looking at the acquisition of human-like psychometric intelligence.

3 A suite of psychometric intelligence tasks

We assemble a suite of cognitively plausible assessments that benchmark PLMs across four abilities of psychometric intelligence. Table 1 summarizes the tasks along with the licensing details for public use. The details of each assessment and their respective operationalization are given below. [1]

3.1 Numeric abilities

Figure 2: Mental Number Line: Organization of magnitude representations in a logarithmically scaled manner.

The question of how humans understand symbolic numbers has been investigated by cognitive scientists for more than half a century. These studies show that people map number symbols to a *mental number line* (MNL, Figure 2) with a log-compressed psychophysical scale (Moyer and Landauer, 1967a).

Prior research on the numerical abilities of PLMs has focused on improving performance on application-driven tasks that require numerical skills in the context of arithmetic equations and word problems (Burns et al., 2021; Amini et al., 2019; Yuan et al., 2023), exact facts (Lin et al., 2020), and measurement estimation (Zhang et al., 2020). *However, these tasks fail to directly implicate the key cognitive construct underlying human numerical understanding, the recruitment of a compressed MNL.*

In a recent study, Shah et al. (2023) found evidence for a human-like MNL in various PLMs. They show that despite lacking explicit neural circuitry to represent numbers, through experience (i.e., vast amounts of training data), PLMs show human-like performance profiles and learn human-like representations for numerical concepts.

We follow Shah et al. (2023) and look for the two behavioral signatures of a compressed number line representation, the distance effect and the ratio

[1] We will add all tasks to a publically available unified language model testing framework, titled *lm-evaluation-harness* (Gao et al., 2023), to support the evaluation of future models on psychometric intelligence assessments.

effect. In humans, these are defined as:

- **Distance effect** (refer to Figure 1A - top): The greater the distance $|x - y|$ between two numbers x and y, the faster they are compared, i.e., the greater (or lesser) number is identified (Moyer and Landauer, 1967b).
- **Ratio effect** (refer to Figure 1A - bottom): The time to compare two numbers x and y is a decreasing function of the ratio of the larger number over the smaller number $\frac{max(x,y)}{min(x,y)}$ (Halberda et al., 2008).

These effects can be mapped to language models by adopting the following linking hypothesis: *the greater the cosine similarity of two number representations in a PLM, the longer it takes to discriminate them, i.e., to judge which one is greater (or lesser).* While we focus on the Distance and Ratio effect, the results for all the effects investigated by Shah et al. (2023) are in Appendix B.1.

Operationalization: We used the same protocol as Shah et al. (2023). For each effect, we test the three formats of number representations of PLMs (mixed-case number words, lower-case number words, and digits). We present the R^2 values for the Distance and Ratio effects, which are averaged across each input representation. The R^2 values for the distance effect in PLMs are obtained by fitting a linear function predicting the cosine similarity of x and y from their distance $|x - y|$. R^2 values for the ratio effect in PLMs are obtained by fitting a negative exponential function predicting the normalized cosine similarity of x and y from their ratio $\frac{\max(x,y)}{\min(x,y)}$. Note: This task requires access to the latent representations of models.

3.2 Linguistic abilities

Language (or verbal) ability is a central component of human cognition and cognitive neuroscience (Hagoort, 2019). At the dawn of the cognitive revolution, it was conceptualized as a largely innate ability, and language acquisition was understood as requiring relatively little learning from experience (Fodor, 1985; Chomsky, 2014). More recently, cognitive developmentalists have shown that infants can learn language through exposure to the statistical regularities of the linguistic environment (Saffran et al., 1996; Siegelman, 2020). These findings have been modeled using multi-layer perceptrons (Elman, 1996) and, more recently, PLMs (Lake and Murphy, 2023).

We use BLiMP: The Benchmark of Linguistic Minimal Pairs for English (Warstadt et al., 2020) to evaluate the linguistic abilities of each PLM under consideration. BLiMP consists of 67 datasets of 1000 pairs of minimally different sentences which vary in acceptability and span 12 phenomena at three levels of language: *morphology*, *syntax*, and *semantics*. The 12 phenomena are described in Appendix B.2. Each pair consists of one acceptable sentence and one unacceptable sentence which otherwise differ minimally. BLiMP evaluates the models by measuring if they assign a higher probability to the acceptable vs. unacceptable sentence of each pair. Figure 1B shows two examples of minimal pairs.

Operationalization: We use the LM-eval-harness (Gao et al., 2023) benchmarking suite to test our models on the BLiMP tasks. We evaluate if a model assigns a higher sequential probability to the acceptable sentence. Note: This requires models that can generate probabilities of tokens.

3.3 Concept understanding

On encountering a new stimulus, humans categorize it – assign it to a known concept – in order to make inferences about its unobservable properties (Murphy, 2002). A striking finding is that not all members of a category are equal (Rosch, 1975). Rather, some members (e.g., pigeon) are more typical of a category (e.g., Bird) than other members (e.g., ostrich). This phenomenon, known as the *Typicality Effect*, is a central feature of human categorization (Lakoff, 2008).

Typicality gradients in humans can be measured using the production task, where participants are given a category label (e.g., Bird) and asked to list as many members of the category as they can in a limited time (Battig and Montague, 1969; Van Overschelde et al., 2004; Castro et al., 2021). The typicality of an item is defined as the proportion of participants who produce it.

Language models have shown some evidence of human-like typicality gradients. Heyman and Heyman (2019) used word2vec embeddings to predict the category typicality norms released by De Deyne et al. (2008). More recent work by Misra et al. (2021) and Bhatia and Richie (2022) has looked at correlations of PLMs like BERT, RoBERTa, and GPT-2 to the Rosch (1975) typicality norms for ten categories. Vemuri et al. (2024) performed the most comprehensive study of the alignment of concept understanding in the latent representations of PLMs. We expand upon their task setup to evaluate human-like concept understanding in the PLMs that are the focus here.

Operationalization: For each model, we calculate the representativeness of a member to its category in three possible ways:

- Closeness judgment problem: Calculate the cosine similarity between the obtained latent representations for the member and the category. This requires models where the latent representations are readily available.
- Surprisal values: For each member in a category, the probability of the sequence *a "member" (eg. pigeon) is a "category" (eg. bird)*. This method requires access to the probability of each token in a sequence.
- Prompting: Prompt the models with the following design: Guidelines, Query, and Options. The Guideline highlights the task of re-ranking the members given in the Options based on appropriateness with the Query. The Query consists of the in-filling task: *A ___ is a [category name]*. The Options are each of the possible members of the category. Given the complexity of the prompting, usable outputs are only obtained from models that are larger than 30 billion parameters.

For the two in-filling problems (i.e., based on surprisal values and prompting), we also evaluate models on zero to three exemplars as context. The details of the experiments on these different exemplar contexts are given in Appendix B.3.

3.4 Fluid reasoning

Humans can logically parse information and detect patterns in novel stimuli without having to rely on prior experiences or learned information. This ability is called Fluid Reasoning (Cattell, 1963).

We focus on the dominant measure of fluid reasoning, the Ravens Progressive Matrices (RPM) test (Raven, 2003). An example Ravens-like problem is given in Figures 1D and 3. An RPM item consists of a 3x3 matrix of cells with one empty cell. Participants must induce the underlying, abstract patterns that hold across the rows and columns of the matrix, and apply these to infer the image in the empty cell from a given set of options. These images vary in visual attributes like shape and color, along with more abstract qualities. The RPM is the standard measure of fluid reasoning (Snow et al., 1984) and is highly correlated with analogical reason (Goswami, 1986; Webb et al., 2023).

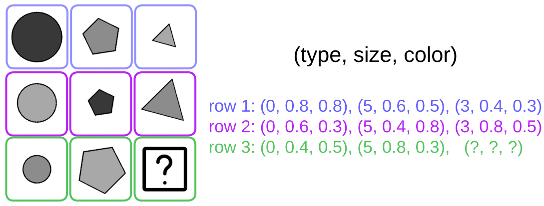

Figure 3: Example adaptation of visual RPM problems to the textual format. Each image is decomposed into tuples of (type, size, color). Type indicates the shape of the image.

Given the visual nature of the RPM, previous work by Hu et al. (2021, 2023) and Webb et al. (2023) mapped the Raven-10000 dataset to a textual format to facilitate the testing of PLMs. The mapping involves reformulating visual elements into text-based numerical tuples to form the I-Raven dataset, representing attributes like shape, size, and color textually, as illustrated in Figure 3. We use their approach with a focus on the "Center Single Alignment" sub-task, which features a single shape per matrix cell. We differ from their work by evaluating a broader set of models.

Operationalization: We determine the model's preferred answer for a problem by comparing the surprisal values of the whole sequence (instruction, question, candidate tuple) for each of the candidate options, i.e. the probability of each completed digit representation of a matrix. For the example given in Figure 3, this would be checking the probability of this sequence (summation of token probabilities) with the correct answer **(3, 0.6, 0.8)** to the other candidates. A comprehensive list of the prompts used in this paper is given in Appendix B.4.

4 Models under consideration

We evaluate a wide range of language model families, shown in Table 2. These models are selected based on the following criteria:

Public availability: Open-source models allow us to perform a thorough analysis by accessing the latent representation and the token probability during generation. We follow Holt et al. (2024) while choosing PLMs. Although most models in this study are publicly available and open-source, we use three state-of-art commercial PLMs that are

Table 2: List of language model families under consideration with their statistics.

Models	Source	Latent rep.	Token prob.	Multiple sizes	Intermediate checkpoints	Known training order
Amber	(Liu et al., 2023)	✓	✓	✗	✓	✓
Falcon	(Almazrouei et al., 2023)	✓	✓	✗	✗	✗
Starling	(Zhu et al., 2023)	✓	✓	✗	✗	✗
Llama-2	(Touvron et al., 2023)	✓	✓	✓	✗	✗
Mistral	(Jiang et al., 2023)	✓	✓	✗	✗	✗
Qwen	(Bai et al., 2023)	✓	✓	✓	✗	✗
Pythia	(Biderman et al., 2023)	✓	✓	✓	✓	✓
Gemini	(Team et al., 2023)	✗	✗	✗	✗	✗
GPT-3.5-Turbo	(OpenAI, 2023b)	✗	✓	✗	✗	✗
GPT 4	(OpenAI, 2023a)	✗	✓	✗	✗	✗

Table 3: Performance of Pre-trained Language Models on the tasks. Distance Effect: Averaged R^2 values of different LLMs when fitting a linear function on the cosine-similarity vs. distance plot. Ratio Effect: Averaged R^2 values of different LLMs when fitting a negative exponential function on the cosine-similarity vs. ratio plot. Note: Each value is averaged across all three input types and all model layers to produce one generalizable score. Latent Rep: Average Spearman's Correlation when using the cosine similarity and latent representation-based approach (Note: * refers to the prompting approaches for select models which are gated by APIs and not the latent representation-based approach), Zero-Shot: Average Spearman's Correlation when using the zero-shot surprisal values, BLiMP: The Benchmark of Linguistic Minimal Pairs for English, RPM: Raven's Progressive Matrices

Model	Numeric Abilities		Linguistic Abilities	Conceptual Understanding		Fluid reasoning
	Distance Effect (R^2)	Ratio Effect (R^2)	BLiMP (Acc.)	Latent Rep. (Average Spearman's Correlation)	Zero Shot	RPM (Acc.)
Amber-7B	0.913	0.591	0.794	0.083	0.250	0.654
Falcon-7B	0.928	0.838	0.817	-0.116	0.180	0.730
Starling-LM-7B-alpha	0.522	0.187	0.827	-0.003	0.258	0.730
Llama-2-7B	0.670	0.614	0.818	-0.065	0.238	0.752
Llama-2-13B	0.672	0.263	0.793	0.076	0.247	0.756
Mistral-7B	0.641	0.233	0.829	-0.025	0.245	0.756
Mistral-7B-Instruct	0.637	0.543	0.834	0.033	0.255	0.674
Qwen-0.5B	0.833	0.553	0.785	0.072	0.282	0.684
Qwen-1.8B	0.878	0.301	0.792	0.114	0.235	0.746
Qwen-4B	0.881	0.264	0.730	0.001	0.246	0.770
Qwen-7B	0.858	0.616	0.789	0.006	0.229	0.766
Qwen-14B	0.783	0.507	0.792	-0.140	0.249	0.776
Pythia-70M	0.829	0.429	0.723	0.005	0.211	0.194
Pythia-160M	0.947	0.665	0.749	0.067	0.260	0.448
Pythia-410M	0.926	0.679	0.815	0.126	0.284	0.608
Pythia-1B	0.944	0.702	0.806	0.090	0.280	0.674
Pythia-1.4B	0.933	0.764	0.819	0.074	0.283	0.730
Pythia-2.8B	0.961	0.723	0.827	0.221	0.273	0.760
Pythia-6.9B	0.909	0.713	0.809	0.105	0.280	0.716
Pythia-12B	0.846	0.595	0.829	0.184	0.291	0.756
Gemini	NA	NA	NA	0.311 *	NA	NA
GPT-3.5-Turbo	NA	NA	0.825	0.242 *	0.231	0.792
GPT-4	NA	NA	0.849	0.559 *	0.428	0.822

gated behind API calls; GPT-3.5-Turbo (pointing to gpt-3.5-turbo-0613 on the OpenAI platform), GPT-4 (pointing to gpt-4-1106 on the OpenAI platform), and Gemini (also referred to as Gemini-1-Pro at the time of writing). The GPT-x model APIs provide token probabilities of the response, allowing us to calculate surprisal, while Gemini does not.

Availability of multiple sizes: The availability of model sizes for the same architecture and training paradigms allows us to evaluate the emergent cognitive abilities of the models. We have multiple sizes available for the LLama-2, Qwen, and the Pythia family of models.

Availability of intermediate training checkpoints: This allows us to evaluate the effects of pre-training on the model outputs. Together, the availability of multiple model sizes and intermediate training checkpoints allow us to best evaluate the developmental alignment of PLMs. Amber and Pythia's family of models have available intermediate training checkpoints. While Amber has 360 intermediate checkpoints, the checkpoints are at 4 Billion tokens each and are not at the required granularity.

Pythia Family of models: Pythia (Biderman et al., 2023) is one of the first open-source projects with the goal of scientific and transparent model development. It has 8 model sizes ranging from 70 Million to 12 Billion parameters, with each model trained on 286 Billion tokens. The models in the suite are equivalent (in size) to popular decoder architectures like GPT-Neo-(125M, 1.3B, 2.7B) and OPT-(125M, 350M, 1.3B, 2.7B, 6.7B), but with the added benefits of training on a known de-duplicated corpus (Gao et al., 2020), using the same training order for each model size, and having 154 intermediate checkpoints to study the learning trajectories of PLMs. Thus, the Pythia suite of models is ideal for studying the psychometric and developmental alignment of PLMs to humans.

All open-source models are obtained from Huggingface (Wolf et al., 2020), while the gated models are obtained from their respective platforms through API calls. For each model in the Pythia suite, the following intermediate checkpoints are available: [1, 2, 4, 8, ... 512; 1000, 2000, 3000 ... 143000 (exponential increase in checkpoint number until the 512th checkpoint and subsequent progression of 1000 steps until the last checkpoint)], with each checkpoint representing 2 Million tokens seen. *Overall, we test 1232 intermediate checkpoints of the Pythia suite of models across all the tasks.*

5 Cognitive and developmental alignment of PLMs

The suite of tasks enables comprehensive evaluation of a variety of PLMs on their cognitive alignment to humans across four domains of psychometric intelligence: numeric abilities, linguistic abilities, concept understanding, and fluid reasoning. Table 3 highlights the key results of this evaluation. For the evaluation of conceptual understanding in PLMs, we only report the results for the zero-shot surprisal values and latent representations. This is because we see similar results for zero-shot and few-shot surprisal value-based methods (see comprehensive results in Appendix B.3).

The *cognitive alignment* of PLMs on psychometrics assessments is summarized below:

- *Numeric abilities*: All PLMs show a human-like distance effect but weakly show a human-like ratio effect. We do not observe any notable changes in alignment with model scaling, indicating the need for the evaluation of future models on this task.
- *Linguistic abilities*: The accuracy of the PLMs on the BLiMP linguistic acceptability tasks improves upon increasing the number of parameters. Furthermore, we find that all PLMs are substantially more accurate on morphological tasks over syntactic and semantic tasks (*Accuracy: semantic* $<$ *syntax* $<<$ *morphology*; see Appendix Table 5, Figure 7). Morphological performance develops first followed by syntax and then semantics.
- *Concept understanding*: Prompting methods in commercial models perform substantially better than other methods – closeness judgment and surprisal values – on all open-source models. In the Pythia suite, we observe that larger models outperform smaller counterparts on the same training data.
- *Fluid reasoning*: For all PLM architecture types, larger models outperform their smaller equivalent models.
- Despite differences in PLM architecture type, all models of an approximate size of 7 Billion parameters perform comparably.

The *developmental alignment* of the PLMs on the tasks is shown in Figure 4. We make the following key observations:

- *Training endows the "blank slate" with requisite structure*: In each assessment, the model "warm-ups" in training on a few million/ billion tokens, moving from a "blank slate" to possessing the requisite structure. This structure can be thought of as the child's endowment at birth. Development of the four abilities begins only after reaching this state.
- *Training shows a region of development*: For all four tasks, we see a window of monotonic development, in which all models gain the respective cognitive abilities.
- *After development, training appears to serve an engineering goal*: After the window of development, training appears to only serve the engineering goal of loss reduction. This observation is especially pronounced for numeric abilities and conceptual understanding.
- *Assessments for Fluid Reasoning and Linguistic Abilities show significant gains in scaling and greater pre-training*: For the Fluid Reasoning and Linguistic Abilities assessments, we see that the alignment score continues to increase as the PLMs are trained on a greater number of tokens. Furthermore, for these abilities, models also show scaling effects, with larger models outperforming smaller ones.

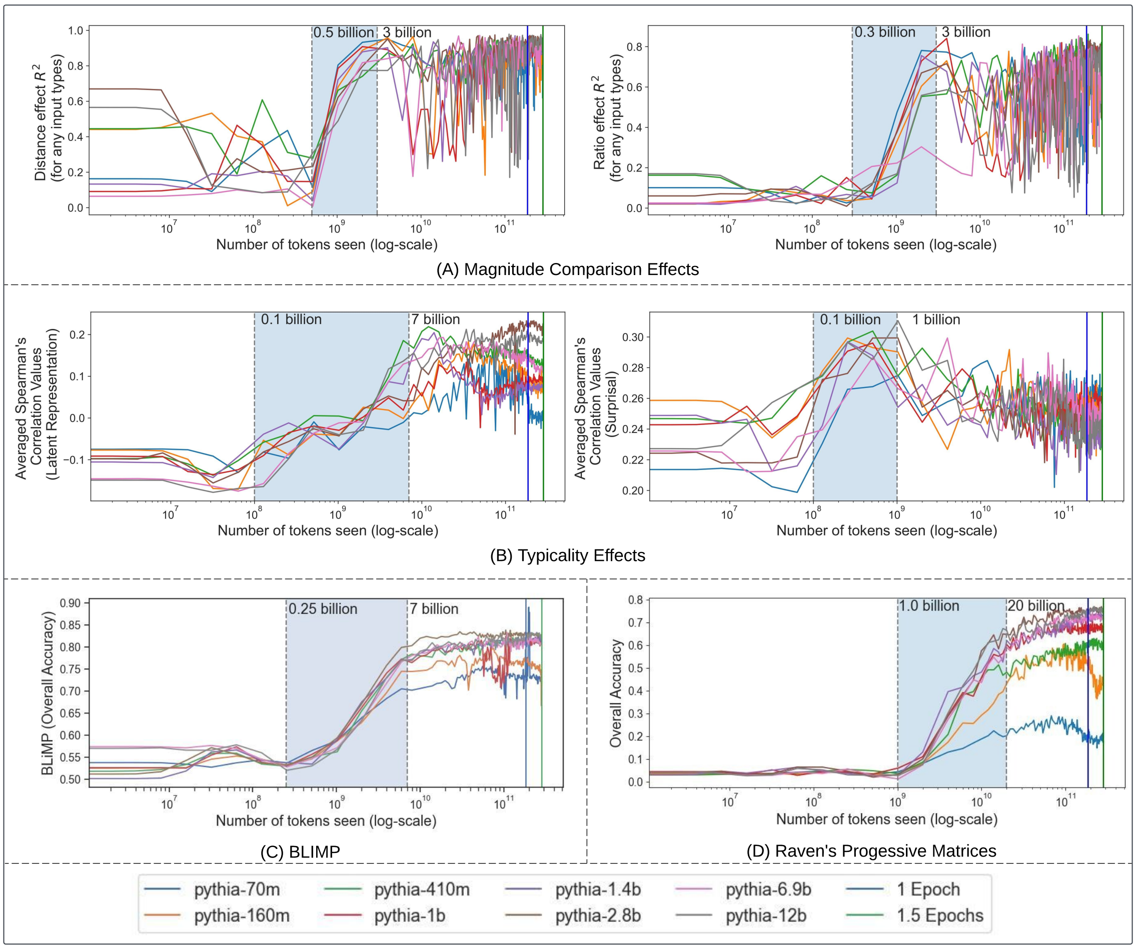

Figure 4: Developmental trajectory of the Pythia suite of models on the psychometric intelligence tasks as a function number of tokens seen. We display the x-axis in a log-scaled manner as maximal development occurs in the range of 100 Million to 20 Billion tokens seen for all tasks. The windows of maximal development are illustrated by the blue shading.

- *The windows weakly align with human ages of development*: Variation in the onsets of windows replicates what is known of cognitive development. For example, children acquire language early (i.e., during the preschool years), whereas the onset of improving fluid reasoning is later, when children enter elementary school, and continues for longer, throughout adolescence. Correspondingly, the models significantly develop linguistic abilities while training on 250 Million to 7 Billion tokens, whereas they acquire fluid reasoning abilities later, while training on 1 to 20 Billion tokens.

6 Conclusions

This paper investigates the evidence appropriateness of using PLMs for human cognitive and developmental modeling with the help of adapted psychometric intelligence assessments. It uses representative assessments of four facets of human intelligence: numeric abilities, linguistic abilities, conceptual understanding, and fluid reasoning. Our experiments show that PLMs develop cognitive abilities purely through their experience in the world, indicating that cognitive abilities in humans may not be innate, but rather learned similarly through the world. Most significantly, we find a window of monotonic development in which all models improve approximately linearly on the four cognitive abilities. Before that window, we interpret training as endowing "blank slate" models with the requisite structure for rapid learning. Also notable is the finding of PLM scaling effects for the assessments of linguistic abilities and fluid reasoning. We propose evaluation against these tasks as a *prerequisite* before treating PLMs as models of human cognition and its development.

7 Limitations

Some limitations of the work are as follows: (1) We use an aggregation of psychometric tests for PLMs. The limitations of each test are inherited in the suite of tasks. (2) The alignment scores may be wrongly interpreted when evaluating PLMs with these tasks. Alignment scores show the similarity of PLM outputs to human outputs on psychometric tests and indicate that PLMs do not need explicit neural circuitry for these intelligence tests. We do not suggest these models as proxies for humans in any manner and recommend further testing before use. (3) The developmental alignment of the models points towards the acquisition of human-like performance on the four psychometric assessments in the range of 100 Million to 20 Billion training tokens. This conclusion has two limitations: Pythia is the only suite of models with available intermediate checkpoints and, while unlikely, the observed developmental trajectories might be artifacts of the pre-training order. (4) The psychometric assessments for PLMs are adapted from similar human psychometric tests. Different ways of adaptation may lead to different results. Furthermore, while representative, these assessments are not exhaustive tests of human intelligence. Future work can expand to other tests like spatial and commonsense reasoning. (5) Some open source models like Llama-2 have larger 70 Billion parameter variants but we lack the compute resources to evaluate them. Large open-source models would lead to appropriate comparisons of performance with commercial models like GPT-4. (6) While our work evaluates changes in cognitive alignment with an increase in model size and the number of pre-training tokens, we do not control for different tuning methodologies like instruction tuning and reinforcement learning with human or artificial intelligence feedback. Accounting for different tuning methods is computationally intensive for the 1200+ model checkpoints across 10 architectures.

8 Ethical Considerations

All tasks and corresponding datasets have low ethical risks and none expose sensitive information. Additionally, we obtain approval from the authors of each dataset for their use and release. There are no major risks associated with conducting this research beyond those associated with working with PLMs. There may be risks in misinterpreting the alignment scores when evaluating with the tests. The psychometric analysis of this study is one-way: we look for human performance characteristics and behaviors in PLMs. PLMs are experimental technologies and future work using this research should proceed with caution. Assessment of the tasks indicates PLM alignment – or the lack thereof – to human cognitive behavior. Indications of higher human alignment do not indicate an absolute proxy for humans. The goal of tasks in this work is a pre-cursor assessment of PLMs on their ability to act as cognitive models. Therefore, researchers and users should perform more tests before use. .

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*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

W. F. Battig and W. E. Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology Monographs*, 80:1–46.

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.

S. Bhatia and R. Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

I. Borg and P.J.F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer.

Gregory C. Burgess, Jeremy R. Gray, Andrew R. A. Conway, and Todd Samuel Braver. 2011. Journal of experimental psychology : General neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span.

Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874.

John B Carroll. 1993. *Human cognitive abilities: A survey of factor-analytic studies*. 1. Cambridge University Press.

Nichol Castro, Taylor Curley, and Christopher Hertzog. 2021. Category norms with a cross-sectional sample of adults in the united states: Consideration of cohort, age, and historical effects on semantic categories. *Behavior research methods*, 53(2):898–917.

Raymond B Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1.

Raymond Bernard Cattell. 1987. *Intelligence: Its structure, growth and action*. Elsevier.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Andrew R. A. Conway, Nelson Cowan, Michael F. Bunting, David J. Therriault, and Scott R. B. Minkoff. 2002. A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30:163–183.

S. De Deyne, S. Verheyen, E. Ameel, W. Vanpaemel, M. J. Dry, W. Voorspoels, and G. Storms. 2008. Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, 40:1030–1048.

Cody S. Ding. 2018. *Fundamentals of Applied Multidimensional Scaling for Educational and Psychological Research*. Springer International Publishing.

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*.

Qixiang Fang, Daniel L Oberski, and Dong Nguyen. 2024. Patch–psychometrics-assisted benchmarking of large language models: A case study of mathematics proficiency. *arXiv preprint arXiv:2404.01799*.

Gustav Theodor Fechner. 1860. Elements of psychophysics. 1.

Jerry A Fodor. 1985. Precis of the modularity of mind. *Behavioral and brain sciences*, 8(1):1–5.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Ben Goertzel. 2023. Generative ai vs. agi: The cognitive strengths and weaknesses of modern llms. *arXiv preprint arXiv:2309.10371*.

Usha Goswami. 1986. Children's use of analogy in learning to read: A developmental study. *Journal of Experimental Child Psychology*, 42(1):73–83.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

Peter Hagoort. 2019. *Human language: From genes and brains to behavior*. MIT Press.

Richard J Haier. 2023. *The neuroscience of intelligence*. Cambridge University Press.

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

Joshua K. Hartshorne and Laura T. Germine. 2015. When does cognitive functioning peak? the asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, 26:433 – 443.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8 - multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Tom Heyman and Gert Heyman. 2019. Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology*, 72:2084–2109.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Faye Holt, William Held, and Diyi Yang. 2024. Perceptions of language technology failures from south asian english speakers.

Keith J Holyoak, Ellen N Junn, and Dorrit O Billman. 1984. Development of analogical problem-solving skill. *Child development*, pages 2042–2055.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. 2021. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 1567–1574.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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.

Brenden M Lake and Gregory L Murphy. 2023. Word meaning in minds and machines. *Psychological review*, 130(2):401.

George Lakoff. 2008. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023. Llm360: Towards fully transparent open-source llms.

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.

Kevin S McGrew. 2009. Chc theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research.

K. Misra, A. Ettinger, and J. T. Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

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

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

G. Murphy. 2002. *The Big Book of Concepts*. MIT press.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. New and improved embedding model. Accessed: 2023-08-14.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

John M. Parkman. 1971. Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, 91(2):191–205.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Inc. Pearson. 2021. Miller's analogy test preparation.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 0(0):17456916231214460. PMID: 38165766.

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.

Jean Raven. 2003. Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence transformers: Multilingual sentence embeddings using bert / roberta / xlm-roberta & co. with pytorch. Accessed: 2023-08-14.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

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.

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

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

Noam Siegelman. 2020. Statistical learning abilities and their relation to language. *Language and Linguistics Compass*, 14(3):e12365.

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.

Robert J Sternberg. 2000. *Handbook of intelligence*. Cambridge University Press.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Louis Leon Thurstone. 1938. Primary mental abilities. *Psychometric monographs*.

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.

Peter D Turney. 2005. Measuring semantic similarity by latent relational analysis. *arXiv preprint cs/0508053*.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

J. P. Van Overschelde, K. A. Rawson, and J. Dunlosky. 2004. Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, 50:289–335.

Siddhartha Vemuri, Raj Sanjay Shah, and Sashank Varma. 2024. Evaluating typicality in combined language and vision model concept representations. In *Under review*.

Alex Warstadt and Samuel R. Bowman. 2024. What artificial neural networks can tell us about human language acquisition.

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. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks?

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness and harmlessness with rlaif.

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*.

A Computational Resources

The models are evaluated on Nvidia A100 GPUs with 80 GB RAM. The evaluation in this paper cumulatively takes 1600 GPU hours. We use the provided APIs by OpenAI and Google for models of the GPT-X family and Gemini respectively.

B Extended set of experiments

B.1 Numeric abilities: Magnitude comparison effects

Physical quantities in the world are encoded as logarithmically scaled magnitude representations (Fechner, 1860). While the distance and the ratio effects are the biggest indicators of the presence of such log-scaled magnitude representations and the numerical precision in humans, other human effects also explain the mental number line. These effects are as follows:

- **Distance effect** (refer to figure 1 (A) top): The greater the distance |x-y| between two numbers (x, y), the faster the comparison in humans (Moyer and Landauer, 1967b).

Table 4: Magnitude Comparison effects. Distance Effect: Averaged R^2 values of different LLMs when fitting a linear function on the cosine-similarity vs distance plot. Size Effect: Averaged R^2 values of different LLMs when fitting a linear function on the cosine-similarity vs size-difference plot. Ratio Effect: Averaged R^2 values of different LLMs when fitting a negative exponential function on the cosine-similarity vs ratio plot. Note: Each value is averaged across all three input types and all model layers to produce one generalizable score. MDS Stress: The stress value is a measure of how well the distances between the points in the multidimensional space represent the dissimilarities of the original data points (lower is better). MDS Correlation: Correlation between the MDS solutions and the expected values of human MNL. Range (Sim): This indicates the range of the cosine-similarities. Max (sim): This indicates the maximum similarity between any two numbers. Range and Max (sim) describe the y-axis.

Model	Distance Effect	Ratio Effect	Size Effect	MDS Stress	MDS Correlation	Range (Sim)	Max (Sim)
Amber-7B	0.913	0.591	0.607	0.157	0.572	0.008	0.995
Falcon-7B	0.928	0.838	0.725	0.183	0.655	0.286	0.779
Starling-LM-7B-alpha	0.522	0.187	0.494	0.320	0.305	0.001	0.995
Llama-2-7B	0.670	0.614	0.535	0.122	0.547	0.016	0.983
Llama-2-13B	0.672	0.263	0.421	0.234	0.372	0.002	0.999
Mistral-7B	0.641	0.233	0.244	0.287	0.425	0.001	0.996
Mistral-7B-Instruct	0.637	0.543	0.182	0.317	0.512	0.001	0.992
Qwen-0.5B	0.833	0.553	0.215	0.246	0.679	0.064	0.911
Qwen-1.8B	0.878	0.301	0.330	0.198	0.328	0.107	0.902
Qwen-4B	0.881	0.264	0.330	0.215	0.581	0.160	0.763
Qwen-7B	0.858	0.616	0.257	0.153	0.636	0.129	0.734
Qwen-14B	0.783	0.507	0.206	0.248	0.369	0.138	0.710
Pythia-70M	0.829	0.429	0.418	0.204	0.463	0.060	0.949
Pythia-160M	0.947	0.665	0.382	0.231	0.715	0.042	0.970
Pythia-410M	0.926	0.679	0.393	0.210	0.710	0.041	0.972
Pythia-1B	0.944	0.702	0.470	0.196	0.725	0.037	0.973
Pythia-1.4B	0.933	0.764	0.600	0.203	0.658	0.022	0.983
Pythia-2.8B	0.961	0.723	0.459	0.256	0.737	0.009	0.993
Pythia-6.9B	0.909	0.713	0.535	0.195	0.663	0.013	0.990
Pythia-12B	0.846	0.595	0.540	0.189	0.620	0.007	0.993

- **Size effect**: Given two comparisons of the same distance (i.e., of the same value for |x - y|), the smaller the numbers, the faster the comparison (Parkman, 1971).

- **Ratio effect** (refer to figure 1 (A) bottom): The time taken by humans to compare two numbers (x,y) is a decreasing function of the ratio of the larger number over the smaller number $\frac{max(x,y)}{min(x,y)}$ (Halberda et al., 2008).

- **Multidimensional scaling:** Along with the three effects, we investigate the consistency of the latent number representations of PLMs with the human MNL using multidimensional scaling (Borg and Groenen, 2005; Ding, 2018). MDS recovers the latent representation from the cosine (dis)similarities between the vector representations of all pairs of numbers (for a given LLM, layer, and number format). This is evaluated by the correlation between the positions of the numbers 1 to 9 in the MDS solution and the expected values (log(1) to log (9)) of the human MNL (refer to the correlation value in table 4).

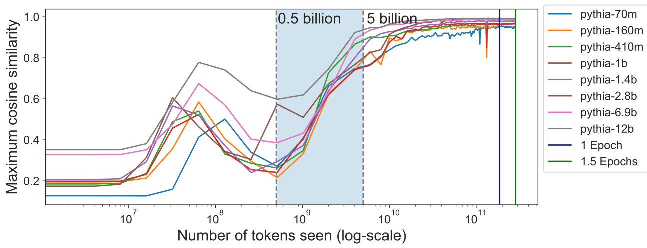

Figure 5: Development of the idea of "numbers" in Pythia. The y-axis indicates the maximum cosine similarity between the latent representations of any two number words/ digits.

Beyond these effects, we investigate the development of the latent understanding of the concept of "numbers" in the PLMs. As PLMs see more data, the average values of the similarity become larger, indicating that models learn the distinctions among numbers better (refer to figure 5). This is further substantiated by figure 6, where the similarities between number words develop to be greater than the similarity between (number, non-number) words and (non-number, non-number) words.

B.2 Linguistic Abilities

The 12 phenomena tested by BLiMP are as follows:

Table 5: Accuracy of different language models on the BLiMP linguistic acceptability tasks.

Model	BLiMP	Syntax	Semantic	Morphology
Amber-7B	0.794 (± 0.174)	0.779 (± 0.011)	0.736 (± 0.011)	0.888 (± 0.009)
Falcon-7B	0.817 (± 0.173)	0.797 (± 0.011)	0.758 (± 0.011)	0.917 (± 0.008)
Starling-LM-7B-alpha	0.827 (± 0.161)	0.799 (± 0.011)	0.788 (± 0.011)	0.938 (± 0.007)
Llama-2-7B	0.818 (± 0.165)	0.792 (± 0.011)	0.782 (± 0.011)	0.917 (± 0.008)
Llama-2-13B	0.793 (± 0.184)	0.757 (± 0.011)	0.767 (± 0.011)	0.898 (± 0.008)
Mistral-7B	0.829 (± 0.174)	0.801 (± 0.011)	0.780 (± 0.010)	0.940 (± 0.007)
Mistral-7B-Instruct	0.834 (± 0.149)	0.808 (± 0.011)	0.788 (± 0.011)	0.931 (± 0.008)
Qwen-0.5B	0.785 (± 0.176)	0.759 (± 0.012)	0.718 (± 0.012)	0.907 (± 0.008)
Qwen-1.8B	0.792 (± 0.162)	0.777 (± 0.012)	0.764 (± 0.011)	0.875 (± 0.010)
Qwen-4B	0.730 (± 0.154)	0.694 (± 0.013)	0.728 (± 0.013)	0.814 (± 0.012)
Qwen-7B	0.789 (± 0.156)	0.769 (± 0.012)	0.736 (± 0.012)	0.885 (± 0.010)
Qwen-14B	0.792 (± 0.144)	0.775 (± 0.012)	0.747 (± 0.012)	0.881 (± 0.010)
Pythia-70M	0.723 (± 0.210)	0.701 (± 0.012)	0.628 (± 0.012)	0.872 (± 0.010)
Pythia-160M	0.749 (± 0.207)	0.717 (± 0.012)	0.718 (± 0.011)	0.864 (± 0.010)
Pythia-410M	0.815 (± 0.169)	0.785 (± 0.011)	0.752 (± 0.011)	0.935 (± 0.007)
Pythia-1B	0.806 (± 0.198)	0.782 (± 0.011)	0.728 (± 0.011)	0.935 (± 0.007)
Pythia-1.4B	0.819 (± 0.173)	0.792 (± 0.011)	0.768 (± 0.011)	0.931 (± 0.008)
Pythia-2.8B	0.827 (± 0.156)	0.800 (± 0.011)	0.782 (± 0.011)	0.925 (± 0.007)
Pythia-6.9B	0.809 (± 0.179)	0.792 (± 0.011)	0.750 (± 0.011)	0.913 (± 0.008)
Pythia-12B	0.829 (± 0.158)	0.804 (± 0.011)	0.778 (± 0.011)	0.932 (± 0.007)
Gemini	NA	NA	NA	NA
GPT-3.5-Turbo	0.825 (± 0.166)	0.818 (± 0.010)	0.781 (± 0.011)	0.931 (± 0.007)
GPT-4	0.849 (± 0.120)	0.797 (± 0.010)	0.801 (± 0.009)	0.941 (± 0.007)

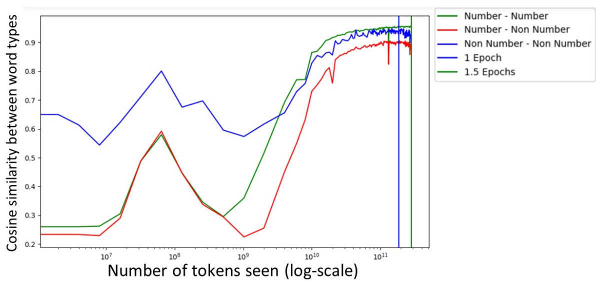

Figure 6: Development of the idea of "numbers" in Pythia. The y-axis shows the cosine similarity between word types. The cosine similarity values are averaged over all input types, all model layers, and all model sizes.

- Anaphor agreement (morphology): This linguistic phenomenon tests if an anaphor (pronoun) adheres to the antecedent (noun or phrase it refers to) in terms of gender, number, or person.

- Argument Structure (syntax): The argument structure tests the relationship between a verb and its arguments (such as nouns or noun phrases).

- Binding (syntax, semantics): This tests the structural relationship between an anaphor (pronoun) and its antecedent (noun or phrase it refers to).

- Control/ Raising (syntax, semantics): These structures test how semantics differ by syntactical variations of subjects/verbs in subordinate and main clauses.

- Determiner-noun agreement (morphology): This tests the agreements of the determiners with the corresponding nouns in number (singular or plural) and sometimes gender (e.g., "his" for masculine nouns, "her" for feminine nouns).

- Ellipsis (syntax): This refers to the omission of words from a sentence that can be understood from the context.

- Filler-gap (syntax): This tests the syntactic structure of sentences that include phrasal movements (wh-questions, relative clauses).

- Irregular forms (morphology): Forms in language that do not follow regular patterns and may need to be memorized. For example, the superlative of good is better, best, and not gooder, goodest.

- Island effects (syntax): These test the constraints on syntactic environments where the gap in a filler-gap dependency can occur.

- NPI licensing (semantics): This phenomenon tests the constrained situations where negative polarity items like any and ever are limited to the scope of negation.

- Quantifiers (semantics): This phenomenon tests the constraints regarding the placement of quantifiers. Specifically, BLiMP looks at superlative quantifiers (such as "at least") that cannot occur within negation, and definite quantifiers and determiners cannot function as subjects in existential "there" constructions.

- Subject-verb agreement (morphology): The subject and tense forms of the verb must agree on the number, for example, singular vs plural.

Table 5 shows that the PLMs are more accurate in morphology than in language syntax and semantics. Most models also perform better on syntactic language features than semantic language features.

B.3 Conceptual Understanding

Table 7 shows the human alignment of PLMs on their concept understanding for different operationalization methods. We see that Gemini, GPT-3.5-Turbo, and GPT-4 perform better than other models. Furthermore, Surprisal and Prompting-based methods are stronger techniques for evaluating conceptual understanding of models than representation-based methods. Given the higher performance of Prompting methods on three API-based models, we only show the category-wise results for those models. The final prompt design is given in section B.3.1 and table 11. Tables 8, 9, and 10 show Spearman's correlation on the categories along with the standard deviation, the minimum correlation, and the maximum correlation. We perform the same infilling tasks 50 times for each category to account for variations in generations. We note that the models often failed to return all the options in the in-filling task. We discard such situations in our analysis.

Note: Under the closeness judgment protocol, our experiments fail to match up to the performance of the models used by Vemuri et al. (2024). This is because our choice of open-source models only provides token representations, on which we later perform an aggregation operation. This aggregation operation leads to a loss of information. In contrast, Vemuri et al. (2024) use sentence-transformer models (Reimers and Gurevych, 2019), which provide singular latent representation for longer text. This variation in experimentation leads to the difference in alignment scores.

Table 6: Typicality effects: Comparing Average Spearman's correlation score across categories from tables 8, 9, and 10.

Categories	GPT 3.5	GPT 4	Gemini
bird	0.183	0.536	0.353
carpenters tool	0.418	0.679	0.610
clothing	0.022	0.594	0.155
color	-0.016	0.882	0.569
dwelling	0.208	0.335	0.340
earth formation	0.251	0.496	0.155
fabric	0.48	0.708	0.504
fish	0.183	0.643	0.247
flower	0.48	0.772	0.515
flying thing	0.07	0.249	0.184
footwear	0.118	0.521	0.218
four-legged animal	0.435	0.818	0.537
fruit	0.465	0.726	0.508
furniture	0.069	0.525	0.147
gardeners tool	0.355	0.557	0.507
green thing	0.196	0.572	0.335
insect	0.18	0.629	0.286
instrument	0.194	0.709	0.450
kitchen utensil	0.384	0.624	0.252
ship	0.104	0.233	-0.078
snake	0.177	0.419	0.328
toy	0.299	0.480	0.169
tree	0.333	0.557	0.445
vegetable	0.096	0.783	0.121
vehicle	0.17	0.381	0.033
weapon	0.348	0.421	0.239
weather	0.333	0.255	0.274
Average	0.242	0.559	0.311

Table 7: Results for the typicality effects using the three methods

Model	Latent Representations	Surprisal Values				Prompting
		Zero-shot	One-shot	Two-shot	Three-shot	
Amber-7B	0.083	0.250	0.227	0.261	0.247	NA
Falcon-7B	-0.116	0.180	0.215	0.242	0.200	NA
Starling-LM-7B-alpha	-0.003	0.258	0.211	0.215	0.235	NA
Llama-2-7B	-0.065	0.238	0.213	0.202	0.207	NA
Llama-2-13B	0.076	0.247	0.163	0.183	0.170	NA
Mistral-7B	-0.025	0.245	0.219	0.261	0.257	NA
Mistral-7B-Instruct	0.033	0.255	0.192	0.204	0.235	NA
Qwen-0.5B	0.072	0.282	0.264	0.288	0.250	NA
Qwen-1.8B	0.114	0.235	0.246	0.251	0.215	NA
Qwen-4B	0.001	0.246	0.217	0.252	0.193	NA
Qwen-7B	0.006	0.229	0.203	0.220	0.220	NA
Qwen-14B	-0.140	0.249	0.224	0.207	0.199	NA
Pythia-70M	0.005	0.211	0.266	0.291	0.285	NA
Pythia-160M	0.067	0.260	0.263	0.276	0.264	NA
Pythia-410M	0.126	0.284	0.235	0.282	0.242	NA
Pythia-1B	0.090	0.280	0.309	0.287	0.264	NA
Pythia-1.4B	0.074	0.283	0.249	0.267	0.235	NA
Pythia-2.8B	0.221	0.273	0.286	0.267	0.236	NA
Pythia-6.9B	0.105	0.280	0.264	0.250	0.220	NA
Pythia-12B	0.184	0.291	0.248	0.274	0.270	NA
Gemini	NA	NA	NA	NA	NA	0.311
GPT-3.5-Turbo	NA	0.231	0.248	0.299	0.270	0.242
GPT-4	NA	0.428	0.471	0.399	0.402	0.559

Table 8: Average Spearman's correlation score for each category on 50 runs of each in-filling experiment on the Gemini-Pro model.

Categories	Average SpearmanR	Minimum Values	Maximum Values	Std Dev
bird	0.353	-0.156	0.582	0.144
carpenters tool	0.610	0.417	0.885	0.104
clothing	0.155	-0.104	0.523	0.141
color	0.569	-0.147	0.916	0.260
dwelling	0.340	0.140	0.499	0.086
earth formation	0.155	-0.449	0.494	0.191
fabric	0.504	0.125	0.811	0.168
fish	0.247	-0.505	0.611	0.265
flower	0.515	-0.183	0.779	0.208
flying thing	0.184	-0.068	0.602	0.193
footwear	0.218	-0.340	0.569	0.215
four-legged animal	0.537	0.225	0.689	0.099
fruit	0.508	-0.019	0.802	0.222
furniture	0.147	-0.479	0.663	0.310
gardeners tool	0.507	0.025	0.771	0.151
green thing	0.335	0.037	0.535	0.117
insect	0.286	-0.121	0.635	0.193
instrument	0.450	0.092	0.832	0.175
kitchen utensil	0.252	-0.164	0.691	0.243
ship	-0.078	-0.414	0.277	0.179
snake	0.328	-0.156	0.596	0.147
toy	0.169	-0.203	0.526	0.174
tree	0.445	0.257	0.585	0.073
vegetable	0.121	-0.322	0.596	0.184
vehicle	0.033	-0.053	0.236	0.055
weapon	0.239	-0.173	0.577	0.193
weather	0.274	-0.029	0.591	0.147

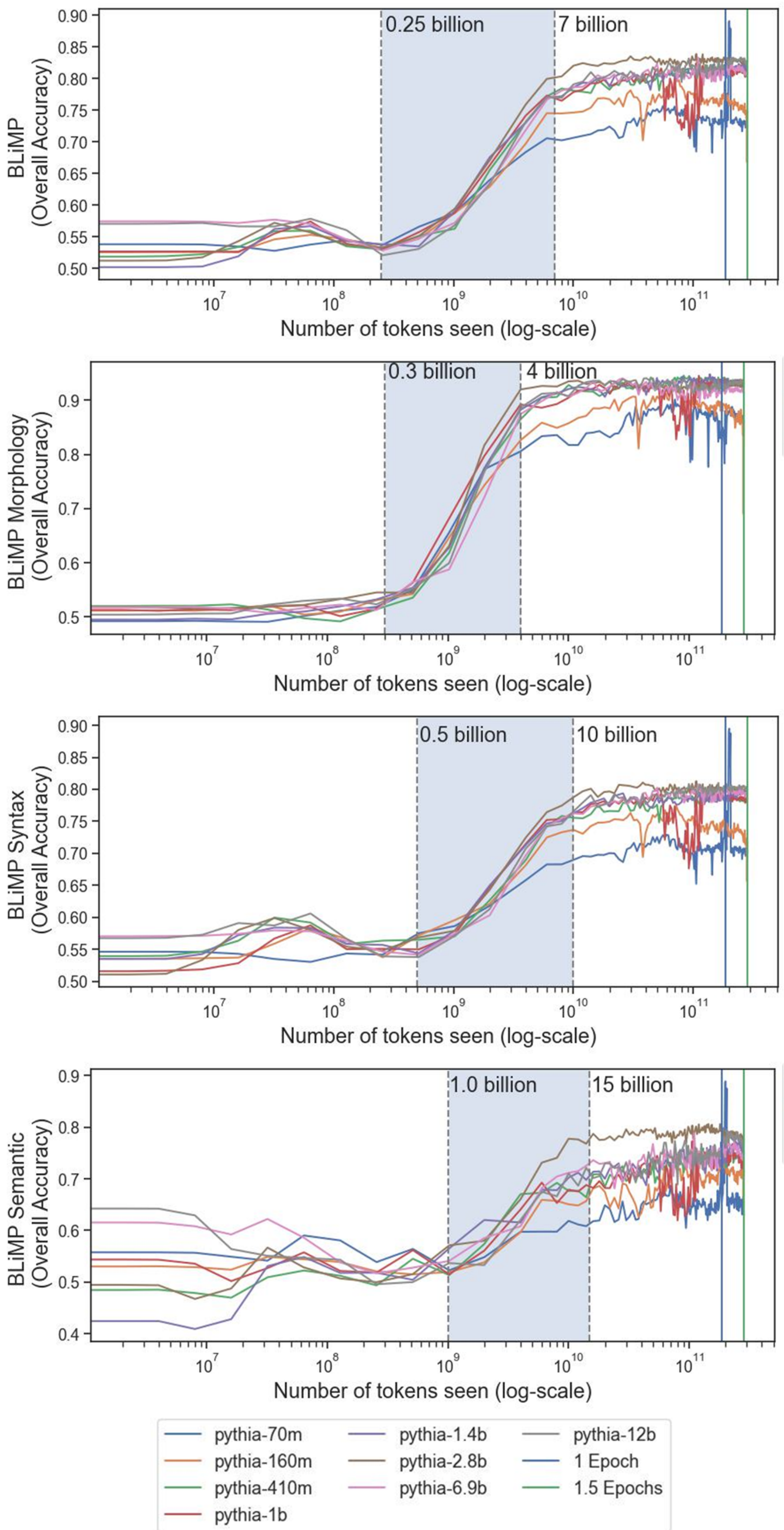

Figure 7: Developmental trajectory of the Pythia suite of models on the BLiMP linguistic acceptability tasks.

Table 9: Average Spearman's correlation score for each category on 50 runs of each in-filling experiment on the GPT-3.5-Turbo model.

Categories	Average SpearmanR	Minimum Values	Maximum Values	Std Dev
bird	0.183	-0.209	0.552	0.209
carpenters tool	0.418	-0.162	0.858	0.282
clothing	0.022	-0.321	0.540	0.192
color	-0.016	-0.596	0.564	0.261
dwelling	0.208	-0.053	0.400	0.123
earth formation	0.251	-0.296	0.562	0.217
fabric	0.480	-0.044	0.767	0.233
fish	0.183	-0.326	0.690	0.280
flower	0.480	-0.301	0.800	0.269
flying thing	0.070	-0.181	0.377	0.149
footwear	0.118	-0.439	0.581	0.241
four-legged animal	0.435	-0.264	0.869	0.292
fruit	0.465	-0.006	0.868	0.241
furniture	0.069	-0.325	0.447	0.195
gardeners tool	0.355	-0.311	0.796	0.294
green thing	0.196	-0.337	0.572	0.211
insect	0.180	-0.248	0.503	0.201
instrument	0.194	-0.242	0.466	0.191
kitchen utensil	0.384	-0.610	0.797	0.334
ship	0.104	-0.314	0.599	0.250
snake	0.177	-0.244	0.591	0.196
toy	0.299	-0.210	0.603	0.180
tree	0.333	-0.199	0.731	0.289
vegetable	0.096	-0.191	0.542	0.172
vehicle	0.170	-0.381	0.381	0.201
weapon	0.348	-0.058	0.609	0.156
weather	0.333	-0.425	0.662	0.236

Table 10: Average Spearman's correlation score for each category on 50 runs of each in-filling experiment on the GPT-4 model.

Categories	Average SpearmanR	Minimum Values	Maximum Values	Std Dev
bird	0.536	0.355	0.756	0.098
carpenters tool	0.679	0.549	0.843	0.078
clothing	0.594	0.350	0.751	0.100
color	0.882	0.813	0.952	0.035
dwelling	0.335	0.183	0.497	0.070
earth formation	0.496	0.373	0.628	0.061
fabric	0.708	0.583	0.801	0.052
fish	0.643	-0.237	0.817	0.218
flower	0.772	0.629	0.869	0.057
flying thing	0.249	-0.118	0.704	0.221
footwear	0.521	0.191	0.721	0.112
four-legged animal	0.818	0.634	0.906	0.056
fruit	0.726	0.567	0.868	0.069
furniture	0.525	0.381	0.605	0.055
gardeners tool	0.557	0.314	0.757	0.098
green thing	0.572	0.444	0.709	0.050
insect	0.629	0.451	0.871	0.103
instrument	0.709	0.585	0.885	0.064
kitchen utensil	0.624	0.358	0.750	0.075
ship	0.233	-0.346	0.618	0.232
snake	0.419	0.002	0.575	0.108
toy	0.480	0.277	0.675	0.111
tree	0.557	0.300	0.781	0.106
vegetable	0.783	0.413	0.892	0.102
vehicle	0.381	0.166	0.699	0.119
weapon	0.421	0.268	0.650	0.082
weather	0.255	0.122	0.357	0.061

Table 11: Prompt design for evaluating typicality effects in models bigger than 30 billion parameters.

Prompt region	Description	Actual prompt
Guidelines	Describe the overall idea of typicality to the model and the task guidelines	Appendix B.3.1
Query	This is the actual fill-in-the-blanks task	The ___ is a "Category-Name"
Options	List of items in a randomized order and separated by a new line	—

B.3.1 Conceptual Understanding - Final Prompt

Typicality effects refer to the influence of the typicality or prototypicality of an object or category on various cognitive processes, including perception, categorization, and memory. The concept of typicality stems from the prototype theory, which suggests that our mental representations of categories are based on prototypes or typical examples.

In the context of perception, typicality effects can influence how we perceive and recognize objects. Objects that are more prototypical or representative of a category are typically perceived more quickly and accurately than atypical objects. For example, when shown a series of pictures of birds, a typical bird like a robin would be recognize faster than a less typical bird like a penguin.

In categorization tasks, typicality effects can influence how we classify objects into categories. Prototypical or highly typical objects are more likely to be assigned to their corresponding category than atypical objects. For instance, when asked to categorize fruits, an apple, being a highly typical fruit, is more likely to be classified as a fruit compared to a less typical fruit like a durian.

Typicality effects also impact memory processes. Prototypical objects are typically better remembered than atypical objects. When asked to recall a list of animals, participants are more likely to remember prototypical animals such as dogs or cats compared to less typical animals like lemurs or armadillos.

Overall, typicality effects demonstrate how the typicality or prototypicality of objects within a category influences our perception, categorization, and memory processes, highlighting the role of prototypes in cognitive functioning.

Based on the typicality effect definitions, give rankings for filling the blank task without any description from the following options.

Make sure to include all the items from the options. Please return items in the following manner:

1. item1

2. item2

3. item3

Also make sure to use the same items as given in the options.

Query:

A ___ is a [Category Name]

Options:

[A]

[B]

B.4 Fluid Reasoning

Humans cannot completely operate without relying on prior experience. The pervasive role of prior knowledge in shaping cognition is a foundational tenet of the cognitive revolution. However, "Fluid intelligence" is the ability to solve novel and abstract problems (Raven, 2003). It is a core cognitive ability, closely related to other domain-general cognitive abilities like working memory, and executive function, both correlationally (Conway et al., 2002) and in terms of the underlying neural correlates (i.e., in the prefrontal cortex) (Burgess et al., 2011). It is distinguished from crystallized intelligence, which is composed of the domain-specific knowledge and skills one acquires through one's lifetime (Hartshorne and Germine, 2015). This distinction is a classic one in psychology (Carroll, 1993).

B.4.1 Scholastic Assessment Test analogy questions

Previous work has shown that fluid reasoning correlates with analogical reasoning (Goswami, 1986; Snow et al., 1984; Cattell, 1987). AI, ML, and NLP research has focused on analogical reasoning because this requires many componential abilities: syntactic parsing, semantic understanding, categorization, inductive reasoning, mathematical reasoning, and so on (Pearson, 2021). Research on the cognitive alignment of PLMs has focused on performance on the 374 Scholastic Assessment Tests (SAT) analogy questions by Turney (2005). Despite being broadly used in literature (Turney, 2005; Turney and Pantel, 2010; Hendrickx et al., 2019; Webb et al., 2023), our pilot experiments show that PLMs like GPT-3.5-Turbo, GPT-4, and Gemini perform nearly at ceiling on this test, while other open source models perform poorly on the same test. This hints that the set of questions in the test may be part of the GPT-X/ Gemini training or tuning data.

Operationalization: Each problem is of the form A:B::?, with answer choices containing candidates for C:D. We evaluate the performance of models in three ways:

- Closeness judgment problem: Calculate the cosine similarity between the obtained latent representations for the member and the category. This requires models where the latent representations are readily available. These cosine similarities are calculated in different ways:
  - 3-cos-add: cos(vector(D),vector(C) - vector(A) + vector(B))
  - 3-cos-mul: cos(vector(D), vector(B))*cos(vector(D), vector(C))/(cos(vector(D), vector(A))+ e); e is a small constant to prevent overflow.
  - Concat-cos: cos([vector(A) || vector(B)] , [vector(C) || vector(D)])
- Surprisal values: Calculating the summation of probabilities for each token with the as=to relationship; forming the sequence A is to B as C is to D.
- Prompting: Prompt the models with the following design: Guidelines, Query, and Options. The Guideline highlights the task of solving the analogy problem. The Query consists of A:B. The options are the candidate pairs C:D.

B.4.2 Raven's Progressive Matrices - list of prompts used in experiments

1. "Solve the following Raven Progressive Matrix problem by identifying the pattern in the sequences. Select the correct choice for the missing element.

2. "Identify the correct option to complete the Raven Progressive Matrix. Consider the patterns in numeric and fractional values across the rows to solve the problem."

3. "Solve the Raven Progressive Matrix problem"

4. "Solve the Raven Progressive Matrix problem. Select the correct choice for the missing element in row 3."

5. "Complete the pattern in the Raven Progressive Matrices problem"

6. "Apply abstract reasoning to solve the following Raven Progressive Matrices problem:"

7. "Solve the Raven Progressive Matrices by identifying patterns and drawing analogies. Select the correct choice for the missing element in row 3."

8. "Select the correct choice for row 3, using the patterns and analogies from rows 1 and 2."

row1: (2,0.5,100), (4,0.5,100), (3,0.5,100)

row2: (3,0.7,50), (2,0.7,50), (4,0.7,50)

row3: (4,0.2,70), (3,0.2,70), ?

Choices: (1,0.2,70), (5,0.2,30), (5,0.2,70), (2,0.2,70), (5,0.2,110), (4,0.2,70), (3,0.5,70), (2,0.2,90)"