Masked Diffusion Language Models with Frequency-Informed Training

Despoina Kosmopoulou[1,2] **Efthymios Georgiou**[3] **Vaggelis Dorovatas**[2]
Georgios Paraskevopoulos[4] **Alexandros Potamianos**[1,2]

[1] National Technical University of Athens
[2] Archimedes RU, Athena RC
[3] University of Bern
[4] Institute of Language and Signal Processing, Athena RC
despoinakkosmopoulou@gmail.com efthymios.georgiou@unibe.ch

Abstract

We present a masked diffusion language modeling framework for data-efficient training for the *BabyLM 2025 Challenge*. Our approach applies diffusion training objectives to language modeling under strict data constraints, incorporating frequency-informed masking that prioritizes learning from rare tokens while maintaining theoretical validity. We explore multiple noise scheduling strategies, including two-mode approaches, and investigate different noise weighting schemes within the Negative Evidence Lower Bound (NELBO) objective. We evaluate our method on the BabyLM benchmark suite, measuring linguistic competence, world knowledge, and human-likeness. Results show performance competitive to state-of-the-art hybrid autoregressive-masked baselines, demonstrating that diffusion-based training offers a viable alternative for data-restricted language learning.

1 Introduction

By the age of 12, human children are typically exposed to fewer than 100 million words (Gilkerson et al., 2017). In contrast, state-of-the-art language models (LMs) (Touvron et al., 2023; Qwen et al., 2025) are trained on trillions of tokens. The BabyLM Challenge (Warstadt et al., 2023a) was introduced to address this striking efficiency gap by encouraging research on more data-efficient pretraining strategies. The *2025 strict track* constrains participants to train models for up to 10 epochs on a 100M-word corpus (Charpentier et al., 2025).

A prominent recent approach, winning the 2024 iteration of the BabyLM Challenge, `GPT-BERT`, combined a Masked Language Modeling (MLM) and Next Token Prediction (NTP) objective during pretraining (Charpentier and Samuel, 2024). The MLM objective has limited learning (gradient signal) efficiency, utilizing only $\sim$15% of corpus tokens per epoch (Devlin et al., 2019), while NTP learns from all tokens; as a result, NTP-based autoregressive (AR) generative models dominate the landscape of state-of-the-art language modeling (Brown et al., 2020). However, AR models typically use causal attention, only attending to previous tokens, which limits their bidirectional understanding and expressive ability (Devlin et al., 2019).

Recent advances in diffusion models have enabled their application to discrete text generation, with masked diffusion language models (`MDLMs`) emerging as a promising approach that combines bidirectional context modeling with generative training (Sahoo et al., 2024). `MDLMs` are masked language models with "parallel" generative capabilities, offering a compelling middle ground between the bidirectional understanding of MLMs and the generative efficiency of AR models. Unlike traditional MLM where a fixed percentage of tokens is masked at each step, `MDLMs` employ a diffusion process that varies masking rates across training, potentially leading to more efficient learning dynamics. This creates a natural curriculum where the model learns to reconstruct text under varying levels of corruption.

Recent work has shown that `MDLMs` can achieve competitive performance with AR models, while maintaining the bidirectional context benefits of masked models (Sahoo et al., 2024; Shi et al., 2025). However, diffusion models face challenges in data-sparse settings, with their multi-step training process potentially amplifying overfitting issues, an area that remains relatively unexplored in language modeling. Specifically, `MDLMs`' effectiveness in data-constrained settings remains unknown. In this work, we explore whether `MDLMs` trained for just 10 epochs over a 100M word corpus can match or surpass hybrid state-of-the-art approaches like `GPT-BERT`.

We *hypothesize* that the principled diffusion training objective of `MDLMs`, combined with strate-

Proceedings of the First BabyLM Workshop, pages 531–539

gic masking approaches, can achieve more sample-efficient learning compared to fixed-rate MLM or purely autoregressive training. To test this hypothesis, we implement a masked diffusion language modeling framework and explore multiple noise scheduling strategies, including two-mode approaches, while investigating different noise weighting schemes within the Negative Evidence Lower Bound (NELBO) objective. We further introduce frequency-informed masking that progressively prioritizes learning from rare tokens during the diffusion process, directing the model's attention toward more informative and challenging aspects of language while preserving the theoretical validity of the diffusion objective.

Our contributions are threefold: 1) we adapt masked diffusion language modeling for data-restricted settings, exploring multiple noise scheduling strategies including two-mode approaches and different NELBO weighting schemes, 2) we introduce a frequency-informed masking strategy that seamlessly integrates into the diffusion objective while preserving theoretical validity, and 3) we provide comprehensive evaluation on the BabyLM benchmark demonstrating that diffusion-based training achieves competitive performance with established baselines. Our code and weights are made available[1].

2 Related Work

Masked Diffusion Language Modeling Inspired by continuous-time diffusion models (Sohl-Dickstein et al., 2015), diffusion frameworks have emerged as a powerful paradigm for discrete text generation. Austin et al. (2023) introduced `D3PM`, establishing the theoretical foundation for applying diffusion to text, with concurrent work by Hoogeboom et al. (2021) and Campbell et al. (2022) developing discrete and continuous-time formulations. The intersection of diffusion with masked language modeling proved particularly promising. Masked diffusion modeling formulates discrete diffusion as a Markov process with an absorbing state, where tokens replaced by `[MASK]` remain masked in subsequent steps, and the reverse process reconstructs original data from progressively corrupted representations. Sahoo et al. (2024) introduced simplified `MDLMs`, unifying masked language modeling and diffusion through a simplified NELBO expression. This combines bidirectional context benefits with generative training in a unified objective. Similar simplified formulations by Shi et al. (2025) and Ou et al. (2025) demonstrated improved efficiency, with recent work by Sahoo et al. (2025) bridging discrete and Gaussian diffusion for enhanced training techniques.

Masking Strategies for MLMs Several approaches have extended `BERT`'s 15% random token masking (Devlin et al., 2019) with more structured strategies. `SpanBERT` masks contiguous random spans rather than individual tokens and introduces a span boundary objective to predict entire masked spans (Joshi et al., 2020), achieving substantial improvements on span selection tasks. `ELECTRA` replaces tokens with plausible alternatives using a generator-discriminator framework, moving beyond simple masking to token replacement detection (Clark et al., 2020). `RoBERTa` introduces dynamic masking where different tokens are masked across training epochs, in contrast to BERT's static masking approach (Liu et al., 2019). `PMI-Masking` proposes a principled approach based on Pointwise Mutual Information, jointly masking token n-grams with high collocation scores over the corpus (Levine et al., 2020).

Diffusion Models in Data-Sparse Settings Diffusion models face significant challenges when applied to data-constrained image scenarios. Zhu et al. (2022) demonstrated that standard diffusion models suffer from diversity degradation in few-shot settings, leading to overfitting on limited training samples. Wang et al. (2024) identified that image-agnostic Gaussian noise creates uneven adaptation effects and proposed adversarial noise selection for more balanced transfer learning. Lu et al. (2023) showed efficient adaptation through fine-tuning specific attention layers, while Kulikov et al. (2023) explored single-image learning by modeling internal patch distributions. However, these findings focus on vision tasks, leaving diffusion models in data-constrained LM underexplored.

Token Frequency, Weighting and Masking Frequency-based training strategies have emerged to address the imbalance of Zipfian distributions of language tokens. Platanios et al. (2019) demonstrated that curriculum learning based on word frequency can improve sample efficiency in neural machine translation. Bengio et al. (2009) showed that gradually increasing task difficulty, *i.e.*, from

[1] https://github.com/DespoinaKK/babylm-diffusion

frequent to rare tokens, can lead to better convergence and generalization. Importance sampling approaches have been developed to reweight training examples based on token loss (Lin et al., 2024). Recent work has explored adaptive masking strategies that prioritize more salient tokens during training (Choi et al., 2024). However, the application of frequency-based weighting specifically to diffusion models remains unexplored, particularly in data-constrained settings where efficient learning from rare tokens becomes critical.

3 Methodology

3.1 Pretraining

Architecture Our model architecture is a Transformer (Vaswani et al., 2023), based on the `LTG-BERT` model (Samuel et al., 2023), with the attention-gating modifications from (Georges Gabriel Charpentier and Samuel, 2023). To time-condition this model for the diffusion process, we use a timestep embedding and incorporate it with Adaptive Layer Normalization (AdaLN) modulation, following (Peebles and Xie, 2023). This approach enables the model to condition its predictions on the current masking level at diffusion timestep t, allowing it to adapt its behavior across different stages (masking rates) of the diffusion process.

Diffusion Objective Our approach is inspired by both last year's winning `GPT-BERT` method and recent advances in `MDLMs` (Sahoo et al., 2024, Shi et al., 2025). While `GPT-BERT` demonstrates the effectiveness of combining encoding and generative objectives through joint training with MLM and NTP, `MDLMs` results reveal that a single principled diffusion objective can achieve similar dual-purpose training. We adopt the `MDLMs` framework to explore whether this unified approach can be effective in the data-restricted BabyLM setting.

Following (Sahoo et al., 2024), at every training step, a masking rate $1 - \alpha_t$ is sampled from a distribution over $(0, 1)$ for each sequence. Only masked tokens contribute to the cross-entropy loss, and the total objective is a weighted average of MLM losses across different masking levels.

Specifically, in expectation, we optimize the simplified continuous-time NELBO objective from `MDLMs` (Sahoo et al., 2024):

$$\mathcal{L} = \mathbb{E}_q \int_{t=0}^{t=1} \frac{\alpha'_t}{1 - \alpha_t} \sum_{\ell=1}^{L} \log \langle \mathbf{x}_\theta^\ell(\mathbf{Z}_t), \mathbf{x}^\ell \rangle \, dt \quad (1)$$

where α'_t denotes the time derivative of the noise schedule α_t, $\mathbf{Z}_t$ represents the masked sequence at time t, $\mathbf{x}^\ell$ the token at position ℓ, $\mathbf{x}_\theta^\ell(\mathbf{z}_t)$ is the model's prediction at that position, and θ the learnable parameters. This formulation provides a principled objective that naturally weights different masking rates according to the diffusion schedule, and involves maximum-likelihood optimization.

Frequency Informed Masking We propose frequency-informed masking that assigns *higher masking probabilities to rare tokens*. This approach prioritizes learning from infrequent but semantically rich tokens rather than common function words. For a given sequence of tokens $\mathbf{Z} = [\mathbf{x}^1, \ldots, \mathbf{x}^L]$ with a pre-assigned masking rate of $1 - \alpha_t$, we follow *a two-step process* to determine the masking probability for each token. First (step-1), we rank tokens based on their global frequency, with rarer tokens receiving higher ranks. These ranks are min-max normalized to produce initial per-token weights $w^\ell \in (0, 1)$, constructing per-sequence weights $\mathbf{w}$. To prevent an over-emphasis on extremely rare tokens, these weights are "softened" by being raised to a power $p < 1$. Our goal is to scale the weights so that they correspond to the tokens' sampling probability. Next (step-2), we apply conditional scaling to these weights to ensure their mean equals the target probability $1 - \alpha_t$.

$$\mathbf{w}_{\text{new}} = \begin{cases} \mathbf{w}^p \, \frac{1-\alpha_t}{\mu} & \text{if } \mu > 1 - \alpha_t \\ -(1 - \mathbf{w}^p) \frac{\alpha_t}{1-\mu} + 1 & \text{otherwise} \end{cases} \quad (2)$$

Each token $\mathbf{x}^\ell$ is then masked with a probability equal to its new weight, w^ℓ_{new}.

This weighting scheme can be naturally extended to a form of curriculum learning (Bengio et al., 2009) by gradually increasing the softening power p from 0 to a value < 1 across training. This process makes the distribution of masking probabilities sharper over time, which forces the model to progressively focus on predicting rarer and more challenging tokens. We note that frequency is only one option for the relative ranking of tokens. In our proposed `MDLMs` framework, any masking strategy can be *flexibly and seamlessly* incorporated.

3.2 Evaluation

We evaluate our framework using the BabyLM Challenge evaluation pipeline, assessing models across linguistic competence, world knowledge,

human-likeness measures, and standard Natural Language Understanding (NLU) tasks. This suite tests both the quality of learned representations and their alignment with human language acquisition.

Zero-Shot Evaluation We evaluate our models on tasks focusing on linguistic performance and understanding, such as BLiMP (Warstadt et al., 2023b) and Blimp Supplement (Warstadt et al., 2023b). Another linguistic test, targeting grammatical generalization is the Derivational Morphology Test (Hofmann et al., 2024), namely the WUG Adjective Nominalization Test, along with a prior contribution, the WUG Past Tense Test (Weissweiler et al., 2023). EWoK (Ivanova et al., 2025) tests the model's *understanding* of the world, including physical concepts and causal relationships. In a similar minimal pair setting, COMPS (Misra et al., 2023) tests inheritance of properties between hierarchical concepts. Entity Tracking (Kim and Schuster, 2023) tests the model's state tracking abilities. In the zero-shot setting, the goal is for the model to assign higher likelihood to the correct sentence, from a group of sentences.

Finetuning Our pretrained model is further finetuned and evaluated on a subset of GLUE (Wang et al., 2019) and SuperGLUE (Wang et al., 2020), testing NLU.

Human-Likeness Alignment with human acquisition is of special interest when training in developmentally plausible settings. We evaluate on a Reading task using data from (de Varda et al., 2024) and on Age of Acquisition (Chang and Bergen, 2022). The derivational morphology tests (Hofmann et al., 2024), (Weissweiler et al., 2023) provide human annotator data, and the higher model with human correlation is favorable.

Evaluation Backend We use the provided MLM backend to estimate pseudo-likelihoods of sentences (Salazar et al., 2020). MDLM can be evaluated with or without time conditioning. Without time conditioning, we set the masking rate to 0, which corresponds to a fully denoised sequence. With time conditioning, we set the masking rate to $1/L$ for a sequence of length L, which matches the expected masking rate when evaluating one token at a time.

3.3 On the MLM evaluation backend

We argue that for MDLMs, the MLM evaluation backend is a rather myopic view of likelihood estimation, as it only focuses on the very last denoising (unmasking) steps, ignoring previous ones. In theoretical contrast to MLMs, MDLMs are generative language models. For MDLMs, perplexity estimation can be viewed as a Monte-Carlo approximation of the diffusion denoising process (Sahoo et al., 2024).

We suggest that a more appropriate evaluation backend would accommodate for the various possible generation trajectories of the same phrase, and thus provide an estimation better aligned with the native diffusion training objective. This approach would require either exhaustive computation, at the expense of exponential compute time, or Monte-Carlo approximation. The latter is practical for perplexity estimation in large texts, but the accompanying non-determinism proves unsuitable for capturing nuances between similar, short sentences. Nonetheless, for the purposes of the BabyLM Challenge, the MLM pseudo-likelihood estimation, utilized for relatively short sentences, offers the advantage of efficient computation, sufficiently good performance, and determinism.

4 Experiments

We briefly describe the training setup and proceed with a series of experiments, ablations, and evaluations which explore different components of the proposed framework and validate the soundness of our method. First, we test different *noise scheduling* options, *e.g.*, uniform and cosine, naturally motivating our submission's adopted approach. We also include an exploration of experimental unimodal and *bimodal gaussian schedules*, ultimately aiming to design a noise schedule that balances the advantages of AR and MLM approaches. Next, we conduct ablation experiments, establishing the benefits of the proposed frequency informed masking method. Finally, we focus on our submission to the BabyLM Challenge, providing implementation details and the full evaluation results.

4.1 Training Setup

Our architecture follows (Charpentier and Samuel, 2024). We use the same tokenization process and optimizer hyperparameters. The training objective aligns with MDLMs' as in Eq. (1). We train our models for 10 epochs on the BabyLM corpus, with a constant sequence length of 128 for ablation studies, and 512 for the submission model.

SCHEDULE	EWoK (↑)	BLiMP (↑)	BLiMP Sup. (↑)
Eval. w/o Time Conditioning			
Uniform	51.98±0.12	77.91±1.35	67.63±3.64
Cosine	52.44±0.24	79.05±0.28	70.74±1.35
Eval. with Time Conditioning			
Uniform	52.16±0.51	77.55±0.55	67.23±0.98
Cosine	52.39±0.48	78.55±0.70	69.41±0.93

Table 1: Performance comparison across different noise schedules, over 5 random seeds. Reported accuracies are field averages. Likelihoods are estimated with the standard MLM Backend.

4.2 Experiments and Ablations

Noise Schedules Table 1 illustrates a comparison between uniform and cosine masking probability schedules. Additionally, we evaluate them with and without time conditioning. We report the zero-shot results for the four configurations.

With the uniform noise schedule all masking rates are treated equally in the loss calculation, which leads to weak results. The cosine schedule focuses on lower masking rates, with an average masking rate of *0.36* compared to the uniform schedule's *0.5*. Our experiments show that the cosine schedule's lower masking rates consistently improve the model's performance in zero-shot likelihood estimation tasks, as they provide more fine-grained focus.

Gaussian schedules In the context of finding a noise schedule that more effectively unifies the benefits of MLM and AR modeling within the masked diffusion framework, we experiment with unimodal and bimodal Gaussian noise schedules. This means that the distribution of $1 - \alpha_t$ is normal (or a Gaussian mixture) when t is sampled uniformly. Table 2 presents results of a qualitative comparison of training with a unimodal and a bimodal noise schedule with similar expected masking rates across training. *Unimodal*, is a unimodal gaussian masking strategy, with masking rates coming from a $\mathcal{N}(0.3, 0.1)$ distribution. *Bimodal*, is a mixture distribution $w_1\mathcal{N}(\mu_1, \sigma_1^2) + (1-w_1)\mathcal{N}(\mu_2(\tau), \sigma_2^2)$ where the right mode progresses to higher values over time. In this experiment, the left mode has weight $w_1 = 0.6$, mean $\mu_1 = 0.12$, and standard deviation $\sigma_1 = 0.02$. The right mode has time-varying mean $\mu_2(\tau) = 0.4 + (0.85 - 0.4)(1 - e^{-\tau})$ and standard deviation $\sigma_2 = 0.08$, with τ representing the training progress.

The importance of scaling α'_t Table 2 shows that using the full derivative term α'_t ($\gamma = 1.0$) in the NELBO optimization leads to poor zero-shot results. However, performance improves significantly when we scale down the derivatives with a small power of γ or remove them completely ($\gamma = 0.0$). The Unimodal schedule shows modest improvement, while the Bimodal schedule shows dramatic gains, nearly matching top baseline scores when derivatives are softened. These results demonstrate that scaling the derivative term is essential when training with Gaussian schedules.

SCHEDULE (γ)	EWoK (↑)	BLiMP (↑)	BLiMP Sup. (↑)
Unimodal(1.0)	50.24	55.70	51.92
Bimodal(1.0)	51.10	68.13	63.0
Unimodal(0.1)	50.65	64.34	59.32
Bimodal(0.1)	52.46	79.49	72.81
Unimodal(0.0)	50.34	65.34	58.76
Bimodal(0.0)	52.95	78.28	73.13

Table 2: Qualitative performance comparison across different noise schedules. Reported accuracies are field averages. Likelihoods are estimated with the standard MLM Backend. (γ) denotes the softening power for the derivative factor. Results are run over 1 random seed.

Frequency Informed Masking Table 3 compares our method's performance across two distinct configurations:

- *No Frequency Weighting*: A baseline where tokens are masked with equal probabilities.
- *Frequency Weighting* (FW): Our frequency-informed method is applied with a softening power of $p = 0.02$, progressively (linearly) reaching this value across epochs.

We inspect the performance of these configurations on EWoK, BLiMP, and BLiMP Supplement, and report on the accuracy of the Adjective Nominalization test. All models were trained on a cosine noise schedule, with sequence length 128.

The frequency informed masking in general preserves or boosts performance across tasks, **improving performance on BLiMP Sup. by an absolute 1% point** consistently. On the **Adjective Nominalization** test, we observed high variance across random seeds, so we conducted a paired comparison, measuring the accuracy difference between models of different configurations trained with the same seeds. The FW configuration, evaluated with time conditioning, enhances performance, **improving it by an average of 7.5 percentage points**.

CONFIG.	EWoK (↑)	BLiMP (↑)	BLiMP Sup. (↑)
Eval. w/o Time Conditioning			
`Cosine`	52.44±0.24	79.05±0.28	70.74±1.35
`Cosine + FW`	52.63±0.36	78.92±0.34	71.77±0.86
Eval. with Time Conditioning			
`Cosine`	52.39±0.48	78.55±0.70	69.41±0.93
`Cosine + FW`	52.21±0.47	78.90±0.37	70.65±1.87

Table 3: Performance comparison across different token frequency weighting configurations, over 5 random seeds. The FW configuration uses weights softened by raising the frequency distribution to power $p = 0.02$ before scaling. Likelihoods are estimated with the standard MLM Backend.

4.3 Submission Model

Implementation *Training Recipe:* A BPE tokenizer (Gage, 1994) was trained with a vocabulary of 16384 tokens. The submission models have size equal to 126.6 M parameters and were trained with a fixed sequence length of 512. The batch size was set to 512, and sequences were not packed. Documents exceeding this length were divided into independent segments. The total training duration was 10 epochs, or 7530 training steps.

Diffusion Model: For our submission to the leaderboard we employed a cosine masking schedule, with $a_t = cos(\frac{\pi}{2}(1-t))$. Timestep embedding dimension was set to 128. For the frequency informed masking, we used $p = 0.02$, starting from 0 at epoch 0 and linearly reaching p at the last epoch.

Evaluation We provide[2] the submission's internal evaluation results, comparing them with the scores of the baseline with the maximum average score under the name `Baseline-gpt-bert-base-mixed (mntp)`. Zero-shot results were computed evaluating with the standard MLM backend, without time conditioning.

Our model is competitive with the baseline models, particularly in the Finetuning evaluation suite, where it performs especially well on the MRPC and RTE tasks (Table 5). On certain zero-shot evaluation tasks, the model slightly underperforms the top-scoring baseline (*e.g.* BLiMP Sup., EWoK), while it achieves better performance in Entity Tracking (Table 4). In terms of human likeness measures, the submission outperforms the top baseline in Reading and on the Adjective Nominalization Test (Table 6).

[2]We will further update our results with the stronger bimodal gaussian schedule in our code release.

TASK	TOP BASELINE	SUBMISSION†
Linguistics		
BLiMP	80.5	76.9
BLiMP Sup.	73.0	72.4
World Understanding		
EWoK	52.4	51.8
COMPS	59.7	56.4
Entity Tracking	39.9	40.8

Table 4: Evaluation results for Linguistics and World Understanding tasks; †: results refer to cosine schedule

Natural Language Understanding (Finetuning)		
TASK	TOP BASELINE	SUBMISSION†
BoolQ	73.4	72.2
MNLI	63.4	63.8
MRPC	85.8	88.7
MultiRC	69.8	69.0
QQP	81.2	79.2
RTE	59.0	64.7
WSC	63.5	65.4

Table 5: Evaluation results for Natural Language Understanding tasks; †: results refer to cosine schedule

Human Alignment		
TASK	TOP BASELINE	SUBMISSION†
Reading	6.3	7.4
WUG Adj. N.	41.2	49.6
WUG Past T.	27.1	15.4
AoA	22.3	-22.0

Table 6: Evaluation results for Human Likeness tasks; †: results refer to cosine schedule

5 Conclusions

`MDLMs` emerge as a compelling pretraining paradigm for data-constrained LM environments, demonstrating competitive performance against state-of-the-art baselines. Our findings reveal that the choice of masking strategy and its induced objective weighting critically determines model effectiveness. Specifically, we demonstrate that cosine noise schedules yield substantial performance gains over uniform schedules, while bimodal approaches unlock even greater potential, but may require special weighting in the NELBO. Furthermore, we establish a principled framework for integrating intra-token masking strategies within the diffusion paradigm, maintaining theoretical coherence while

expanding practical applicability. These results position masked diffusion as a viable path forward for efficient language model pretraining, particularly valuable when computational resources or training data are limited.

Limitations

This work represents a conceptual integration of MDLMs into the LTG-BERT model family, doing minimal architectural modifications. Standard implementations of MDLMs often incorporate additional optimizations that can substantially impact performance; such optimizations are not explored here. Furthermore, accurately and efficiently estimating likelihoods for zero-shot tasks with short sequences using conventional diffusion approaches while maintaining low variance remains an open challenge. We hypothesize that, while the current MLM-based likelihood estimation approach captures relative trends well, it may be suboptimal, further undermining the MDLMs performance.

Acknowledgments

This work has been supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. We acknowledge EuroHPC JU for awarding the project ID EHPC-AI-2024A04-051 access to the EuroHPC supercomputer LEONARDO hosted by CINECA (Italy).

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2023. Structured denoising diffusion models in discrete state-spaces. *Preprint*, arXiv:2107.03006.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. 2022. A continuous time framework for discrete denoising models. *Preprint*, arXiv:2205.14987.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *Preprint*, arXiv:2502.10645.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Hyesong Choi, Hyejin Park, Kwang Moo Yi, Sungmin Cha, and Dongbo Min. 2024. Salience-based adaptive masking: Revisiting token dynamics for enhanced pre-training. In *European Conference on Computer Vision (ECCV)*, pages 343–359. Springer.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, and 1 others. 2017. Mapping the early language environment using all-day recordings and automated analysis. 26(2):248–265.

Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet Pierrehumbert. 2024. Derivational morphology reveals analogical generalization in large language models. *Preprint*, arXiv:2411.07990.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Preprint*, arXiv:2102.05379.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Preprint*, arXiv:2405.09605.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. 2023. Sinddm: A single image denoising diffusion model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 17920–17930. PMLR.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2020. Pmi-masking: Principled masking of correlated spans. *Preprint*, arXiv:2010.01825.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. 2023. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14267–14276.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2025. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *Preprint*, arXiv:2406.03736.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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 1162–1172.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Preprint*, arXiv:2406.07524.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. 2025. The diffusion duality. *Preprint*, arXiv:2506.10892.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: Bert meets british national corpus. *Preprint*, arXiv:2303.09859.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. 2025. Simplified and generalized masked diffusion for discrete data. *Preprint*, arXiv:2406.04329.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *Preprint*, arXiv:1503.03585.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Xiyu Wang, Baijiong Lin, Daochang Liu, Ying-Cong Chen, and Chang Xu. 2024. Bridging data gaps in diffusion models with adversarial noise-based transfer learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 1–11. PMLR.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2301.11796.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023b. Blimp: The benchmark of linguistic minimal pairs for english. *Preprint*, arXiv:1912.00582.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schütze, Kemal Oflazer, and David R. Mortensen. 2023. Counting the bugs in chatgpt's wugs: A multilingual investigation into the morphological capabilities of a large language model. *Preprint*, arXiv:2310.15113.

Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. 2022. Few-shot image generation with diffusion models. *arXiv preprint arXiv:2211.03264*.