

The Impact of Preference Optimization Data Selection: Teaching Practical Life Skills to BabyLM

Barbora Štěpánková¹² and Vojtěch Lanz¹ and Jakub Szymanik²

¹Institute of Formal and Applied Linguistics, Charles University

²Center for Mind/Brain Sciences, University of Trento

barbora.stepankova320@student.cuni.cz,

lanz@ufal.mff.cuni.cz, jakub.szymanik@unitn.it

Abstract

This work investigates how the choice of preference optimization data affects small language models in the BabyLM setting. We apply SimPO to a GPT-2 model pretrained on 9M words, using 21 different 1M-word preference datasets, including UltraFeedback, which we cluster by topic, grammatical, and Reddit-based data. We show that preference datasets of the same size lead to different downstream performance. A combination of Business Content Creation, Food and Cooking, and Math and Calculations, which we refer to as the Practical Life Skills dataset, consistently improves Entity Tracking, MultiRC, and WSC tasks. The Entity Tracking gains are largely related to the model’s ability to handle cases where “nothing.” is a valid option, raising the question of whether models’ ability to abstain from selecting an answer in such cases or explicitly express uncertainty should receive more attention. Overall, we show that under strict data constraints, allocating the final 1M words to a different training objective can provide larger benefits than using them for additional pretraining.

1 Introduction

Approaches to training language models are increasingly inspired by principles of human learning and cognition (Apidianaki et al., 2024). The BabyLM initiative is one example of this trend, investigating the training process of language models under developmentally motivated constraints inspired by human language acquisition (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025b). Language models are oftentimes pretrained by predicting the next token from large collections of text. While this objective has proven highly effective, it captures only one aspect of human language acquisition. Children do not learn language only by passive exposure, but also

through communicative and social interactions (Tomasello, 2005; Kuhl, 2007). These interactions may accelerate the child’s learning by presenting known structures in surprising context (Fazekas et al., 2020), providing them with adult reformulations in a form of recasts (Chouinard and Clark, 2003), or other social signals that indicate which behaviors are more appropriate in a given context (Nikolaus and Fourtassi, 2023). The Competition Model (Macwhinney, 1987) argues that language understanding emerges through competition between alternative linguistic cues, whose relative importance is gradually shaped by experience

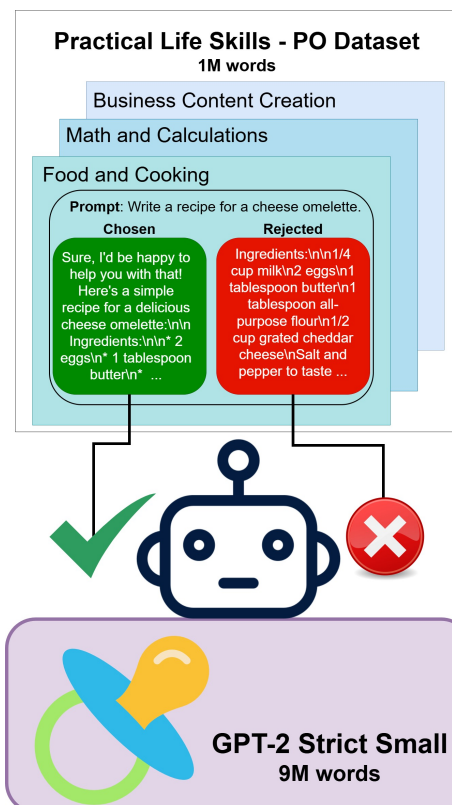


Figure 1: Training pipeline of our Practical Life Skills model. The 1M-word PLS dataset is used for preference optimization of a GPT-2 model pretrained on 9M words.

and communicative feedback. [Young et al. \(2012\)](#) show that prompting children to consider alternative explanations can lead the children to drawing stronger, more accurate inferences from the same evidence. More generally, developmental theories of learning suggest that children actively generate and evaluate alternative explanations, gradually revising their beliefs as new evidence becomes available ([Gopnik, 2012](#); [Bonawitz and Lombrozo, 2012](#); [Walker et al., 2017](#)).

Preference optimization introduces an analogous comparison-based learning signal for language models ([Rafailov et al., 2023](#); [Meng et al., 2024](#)). During training, the model compares alternative responses: rather than learning from a single target output, the model is trained to assign higher probability to the preferred response and lower probability to the less preferred.

While preference optimization has shown promising results, it remains unclear which types of preference data are effective and for which specific tasks. Different preference datasets emphasize different aspects of response quality, such as factuality, helpfulness, reasoning, or creativity ([Cui et al., 2024](#)), and each preference dataset defines a different notions of what a “good answer” is. We cluster existing preference data into coherent categories and further extend them with an additional synthetic grammatical category created by us and a separate category derived from Reddit’s *Explain Like I’m Five* discussions ([Ethayarajh et al., 2022](#)). We then investigate how different preference data categories affect language learning in the BabyLM-strict-small Challenge setting, where models are trained with a limited 10M-word budget.

We show that preference optimization on 1M words, despite being applied after only a small 9M-word pretraining phase, can improve model performance on some tasks. However, the choice of preference data category can have a substantial impact on the resulting capabilities. We find that consistent improvements are achieved by the *Business Content Creation*, *Food and Cooking*, and *Math and Calculations* categories, primarily on the Entity Tracking task ([Kim and Schuster, 2023](#)), and coreference resolution (WSC) and MultiRC subtasks of the SuperGLUE benchmark ([Wang et al., 2019](#)). Based on these findings, we propose a new Practical Life Skills (PLS) Preference Optimization dataset by combining these three categories and randomly sampling preference pairs to satisfy the remaining 1M-word constraint of the BabyLM

Strict-Small setting (see Figure 1).

Compared to full pretraining on 10M words without preference optimization, PLS doubles the Entity Tracking score and increases WSC and MultiRC performance, while not influencing the other tasks much. Interestingly, when removing all Entity Tracking examples where the model can answer “nothing.” from the available options, the improvement from PLS disappears. This suggests that a major contribution of PLS is improving the model’s ability to recognize when none of the provided options applies. We argue that this capability is an important aspect of language understanding that should receive more attention in future BabyLM evaluations: models should not only select among predefined options, but also recognize when the correct response is, for instance, “there is nothing”, “I do not want any of these”, or “I do not know”.

2 Related Work

Preference optimization has become a common approach for adapting language models after pretraining, with its usage rapidly expanding into diverse alternative directions ([Park et al., 2023](#); [Lee et al., 2024](#); [Lanz et al., 2026](#)). While there has been work carried out comparing the various preference optimization algorithms ([Xiao et al., 2026](#); [Meng et al., 2024](#)), relatively little attention has been paid to the role of the preference data itself. Recent work has shown that preference tuning is sensitive to domain shift, suggesting that the domain of the preference data can affect downstream performance ([Karouzos et al., 2026](#)). Our work complements these findings by systematically studying how different preference datasets influence linguistic and reasoning abilities in data-constrained BabyLM models while keeping the optimization algorithm fixed.

Several recent BabyLM papers have explored model alignment methods. [Zhao et al. \(2023\)](#) applied reinforcement learning from human feedback to improve story generation, while [Padovani et al. \(2025\)](#) investigated dialogue-based learning together with preference optimization and reinforcement learning to improve communicative abilities. Recently, [Charpentier et al. \(2025a\)](#) proposed an interactive training framework combining teacher corrections with SimPO ([Meng et al., 2024](#)), and [Mayer Martins et al. \(2025\)](#) explored interactive feedback for storytelling.

Dataset	All Examples	Sampled Examples	Avg Prompt	Avg Response	Jaccard
Grammatical	23403	23403	0.00	21.36	0.81
r/explainlikeimfive	2484	2484	38.95	159.51	0.10
Algorithms and Programming	2152	2040	90.70	154.31	0.17
Business and Marketing	2956	1322	113.80	264.26	0.20
Business Content Creation	2683	1320	84.24	294.32	0.18
Creative Storytelling	2529	1280	99.83	290.51	0.14
Creative Writing and Arts	1895	1812	81.97	193.93	0.16
Environment Sustainability	1852	1274	85.81	306.65	0.19
Food and Cooking	1881	1571	104.02	214.23	0.20
General Knowledge Questions	3252	2654	39.85	148.54	0.18
Health and Wellbeing	2008	1396	107.29	250.68	0.19
Language Understanding Tasks	4050	2188	171.86	56.65	0.14
Math and Calculations	2413	1716	84.46	206.86	0.19
Multilingual Translation Tasks	5461	3216	93.92	61.52	0.11
Natural Language Inference	4262	1631	98.30	208.20	0.19
QA and Reasoning Tasks	2806	2163	143.84	87.15	0.16
Reading Comprehension Tasks	3388	1281	273.03	117.02	0.19
Software Engineering	7815	1609	80.72	229.83	0.19
Technology and AI Concepts	3054	1408	73.15	281.94	0.19
Text Classification Tasks	4587	2238	113.48	109.83	0.15
Writing and Essays	2091	1172	90.56	336.04	0.18

Table 1: Preference dataset statistics. *All Examples* refers to the total number of preference pairs before sampling, while *Sampled Examples* denotes the number of pairs used for training after applying the 1M-word budget. *Avg Prompt* and *Avg Response* report the average number of words in the prompt and in each response, respectively. *Jaccard* measures the word-level similarity between the chosen and rejected responses (0 = no overlap, 1 = identical).

3 Data

For pretraining, we use a randomly sampled subset of the provided BabyLM-Strict-Small dataset (Choshen et al., 2026). For preference optimization, we construct several preference datasets that differ in the type of preference signal they provide, and the resulting statistics are reported in Table 1.

Ultrafeedback Subsets UltraFeedback¹ is a large-scale preference dataset containing instruction-response pairs covering a broad range of topics and tasks (Cui et al., 2024). We use this dataset to investigate whether the topic of preference data influences downstream model abilities.

To create preference subsets covering individual task types and topics, we cluster examples based on their prompts. Each prompt is encoded using the all-MiniLM-L6-v2 sentence transformer (Reimers and Gurevych, 2019) and clustered using the mini-batch k-means algorithm (Sculley, 2010). We set $k = 25$ to obtain diverse clusters while maintaining sufficient data size. After inspecting the resulting clusters, we merge underrepresented and semantically overlapping clusters with their

closest counterparts. This results in 19 final preference subsets. Cluster names were assigned by GPT-5.5-mini (OpenAI, 2026) based on sampled prompts from each cluster and are used only for readability and do not affect training.

The resulting subsets differ in topic, as well as in average prompt and response length, where the prompt length ranges from 39 words on average in *General Knowledge Questions* to 273 words per prompt for *Reading Comprehension Tasks*.

Synthetic Grammatical Preference Dataset Existing preference datasets primarily focus on instruction following and helpfulness, while providing relatively little supervision targeted specifically at grammatical correctness. We therefore construct a synthetic preference dataset designed to provide an explicit grammatical preference signal.

We sample passages from BookSum (Kryściński et al., 2021) and create corrupted versions using one of three randomly selected operations: insertion, deletion, or swapping of 1–5 words. The original passage is used as the preferred response and the corrupted version as the dispreferred response. The prompt is left empty.

This corruption-based construction follows previous work showing that synthetic perturbations

¹https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

can improve performance on grammaticality-related benchmarks (Wang et al., 2026). While the generated errors are artificial, they provide a controlled comparison between linguistically plausible and corrupted alternatives.

The dataset contains over 23 thousand short preference pairs. Compared to other datasets, the chosen and rejected responses are highly similar, with an average Jaccard similarity of 0.81.

r/ExplainLikeImFive We additionally use a subset of the Stanford Human Preferences Dataset (Ethayarajh et al., 2022), which contains naturally occurring Reddit questions paired with human preference judgments over responses.

We select examples from the *Explain Like I’m Five* subreddit because they provide preference data for explanations intended to be simple and understandable. Unlike synthetic preference datasets, these examples reflect naturally occurring human judgments about response quality.

We keep only examples where the preferred response received at least twenty times more upvotes than the rejected response. We also remove duplicate responses appearing in multiple preference pairs. Both responses remain plausible answers to the same question, requiring the model to learn which type of explanation is preferred rather than simply distinguishing correct from incorrect answers.

4 Experimental Setup

In our experiments, we first train a small language model on a limited BabyLM corpus (9M words) and then apply preference optimization using different preference datasets (1M words, counting the chosen response, rejected response, and the prompt twice for each preference pair) described in Section 3.² To adhere to the BabyLM Strict-Small challenge rules, the total amount of training data seen by the model is limited to 10M words.

To evaluate the impact of different preference optimization data, we account for training variability by using five random seeds. Each preference optimization dataset is applied to all five independently trained base models, and the reported results are averaged across the resulting runs.

Base Model and Pretraining Following Tapaninaho (2025), we use a compact GPT-2-style

decoder-only architecture (Radford et al., 2019) designed for efficient small-scale language modeling under limited data and compute budgets. Rather than adopting the larger GPT-2 configuration from the BabyLM shared-task baseline, we use this reduced configuration to enable efficient experimentation while maintaining the core GPT-2 architecture and similar performance (Tapaninaho, 2025). The model is initialized from scratch and consists of 12 Transformer layers with a hidden size of 384 and 6 attention heads.

The model is trained on a randomly sampled subset of the BabyLM-Strict-Small dataset containing 9M words. We train a Byte-Level BPE tokenizer from the training corpus with a vocabulary size of 16,384 tokens. Training uses the AdamW optimizer with a learning rate of 3×10^{-4} , weight decay of 0.1, cosine learning-rate decay, and 800 warmup steps. The batch size is 16, and training is performed for 10 epochs.

Preference Optimization After pretraining on the 9M words, we apply SimPO (Meng et al., 2024) as the preference optimization method. SimPO directly optimizes preference pairs without requiring a separately trained reward model, making it suitable for controlled comparisons across multiple preference datasets.

For each preference dataset, we randomly sample preference pairs without replacement until reaching 1M words. Both prompts and responses are included in this word count, although the model weights are updated only based on the response. All datasets are trained with identical SimPO hyperparameters: 10 epochs, effective batch size of 16 (batch size 4 with gradient accumulation over 4 steps), learning rate of 5×10^{-6} , and maximum sequence length of 512 tokens. The maximum prompt length is limited to 256 tokens. We use the SimPO loss with $\beta = 2.0$ and target margin $\gamma = 0.5$.

5 Results

We compare the 9M-word pretrained GPT-2 base model with the same model after SimPO training on different 1M-word preference datasets. Results are averaged across five random seeds, with standard deviations reported in the tables. Results on zero-shot benchmarks are shown in Table 2, while results on fine-tuning benchmarks are reported in Table 3.

²https://github.com/stepankovab/BabyLM_2026

	BLiMP	BLiMP Supplement	COMPS	ET	ET w/o nothing.	EWOK
Base (9M words)	67.57 $\pm$ 0.31	54.28 $\pm$ 1.10	51.51 $\pm$ 0.18	21.84 $\pm$ 5.21	19.26 $\pm$ 0.15	50.79 $\pm$ 0.87
Grammatical	63.12 $\pm$ 0.72	56.28 $\pm$ 1.26	50.38 $\pm$ 0.18	31.40 $\pm$ 10.54	17.22 $\pm$ 0.40	50.25 $\pm$ 0.15
r/explainlikeimfive	67.12 $\pm$ 0.48	52.11 $\pm$ 1.92	50.94 $\pm$ 0.27	29.40 $\pm$ 3.94	18.43 $\pm$ 0.69	50.00 $\pm$ 0.81
Algorithms and Programming	66.11 $\pm$ 0.39	55.15 $\pm$ 1.49	51.52 $\pm$ 0.33	34.05 $\pm$ 9.94	18.48 $\pm$ 0.37	50.36 $\pm$ 0.52
Business and Marketing	67.79 $\pm$ 0.29	54.41 $\pm$ 2.01	51.12 $\pm$ 0.22	34.14 $\pm$ 4.92	19.18 $\pm$ 0.23	50.43 $\pm$ 0.77
Business Content Creation	67.45 $\pm$ 0.41	55.38 $\pm$ 2.00	50.99 $\pm$ 0.33	40.68 $\pm$ 1.28	18.65 $\pm$ 0.42	50.19 $\pm$ 0.54
Creative Storytelling	66.83 $\pm$ 1.08	54.59 $\pm$ 1.12	51.01 $\pm$ 0.40	34.42 $\pm$ 8.03	18.58 $\pm$ 0.73	50.57 $\pm$ 0.32
Creative Writing and Arts	65.94 $\pm$ 0.97	53.89 $\pm$ 0.62	51.00 $\pm$ 0.50	35.62 $\pm$ 4.89	17.75 $\pm$ 0.56	50.59 $\pm$ 0.34
Environment Sustainability	65.74 $\pm$ 0.40	55.81 $\pm$ 1.70	51.47 $\pm$ 0.30	33.95 $\pm$ 7.89	18.32 $\pm$ 0.50	50.28 $\pm$ 0.73
Food and Cooking	67.07 $\pm$ 0.45	54.77 $\pm$ 1.75	51.36 $\pm$ 0.29	40.06 $\pm$ 2.88	18.28 $\pm$ 0.79	50.00 $\pm$ 0.77
General Knowledge Questions	65.51 $\pm$ 1.11	54.29 $\pm$ 1.56	51.13 $\pm$ 0.22	37.47 $\pm$ 3.28	17.76 $\pm$ 0.48	50.12 $\pm$ 0.41
Health and Wellbeing	66.96 $\pm$ 0.31	55.02 $\pm$ 1.44	51.23 $\pm$ 0.41	34.46 $\pm$ 6.54	18.57 $\pm$ 0.73	50.31 $\pm$ 0.63
Language Understanding Tasks	66.55 $\pm$ 0.35	54.82 $\pm$ 0.68	51.55 $\pm$ 0.25	36.70 $\pm$ 8.22	18.36 $\pm$ 0.26	50.16 $\pm$ 0.69
Math and Calculations	67.26 $\pm$ 0.50	54.84 $\pm$ 1.21	51.14 $\pm$ 0.30	40.45 $\pm$ 2.60	18.63 $\pm$ 0.56	50.05 $\pm$ 0.22
Multilingual Translation Tasks	66.73 $\pm$ 0.42	55.89 $\pm$ 0.76	51.36 $\pm$ 0.23	37.89 $\pm$ 4.75	19.05 $\pm$ 0.49	50.01 $\pm$ 0.95
Natural Language Inference	67.88 $\pm$ 0.51	55.63 $\pm$ 1.74	51.39 $\pm$ 0.30	37.95 $\pm$ 5.64	18.78 $\pm$ 0.67	50.60 $\pm$ 0.21
QA and Reasoning Tasks	66.44 $\pm$ 0.49	53.72 $\pm$ 1.00	51.76 $\pm$ 0.14	37.87 $\pm$ 6.20	18.12 $\pm$ 0.42	50.20 $\pm$ 0.56
Reading Comprehension Tasks	66.37 $\pm$ 0.53	54.51 $\pm$ 1.14	51.42 $\pm$ 0.25	34.19 $\pm$ 10.64	18.79 $\pm$ 0.82	50.54 $\pm$ 0.43
Software Engineering	66.83 $\pm$ 0.40	54.19 $\pm$ 1.36	51.27 $\pm$ 0.42	39.62 $\pm$ 2.75	18.58 $\pm$ 0.80	50.30 $\pm$ 0.65
Technology and AI Concepts	67.03 $\pm$ 0.29	56.47 $\pm$ 1.70	51.00 $\pm$ 0.40	33.99 $\pm$ 6.29	18.71 $\pm$ 0.48	50.46 $\pm$ 0.51
Text Classification Tasks	65.26 $\pm$ 0.95	53.47 $\pm$ 0.81	51.56 $\pm$ 0.34	27.92 $\pm$ 11.95	18.29 $\pm$ 0.71	50.44 $\pm$ 0.46
Writing and Essays	66.13 $\pm$ 1.29	53.92 $\pm$ 1.21	51.34 $\pm$ 0.21	28.05 $\pm$ 4.07	17.75 $\pm$ 0.60	50.20 $\pm$ 0.67

Table 2: Zero-shot benchmark performance after preference optimization with different datasets. Values represent mean performance across five runs with standard deviation.

Zero-Shot Benchmarks Most zero-shot benchmarks remain close to the pretrained baseline regardless of the preference optimization dataset.

On BLiMP (Warstadt et al., 2020), most datasets change the overall score by less than one point. The main exception is the *Grammatical* subset, which decreases the score from 67.57 to 63.12. Interestingly, the opposite trend is observed on BLiMP Supplement, where the same subset improves performance from 54.28 to 56.28. COMPS (Misra et al., 2023) and EWOK (Ivanova et al., 2025) remain nearly unchanged across all evaluated datasets.

The largest differences are observed on the Entity Tracking (ET) benchmark. The pretrained baseline achieves 21.84 points, while almost every preference optimization dataset improves this score by at least six points. The largest improvements are obtained using the *Business Content Creation*, *Food and Cooking*, and *Math and Calculations* subsets, all exceeding 40 points. We also observe larger standard deviations for lower-performing datasets, whereas the highest-scoring datasets remain consistently strong across all five runs.

The Entity Tracking benchmark contains a substantial number of instances where the option “nothing.” is available among the possible answers (2703 out of 9483 instances). This allows the model to indicate that none of the provided entities applies, reducing potential hallucinations caused by being forced to select an entity when no valid option

is available. To investigate whether the observed improvements are primarily due to learning this ability or instead reflect improved entity tracking among valid options, we additionally evaluate models on a filtered version of the benchmark where all instances containing the “nothing.” option are removed. This results in the *ET w/o nothing* setting reported in Table 2. On this version of the benchmark, the improvements observed on the original Entity Tracking task disappear, with performance becoming comparable to or slightly below the pretrained baseline.

Fine-Tuning Benchmarks Most of the fine-tuning benchmarks also show relatively small differences across preference optimization datasets.

Performance on BoolQ, MNLI, MRPC and QQP (Wang et al., 2019) remains close to the pretrained baseline for nearly all datasets. MultiRC is similarly stable overall, although the *r/explainlikeimfive*, *Math and Calculations* and *Food and Cooking* datasets improve the average score by more than two points.

RTE performance decreases slightly for most datasets, typically by around one percentage point compared to the pretrained baseline. In contrast, WSC generally improves after preference optimization, with most datasets gaining between two and three points. The largest improvement is achieved by the *General Knowledge Questions* subset, which improves the average score by almost five points.

	BoolQ (acc)	MNLI (acc)	MRPC (f1)	MultiRC (acc)	QQP (f1)	RTE (acc)	WSC (acc)
Base (9M words)	66.92 $\pm$ 0.84	50.30 $\pm$ 0.93	83.02 $\pm$ 0.41	60.29 $\pm$ 2.40	62.25 $\pm$ 0.93	59.42 $\pm$ 1.09	62.69 $\pm$ 2.92
Grammatical	67.21 $\pm$ 0.66	50.38 $\pm$ 0.54	82.73 $\pm$ 0.19	59.44 $\pm$ 1.92	61.83 $\pm$ 0.33	57.84 $\pm$ 2.47	66.54 $\pm$ 4.21
r/explainlikeimfive	66.96 $\pm$ 0.58	50.74 $\pm$ 0.47	82.61 $\pm$ 0.20	62.52 $\pm$ 3.16	62.08 $\pm$ 0.71	58.56 $\pm$ 2.25	63.46 $\pm$ 1.36
Algorithms and Programming	67.74 $\pm$ 0.85	50.47 $\pm$ 0.59	83.06 $\pm$ 0.50	61.59 $\pm$ 3.21	62.37 $\pm$ 0.54	58.13 $\pm$ 2.56	65.38 $\pm$ 4.51
Business and Marketing	66.66 $\pm$ 0.92	50.63 $\pm$ 0.73	82.81 $\pm$ 0.61	60.49 $\pm$ 3.08	62.17 $\pm$ 1.08	56.12 $\pm$ 1.13	63.08 $\pm$ 2.50
Business Content Creation	67.47 $\pm$ 0.74	50.44 $\pm$ 0.61	82.96 $\pm$ 0.30	60.98 $\pm$ 2.27	61.96 $\pm$ 0.82	58.42 $\pm$ 3.43	64.23 $\pm$ 2.19
Creative Storytelling	67.08 $\pm$ 1.17	50.40 $\pm$ 0.37	82.92 $\pm$ 0.41	61.69 $\pm$ 2.79	62.50 $\pm$ 0.39	59.13 $\pm$ 1.56	65.00 $\pm$ 3.94
Creative Writing and Arts	67.01 $\pm$ 0.50	50.24 $\pm$ 0.60	82.72 $\pm$ 0.38	61.59 $\pm$ 2.40	62.27 $\pm$ 0.76	60.29 $\pm$ 2.12	65.00 $\pm$ 5.16
Environment Sustainability	66.80 $\pm$ 0.71	49.99 $\pm$ 0.57	82.87 $\pm$ 0.60	59.95 $\pm$ 2.00	61.90 $\pm$ 0.54	57.84 $\pm$ 1.30	64.23 $\pm$ 1.72
Food and Cooking	66.53 $\pm$ 0.93	50.49 $\pm$ 0.63	82.78 $\pm$ 0.46	62.69 $\pm$ 3.39	61.97 $\pm$ 0.86	57.98 $\pm$ 2.99	64.61 $\pm$ 2.58
General Knowledge Questions	66.80 $\pm$ 1.04	50.58 $\pm$ 0.16	82.59 $\pm$ 0.40	61.24 $\pm$ 3.57	62.06 $\pm$ 0.76	58.13 $\pm$ 2.18	67.31 $\pm$ 4.90
Health and Wellbeing	66.39 $\pm$ 0.57	50.57 $\pm$ 0.39	82.53 $\pm$ 0.40	60.85 $\pm$ 2.69	62.15 $\pm$ 0.51	58.42 $\pm$ 2.98	65.00 $\pm$ 2.85
Language Understanding Tasks	66.35 $\pm$ 1.25	50.22 $\pm$ 0.60	82.45 $\pm$ 0.22	60.44 $\pm$ 2.80	62.11 $\pm$ 0.45	59.71 $\pm$ 3.94	66.15 $\pm$ 3.22
Math and Calculations	66.71 $\pm$ 0.87	50.23 $\pm$ 0.43	83.13 $\pm$ 0.65	62.94 $\pm$ 2.97	62.20 $\pm$ 0.81	58.56 $\pm$ 0.97	65.00 $\pm$ 3.16
Multilingual Translation Tasks	66.16 $\pm$ 0.73	49.89 $\pm$ 0.37	82.16 $\pm$ 0.44	60.24 $\pm$ 3.37	61.64 $\pm$ 0.79	57.41 $\pm$ 1.71	65.00 $\pm$ 3.16
Natural Language Inference	66.79 $\pm$ 1.25	50.40 $\pm$ 0.68	82.77 $\pm$ 0.37	60.77 $\pm$ 3.06	62.20 $\pm$ 0.24	58.70 $\pm$ 1.31	64.23 $\pm$ 1.72
QA and Reasoning Tasks	66.80 $\pm$ 0.59	50.35 $\pm$ 0.23	82.63 $\pm$ 0.83	59.57 $\pm$ 2.92	62.22 $\pm$ 0.45	58.56 $\pm$ 1.31	65.77 $\pm$ 3.94
Reading Comprehension Tasks	66.57 $\pm$ 1.09	50.37 $\pm$ 0.47	82.22 $\pm$ 0.44	58.63 $\pm$ 0.23	62.56 $\pm$ 0.53	57.70 $\pm$ 3.15	64.23 $\pm$ 2.58
Software Engineering	65.78 $\pm$ 0.54	50.32 $\pm$ 0.21	82.67 $\pm$ 0.87	59.73 $\pm$ 2.95	62.33 $\pm$ 0.36	57.12 $\pm$ 1.49	64.23 $\pm$ 2.19
Technology and AI Concepts	67.31 $\pm$ 0.56	50.57 $\pm$ 0.57	82.94 $\pm$ 0.29	59.93 $\pm$ 2.85	62.22 $\pm$ 0.28	57.98 $\pm$ 1.49	63.46 $\pm$ 2.35
Text Classification Tasks	66.77 $\pm$ 0.81	50.10 $\pm$ 0.23	82.60 $\pm$ 0.49	60.63 $\pm$ 2.16	62.48 $\pm$ 0.42	57.84 $\pm$ 0.82	63.85 $\pm$ 2.85
Writing and Essays	66.55 $\pm$ 0.93	50.27 $\pm$ 0.65	82.73 $\pm$ 0.26	59.77 $\pm$ 1.03	62.02 $\pm$ 0.49	61.15 $\pm$ 2.55	66.15 $\pm$ 2.19

Table 3: Fine-tuning benchmark performance after preference optimization with different datasets. Values represent mean performance across five runs with standard deviation.

Benchmark	Baseline	10M	9M	PLS
BLiMP	65.22	67.98	67.85	66.86
BLiMP Sup.	57.25	55.85	53.93	55.50
COMPS	51.81	51.18	51.38	51.57
ET	21.07	20.69	21.87	41.92
ET w/o nothing.	19.10	19.79	19.20	18.97
EWOK	50.63	50.23	50.71	50.39
BoolQ	68.26	67.28	67.58	65.87
MNLI	50.08	50.26	49.74	50.12
MRPC	81.93	82.53	82.74	82.39
MultiRC	65.29	58.50	58.00	65.35
QQP	62.60	63.16	61.07	62.23
RTE	58.27	56.83	59.71	58.27
WSC	61.54	61.54	61.54	65.38

Table 4: Performance of GPT-2-based models across zero-shot and fine-tuning benchmarks, including the official BabyLM-2026-Baseline-GPT2-Strict-Small baseline, 9M- and 10M-word pretrained variants, and our PLS model.

5.1 Final Submission

Based on the benchmark results above, we selected three datasets that performed consistently well across the evaluated benchmarks: *Business Content Creation*, *Food and Cooking*, and *Math and Calculations*. We refer to this combination as the *Practical Life Skills* dataset. We trained our final submission model³ using the PLS dataset and the same SimPO procedure.

Table 4 compares the resulting PLS model with

³<https://huggingface.co/stepanb2/babylm-gpt2-simpo-pls-10m>

our GPT-2 model pretrained on 9M words, the same model pretrained on the full 10M-word BabyLM-Strict-Small dataset, and the official BabyLM-2026-Baseline-GPT2-Strict-Small baseline (Choshen et al., 2026).

Overall, the three pretrained models achieve comparable performance across most benchmarks. In contrast, the PLS model doubles the Entity Tracking score on the original benchmark, while also achieving the highest scores on MultiRC and WSC, increasing WSC performance by 4 points, and MultiRC performance by 7 points compared to both the 9M and 10M-word pre-trained base model. Performance on the remaining benchmarks is almost unchanged or only slightly decreased.

6 Discussion

Small pretraining improvements The 9M-word and 10M-word pretrained models achieve comparable performance, showing that the additional 1M words used for pretraining provide only limited and inconsistent gains across downstream benchmarks, such as small improvements on some tasks (e.g., BLiMP Supplement, ET without “nothing.”, MNLI, and QQP) and decreases on others (e.g., Entity Tracking, EWOK, BoolQ, and RTE). In contrast, using the same 1M-word budget for preference optimization leads to larger improvements, particularly on Entity Tracking, MultiRC, and WSC. This suggests that under strict data constraints, allocating part of the training budget to a different optimiza-

tion objective can be more effective than using the entire budget for additional pretraining.

Dataset content matters Our results suggest that the topic of the preference optimization data matters, although its effect depends strongly on the downstream task. While most evaluated benchmarks remain stable across different preference datasets, some datasets consistently perform above or below average. In particular, the *Food and Cooking* or *Math and Calculations* datasets repeatedly rank among the strongest performers, while datasets such as *Multilingual Translation Tasks*, *Writing and Essays*, and *Text Classification Tasks* consistently achieve lower scores.

We also investigated whether the observed differences could be explained by simple dataset statistics (Table 1). We found a small number of statistically significant correlations, for example between average prompt length and COMPS performance (Spearman $\rho = 0.62$, $p = 0.003$), and between average response length and MRPC (Spearman $\rho = 0.55$, $p = 0.009$). However, these benchmarks vary by less than one point across almost all preference datasets, suggesting that these correlations do not have a substantial impact on downstream performance. Overall, these correlations are unlikely to explain the main differences observed in our experiments. In particular, we found no meaningful relationship between the examined dataset statistics and performance on Entity Tracking, MultiRC, or WSC, where the largest gains were observed.

Practical Life Skills The three datasets selected for our final submission all involve solving practical procedural tasks. Recipes require keeping track of ingredients throughout a sequence of actions, mathematical problems require tracking variables during a computation, and business communication frequently describes requests, documents, or workflows whose state changes over time. While this observation is only qualitative, it motivated our decision to combine these datasets into the *Practical Life Skills* dataset.

Entity Tracking The most striking result of our experiments is the difference between the original Entity Tracking benchmark and its filtered version without instances containing the “nothing.” option. On the original benchmark, the PLS model almost doubles the Entity Tracking score compared to the pretrained baseline. However, this improvement almost completely disappears once examples

whose possible option is “nothing.” are removed. This suggests that the observed gains are primarily driven by the model’s ability to recognize when none of the provided entities applies, reducing hallucinations caused by being forced to select one of the available options, rather than by a general improvement in entity tracking. Since the evaluation is performed in a zero-shot setting, this reflects the model’s ability to recognize such cases rather than overfitting to the benchmark.

One possible explanation is that practical procedural domains naturally contain many situations where objects, ingredients, resources, or quantities become unavailable or are no longer present. Preference optimization on such examples may therefore encourage the model to predict the absence of an entity when appropriate instead of selecting one of the remaining alternatives.

More broadly, this observation highlights an aspect of language understanding that receives relatively little attention in current benchmark design: recognizing when none of the candidate answers applies. In addition to selecting the correct answer, language models should also learn to recognize situations where the correct response is “there is nothing”, “none of the above”, or “I do not know”. We believe future BabyLM evaluations could benefit from including more tasks of this kind.

7 Conclusion

In this work, we studied how the choice of preference optimization data affects small language models in the BabyLM setting. Despite the model being pretrained on only a very limited amount of data, applying preference optimization with a 1M-word dataset led to improvements on several downstream tasks. Moreover, these improvements were larger than those obtained by using the same 1M-word budget for additional pretraining.

We found that preference datasets of the same size can lead to different performance, showing that the content of preference data plays an important role. The *Business Content Creation*, *Food and Cooking*, and *Math and Calculations* datasets consistently performed well and improved several benchmarks when used as a 1M-word preference optimization dataset. We combined these three datasets into a single 1M-word *Practical Life Skills* dataset, which improved performance on MultiRC and WSC, as well as the model’s ability to handle cases where “nothing.” is available as a possible

answer in Entity Tracking. This raises a question for future work: whether the ability of models to abstain from selecting an answer when none of the options applies, or to explicitly state uncertainty (e.g., “I don’t know”), should receive more attention in future evaluations.

Limitations

Our study has several limitations. First, we evaluate only SimPO, leaving open whether the same observations hold for other preference optimization methods. Second, all experiments are conducted in the BabyLM Strict-Small setting using a GPT-2 architecture, and the findings may not directly transfer to larger models. An important direction for future work is investigating whether these findings scale to the Strict (100M) setting and whether post-training a larger model on the same 1M-word corpus leads to similar observations. Finally, our interpretation of the Entity Tracking results is based on a single benchmark. Evaluating the models on additional benchmarks that require recognizing the absence of a correct answer, uncertainty, or hallucinations would help determine whether the observed behavior generalizes beyond Entity Tracking.

Acknowledgments

This research was partially supported by the SVV project number 260 821.

References

- Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. 2024. [Language learning, representation, and processing in humans and machines: Introduction to the special issue](#). *Computational Linguistics*, 50(4):1201–1210.
- Elizabeth Baraff Bonawitz and Tania Lombrozo. 2012. [Occam’s rattle: Children’s use of simplicity and probability to constrain inference](#). *Developmental Psychology*, 48(4):1156–1164.
- 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. 2025a. [BabyLM turns 3: Call for papers for the 2025 babyLM workshop](#). *Preprint*, arXiv:2502.10645.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gottlieb Wilcox, and Adina Williams. 2025b. [Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data](#). In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gottlieb Wilcox. 2026. [BabyLM turns 4 and goes multilingual: Call for papers for the 2026 babyLM workshop](#). *Preprint*, arXiv:2602.20092.
- Michelle M. Chouinard and Eve V. Clark. 2003. [Adult reformulations of child errors as negative evidence](#). *Journal of Child Language*, 30(3):637–669.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Qiang Yue, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2024. [Ultrafeedback: Boosting language models with high-quality feedback](#).
- Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. [Understanding dataset difficulty with V-usable information](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.
- Judit Fazekas, Andrew Jessop, Julian Pine, and Caroline Rowland. 2020. [Do children learn from their prediction mistakes? a registered report evaluating error-based theories of language acquisition](#). *Royal Society Open Science*, 7(11):180877.
- Alison Gopnik. 2012. [Scientific thinking in young children: Theoretical advances, empirical research, and policy implications](#). *Science*, 337(6102):1623–1627.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gottlieb Wilcox. 2024. [Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.
- Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U. Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian C. Paulun, Maria Ryskina, Ekin Akyürek, Ethan G. Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. [Elements of world knowledge \(EWOKE\): A cognition-inspired framework for evaluating basic world knowledge in language models](#). *Transactions of the Association for Computational Linguistics*, 13:1245–1270.
- Constantinos Karouzos, Xingwei Tan, and Nikolaos Aletras. 2026. [An empirical study on preference tuning generalization and diversity under domain shift](#). *Preprint*, arXiv:2601.05882.

- 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.
- Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. [Booksum: A collection of datasets for long-form narrative summarization](#). *Preprint*, arXiv:2105.08209.
- Patricia K Kuhl. 2007. Is speech learning ‘gated’ by the social brain? *Developmental science*, 10(1):110–120.
- Vojtěch Lanz, Yufei Cui, and Prasanna Parthasarathi. 2026. Dimpo: Dimensionality reduction for attention using preference optimization.
- Janghwan Lee, Seongmin Park, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. 2024. [Improving conversational abilities of quantized large language models via direct preference alignment](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11346–11364, Bangkok, Thailand. Association for Computational Linguistics.
- Brian Macwhinney. 1987. *The Competition Model*, pages 249–308. Psychology Press.
- Jonas Mayer Martins, Ali Hamza Bashir, Muhammad Rehan Khalid, and Lisa Beinborn. 2025. [Once upon a time: Interactive learning for storytelling with small language models](#). In *Proceedings of the First BabyLM Workshop*, pages 454–468, Suzhou, China. Association for Computational Linguistics.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. [SimPO: Simple preference optimization with a reference-free reward](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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.
- Mitja Nikolaus and Abdellah Fourtassi. 2023. [Communicative feedback in language acquisition](#). *New Ideas in Psychology*, 68:100985.
- OpenAI. 2026. Gpt-5.5-mini. <https://openai.com/>.
- Francesca Padovani, Bastian Bunzeck, Manar Ali, Omar Momen, Arianna Bisazza, Hendrik Buschmeier, and Sina Zarriß. 2025. [Dialogue is not enough to make a communicative BabyLM \(but neither is developmentally inspired reinforcement learning\)](#). In *Proceedings of the First BabyLM Workshop*, pages 421–435, Suzhou, China. Association for Computational Linguistics.
- Ryan Park, Ryan Theisen, Navriti Sahni, Marcel Patek, Anna Cichońska, and Rayees Rahman. 2023. [Preference optimization for molecular language models](#). *Preprint*, arXiv:2310.12304.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- D. Sculley. 2010. [Web-scale k-means clustering](#). In *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, page 1177–1178, New York, NY, USA. Association for Computing Machinery.
- Joonas Tapaninaho. 2025. [MoEP: Modular expert paths for sample-efficient language modeling](#). In *Proceedings of the First BabyLM Workshop*, pages 540–547, Suzhou, China. Association for Computational Linguistics.
- Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
- Caren M. Walker, Tania Lombrozo, Joseph J. Williams, Anna N. Rafferty, and Alison Gopnik. 2017. [Explaining constrains causal learning in childhood](#). *Child development*, 88(1):229–246. Publisher Copyright: © 2016 The Authors. Child Development © 2016 Society for Research in Child Development, Inc.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Yingshan Susan Wang, Linlu Qiu, Zhaofeng Wu, Roger P. Levy, and Yoon Kim. 2026. [Implicit representations of grammaticality in language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15035–15055, San Diego, California, United States. 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. 2023. [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, 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.
- Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongyue Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chan, Hao Jiang, Zhou Zhao, and Fei Wu. 2026. [A survey of direct preference optimization: Datasets, theories, variants, and applications](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.
- Andrew G. Young, Martha W. Alibali, and Charles W. Kalish. 2012. Disagreement and causal learning: Others’ hypotheses affect children’s evaluations of evidence. *Developmental Psychology*, pages 1242–1253.
- Xingmeng Zhao, Tongnian Wang, Sheri Osborn, and Anthony Rios. 2023. [BabyStories: Can reinforcement learning teach baby language models to write better stories?](#) In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 186–197, Singapore. Association for Computational Linguistics.