

Dreaming Out Loud: A Self-Synthesis Approach For Training Vision-Language Models With Developmentally Plausible Data

Badr AlKhamissi* Yingtian Tang* Abdülkadir Gökçe*
Johannes Mehrer† Martin Schrimpf†
EPFL

Abstract

While today’s large language models exhibit impressive abilities in generating human-like text, they require massive amounts of data during training. We here take inspiration from human cognitive development to train models in limited data conditions. Specifically we present a self-synthesis approach that iterates through four phases: Phase 1 sets up fundamental language abilities, training the model from scratch on a small corpus. Language is then associated with the visual environment in phase 2, integrating the model with a vision encoder to generate descriptive captions from labeled images. In the “self-synthesis” phase 3, the model generates captions for unlabeled images, that it then uses to further train its language component with a mix of synthetic, and previous real-world text. This phase is meant to expand the model’s linguistic repertoire, similar to humans self-annotating new experiences. Finally, phase 4 develops advanced cognitive skills, by training the model on specific tasks such as visual question answering and reasoning. Our approach offers a proof of concept for training a multimodal model using a developmentally plausible amount of data.

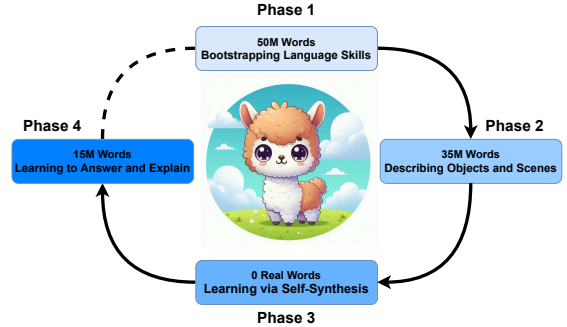


Figure 1: **Self-Synthesis Training Framework.** Our model BabyLLaMA is trained in four phases that connect fundamental language abilities to vision by learning to describe unlabeled visual experiences. We divided our approach in 4 phases, each feeding its best snapshot in terms of validation loss to the next phase. Phase 1 is concerned with fundamental language skill acquisition using 50M words. Phase 2 combines visual and text data (35 M words) to learn to describe objects and scenes. In phase 3 - making our approach one revolving around self-synthesis - we generate captions from images and use this synthesized text (i.e., 0 words from real-world corpora) to further train the model’s language decoder. Phase 4 closes the loop using 15M words to develop skills for advanced visuo-linguistic tasks such as question answering and reasoning about the environment.

1 Introduction

Recent advances in machine learning have produced large language models (LLMs) that, after training on massive text corpora, are capable of generating human-like text. However, when comparing LLM training to human development, the amount of data required for successful model training far exceeds the quantities that humans learn from during their development (Warstadt et al., 2023a). The human brain is thus often seen as a more sample-efficient learning machine than contemporary artificial neural network approaches (Marcus, 2020).

In this work, we take inspiration from human cognitive development to build new models under limited data conditions that more closely resemble the language experience of humans. Specifically, humans learn language in combination with other senses, and use it to reflect on their experiences. We implement this idea via a *self-synthesis* approach that combines vision and language such that the model learns on external (real-world) text as well as its own (synthetic) description of unlabeled visual experiences (Figure 1). Self-synthesis can also be seen as analogous to the process of dreaming, which neuroscience research suggests functions as providing “augmented samples of waking experiences,” helping to shape neural representations

*Equal Contribution

†Equal Supervision

and prevent overfitting to those experiences (Hoel, 2021; Prince and Richards, 2021).

2 Dataset Selection

In line with the BabyLM challenge requirements (Warstadt et al., 2023b), we restrict our training data to 100 million words, which approximates the maximum number of words a 13-year-old would encounter in their lifetime (Gilkerson et al., 2017). In contrast, the latest LLaMA-3-8B model was trained on 15 trillion tokens (Dubey et al., 2024), which is 150,000 times larger than our training budget. We created our own dataset of 100 million words, emphasizing diversity and quality. This word budget is split evenly between a text-only corpus and a multimodal image-text corpus.

Text-Only Data Our text corpus comprises 50 million words selected from the top-scoring sentences of FineWeb-Edu’s October 2024 Common-Crawl snapshot (Lozhkov et al., 2024), based on their educational quality. FineWeb-Edu is a subset of the FineWeb dataset (Penedo et al., 2024), which is created using scalable, automated annotations to assess educational value. The educational scores were assigned by LLaMA-3-70B-Instruct, which rated 500,000 samples on a scale from 0 to 5 for their educational quality (Penedo et al., 2024). Models trained on this dataset have surpassed all other publicly available web datasets on several educational benchmarks, including MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018).

Image-Text Data Our image-text corpus consists of two groups: (1) image-caption data used for visual experience training (“phase 3” Section 5.3); (2) multi-task image-text data used for finetuning the model towards advanced reasoning (“phase 4”, Section 5.4), which include captioning, VQA, and visual reasoning. For the images with captions used for visual experience training, we select subsets from WIT (Srinivasan et al., 2021), obelics (Laurençon et al., 2024), and LAION (Schuhmann et al., 2021). These datasets include diverse image descriptions such as wikipedia paragraphs, news, and also simple short captions. We sampled 27 million, 5 million, and 3 million words respectively from the 3 datasets. For the multi-task image-text data, we used M3IT (Li et al., 2023), a dataset curated for multi-lingual instruction tuning and sampled 15 million words from it. The goal is

to enhance the model’s ability to follow instructions as well as gain more advanced skills such as visual-reasoning, such that it can utilize its acquired knowledge more effectively. Taken together, the two groups of image-text data make up a total of 50 million words. The selection of these datasets was not arbitrary; it resulted from multiple iterations aimed at ensuring both diversity and quality.

3 Benchmarks

We evaluate our model across six benchmarks: three focused on language-only tasks and three on vision-language tasks. Except for GLUE, where we fine-tune the model on each subtask using LoRA (Hu et al., 2022), all benchmarks are evaluated in a zero-shot setting.

3.1 Language-Only Benchmarks

BLiMP BLiMP is a benchmark that evaluates key grammatical phenomena in English. It is composed of 67 sub-datasets, each containing 1,000 minimal pairs designed to highlight specific contrasts in syntax, morphology, or semantics. The data is automatically generated based on grammars developed by experts (Warstadt et al., 2019).

Elements of World Knowledge (EWoK) EWoK is a benchmark that evaluates the world modeling abilities of language models. It covers 11 key domains of world knowledge essential for human-like world modeling. These domains range from reasoning about spatial relations to understanding social interactions (Ivanova et al., 2024).

GLUE The General Language Understanding Evaluation (GLUE) benchmark is a comprehensive suite of resources designed to train, evaluate, and analyze natural language understanding models. It includes nine diverse tasks focused on sentence or sentence-pair understanding, drawn from well-established datasets. These tasks vary in dataset size, text genre, and complexity, providing a broad assessment of language understanding capabilities (Wang et al., 2018). In our experiments, we utilize LoRA (Hu et al., 2022), a parameter efficient fine-tuning method, in order to tune our model to the GLUE tasks.

3.2 Vision-Language Benchmarks

VQA We use the second version of the Visual Question Answering (VQA) benchmark that builds upon the original VQA (Zhang et al., 2015) by incorporating complementary images. In this dataset,

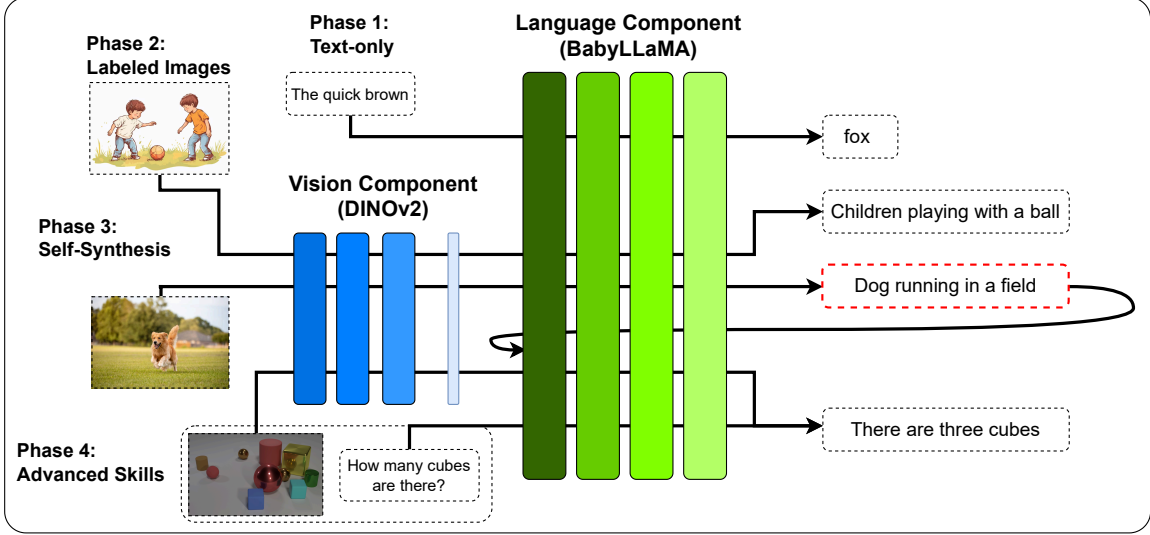


Figure 2: Overview diagram illustrating the four phases of training. Starting from training on text only (phase 1), language capabilities are connected to images (phase 2). The model then self-synthesizes text (red border) on unseen images, and uses this text to continue training the language component (phase 3), which is further refined for e.g. question answering (phase 4). Sizes of model components do not reflect number of parameters.

each question is linked to a pair of similar images, each yielding a distinct answer, thus increasing the challenge. For the model to answer these questions, it requires a grasp of vision, language, and commonsense knowledge (Goyal et al., 2016).

Winoground Winoground is a challenging task and dataset designed to assess the visio-linguistic compositional reasoning abilities of vision-language models. The objective is to correctly match two images with two captions, where both captions use the exact same words or morphemes but arranged in different orders. Expert annotators carefully curated the dataset, providing fine-grained tags to facilitate a detailed analysis of model performance (Thrush et al., 2022).

DevBench This benchmark contains 7 tasks across lexical, syntactic, and semantic domains, each accompanied by human response data at the item level, allowing for detailed comparisons between model scores and human response distributions. The lexical tasks evaluate vocabulary knowledge by assessing the model’s ability to correctly identify the visual referent of a given noun. Syntactic tasks test grammatical understanding, requiring the model to choose the correct scene that aligns with a provided sentence. Semantic tasks measure the model’s ability to represent conceptual similarity, either visually or linguistically, by comparing representational similarity scores (Tan et al., 2024).

4 Model Details

We use the same model architecture provided by the BabyLM Challenge organizers, called BabyLLaMA, which consists of a reduced LLaMA architecture, and we equip it with the DINOv2_{Large} vision encoder to be able to handle visual inputs.

Tokenizer We train a BPE tokenizer with a vocabulary size of 16,000 on the text data from the curated dataset described in Section 2.

Language Model The language models employs the same architectural components as the LLaMA model (Dubey et al., 2024), but with only 16 Transformer layers and a reduced hidden dimension size of 512. The intermediate size in the MLP is 1,024, resulting in a total model size of 58 million parameters. The attention mechanism uses 8 attention heads in each layer.

Vision Encoder To equip our language model with visual capabilities, we incorporate the pre-trained DINOv2_{Large} vision encoder (Oquab et al., 2023). DINOv2 was trained on a large-scale, unlabeled image dataset. It is built on the ViT architecture (Dosovitskiy et al., 2020) and generates 256 vision tokens per image. The vision encoder remains frozen during all experiments to preserve its pretrained features.

Projection Module The projection module serves as the bridge between the vision encoder




Image	Synthetic Description
	Hot off the field at the Ravensboro Golf Club in Ravensboro, IL. I am looking forward to the win of the season. I love the game and hopefully the games are really going to be a big thing...
	The first section of the East End of London's West End was dedicated to the Holy Spirit. The West End of London's West End was the last part of the East End of London...
	The airport is in the midst of a multi-year, \$10 billion contract with the U.S. Navy, which is expected to be operational over the next few years. The agreement is expected to be signed by the United States, Canada, and the United States...

Table 1: Synthetic descriptions generated by the model for the images shown. This table illustrates the model’s ability to associate visual cues with corresponding textual representations.

and the language model. It comprises a two-layer MLP with a GeLU activation function in between. This module projects the concatenated image tokens to match the dimensionality of the language model and is learnable throughout the training process.

5 Self-Synthesis Training Phases

Our framework trains the model in four phases. In each phase, we record the model checkpoint with the lowest validation loss and use it as a starting point for the following phase. For all phases, we use the AdamW optimizer combined with a cosine learning rate scheduler and a batch-size of 256. The learning rate begins with a linear warm-up phase and then gradually decreases to zero over the course of the training.

5.1 Phase 1: Bootstrapping Language Skills

Similar to how children learn a fundamental linguistic repertoire with supervision from their environment, the language component of our model is first trained from scratch on a text-only corpus. Specifically, we train BabyLLaMA for 15 epochs on fewer than 50 million words, using the top-scoring sentences from FineWeb-Edu based on their educational quality. Rather than concatenating and chunking the entire corpus into the maximum sequence length, as is common in language model pretraining, we divided each document from the FineWeb-Edu snapshot into individual sentences. Each sentence was truncated to have a maximum of 256 tokens and a minimum of 10 tokens. We

found that training on shorter sequences by segmenting documents in this way resulted in better performance on the BLiMP benchmark (Warstadt et al., 2019) compared to training with fixed long sequences. The model was trained with a peak learning rate of $1e-4$ and a linear warm-up for the first 5,000 optimization steps. (Learning rates $1e-4$, $5e-5$, $1e-5$ were tried and the one with the lowest validation error was chosen. We did not conduct other hyperparameter selections due to the limited resources. This also applies to other phases.)

5.2 Phase 2: Learning to Associate Language and Vision

Inspired by children learning to associate words with the objects they encounter daily, this training phase integrates a DINOv2_{Large} vision encoder into the model to link visual inputs with language. The model is trained on image-text pairs, keeping the weights of the vision encoder frozen. We first divide each image into 16x16 patches. These 256 tokens are then transformed into feature embeddings by the model. We concatenate every 4 consecutive tokens together to form one embedding to reduce the number of tokens from 256 to 64 before passing them to the projection module. Training involves an autoregressive loss applied exclusively to the text tokens, conditioned on the corresponding image embeddings. In this setup, the projected image embeddings are concatenated with the text embeddings $\mathbf{t}_{1:s}$ before being passed through the language model. This allows the model

Phase	Language-Only Benchmarks				Vision-Language Benchmarks		
	BLiMP	BLiMP Supp.	EWoK	GLUE	VQA	Winoground	DevBench
Phase 1	0.723	0.533	0.500	0.651	-	-	-
Phase 2	0.728	0.561	0.504	0.650	0.395	0.507	0.242
Phase 3	0.736	0.556	0.514	0.647	0.380	0.507	0.350
Phase 4	0.729	0.542	0.502	0.659	0.420	0.509	0.228

Table 2: Performance comparison of the model across different phases of training on various benchmarks. The results show accuracy scores on language-only benchmarks (BLiMP, BLiMP Supp., EWoK, GLUE) and multimodal tasks (VQA, Winoground, DevBench). All benchmarks are evaluated in a zeroshot manner, except for GLUE, which is first finetuned using LoRA for each of its tasks separately. The best result across phases is highlighted in **bold**.

to learn a joint representation that conditions the text generation on the visual context provided by the image.

Formally, let $\mathbf{i} = \{i_1, i_2, \dots, i_{64}\}$ be the set of image embeddings produced by the vision encoder for a given image, and $\mathbf{t} = \{t_1, t_2, \dots, t_s\}$ be the sequence of text tokens associated with that image, where $s \leq 512$. The training objective is to maximize the conditional likelihood of the next text token t_{s+1} given the projected image embeddings and the preceding text tokens, where f is the projection module. This can be formulated as:

$$\max_{\theta, \phi} \sum_{n=1}^N \sum_{s=1}^{|\mathbf{t}_n|} \log p_{\theta, \phi}(t_{n,s+1} \mid [f(\mathbf{i}_n); \mathbf{t}_{n,1:s}])$$

where: $p_{\theta, \phi}(\cdot)$ is the probability distribution generated by the combined model, $f(\mathbf{i}_n)$ represents the image embeddings processed through the projection module, $\mathbf{t}_n = \{t_1, t_2, \dots, t_s\}$ are the text tokens for the n -th image-text pair, N is the total number of training examples, and $|\mathbf{t}_n|$ is the length of the n -th text sequence.

Therefore, just as children learn to describe their visual environment based on supervisory signals (e.g. parents describing the surroundings), the model learns to generate captions for images, articulating what it “sees.” To achieve this, we train the model to produce detailed descriptions across a diverse range of images. Consequently, we balanced the datasets to include samples with detailed descriptions (from WIT and obelics; 35842 samples / 6M words, 135393 samples / 21M words) alongside those with concise captions (from LAION; 323929 samples / 3M words). It is worth noting that although LAION contains only 3 million words, it accounts for more than half of the images due to its short captions. In this phase, we train the model for

5 epochs, with a learning rate that linearly warms-up to 10^{-5} for 250 steps, then decreases to zero throughout training.

5.3 Phase 3: Learning via Self-Synthesis

Self-Synthesis Using Images in the Wild. Beyond supervised learning on images, children also imagine and narrate stories about what they have seen. We implement this idea by having the model generate text from a set of unlabeled images and synthesizing captions that are then used to further train the language component with more diverse text. Concretely, we collected 1.1 million images from obelics that were not used during training. Using nucleus sampling ($p=0.95$) and top-k sampling ($k=50$) with a temperature of 0.7, we generated a total of 42 million words. For each image, a maximum token length between 32 and 64 was uniformly sampled. Table 1 shows a few examples of images and their corresponding text generated by our model. To avoid repetition in the generated text, we limit the maximal number of generated tokens to be 256. Note that some descriptions do not perfectly match the content of the images. This is insofar not an issue, as grammatically and vocabulary-rich text suffices for our purpose.

Continuing Pretraining Inspired by humans mixing real and imagined experiences to enhance their understanding, we train BabyLLaMA on a mixture of self-synthesized text and previously seen “real-world” data to deepen its language abilities. Specifically, we transition back from image-text training to text-only training, combining all the text data we have gathered thus far. This results in a total of 85 million real words and 42 million synthetic words. Our model is trained for just 2 epochs, with a learning rate that linearly warms up to $1e-5$ over 500 optimization steps then decreases towards zero.

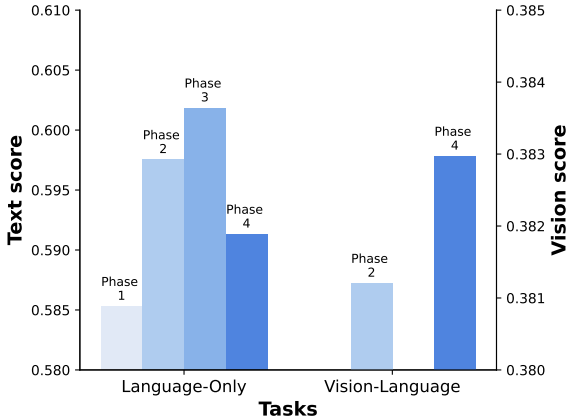


Figure 3: Average performance on all language-only (left) and vision-language-benchmarks (right) across training phases. Each phase yields a small boost for its respective training objective.

To assess the contribution of the self-synthesized text, we train another model version using only the 85 million real words and report the results on the text benchmarks in Section 6.1.

5.4 Phase 4: Learning to Answer and Explain

Equipped with fundamental language skills and the ability to describe their surroundings, human cognitive development includes answering questions and reasoning about their environment. Similarly, we train BabyLLaMA to handle complex visual-linguistic tasks: We finetune the language model along with the projection layer on M3IT. We set the learning rate to 10^{-5} with 250 warm-up updates. The model is trained for 2 epochs.

The division in 4 training phases is inspired by language acquisition in human infants. However, we do not suggest that the exact same phases accurately describe human linguistic development. For example, humans are unlikely to establish fundamental language skills (phase 1) without concurrent visual input that our model only encounters in phase 2.

6 Results

Table 2 presents the performance across various benchmarks, including both language-only and vision-language datasets. For language-only benchmarks, the phase 3 model significantly outperforms earlier models on BLiMP and EWoK, while the phase 4 model achieves the best results on GLUE. Notably, the phase 2 model delivers the highest performance on BLiMP Supplement, which is a smaller dataset compared to BLiMP. In vision-

Benchmark	+ Synth	- Synth
BLiMP	0.736	0.736
BLiMP Supp.	0.556	0.550
EWoK	0.514	0.510

Table 3: Results of the ablation study on language-only benchmarks, comparing the performance of the model trained solely on real-world text (-Synth) against the model trained on a combination of real and synthetic data (+Synth). All benchmarks were evaluated in a zero-shot manner, illustrating the contribution of synthetic data to overall model performance.

language benchmarks, the phase 4 model surpasses the phase 3 model on VQA and Winoground but underperforms on DevBench. Overall, models after phase 3 achieve the highest scores across most benchmarks. To emphasize performance differences across training phases, Figure 3 illustrates the average scores on various benchmarks. For language-only tasks, the phase 3 model shows a substantial improvement over models from phases 1 and 2. However, the phase 4 model lags slightly, likely due to fine-tuning on question-answer datasets, which shifts its focus away from general text modeling. Table 1 provides examples of synthetic descriptions generated by the phase 2 model conditioned on different images. The model accurately captures key elements in the images and produces varied syntactic and content-rich descriptions. However, there are occasional issues with logical consistency, such as the repetition of "United States" in the third example.

6.1 Ablation Study

To measure the contribution of the synthetic data, we train a separate phase 3 model using only real-world text, excluding any generated text, and compare its performance with the model trained on a mixture of both real and synthetic data. Table 3 presents the results on the language-only benchmarks, all evaluated in a zero-shot manner. The findings demonstrate that incorporating synthetic data either enhances or maintains performance across benchmarks, highlighting the potential of scaling self-synthesis with larger datasets.

7 Conclusion

This work proposes a novel self-synthesis approach to training vision-language models in a data-efficient manner inspired by human cognitive

development. By structuring the learning process into four distinct phases—beginning with foundational language abilities, integrating vision and language, generating synthetic data through unlabeled image captioning, and advancing cognitive tasks—the resulting model is able to solve both vision-language and language only benchmarks using a limited amount of data in a unified manner.

While we observed improved performance from each phase of training, these improvements were comparatively small. Curriculum learning methods or architectural modifications might further improve the model’s learning efficiency within the proposed framework. For instance, the phases could be ran repeatedly, such that the model iteratively trains on a mix of real-world text and continuously improving self-synthesized text. A layer-fusion approach could better utilize intermediate layer representations, which has been shown to enhance training in data-limited settings (ElNokrashy et al., 2024). These efforts could close the performance gap while maintaining the developmental plausibility of the training setup. In summary, results presented here suggest that self-synthesis can make effective use of information across modalities, and might help to train performant models with developmentally plausible data regimes.

References

- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *ArXiv*, abs/1803.05457.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. [An image is worth 16x16 words: Transformers for image recognition at scale](#). *ArXiv*, abs/2010.11929.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. 2024. [The llama 3 herd of models](#). *ArXiv*, abs/2407.21783.
- Muhammad ElNokrashy, Badr AlKhamissi, and Mona Diab. 2024. [Depth-wise attention \(DWAtt\): A layer fusion method for data-efficient classification](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4665–4674, Torino, Italia. ELRA and ICCL.
- Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. [Mapping the early language environment using all-day recordings and automated analysis](#). *American Journal of Speech-Language Pathology*, 26(2):248–265.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. [Making the v in vqa matter: Elevating the role of image understanding in visual question answering](#). *International Journal of Computer Vision*, 127:398 – 414.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Erik Hoel. 2021. [The overfitted brain: Dreams evolved to assist generalization](#). *Patterns*, 2(5):100244.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Anna Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, Pramod RT, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Josh Tenenbaum, and Jacob Andreas. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv*.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.
- Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023. M³ it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.
- Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. [Fineweb-edu](#).
- Gary F. Marcus. 2020. [The next decade in ai: Four steps towards robust artificial intelligence](#). *ArXiv*, abs/2002.06177.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

- pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. [Dinov2: Learning robust visual features without supervision](#). *ArXiv*, abs/2304.07193.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). *Preprint*, arXiv:2406.17557.
- Luke Y. Prince and Blake A. Richards. 2021. [The over-fitted brain hypothesis](#). *Patterns*, 2(5):100268.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.
- Alvin Wei Ming Tan, Sunny Yu, Bria Long, Wan-jing Anya Ma, Tonya Murray, Rebecca D. Silverman, Jason D. Yeatman, and Michael C. Frank. 2024. [Devbench: A multimodal developmental benchmark for language learning](#). *ArXiv*, abs/2406.10215.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023a. [Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023b. [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. 2019. [Blimp: The benchmark of linguistic minimal pairs for english](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2015. [Yin and yang: Balancing and answering binary visual questions](#). *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022.