Resource Efficient Synthetic Data Generation for Preference Optimization

EuiYul Song
Samsung Electronics
euiyul.song@samsung.com

Abstract

Mistral-7b outperforms Llama2-13b across all benchmarks. However, the pre-training data distribution of Mistral-7b remains unclear. The small size of the AI2 Reasoning Challenge training dataset poses a risk of overfitting during adaptive pre-training. Additionally, the AI2 Reasoning Challenge corpus contains significant amounts of uninformative data, which could lead to catastrophic forgetting and deteriorating model performance due to heteroscedastic noise. To address this, we filter the corpus using Named Entity Recognition, hierarchical Gaussian Mixture Model clustering, and Nearest Neighbor Search. We then perform continual pre-training with a subset of the filtered corpus using the SciQ dataset. Finally, we train it with our synthetic data acquired from the filtered corpus with GPT-3.5 Turbo, aligning it with a training dataset using Odds Ratio Preference Optimization. Our synthetic data generation strategy reduces time and resource usage by over three times compared to the self-critic methods with a 1.27% increase in normalized accuracy compared to Supervised Fine-tuning.

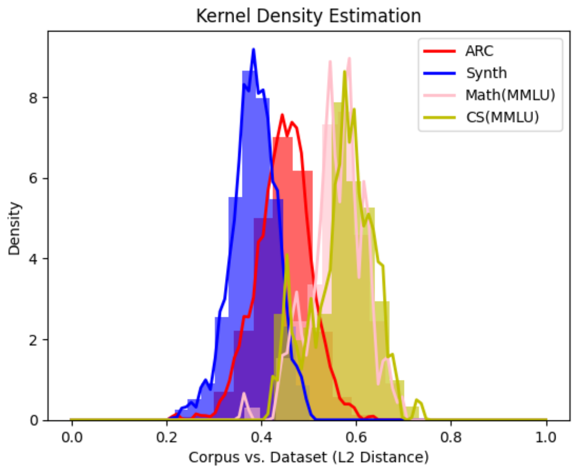

Figure 1: The probability density functions of cosine similarities between a corpus and datasets reveal that our synthetic dataset (Synth) (1,000 data) and the AI2 Challenge dataset (1,000) exhibit similar patterns, which are even more closely relevant to the corpus. In contrast, these patterns differ significantly when compared to the College Math (MATH(MMLU)) and Computer Science (CS(MMLU)) subsets of the Massive Multitask Language Understanding (MMLU) dataset.

1 Introduction

Pre-trained Language Models (PLMs), such as GPT-4 (OpenAI et al., 2024), Gemini 1.5 (Team et al., 2024), Mixtral (Jiang et al., 2024), and DeepSeek (DeepSeek-AI, 2024), are trained on extensive corpora of text data, capturing broad knowledge about language and context. Further pre-training (Gururangan et al., 2020) can enhance performance on specific tasks by adapting to a task corpus and a particular domain. However, Domain Adaptive Pre-Training (DAPT) can shift the model's data distribution, lack overlapping features, and carry outdated information, increasing the risk of catastrophic forgetting (Ke et al., 2023). Additionally, Task Adaptive Pre-Training (TAPT) is resource-intensive, requiring extensive tuning on training candidate selection and potentially misaligning the agent's goals if trained with too many or uninformative candidates (Ladkat et al., 2022).

Crowdsourcing and adversarial training are powerful strategies for enhancing model performance and mitigating the risks of bias propagation in pre-training tasks. However, data annotation requires significant investments in time and resources for project management, data labeling, and validation (Thorne et al., 2018; Schuster et al., 2019; Smit et al., 2020; Wang et al., 2023). In addition, adversarial training (Ribeiro et al., 2018, 2020; Chen et al., 2020; Ivgi and Berant, 2021; Qi et al., 2021; Miyato et al., 2021; Perez et al., 2022; Shi et al., 2023) increases computational costs due to perturbation processes and carries the risk of overfitting to adversarial data.

Generating synthetic data for pre-training can help reduce the expenses related to training, validation, and human annotation. Nonetheless, creating synthetic data from a heteroscedastic noisy corpus poses significant challenges. Such data may exhibit biases towards specific patterns and inaccuracies of the model (Hao et al., 2024). Moreover, this process necessitates self-critique and self-refinement by a Large Language Model, which triples the time and cost required to ensure the accuracy of the synthetic data (Bai et al., 2022).

To mitigate the potential risks of training a model using synthetic data generated from a noisy corpus without manual evaluation, we propose a simple yet effective data generation method:

- We apply Named Entity Recognition to filter the corpus. Then, we use Sentence Transformers to embed the corpus, followed by hierarchical Gaussian Mixture Model clustering to remove outliers from the embedded data.

- We generate multiple-choice questions from the corpus in few-shot settings. We then filter the generated questions with Named Entity Recognition and perform a Nearest Neighbor Search between the fine-tuning and the generated dataset to select the most reliable candidates for training.

- We perform Continual Pre-training consisting of DAPT and TAPT, followed by synthetic data pre-training. We then optimize our model with Odds Ratio Preference Optimization (ORPO) with the training dataset.

We propose a synthetic data generation method derived from a noisy corpus that bridges the gap between TAPT and DAPT without requiring iterative training and evaluation steps during the pre-training candidate selection stage. Our method is tested with 1,000 synthetic data samples from a corpus of 13,000,000 unlabeled samples. It effectively prevents performance degradation caused by hallucinations in the synthetic data. Additionally, our strategy reduces time and resource usage by over three times compared to self-critic and self-refine methods, while maintaining equivalent precision.

2 Related Works

2.1 Synthetic Data Generation

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have become a cornerstone in computer vision applications due to their ability to produce highly realistic images through adversarial training (Radford et al., 2016; Mirza and Osindero, 2014; Karras et al., 2019). Emerging models such as diffusion models (Ho et al., 2020), transformers (Vaswani et al., 2023), and Mamba (Gu and Dao, 2023) have extended the capabilities of synthetic data generation beyond computer vision to include text and music generation. Additionally, privacy-preserving data generation often employs models like Markov chains (Gilks et al., 1996; Juang, 2003; Lafferty et al., 2001; Brooks et al., 2011; Nemeth and Fearnhead, 2019) and Bayesian Neural Networks (BNNs) (Williams and Rasmussen, 1995; Garnelo et al., 2018a,b), which maintain statistical properties while ensuring individual data points cannot be traced back to real individuals. A significant challenge in this field, highlighted by Bauer et al. (2024) the absence of standardized evaluation metrics and datasets, which complicates model comparisons. To minimize human labor in evaluating the honestness of synthetic data, we propose hierarchical GMM (Reynolds, 2018) clustering with NER (Chinchor and Robinson, 1998) and NNS (Malkov and Yashunin, 2018) to generate and select generated synthetic data for pre-training.

2.2 Continual Learning

Adapting to task-specific unlabeled data improves performance even after DAPT, with effective alternatives being task corpus augmentation using data selection strategies (Gururangan et al., 2020). Integrating recent information into LLMs is crucial, supported by frameworks like ERNIE 2.0 (Sun et al., 2019), which help incrementally update temporal knowledge, reduce forgetting, and ensure efficient updates. Domain-incremental pre-training and domain-specific continual learning enhance LLMs in areas such as finance and e-commerce. Expanding the linguistic range for underrepresented languages is essential (Gogoulou et al., 2024), and advances in programming language understanding improve coding practices (Yadav et al., 2023). Continual Instruction Tuning (He et al., 2023) enhances LLMs' instruction-following abilities, categorized into task-incremental, domain-incremental, and tool-incremental types, using techniques like TAPT (Gururangan et al., 2020), ConPET (Song et al., 2023), and PlugLM (Cheng et al., 2023) to mitigate forgetting and optimize performance. Continual value alignment incorporates ethical guidelines and adapts to cultural sensitivities (Yao et al.,

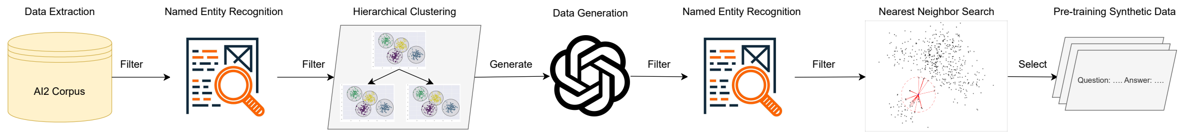

Figure 2: Our process for generating and filtering synthetic data involves several steps. Initially, we clean the AI2 corpus by applying Named Entity Recognition and utilizing hierarchical clustering with Gaussian Mixture Models. We then create multiple-choice questions and answers from the AI2 Corpus using GPT-3.5 Turbo. For the final selection of our dataset, which will be used for pre-training, we employ Named Entity Recognition again and conduct Nearest Neighbor Searches between the generated data and the AI2 Challenge dataset to ensure quality.

2023), with approaches like Continual Proximal Policy Optimization (Xuan et al., 2023) balancing policy learning and knowledge retention. Among these methods, we utilize one-step Continual Pre-training with a mixture of TAPT and DAPT data.

2.3 Human Alignment

Reinforcement learning with human feedback (RLHF) (Christiano et al., 2023) frequently employs the Bradley-Terry model (Bradley and Terry, 1952) trains models to optimize the reward model's score for selected responses, aligning language models with human preferences. Alternatives like reinforcement learning from language model feedback (RLAIF) (Bai et al., 2022) have been suggested. Nonetheless, RLHF encounters difficulties due to Proximal Policy Optimization (PPO) (Schulman et al., 2017)'s instability and the sensitivity of reward models. Direct policy optimization (DPO) (Rafailov et al., 2023) integrates the reward modeling stage into preference learning to address these issues. Identity preference optimization (IPO) (Azar et al., 2023) aims to reduce potential overfitting in DPO. Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) and Unified Language Model Alignment (ULMA) (Cai et al., 2024) bypass the need for pairwise preference datasets. Supervised fine-tuning (SFT) using filtered small curated datasets can also be adequate to create human-aligned models (Zhou et al., 2023). Iterative fine-tuning with model-generated outputs after selection has also shown promising results (Li et al., 2024; Haggerty and Chandra, 2024). Finally, Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024) introduces an odds ratio-based penalty to the negative log-likelihood loss to distinguish between preferred and non-preferred responses. This paper replaces SFT with ORPO to train the training dataset.

3 Methods

We begin by applying Named Entity Recognition (NER) to filter the corpus and using Sentence Transformers to embed the data. Hierarchical Gaussian Mixture Model (GMM) clustering is then employed to remove outliers from the embedded data. Next, we generate multiple-choice questions from the corpus using few-shot examples from the publicly available dataset with GPT-3.5 Turbo (Brown et al., 2020) and enhance their precision through filtering with NER. We perform a Nearest Neighbor Search (NNS) between the training dataset and the generated dataset to select the most reliable candidates for training. Finally, we perform one-step Continual Pre-training consisting of TAPT and DAPT datasets to shift our model in distribution, pre-train with the synthetic data, and optimize our model using ORPO with training datasets. We ablate our method with a combination of each stage.

3.1 Synthetic Data Generation

Preprocessing Before preprocessing data, we randomly sample 20,00 data points from the corpus for resource efficiency. We then remove the hyperlinks from the AI2 Corpus (Clark et al., 2018) and preprocess the corpus using Named Entity Recognition to remove sentences lacking objects, subjects that are not pronouns, and verbs, utilizing spaCy[1]. These sentences are embedded using the Sentence Transformer (Reimers and Gurevych, 2019), specifically employing the M3 (Bai et al., 2024).

3.2 Hierarchical Clustering

We use a GMM to detect and eliminate outliers, selecting the most informative candidate from each cluster to enhance data utility in the hierarchical clustering of embedded documents. The cluster size is set to $\sqrt{|D|}/2$, where $|D|$ represents the number of documents. The GMM assumes the data originates from multiple Gaussian distributions, each

[1] https://spacy.io

```
Here are examples of generating multiple-choice datasets:

Generate the multiple choice [DATASET] given a [CONTEXT]. Impute [CONTEXT] to build [DATASET] if [CONTEXT] does not have enough information.\n[CONTEXT]:

[CONTEXT]: Plants and fungi cells both have a cell wall, which animals do not have.
=>
[DATASET]: Question: Plants and fungi cells have what basic cell structure, which animals do not have?\n(A) Cell Membrane\n(B) cell wall\n(C)
Lysosomes\n(D) Cytoskeleton\nAnswer: (B) cell wall

...

Generate the multiple choice [DATASET] given a [CONTEXT]. Impute [CONTEXT] to build [DATASET] if [CONTEXT] does not have enough information.\n[CONTEXT]:

[CONTEXT]: Technology consumes resources which are not necessarily renewable, including living resources, such as forests and populations of fish, and
inanimate resources, such as natural chemicals and minerals.
=>
[DATASET]:
```

```
Question: What type of resources does technology consume that are not necessarily renewable? (A) fossil fuels (B) wind energy (C) solar power (D) nuclear
energy Answer: (A) fossil fuels
```

Figure 3: We generate multiple-choice questions and answers from the AI2 Corpus by utilizing publicly available multiple-choice datasets, with the context integrated using GPT-3.5 Turbo. To address issues of missing information, we compel GPT-3.5 Turbo to impute context.

with distinct means and variances. In Figure 1, the Gaussian distribution of cosine similarity between the embedded corpus and the dataset supports the viability of our method.

In the GMM, the Expectation-Maximization (EM) algorithm (Moon, 1996) iteratively refines parameters through two steps. The Expectation step computes the likelihood of each data point belonging to specific clusters based on current parameters, followed by the Maximization step, which updates these parameters to optimize data fit. This process repeats until minimal changes between iterations.

After convergence, each sentence is assigned to a Gaussian component. Sentences with a log-likelihood below one standard deviation from the mean are filtered out. We repeat this GMM filtering process until we obtain 10,000 data points. We randomly select 4,000 data points for efficiency.

3.3 Data Generation

We generated multiple choice questions and answer pairs from 10,000 filtered contexts using GPT-3.5 Turbo, following a five-shot approach with random contexts from the SciQ dataset (Welbl et al., 2017). Due to the presence of aleatoric and epistemic uncertainties, there are instances where GPT-3.5 Turbo indicates that the provided contexts lacked sufficient information. This absence of information and its systematic link to the biases inherent in GPT-3.5 and the original data can significantly impact performance. This missing information demonstrates characteristics typical of Missing Not At Random (MNAR) data.

Consider a synthetic dataset defined as $D = (x_t^{(i)}, o_t^{(i)})^{T_i} t = 1, {y^{(i)}}^n i = 1$ where x represents a context, o is a missingness indicator, y is relevant synthetic data, n is the number of datasets, T is the number of multiple choice pairs that can be generated from a context, f_θ represents the function by GPT-3.5 Turbo that produces a logit, and k is the corpus. For token classification, the probability is given by $p(y|x_{1:T}, o_{1:T}, \theta) = \frac{e^{f_\theta(k(x_{1:T}, o_{1:T}))1}}{\sum^1 j=0 e^{f_\theta(k(x_{1:T}, o_{1:T}))_j}}$. $p(x, o|\theta) = p(x|\theta)p(o|x, \phi)$ depends on both the existing (x) and non-existing (ϕ) context in the corpus. The probability of missing synthetic data is related to the context.

In such cases, we compelled GPT-3.5 Turbo to infer additional context to ensure it had enough information for question generation. The details of our prompt are shown in Figure 3.

3.4 Data Filtering

Although the data generation method described does not require manual effort in data generation and evaluation, it may produce hallucinated data (Borra et al., 2024). Bai et al. (2022) suggests using self-critique and revision techniques to evaluate and improve the generated responses. However, these methods can triple the time and cost while leaving room for model uncertainty. We employ Named Entity Recognition to eliminate sentences that lack objects or have subjects that are not pronouns, as well as verbs. Additionally, we discard any generated data that does not include the terms "Question:" or "Answer:" and we remove data containing the words "context" and "information" as GPT-3.5 Turbo sometimes creates questions like "What is he doing in the given context?".

We use M3 to embed the entire AI2 Challenge training set (Clark et al., 2018) and then load these embeddings into a FAISS index[2]. Similarly, we

[2]https://github.com/facebookresearch/

embed all filtered synthetic data and conduct an approximate nearest neighbor search. We select the top 1,000 generated datasets for training based on the highest top 1 cosine similarity.

3.5 Training

Continual Pre-training Although Gururangan et al. (2020) suggest that TAPT following DAPT with 500 semantically similar data points to training sets yields the best performance, the AI2 Challenge task presents difficulties due to its outdated nature and high data demands for effective TAPT. Consequently, we supplement our dataset by combining 10,000 data points from the SciQ dataset with another 10,000 from the cleaned corpus detailed in Section 3.1 for pre-training. Specifically, we pre-train the Mistral 7B (Jiang et al., 2023) checkpoint[3] using this mixed dataset to ensure the model adapts to both the corpus and the task. We then further enhance the pre-trained model by incorporating 1,000 synthetic data points.

Preference Optimization We apply rejection sampling to the AI2 Challenge dataset using the M3 embedding approach in Section 3.1. Specifically, we concatenate each question with its corresponding rejected answers from the multiple-choice options using the "[SEP]" token as follows:

> "Which of these is a property of water that allows it to transport materials through the Earth system? [SEP] Answer: (C) It dissolves many substances."

We conduct a Maximum Inner Product Search (MIPS) between embeddings of rejected answers and a gold answer to sample the top 1 semantically similar rejected answer. This process helps us identify appropriate rejection samples. We then utilize these rejection samples to align our pre-trained model using the Odds Ratio Preference Optimization, bypassing Supervised Fine-Tuning.

4 Experiments

4.1 Training

Quantization We utilize Int4 zero-point quantization (Wu et al., 2023), shifting the input distribution to span the entire range [127, 127] by scaling with the normalized dynamic range and then applying a shift by the zero-point. This minimizes quantization errors, particularly for asymmetric distributions.

Low-Rank Adaptation (LoRA) We maintain the pre-trained model weights in a frozen state and introduce trainable rank decomposition matrices into each layer of the Transformer architecture using LoRA (Hu et al., 2021). This approach significantly reduces the number of trainable parameters for downstream tasks.

8-bit Optimizer We eliminate the need for slow transfers to GPU memory or additional temporary memory for quantization and dequantization with the 8-bit optimizers (Dettmers et al., 2022). This feature makes 8-bit optimizers on GPUs faster than 32-bit counterparts.

4.2 Datasets

We use the dataset provided by Allen AI[4], which includes the SciQ dataset and the AI2 Reasoning Challenge (ARC) dataset and corpus. The SciQ dataset is utilized for TAPT and generating few-shot examples to create synthetic data for fine-tuning. The ARC challenge dataset is employed for preference optimization, while the AI2 corpus is used as context for generating synthetic data.

4.3 Evaluation

The customary practice involves using the model selection development set during training. However, we do not evaluate every training step in order to achieve a faster, more universal solution. In this study, we assess our model using the Language Model Evaluation Harness framework[5] with the ARC challenge test set.

Metrics The Language Model Evaluation Harness framework defines $x_{0:m}$ as the initial prompt and $x_{m:n_i}$ as the ith potential continuation, where the token length of this continuation is $n_i - m$.

The score for continuation i is calculated using $\sum_{j=m}^{n_i-1} logP(x_j|x_{0:j})$. This method sums the log probabilities of each token in the continuation, assuming the continuation is sampled from the model following the prompt. This method is employed by the evaluation harness in all multiple-choice tasks and is referred to as **acc**.

faiss

[3] https://huggingface.co/mistralai/Mistral-7B-v0.1

[4] https://allenai.org/

[5] https://github.com/EleutherAI/lm-evaluation-harness

Single-Task									
		-		Synth		CPT		Synth(CPT)	
		acc	acc_n	acc	acc_n	acc	acc_n	acc	acc_n
SFT	μ	56.23	59.22	-	-	55.55	**59.90**	**56.23**	59.56
ORPO	μ	**56.31**	60.24	55.80	59.56	55.63	59.90	55.80	**60.49**
SFT	σ_M	**1.450**	1.436	-	-	1.452	1.432	**1.450**	**1.430**
ORPO	σ_M	**1.449**	**1.430**	1.450	**1.430**	1.450	**1.430**	1.450	**1.430**

Table 1: Performance on a single task fine-tuning, specifically ARC challenge task, with pre-training methods specified in the columns and fine-tuning methods listed in the rows. Pre-training method before each method is indicated in parentheses. Here, μ denotes the mean, and σ_M represents the standard error. "Synth" refers to Synthetic Data Pre-training with 1,000 nearest neighbor synthetic data. "CPT" stands for Continual Pre-training, which involves a combined training of 10k instances from TAPT and 10k instances from DAPT. "ORPO" signifies Odds Ratio Preference Optimization.

The byte-length normalized score for a continuation is computed as follows:

$$\frac{\sum_{j=m}^{n_i-1} \log P(x_j|x_{0:j})}{\sum_{j=m}^{n_i-1} L_{x_j}} \tag{1}$$

where L_{x_j} represents the byte count of token x_j. This normalization method adjusts for the length of continuations by averaging the log probability per byte, thereby making the scoring tokenization agnostic. The evaluation harness uses this scoring method for all multiple-choice tasks as **acc_n**.

4.4 Hyperparameters

This paper does not explore hyperparameter tuning to evaluate model agnosticism. We have used default hyperparameter settings and hardware configurations across all tasks, detailed in Table 7. Refer to Table 8 for specific configurations for each stage.

5 Result

5.1 Continual Pre-training

Task Adaptive Pre-training According to Table 2, the quantity of documents is crucial for conducting task-adaptive pre-training on the Mistral model. A 2-sample t-test comparing the results from models fine-tuned after pre-training with 100k and 20k documents shows a p-value of 0.0233, indicating statistical significance. We hypothesize that the AI2 dataset might be outdated, and the Mistral 7B model may not have been pre-trained on a significant amount of scientific corpora. Due to the distributional shift between single sentences and multiple-choice questions, task adaptive pre-training requires domain adaptation. As indicated in Table 1, this approach underperforms relative to Supervised Fine-tuning without pre-training.

Continual Pre-training					
	Task			Domain	
	|N|	acc	acc_n	acc	acc_n
SFT μ	20k	50.43	53.50	**56.40**	**59.13**
SFT μ	100k	54.27	58.19	**56.40**	**59.13**
SFT σ_M	20k	1.460	1.460	**1.450**	**1.440**
SFT σ_M	100k	1.456	1.441	**1.450**	**1.440**

Table 2: Ablation Study on our Task Adaptive and Domain Adaptive Pre-training after Supervised Finetuning (SFT). We randomly sample |N| particles from the refined corpus via NER and GMM for Task Adaptive Pre-training. For Domain Adaptive Pre-training, we mix SciQ and ARC Easy datasets.

Domain Adaptive Pre-training Table 2 illustrates that the Mistral 7B model already captures complex patterns and representations in multiple-choice question-answering tasks. Adding domain-adaptive pre-training does not improve the performance of Supervised Fine-Tuning without prior pre-training, as shown in Table 1. By examining the outcomes from domain and task adaptive pre-training, we recommend that the Mistral model requires pre-training using the entire cleaned AI2 corpus, enhanced with multiple-choice questions and answers derived from this corpus.

Domain and Task Adaptive Pre-training CPT in Table 1 indicates that the model is pre-trained using 10,000 uniform samples from both the SciQ training set and a cleaned corpus. Despite the lack of statistical significance (p-value of 0.7390) between SFT without pre-training and with CPT, the use of multiple-choice questions and answers from SciQ helps to mitigate distribution shifts during task adaptation in Table 2 caused by heteroscedastic noise and sentence representation in the corpus.

Synthetic Data Selection					
		NER+NNS		Random	
		acc	acc_n	acc	acc_n
ORPO	μ	**55.80**	**60.49**	53.33	55.80
ORPO	σ_M	**1.430**	**1.430**	1.430	1.450

Table 3: Ablation analysis to assess the effectiveness of Nearest Neighbor Search (NNS) in selecting 1,000 synthetic data points with NER post-processing (NER+NNS), compared to selecting 1,000 synthetic data points at random without NER post-processing (Random), for training purposes.

Performing domain-adaptive pre-training with curated synthetic data could enhance performance. However, due to constraints in resources and time, we do not explore pre-training with a large volume of generated multiple-choice science questions and answers alongside a cleaned corpus.

5.2 Synthetic Data Generation

Nearest Neighbor Search Table 3 highlights the importance of selecting pre-training candidates using NER and NNS from generated synthetic data. Randomly sampling 1,000 data points without NER nor NNS for further pre-training deteriorates model performance when fine-tuned. A two-sample t-test between models pre-trained with random and selected candidates yields a p-value of 0.0214, indicating statistical significance. It's important to note differences in the training environments: The NER and NNS method involves pre-training with 10,000 data points each from the SciQ dataset and a cleaned corpus, followed by further pre-training with 1,000 synthetic data points ranked by cosine similarity relative to the ARC Challenge dataset. In contrast, the random selection method uses 2,000 data points each from the ARC Easy dataset, the cleaned corpus, and 1,000 randomly sampled synthetic data points that are not post-processed. Despite these differences, juxtaposing ORPO with ARC_C+ARC_E in Table 5 and ORPO with CPT in Table 1 demonstrates that samples from the cleaned corpus and ARC Easy dataset do not degrade performance, implicating the uncleaned synthetic data as the detrimental factor.

Further Pre-training Continual pre-training of CPT with synthetic data, as shown as Synth(CPT) in Table 1, utilizing our post-processing, enhances performance when preference is optimized with ORPO. Compared to SFT without pre-training, there is an improvement of 1.27% in normalized accuracy. This demonstrates that our synthetic data, processed through our automatic filtering method, is safe for pre-training.

Number of Nearest Neighbor					
|NN|		1,000		3,000	
		acc	acc_n	acc	acc_n
Synth	μ	**55.80**	**59.56**	**55.80**	59.39
Synth(CPT)	μ	55.80	**60.49**	**56.06**	59.90
Synth	σ_M	**1.450**	**1.430**	**1.450**	1.440
Synth(CPT)	σ_M	**1.450**	**1.430**	**1.450**	**1.430**

Table 4: Ablation study on the optimal number of Nearest Neighbor candidates for selecting synthetic candidates for pre-training. Note that the performance above is evaluated after ORPO on each pre-training method. |NN| indicates the number of nearest neighbors to the ARC challenge dataset.

Multi-Task				
		data	acc	acc_n
SFT	μ	SciQ+ARC_C	55.03	58.87
ORPO	μ	ARC_C+ARC_E	**55.38**	**59.04**
SFT	σ_M	SciQ+ARC_C	1.454	**1.438**
ORPO	σ_M	ARC_C+ARC_E	**1.450**	1.440

Table 5: A multi-task training performance without pre-training. "ARC_C" indicates the Challenge task of the ARC, while "ARC_E" is the Easy task.

5.3 Multi-Task Reasoning

The Mistral 7B model is already adept at tasks related to multiple-choice questions and answers. Since there is minimal overlap between the ARC Challenge test set and publicly available multiple-choice science datasets like SciQ and ARC Challenge Easy, multi-task SFT does not outperform single-task SFT.

Preference Optimization				
	DPO		ORPO	
	acc	acc_n	acc	acc_n
μ	56.14	58.96	**56.31**	**60.24**
σ_M	1.450	1.437	**1.449**	**1.430**

Table 6: Preference Optimization performance ablation Study. "DPO" represents Direct Preference Optimization. DPO overfits limited training data.

5.4 Preference Optimization

Table 6 indicates that Direct Preference Optimization (DPO) tends to overfit when only a small amount of fine-tuning data is available. Conversely, according to Table 1, ORPO with pre-training surpasses other combinations.

6 Conclusion

We propose a synthetic data generation method that leverages Named Entity Recognition (NER) for corpus cleaning and a hierarchical Gaussian Mixture Model to select representative sentences. We generate 3,000 synthetic data points from these candidates employing GPT-3.5 Turbo and add necessary context when the information in the corpus is sparse. Post-generation, we refine this data using NER and NNS. Our approach includes task and domain adaptive pre-training followed by synthetic data pre-training. We enhance our model's performance through ORPO with rejection samples acquired with MIPS, achieving a 1.27% increase in normalized accuracy compared to SFT. Although our method uses only 1,000 synthetic data points, our findings indicate that it effectively reduces factually incorrect generations and eliminates the need for labor-intensive human validation. This research paves the way for automated synthetic data generation tailored for preference optimization.

7 Limitation

Due to time and resource constraints, we use a limited volume of data for synthetic data generation and training. Nevertheless, we maintain that our method is generalizable, as it incorporates at least 1,000 data points for pre-training, indicating a statistically significant baseline.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences.

Yang Bai, Anthony Colas, Christan Grant, and Daisy Zhe Wang. 2024. M3: A multi-task mixed-objective learning framework for open-domain multi-hop dense sentence retrieval.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen tau Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers.

Tianchi Cai, Xierui Song, Jiyan Jiang, Fei Teng, Jinjie Gu, and Guannan Zhang. 2024. Ulma: Unified language model alignment with human demonstration and point-wise preference.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022. Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 191–200.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8801–8811, Online. Association for Computational Linguistics.

Xin Cheng, Yankai Lin, Xiuying Chen, Dongyan Zhao, and Rui Yan. 2023. Decouple knowledge from paramters for plug-and-play language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14288–14308, Toronto, Canada. Association for Computational Linguistics.

N. Chinchor and P. Robinson. 1998. Appendix E: MUC-7 named entity task definition (version 3.5). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

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.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization.

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Hazan E. Singer Y. Duchi, J. 2011. Adaptive subgradient methods for online learning and stochastic optimization. journal of machine learning research.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31.

Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. 2018a. Conditional neural processes.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. 2018b. Neural processes.

Walter Gilks, Sylvia Richardson, and D. Spiegelhalter. 1996. Introducing markov chain monte carlo. *Markov Chain Monte Carlo in Practice*.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2024. Continual learning under language shift.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks.

Vipul Gupta, Santiago Akle Serrano, and Dennis DeCoste. 2020. Stochastic weight averaging in parallel: Large-batch training that generalizes well.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Hamish Haggerty and Rohitash Chandra. 2024. Self-supervised learning for skin cancer diagnosis with limited training data.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic data in ai: Challenges, applications, and ethical implications.

Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. 2023. Continual instruction tuning for large multimodal models.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. 2017. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*.

Jabbar Hussain. 2019. Deep learning black box problem.

Maor Ivgi and Jonathan Berant. 2021. Achieving model robustness through discrete adversarial training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1529–1544, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *https://arxiv.org/pdf/2112.09118.pdf*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*, 2208.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018, pages 876–885. Association For Uncertainty in Artificial Intelligence (AUAI). Funding Information: Acknowledgements. This work was supported by NSF IIS-1563887, Samsung Research, Samsung Electronics and Russian Science Foundation grant 17-11-01027. We also thank Vadim Bereznyuk for helpful comments. Funding Information: This work was supported by NSF IIS-1563887, Samsung Research, Samsung Electronics and Russian Science Foundation grant 17-11-01027. We also thank Vadim Bereznyuk for helpful comments. Publisher Copyright: © 34th Conference on Uncertainty in Artificial Intelligence 2018. All rights reserved.; 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018 ; Conference date: 06-08-2018 Through 10-08-2018.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Biing-Hwang Juang. 2003. Hidden markov models.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models.

Diederik P. Kingma and Max Welling. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Arnav Ladkat, Aamir Miyajiwala, Samiksha Jagadale, Rekha A. Kulkarni, and Raviraj Joshi. 2022. Towards simple and efficient task-adaptive pre-training for text classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 320–325, Online only. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2024. Self-alignment with instruction backtranslation.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022. Rethinking attention-model explainability through faithfulness violation test.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu. A. Malkov and D. A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs.

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The jensen-shannon divergence.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2021. Adversarial training methods for semi-supervised text classification.

T.K. Moon. 1996. The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, 13:47 – 60.

Christopher Nemeth and Paul Fearnhead. 2019. Stochastic gradient markov chain monte carlo.

Philhoon Oh and James Thorne. 2023. Detrimental contexts in open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11589–11605, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021a. Kilt: a benchmark for knowledge intensive language tasks.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021b. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Douglas A. Reynolds. 2018. Gaussian mixture models. In *Encyclopedia of Biometrics*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Karthik Raman Hamed Zamani Sebastian Hofstatter, Jiecao Chen. 2022. Fid-light: Efficient and effective retrieval-augmented text generation. *https://arxiv.org/pdf/2209.14290.pdf*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable?

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Chenyang Song, Xu Han, Zheni Zeng, Kuai Li, Chen Chen, Zhiyuan Liu, Maosong Sun, and Tao Yang. 2023. Conpet: Continual parameter-efficient tuning for large language models.

EuiYul Song, Sangryul Kim, Haeju Lee, Joonkee Kim, and James Thorne. 2024. Re3val: Reinforced and reranked generative retrieval.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu,

YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexan-

der Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

James Thorne. 2022. Data-efficient autoregressive document retrieval for fact verification. *arXiv preprint arXiv:2211.09388*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions.

Christopher Williams and Carl Rasmussen. 1995. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256.

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

Xiaoxia Wu, Cheng Li, Reza Yazdani Aminabadi, Zhewei Yao, and Yuxiong He. 2023. Understanding int4 quantization for transformer models: Latency speedup, composability, and failure cases.

Chengbin Xuan, Feng Zhang, Faliang Yin, and Hak-Keung Lam. 2023. Constrained proximal policy optimization.

Prateek Yadav, Qing Sun, Hantian Ding, Xiaopeng Li, Dejiao Zhang, Ming Tan, Xiaofei Ma, Parminder Bhatia, Ramesh Nallapati, Murali Krishna Ramanathan, Mohit Bansal, and Bing Xiang. 2023. Exploring continual learning for code generation models.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2023. A survey of active learning for natural language processing.

Haotian Zhou, Tingkai Liu, Qianli Ma, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Lobass: Gauging learnability in supervised fine-tuning data.

Hyperparameters	
scheduler	constant w/ warmup
max length	256
max prompt length	256
beta	0.1
group by length	True
weight decay	0.001
max grad norm	0.3
max steps	-1
epoch	1
gpu	A100
warmup	0.1

Table 7: The hyperparameter settings and hardware configurations utilized in our study are detailed previously. When not otherwise specified, we employ the default configurations of the Hugging Face Trainer. Please refer to the Hugging Face Trainer[6] for more information documentation.

Hyperparameters		
Stage	Learning Rate	Batch Size
CPT	1e-5	45
TAPT	1e-5	45
SFT	1e-5	45
Synth	1e-5	45
Synth(CPT)	5e-6	45
SFT(TAPT)	1e-5	45
SFT(CPT)	5e-6	45
SFT(Synth)	5e-6	45
SFT(Synth(CPT))	1e-6	45
ORPO(CPT)	5e-6	35
ORPO(Synth)	5e-6	35
ORPO(Synth(CPT))	1e-6	35
DPO(SFT)	1e-6	20

Table 8: Configuration of the learning rate for each stage, with the pre-training method indicated inside parentheses. "Synth" refers to Synthetic Data Pre-training. "CPT" stands for Continual Pre-training, which includes Domain-Adaptive Pre-training (DAPT) and Task-Adaptive Pre-training (TAPT). "DPO" represents Direct Preferences Optimization, while "ORPO" denotes Odds Ratio Preference Optimization.