Design and Analysis of few Million Parameter Transformer-based Language Models trained over a few Million Tokens Dataset

Yen-Che Hsiao[1], Abhishek Dutta[1]

[1]Department of Electrical and Computer Engineering
University of Connecticut
Storrs, 06269, CT, USA

Correspondence: yen-che.hsiao@uconn.edu

Abstract

In this work, we systematically explore training methods and perform hyperparameter tuning to identify key language model parameters upper bounded by 28 million. These models are designed to generate a broad spectrum of basic general knowledge in simple and coherent English with limited generalization ability. We use the Simple English Wikipedia as the training dataset, selecting samples between 64 and 512 words, which provides a high-quality, compressed representation of general knowledge in basic English. Through hyperparameter tuning, we identify the best-performing architecture, yielding the lowest training loss, as a decoder-only Transformer with rotary positional encoding, multi-head attention, root-mean-square normalization, Gaussian error linear unit activation, post-normalization, no interleaved group query attention, an embedding dimension of 512, 8 layers, 8 attention heads, a feedforward dimension of 2048, and zero dropout. Models trained with a learning rate decaying linearly from 10^{-4} to 10^{-5} over 64 epochs achieve a training loss of 0.1, which appears sufficient for reproducing text more effectively than models trained to losses of 0.2 or 0.5. Fine-tuning on rephrased text further demonstrates that the model retains its ability to produce simple and coherent English covering broad basic knowledge, while exhibiting limited generalization capability.

1 Introduction

Several works have developed language models capable of performing a wide range of tasks (Sindhu et al., 2024), including but not limited to code completion (Husein et al., 2025), question answering (Nassiri and Akhloufi, 2023), and text summarization (Zhang et al., 2025). Large language models often contain more than a billion parameters and are typically trained on more than 100 billion tokens (Raiaan et al., 2024). However, such large-scale models present challenges in terms of deployment on local devices, and their substantial carbon footprint underscores the need to reduce computational requirements while maintaining comparable or acceptable performance.

Practitioners have also released language models with less than one billion parameters that are capable of generating grammatically correct and informative text, such as the 0.6B models from Qwen3 (Yang et al., 2025) and the 135M and 360M models from SmolLM (Allal et al., 2025). While some information regarding the model architecture, training hyper-parameters, and training data is often provided, details on how the architecture, hyper-parameters, model checkpoints, or datasets are selected are rarely, if ever, discussed.

In this work, we aim to identify the transformer-based language model with the minimum number of parameters less than 28 million (M) capable of generating broad spectrum of basic general knowledge in simple and coherent English and has limited generalization ability. The Simple English Wikipedia dataset is used for training, as we consider it a high-quality and compact dataset that covers a broad range of general knowledge in basic English, compared to other datasets such as WikiText-2. As a starting point, we set the target of learning a dataset containing 20 M tokens within 10 epochs. Following the result reported in Table 3 of (Hoffmann et al., 2022), we use the empirical observation that the optimal number of model parameters is approximately equal to the total number of training tokens divided by 20. This yields a target of roughly 10 M parameters for our setting (20 M tokens $\times$ 10 epochs $\div$ 20). Our study therefore focuses primarily on models with parameter counts exceeding 10M, up to 28M parameters, for over-parameterization.

We investigate models with varying numbers of decoder blocks, trained under different layers, batch sizes, learning rates, and training strategies. For dataset selection, we compare the Sim-

Proceedings of the First BabyLM Workshop, pages 109–129

ple English Wikipedia dataset with the WikiText-2 dataset. We discard WikiText-2, since it contains non-English characters as some of the most frequent tokens, suggesting that the dataset may not consist of coherent and simple English compared to Simple English Wikipedia. The Simple English Wikipedia dataset is further processed by removing entries with fewer than 64 words and more than 512 words, as shorter entries often consist of incomplete sentences and longer entries exceed the sequence length of the model. The resulting dataset is referred to as the long-context subset. Analysis of the word frequency distribution in Simple English Wikipedia shows that 2,206 words appear more than 500 times, motivating the choice of a vocabulary with 2,048 tokens.

We evaluate models trained on the Simple English Wikipedia dataset under different batch sizes, learning rates, and training strategies. The first strategy, referred to as the 2-stage training method, involves initially training on short-context data with a small maximum sequence length for several epochs, followed by training on long-context data with a larger maximum sequence length. The second strategy, referred to as the interleaved training method, alternates between short- and long-context datasets each epoch, with corresponding adjustments to the maximum sequence length, to mitigate catastrophic forgetting. Our results show that the best-performing model, achieving the lowest training loss within 10 epochs, is a decoder-only Transformer with rotary positional encoding, multi-head attention, root-mean-square normalization, Gaussian error linear unit (GELU) activation, post-normalization, no interleaved group query attention, an embedding dimension of 512, 8 layers, 8 heads, a feedforward dimension of 2048, zero dropout, and a learning rate of 10^{-4} trained with standard mini-batch gradient descent. Analysis of generated text indicates that a training loss of 0.1 is sufficient for the model to reproduce the target text more effectively than models trained to losses of 0.2 or 0.5.

To evaluate limited generalization ability, we fine-tune the 0.1-loss pre-trained model on rephrased text. The fine-tuning dataset is constructed from 100 selected entries of the long-context subset of Simple English Wikipedia. Each entry is split into two parts: the first as the context and the second as the target. The context is rephrased into three variants using ChatGPT-5. The rephrased contexts paired with the original target form new training and evaluation entries: the first and second rephrased contexts are used for the training set, the third rephrased context is used for the test set, and the original context is used for the validation set. After fine-tuning for 64 epochs, we select the model with the lowest validation loss and evaluate it on five entries each from the validation and test sets. The model successfully reproduces the target text for two of five entries in both the validation and test sets, succeeds only in the test set for one entry, succeeds only in the validation set for one entry, and fails on both sets for one entry. These results demonstrate that the fine-tuned model exhibits limited generalization ability, while maintaining the ability to generate simple and coherent English covering a broad spectrum of basic general knowledge.

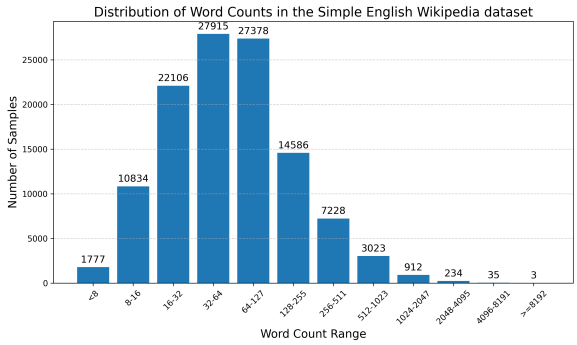
2 Data Preparation

2.1 Dataset Selection

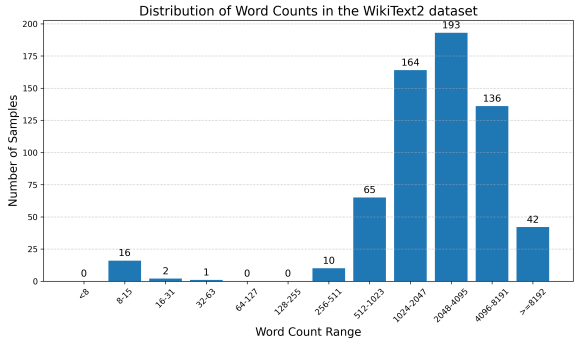
To build a model capable of generating a broad spectrum of basic general knowledge in simple and coherent English, we aim to identify a suitable text dataset for training. We considered two datasets as candidates: the WikiText-2 dataset (Merity et al., 2016) and the Simple English Wikipedia dataset from the 100M training set of the BabyLM Challenge (Charpentier et al., 2025).

To make the datasets suitable for training, we process them to ensure that each data entry corresponds to text related to a single Wikipedia topic. In the Simple English Wikipedia dataset, topics are enclosed by the "= = =" symbol, with the related context in the following lines, and each topic-context pair separated by two newline symbols. We construct the training dataset by extracting each context as a single entry, resulting in 116,031 samples containing a total of 13,707,770 words. For the WikiText-2 dataset, topics are enclosed by a single "=" symbol, while subsections and subsubsections are enclosed by "= =" and "= = =" symbols, respectively. We process the dataset by extracting the text associated with each topic enclosed by a single "=" symbol, continuing until the next topic marker. This yields 629 samples with a total of 2,051,910 words.

To compare the datasets, we first inspect histograms of sample counts versus word counts. A sample corresponds to text associated with one topic, and the word count is computed as the number of consecutive character sequences separated



(a) Simple English Wikipedia Dataset



(b) WikiText-2 Dataset

Figure 1: Histogram of the number of data entries in each word count range. Most data entries contain between 32 and 127 words in the Simple English Wikipedia dataset, while most data entries contain between 1,024 and 8,191 words in the WikiText-2 dataset.

by whitespace (spaces, tabs, or newlines). The histogram in Figure 1a shows that most Simple English Wikipedia data entries contain between 32 and 127 words, while most WikiText-2 data entries contain between 1,024 and 8,191 words, as shown in Figure 1b. We also examine the ten most frequent words in each dataset. In Simple English Wikipedia, these are: “the”, “of”, “in”, “and”, “a”, “is”, “to”, “was”, “The”, and “for”. In contrast, the top-10 words in WikiText-2 are: “the”, “,”, “.”, “of”, “<unk>”, “and”, “in”, “to”, “a”, and “=”. Four of these ten tokens are not actual words, suggesting improper processing during text extraction. Since tokens such as “<unk>” do not appear in natural English, and isolated “,” and “.” are uncommon, we conclude that WikiText-2 may degrade the ability of a language model to learn proper English. Therefore, we discard the WikiText-2 dataset.

2.2 Processing the Simple English Wikipedia Dataset

The statistics of the Simple English Wikipedia dataset from the 100M training set of the BabyLM Challenge (Charpentier et al., 2025) are presented in Table 1, where word counts are computed as the number of consecutive character sequences separated by whitespace.

To further assess dataset quality, we examine entries across different word count ranges. Examples are shown in Figure 7 to Figure 11. Entries with fewer than 8 words (Figure 7) are mostly incomplete sentences and insufficient to describe a topic. In contrast, entries with 8 words or more (Figs. 8 to 11) typically consist of complete sentences and are adequate to describe a topic. Most entries fall within 8 to 511 words, consistent with the distribution shown in Figure 1a.

2.3 64–512 Word Subset of the Simple English Wikipedia Dataset

We construct a 64–512 word subset of the Simple English Wikipedia dataset by excluding entries with fewer than 64 words. The statistics of this subset are reported in the third row of Table 1. Each sample is padded or truncated to a maximum of 1,024 tokens. Note that these statistics are computed on the truncated entries. During tokenization and detokenization, isolated symbols such as “,” and “.” are concatenated to the preceding word, resulting in slightly lower word counts than the predefined minimum.

2.4 Tokenizer

For training on the Simple English Wikipedia Dataset, we construct a tokenizer using GPT-2 style byte-pair encoding (BPE) (Radford et al., 2019) trained on the 64–512 word subset. The tokenizer includes the special tokens `[PAD]`, `[UNK]`, `[MASK]`, and `[EOS]`.

To determine the vocabulary size, we plot the frequency distribution of unique words in the 64–512 word subset (Figure 2). We observe that 2,206 unique words occur more than 500 times, motivating the choice of a 2,048-token vocabulary, which is close to this number. We also examine the effect of vocabulary size on training loss. Results show that a 4-layer model with a larger vocabulary size tends to yield higher training loss (Figure 12 in Appendix B).

3 Language Model Architecture and Hyperparameters

The architecture of the transformer-based language model (Vaswani et al., 2017), including the first

Table 1: Statistics of word counts for the Simple English Wikipedia dataset and its subsets.

Dataset	*Mean*	*Median*	*Maximum*	*Minimum*	*Total words*	*Samples*
Simple English Wikipedia	118.14	58	9,423	1	13,707,770	116,031
64–512 Word Subset	183.29	126	660	61	9,654,100	52,671

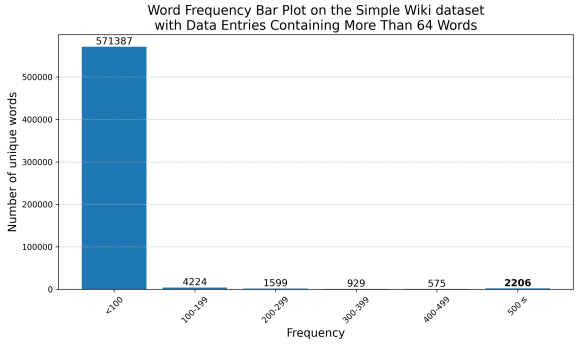


Figure 2: Bar plot of the number of unique words versus their frequency of occurrence in the 64–512 word subset of the Simple English Wikipedia dataset.

decoder block, is shown in Figure 3. The feed-forward network consists of 2,048 neurons, the attention mechanism uses 8 heads, and no dropout is applied during training. Other parameters are selected based on the experiments described in Appendix B. We found that the best-performing configuration is a decoder-only Transformer with rotary positional encoding, multi-head attention, root-mean-square normalization, GELU activation, post-normalization, no interleaved group query attention, an embedding dimension of 512, 8 layers, 8 heads, a feedforward dimension of 2,048, zero dropout, trained with a batch size of 2 and an initial learning rate of 10^{-4}.

We also observe that small models struggle to generate coherent English and reproduce factual knowledge for training samples with long contexts. To address this, we evaluated two training strategies. The first is a two-stage approach, where the model is first trained on short-context data and then on long-context data. Although this strategy produces coherent sentences, it fails to preserve factual knowledge. The second strategy interleaves short- and long-context datasets, alternating maximum sequence lengths each epoch. This mitigates catastrophic forgetting of short-context samples and produces coherent sentences, but similarly fails to preserve factual knowledge in long-context samples. The detailed results and analysis are presented in Appendix C.

4 Training

Given a set of training data $\mathcal{D} = \{d^1, d^2, \ldots, d^N\}$ with N samples, where the i-th sample $d^i = (d^i_1, d^i_2, \ldots, d^i_{k^i})$ contains k^i number of tokens, the loss for each sample is computed by the negative log-likelihood:

$$\mathcal{L}(\theta; d^i) = -\sum_{j=1}^{k^i} log(p_\theta(d^i_j | d^i_{j-1}, \ldots, d^i_1)), \quad (1)$$

where $p_\theta(d^i_j | d^i_{j-1}, \ldots, d^i_1) \in [0, 1]$ is the probability assigned by the transformer-based language model parameterized by θ to token d^i_j, given the preceding tokens $d^i_1, \ldots, d^i_{j-1}$.

The model is trained using standard mini-batch gradient descent as detailed in Algorithm 1.

Algorithm 1 Mini-Batch Gradient Descent

Input: Initial learning rate $\alpha \in \mathbb{R}$, momentum factors $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$, weight decay factor $\lambda \in \mathbb{R}$, $\epsilon \in \mathbb{R}$, batch size b, maximum epochs E, dataset $\mathcal{D} = \{d^i\}_{i=1}^N$, loss function $\mathcal{L}$, schedule multiplier $\eta_{t=0} \in \mathbb{R}$, time step $t \leftarrow 0$
Output: Optimized parameters θ
Initialize parameters θ randomly
for $ep = 1$ to E **do**
  Shuffle dataset $\mathcal{D}$
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ of size b **do**
    $t \leftarrow t + 1$
    $\eta_t \leftarrow$ SetScheduleMultiplier(t)
    Compute gradient:

$$\vec{g} \leftarrow \frac{1}{(\sum_{d^i \in \mathcal{B}} |d^i|)} \sum_{d^i \in \mathcal{B}} \nabla_\theta \mathcal{L}(\theta; d^i)$$

    Update parameters:

$$\theta \leftarrow \text{AdamW}(\vec{g}, \alpha, \beta_1, \beta_2, \epsilon, \lambda, \eta_t)$$

  end for
end for
return θ

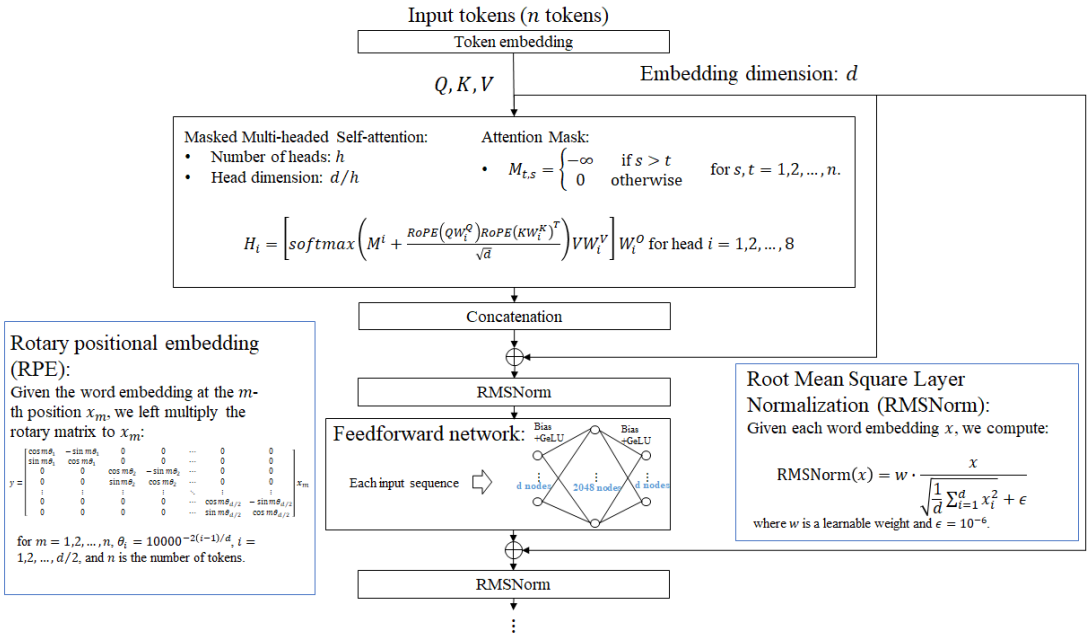


Figure 3: Illustration of the architecture of the initial part of the transformer-based language model used in this study, including the first decoder block.

The common training hyperparameters for all the experiments are $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and $\lambda = 0.01$ for the AdamW optimizer (Loshchilov and Hutter, 2019) and a maximum gradient norm of 1 for gradient clipping.

4.1 Results for Training on the 64–512 Word Subset

We train the 8-layer decoder-only transformer model described in Figure 3 with an embedding dimension of 512, a vocabulary size of 2,048 tokens, and 8 attention heads, using an initial learning rate of 10^{-4} with a linear scheduler. The training loss and learning rate under two different linear scheduling strategies are shown in Figure 4a and Figure 4b, respectively. As shown in Figure 4a, the model trained with a learning rate schedule that decays 10-fold every 64 epochs achieves a training loss below 0.1 at 169 epochs, whereas the model trained with a 10-fold decay every 128 epochs reaches the same threshold at 179 epochs. The blue dotted and dash-dotted lines indicate that the model with a 64-epoch decay reaches training losses below 0.5 and 0.2 at 58 and 94 epochs, respectively.

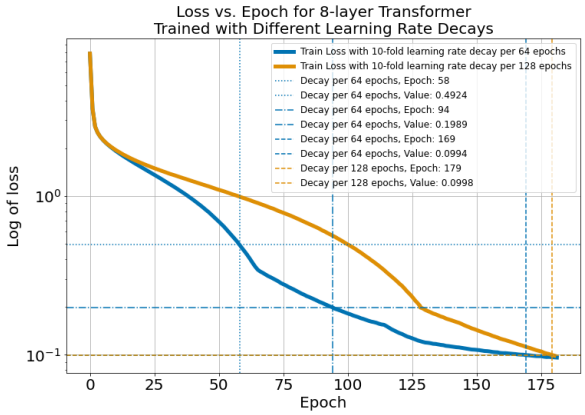
To evaluate the quality of model outputs across training stages, we focus on the model with 10-fold decay every 64 epochs, as it reaches a training loss below 0.1 faster than the model trained with a learning rate of 10-fold decay every 128 epochs. We select checkpoints corresponding to training losses of 0.5, 0.2, and 0.1, and inspect their outputs on selected training samples in Table 7 in Appendix D.

For reproducibility, Table 7 shows that the model trained to a loss of 0.1 successfully reproduces text from all three selected training entries (labeled "seen"). In contrast, the models trained to losses of 0.2 and 0.5 reproduce two of the three entries but fail on the third, suggesting that a threshold of 0.1 may be appropriate for the model to reliably reproduce training text.

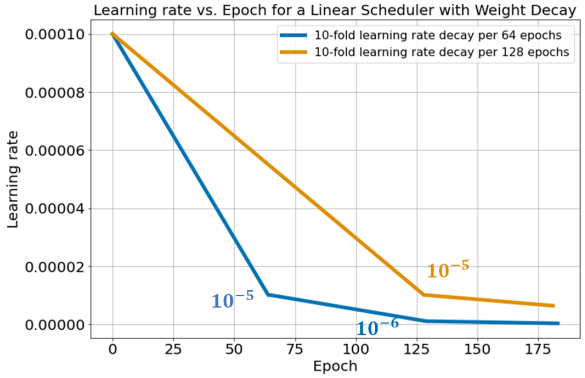
For generalization, the "rephrased" rows in Table 7 show that none of the models fully reproduce the target text when inputs are rephrased versions of the training samples. However, in the second rephrased example, the models trained to losses of 0.1 and 0.2 generate the correct date of Kahn's death, and the model trained to a loss of 0.5 even produces the correct age at death. The rephrased inputs were generated using ChatGPT-5 with the prompt: "Rephrase the following sentences with minimum changes: *input text*." These results suggest that generalization is difficult to achieve during pretraining. In the next section, we attempt fine-tuning on rephrased text to assess whether this improves generalization to unseen rephrasings.

5 Generalization

In this section, we aim to enable the model to generate coherent English for text that is rephrased from, but not present in, the training set. To build a



(a) Training loss on the 64–512 word subset



(b) Learning rate for training on the 64–512 word subset

Figure 4: Training loss and learning rate for an 8-layer transformer trained on the 64–512 word subset of the Simple English Wikipedia dataset. The blue lines correspond to a linear scheduler with an initial learning rate of 10^{-4} and a 10-fold decay every 64 epochs. The orange lines correspond to a 10-fold decay every 128 epochs. Dashed lines mark the epochs and loss values where each model reaches a training loss below 0.1. The dotted and dash-dotted blue lines mark the epochs where the model with 64-epoch decay reaches training losses below 0.5 and 0.2, respectively.

dataset for fine-tuning, we select 100 data entries with word counts between 64 and 127 words from the Simple English Wikipedia. Each entry is manually split into two parts: the first half as the context and the second half as the target. The context of each entry is then rephrased in three different ways using ChatGPT-5 with the following prompt:

> Rephrase the text in the following dictionary in 3 different ways and fill them in textR1, textR2, and textR3. Make minimum change and make sure it can be connected after the target: "text": <context>, "target": <target>, "textR1": "", "textR2": "", "textR3": ""...

For each entry, the first two rephrased contexts (textR1 and textR2) are included in the fine-tuning dataset, resulting in a dataset with a total of 16,242 words. The third rephrased context (textR3) is used as the validation/test set to evaluate generalization.

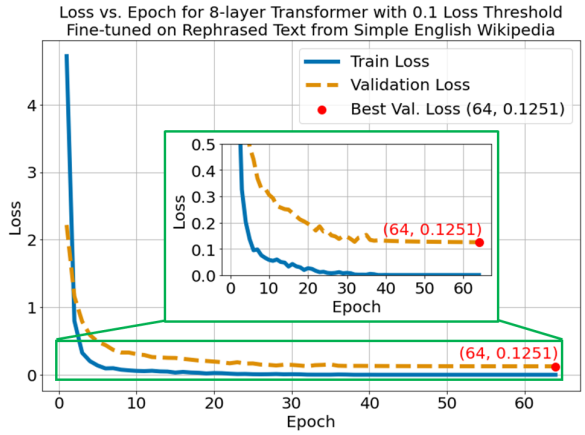
When calculating the loss, we mask out the tokens corresponding to the context so that predictions for context tokens are excluded from the loss computation during fine-tuning. We then fine-tune the 8-layer decoder-only transformer model pretrained with a learning rate schedule that decays 10-fold every 64 epochs and achieves a training loss of 0.1. The hyperparameters for fine-tuning are: AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and $\lambda = 0.01$; a maximum gradient norm of 1.0 for gradient clipping; a batch size of 2; a linear learning rate schedule decaying from 10^{-4} to 10^{-5} over 64 epochs; and a sequence length of 256 tokens. The shorter sequence length is chosen because the fine-tuning dataset contains no more than 128 words per entry (approximately $128 \times 2 = 256$ tokens). The training and validation losses are shown in Figure 5a, and the learning rate is shown in Figure 5b.

The model at epoch 64, which achieves the best validation loss of approximately 0.1251, is selected for evaluating generalization on the test set. To thoroughly evaluate all validation and test data, we compare the target and generated text using ChatGPT-5 with the following prompt as input:

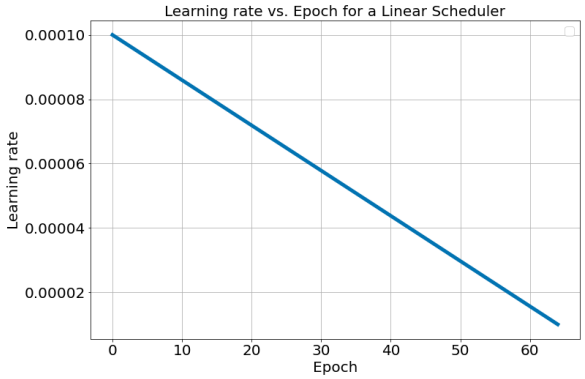
> Check whether the generated text conveys the same meaning as the target text. If it does, indicate the label number corresponding to that match.
> [Sample <sample label>]
> input_len: <number of input tokens>
> target_len: <number of target tokens>
> max_new_tokens: <number of maximum new tokens>
> Input: <input>
> Target: <target>
> Generated: <output>
> ...

The results show that the model can generate text that matches all 200 entries in the training set, and it can generate text that matches the target text in 42 out of 100 entries in the validation/test set. In addition, Table 2 presents the successful cases where the model generates text that matches the target in both the training and validation/test sets.

These results indicate that fine-tuning on rephrased data enables limited generalization in



(a) Training and validation loss for fine-tuning on rephrased text



(b) Learning rate for fine-tuning on rephrased text

Figure 5: Training and validation loss and learning rate for an 8-layer transformer fine-tuned on rephrased data entries (64–127 words) from the Simple English Wikipedia dataset. The model was pretrained to a loss of 0.1 on the 64–512 word subset of Simple English Wikipedia using a learning rate schedule with 10-fold linear decay every 64 epochs (Section 4.1). In subfigure (a), the blue line shows the training loss and the orange dashed line shows the validation loss. The best validation loss is highlighted with a red circle marker at epoch 64, with a value of 0.1251.

the pretrained model, while still producing text that is coherent and consistent with basic English.

6 Conclusion

In this work, we systematically explored several model architectures and training strategies to identify a transformer-based model with the minimum number of parameters upper bounded by 28 M capable of producing a broad spectrum of basic general knowledge in simple and coherent English, while exhibiting limited generalization ability. For dataset selection, we chose the Simple English Wikipedia dataset instead of the WikiText-2 dataset, since the latter contains a higher proportion of non-lexical tokens such as ‘<unk>’, ’=’, ’,’, and ’.’, among its most frequent tokens, suggesting that the text in WikiText-2 may not compose of coherent English. For data cleaning, we removed entries with fewer than 64 words, as they are often incomplete sentences, and entries with more than 512 words, to ensure that each topic description ends within the specified sequence length during training. The resulting dataset is referred to as the 64–512 word subset of the Simple English Wikipedia dataset.

To determine the vocabulary size, we set the number of tokens to approximately match the number of unique words appearing more than 500 times in the dataset. We then performed hyperparameter tuning on models with parameter counts up to 28 million. The best-performing model was found to be a decoder-only transformer with rotary positional encoding, multi-head attention, root-mean-square normalization, Gaussian error linear unit (GELU) activation, post-normalization, no interleaved group query attention, an embedding dimension of 512, 8 layers, 8 heads, a feedforward dimension of 2,048, zero dropout, and an initial learning rate of 10^{-4}. This configuration achieved a lower training loss compared to other tested models. We also evaluated different training strategies, including the two-stage and interleaved methods, but neither resulted in significant improvements in training loss.

We therefore trained the model using standard mini-batch gradient descent on the 64–512 word subset and experimented on linear learning rate schedulers with different decay rates. We found that a scheduler with an initial learning rate of 10^{-4} and a 10-fold decay every 64 epochs achieved a training loss of 0.1 faster than the corresponding scheduler with a 128-epoch decay interval. In generation tests, the model trained to a loss of approximately 0.1 was able to reproduce text from three selected training entries and demonstrated the ability to generate simple and coherent English covering broad basic knowledge, whereas models trained to losses of 0.2 and 0.5 performed worse. However, none of the models were able to reproduce target text when the inputs were rephrased using ChatGPT-5.

To enable limited generalization, we fine-tuned the model pretrained on the 64–512 word subset at a loss of 0.1. Specifically, we selected 100 data entries, splitting each into a context and a target. The context was rephrased in three different ways

Table 2: Generated text from the fine-tuned model with 8 layers, a vocabulary size of 2,048, an embedding dimension of 512, and 8 heads. The model was pretrained on the 64–512 word subset of Simple English Wikipedia using a learning rate schedule with 10-fold linear decay every 64 epochs and an initial learning rate of 10^{-4}. Input text labeled (validation) represents entries in the validation set, while input text labeled (test) represents rephrased entries from the validation set generated by ChatGPT-5. Outputs are generated using greedy decoding with a maximum of 128 new tokens. Common words between the target and generated text are highlighted in yellow.

Label	***Input***	***Target***	***Output***
1 (train)	Salko Hamzic (born 17 September 2006) is a Bosnian goalkeeper who plays for FC Liefering in Austria's second division. \n Career.\n He started his youth career at UFC Siezenheim before joining SV Austria Salzburg in December 2015, later moving to FC Red Bull Salzburg's academy in February 2019.	He then went through all age levels in the academy from the 2020/21 season.\n In May 2023 the goalkeeper was in the squad of the second-class farm team FC Liefering for the first time. For the 2023/24 season he moved into the Liefering squad. He made his debut in the 2nd league on 15 September 2023 when he was in the starting line-up on matchday seven of that season against SV Stripfing.	He then went through all age levels in the academy from the 2020/21 season.\n In May 2023 the goalkeeper was in the squad of the second-class farm team FC Liefering for the first time. For the 2023/24 season he moved into the Liefering squad. He made his debut in the 2nd league on 15 September 2023 when he was in the starting line-up on matchday seven of that season against SV Stripfing.ed to name a place before
1 (validation/test)	Salko Hamzic (born 17 September 2006) is a football goalkeeper from Bosnia who plays in Austria's 2nd league with FC Liefering. \n Career.\n He first played for UFC Siezenheim, then joined SV Austria Salzburg in December 2015, and later moved to FC Red Bull Salzburg's youth academy in February 2019.		He then went through all age levels in the academy from the 2020/21 season.\n In May 2023 the goalkeeper was in the squad of the second-class farm team FC Liefering for the first time. For the 2023/24 season he moved into the Liefering squad. He made his debut in the 2nd league on 15 September 2023 when he was in the starting line-up on matchday seven of that season against SV Stripfing.ed to name areashes
2 (train)	Muhammad Tawhidi, widely recognized online as the Imam of Peace, is a Shiite Imam and public influencer. Born in Qom, Iran, between 1982 and 1983, he became Vice President of the Global Imams Council in Najaf, Iraq in January 2022.	\nViews on Islam. His views on Islam are that Islam needs to be reformed to survive. He believes that terrorism are forbidden in the Quran, and made a speech denouncing the Islamic State of Iraq and Syria along with it's affilates such as Boko Haram.	\nViews on Islam.\n His views on Islam are that Islam needs to be reformed to survive. He believes that terrorism are forbidden in the Quran, and made a speech denouncing the Islamic State of Iraq and Syria along with it's affilates such as Boko Haram.or of this views, while various possible poses to the Islamic worldlying it best.\n Views on
2 (validation/test)	Known on the internet as the Imam of Peace, Muhammad Tawhidi is a Shiite Imam and social influencer. He was born in Qom, Iran, sometime between 1982 and 1983, and by January 2022 he was serving as Vice President of the Global Imams Council in Najaf, Iraq.		\nViews on Islam.\n His views on Islam are that Islam needs to be reformed to survive. He believes that terrorism are forbidden in the Quran, and made a speech denouncing the Islamic State of Iraq and Syria along with it's affilates such as Boko Haram. haram. haram has a creation and a net worship for both the Jews and HumLahr Z
3 (train)	Gearóid Iarla (1335–disappeared 1398), also called Gerald FitzMaurice FitzGerald, served as the 3rd Earl of Desmond, lord of Munster, and was a Norman-Gaelic poet. He vanished in 1398, aged 62 or 63.\n Biography.	\nHe was born in 1335, as the child (son) of Maurice FitzGerald, 1st Earl of Desmond. When he grew up, he became the 3rd Earl of Desmond, lord of Munster, and a Norman-Gaelic poet. He disappeared in 1398, at the age of 62 or 63.	\nHe was born in 1335, as the child (son) of Maurice FitzGerald, 1st Earl of Desmond. When he grew up, he became the 3rd Earl of Desmond, lord of Munster, and a Norman-Gaelic poet. He disappeared in 1398, at the age of 62 or 63.ons frail, and he became the father of Maurice FitzGerald, whose most interest in views on the
3 (validation/test)	Gearóid Iarla (1335–1398, disappeared), also referred to as Gerald FitzMaurice FitzGerald, became the 3rd Earl of Desmond, lord of Munster, and a Norman-Gaelic poet. He went missing in 1398 at the age of 62 or 63.\n Biography.		\nHe was born in 1335, as the child (son) of Maurice FitzGerald, 1st Earl of Desmond. When he grew up, he became the 3rd Earl of Desmond, lord of Munster, and a Norman-Gaelic poet. He disappeared in 1398, at the age of 62 or 63. 1798 or 1598, he became the father of Maurice FitzGerald, the first woman named Margaret

using ChatGPT-5. The fine-tuning dataset was constructed by pairing the rephrased contexts with their original targets: the first two rephrased versions formed the training set, the original context formed the validation set, and the third rephrased version formed the test set. After fine-tuning for 64 epochs, we selected the model with the lowest validation loss for testing. Results on five selected entries showed that the model reproduced the target text in two validation and two test cases, succeeded only in validation or test cases for one entry each, and failed in both validation and test sets for one entry. These results suggest that fine-tuning on rephrased text enables limited generalization ability. Furthermore, the generated outputs were mostly coherent English, achieving our objective. Future research is needed to investigate how the number of rephrased samples or alternative rephrasing methods affect the generalization ability of small language models.

Acknowledgments

The computational work for this project was conducted using resources provided by the Storrs High-Performance Computing (HPC) cluster. We extend our gratitude to the UConn Storrs HPC and its team for their resources and support, which aided in achieving these results.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clé-

mentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737. Unpublished.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Rasha Ahmad Husein, Hala Aburajouh, and Cagatay Catal. 2025. Large language models for code completion: A systematic literature review. *Computer Standards & Interfaces*, 92:103917.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Khalid Nassiri and Moulay Akhloufi. 2023. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9):10602–10635.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

B Sindhu, RP Prathamesh, MB Sameera, and S KumaraSwamy. 2024. The evolution of large language model: Models, applications and challenges. In *2024 international conference on current trends in advanced computing (ICCTAC)*, pages 1–8. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388. Unpublished.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

A Additional Analysis on the Datasets

A.1 Shared topics between the Simple English Wikipedia dataset and the WikiText-2 dataset

To check whether data entries in the WikiText-2 Dataset and Simple English Wikipedia Dataset share the same topics, we searched for each subject name in WikiText-2 across all texts in the Simple English Wikipedia Dataset. If the subject name appears in the Simple English Wikipedia Dataset, we will consider the subject as the shared topic. Out of a total of 629 subjects in WikiText-2, we found 97 subjects also present in the Simple English Wikipedia Dataset, as shown in Figure 6.

A.2 Data entries in the Simple English Wikipedia dataset

Part of the data entries of the Simple English Wikipedia dataset are shown from Figure 7 to Figure 11. From part of the data entries with word count less than 8 in Figure 7, most of the sentences are not complete, indicating low quality of the data entries within this word count range. While for part of the data entries with word count between 8 and 15 in Figure 8, it shows that the first 13 data entries are composed of complete sentences describing a place, but the later 3 data entries are incomplete. For part of the data entries with word count between 16 and 31 in Figure 9, the 15 selected data entries are composed of complete and sufficient sentences for describing a topic, and same of the data entries with word count between 32 and 63 in Figure 10 and data entries with word count between 64 and 127 in Figure 11.

B Determining the Hyperparameters for Model Architecture and Training

A 10M-word subset of the Simple English Wikipedia dataset is created by sampling the dataset with word counts between 8 and 511, stopping before the total word count reaches 10 million

1.John Cullen
2.South of Heaven
3.Tina Fey
4.Elephanta Caves
5.Michael Jordan
6.West End Girls
7.Sholay
8.Antimony
9.Astraeus hygrometricus
10.Paul Thomas Anderson
11.Art Ross
12.Sarnia
13.World War Z
14.Rachel Green
15.The Importance of Being Earnest
16.Ireland
17.Hellblazer
18.2010 Haiti earthquake
19.James Nesbitt
20.Rebbie Jackson
21.Protein
22.Aston Villa F.C.
23.Cadmium
24.Leg before wicket
25.Ceratopsia
26.Super Mario Land
27.Guitar Hero
28.Tintin in the Congo
29.Oldham
30.In Bloom
31.Giacomo Meyerbeer
32.Odaenathus
33.Bob Dylan
34.Mogadishu
35.Charmbracelet
36.Thomas Quiney
37.Transit of Venus
38.Roger Federer
39.The Son Also Draws
40.Humpty Dumpty
41.Welsh National Opera
42.England national rugby union team
43.Maggie Simpson
44.Chasing Vermeer
45.Yoko Shimomura
46.Lisa the Simpson
47.Xenon
48.Eva Perón
49.Kalyanasundara
50.Hoover Dam
51.Nina Simone
52.Harajuku Lovers Tour
53.The Stolen Eagle
54.Laurence Olivier
55.Burn
56.Alice in Chains
57.Cougar
58.Sorraia
59.Haifa
60.Fernando Torres
61.Dota 2
62.Djedkare Isesi
63.Christine Hakim
64.Gregory Helms
65.Amanita muscaria
66.Track and field
67.Isabella Beeton
68.Ed Barrow
69.Kitsune
70.Kakapo
71.Robbie Fowler
72.Erving Goffman
73.Cater 2 U
74.Gold Beach
75.Tautiška giesmė
76.Liu Kang
77.Allah
78.Chagas disease
79.DuMont Television Network
80.Skye
81.Florida Atlantic University
82.The Clean Tech Revolution
83.Missouri River
84.Ælfric of Abingdon
85.Condom
86.Iguanodon
87.Max Mosley
88.Corythosaurus
89.Wales national rugby union team
90.Ace Attorney
91.Varanasi
92.1973 Atlantic hurricane season
93.and
94.Partington
95.The General in His Labyrinth
96.Wilhelm Busch
97.Star

Figure 6: The shared topic names in the WikiText-2 dataset and the Simple English Wikipedia Dataset.

words. The statistics of this subset are reported in the third row of Table 3.

For training on the 10M-word subset, we construct a tokenizer using GPT-2 style byte-pair encoding (BPE) (Radford et al., 2019) trained on this subset. The tokenizer includes the special tokens `[PAD]`, `[UNK]`, `[MASK]`, and `[EOS]`. The total number of vocabularies is varied across experiments.

B.1 Results for Training on the 10M-word Subset

We perform hyperparameter tuning by training a decoder-only transformer-based language model with the architecture of the decoder block describe in Figure 3 by the Mini-Batch Gradient Descent in Algorithm 1 on the 10M-word subset of the Simple English Wikipedia dataset. The 10M-word subset of the Simple English Wikipedia dataset is created by sampling the dataset with word counts between 8 and 511, stopping before the total word count reaches 10 million words. The statistics of this subset are reported in the third row of Table 3. For training on the 10M-word subset, we construct a tokenizer using GPT-2 style byte-pair encoding (BPE) (Radford et al., 2019) trained on this subset. The tokenizer includes the special tokens `[PAD]`, `[UNK]`, `[MASK]`, and `[EOS]`. The total number of vocabularies is varied across experiments.

To determine the vocabulary size, we train three transformer models, each with 4 layers, 8 attention heads, and an embedding dimension of 512, but with vocabulary sizes of 2,048, 4,096, and 8,192, respectively. The learning rate is generated from the cosine scheduler with an initial learning rate of 10^{-4} as shown in the blue solid line in Figure 13(a). As shown in Figure 12, the model with the smallest vocabulary achieves the lowest training loss. Based on this result, we fix the vocabulary size to 2,048 for subsequent experiments.

To determine the optimal embedding dimension, number of layers, and learning rate, we evaluate models with embedding dimensions of 256 and 512, layer counts of $1, 2, 4, 8$, and an initial learning rates of 1×10^{-3}, 5×10^{-4}, 1×10^{-4}, and 5×10^{-5} for the cosine scheduler as shown in the green, orange, blue, and red solid line in Figure 13(a), respectively. Figure 14(a) shows that the 8-layer model achieves the lowest training loss of 1.2919 at epoch 10, and that larger embedding dimensions generally yield lower losses. Figure 14(b) compares the 4- and 8-layer models across different learning rates, revealing that the 8-layer model with an initial learning rate of 1×10^{-4} achieves the lowest loss (1.2919 at epoch 10). Based on these findings, we fix the configuration for subsequent experiments to 8 layers, an embedding dimension of 512, and an initial learning rate of 1×10^{-4}.

The generated text from the best model trained for 10 epochs is presented in Table 4. The model input is a truncated segment from the training dataset, and the expected output is the target text. As shown in Table 4, the model reproduces the target exactly for the first sample but fails to generate the correct number of people in the second sample, produces

1.Frosta can be
2.Pinh\u00e3o can mean:
3.CQD or cqd can be
4.Aral may refer to:
5.Igbo can mean:
6.Kongo can mean:
7.Wolof can mean:
8.Xhosa can mean:
9.Virac can mean:
10.Dido or DIDO may refer to:
11.FSA may refer to:
12.Moriarty may refer to:
13.Hillsborough is the name of several places:
14.Ellendale is the name of several places:
15.Leila can be
16.Stepove can be these settlements in Ukraine
17.Ivanivske can be
18.Lukas can be
19.Jacobsen can be
20.Leonard can be
21.V\u00e1clav is a given name.
22.Gaza War may refer to:
23.New Market may refer to:
24.This is a list of Dutch painters.
25.Vesele or Vesel\u00e9 can be
26.Pirita can be
27.Rauma can be
28.is a city in Osaka Prefecture, Japan.
29.is a town in Osaka Prefecture, Japan.
30.is a city in Osaka Prefecture, Japan.
31.Badajoz is a constituency of Extremadura.
32.C\u00e1ceres is a constituency of Extremadura.

Figure 7: Part of the data entries of the Simple English Wikipedia dataset with word count less than 8.

1.Champmotteux is a commune. It is in \u00cele-de-France in the Essonne department in north France.
2.Chatignonville is a commune. It is in \u00cele-de-France in the Essonne department in north France.
3.Chauffour-l\u00e8s-\u00c9tr\u00e9chy is a commune. It is in \u00cele-de-France in the Essonne department in north France.
4.Cheptainville is a commune. It is in \u00cele-de-France in the Essonne department in north France.
5.Chevannes is a commune. It is in \u00cele-de-France in the Essonne department in north France.
6.Chilly-Mazarin is a commune. It is in \u00cele-de-France in the Essonne department in north France.
7.Congerville-Thionville is a commune. It is in \u00cele-de-France in the Essonne department in north France.
8.Gjerstad is a municipality in Agder county, Norway. In 2022, 2,427 people lived there.
9.Ronago is a \"comune\" in the Province of Como in the Lombardy region in Italy.
10.Corbreuse is a commune. It is in \u00cele-de-France in the Essonne department in north France.
11.Courances is a commune. It is in \u00cele-de-France in the Essonne department in north France.
12.Courdimanche-sur-Essonne is a commune. It is in \u00cele-de-France in the Essonne department in north France.
13.Karachev () is a town in Bryansk Oblast, Russia. In 2010, 19,715 people lived there.
14.Justice is the morally fair treatment of people and things.\nJustice can also mean:
15.Enterprise (or the archaic spelling Enterprize) may refer to:
16.This is a list of Austrian football stadiums.

Figure 8: Part of the data entries of the Simple English Wikipedia dataset with word count between 8 and 15.

entirely incorrect text in the third sample, and correctly generates only 10 words in the fourth sample.

We further train the model for 50 and 100 epochs with an initial learning rate of 10^{-4} for the cosine scheduler to see if the prediction of the target text can be improved. The learning rates are plotted in Figure 13(a) with a purple and a brown solid line for the training of 50 and 100 epochs, respectively. The result in Figure 15 shows that both models achieve similar final training losses, with the 100-epoch model reaching a minimum loss of approximately 0.2259, compared to 0.3939 for the 50-epoch model. The generated outputs in Table 4 indicate that both models can accurately predict text for samples 1–3 when the target contains fewer words, but fail on sample 4, which has a longer target sequence. Additionally, we train the 8-layer model on samples truncated to a maximum sequence length of 512 for 50 epochs to evaluate its impact on training loss. As shown in Figure 15, this setting achieves a slightly lower loss of 0.3390 compared to training with a maximum sequence length of 1024.

C Exploring Training Methods for Alleviating Catastrophic Forgetting

The Simple English Wikipedia dataset is split into a short-context subset (samples with no more than 64 words) and a long-context subset (samples with more than 64 words) for the 2-stage and interleaved training methods. The statistics of these subsets are reported in the fourth and fifth rows of Table 3.

Table 3: Statistics of word counts for the Simple English Wikipedia dataset and its subsets.

Dataset	***Mean***	***Median***	***Maximum***	***Minimum***	***Total words***	***Samples***
Simple English Wikipedia	118.14	58	9,423	1	13,707,770	116,031
10M-word subset	86.47	56	511	8	9,515,583	110,047
Short context subset	31.51	29	64	1	1,996,707	63,360
Long context subset	222.34	126	9,423	65	11,711,063	52,671

Table 4: Generated text from the model with 8 layers, a vocabulary size of 2,048, an embedding dimension of 512, and 8 heads, trained on the 10M-word subset using a cosine scheduler with an initial learning rate of 10^{-4}. Common words between the target and the generated text are highlighted in yellow.

Label	***Input***	***Target***	**Trained epochs**	***Output***
1	Champmotteux is a commune. It is in Ile-de	-France in the Essonne department in north France.	10	-France in the Essonne department in north France.
			50	-France in the Essonne department in north France.
			100	-France in the Essonne department in north France.
2	Mati is a city in the Philippines. It is the capital of the province of Dava	o Oriental. According to the 2020 census, 147,547 people lived there.	10	o Oriental. According to the 2020 census, 104,490 people lived there.
			50	o Oriental. According to the 2020 census, 147,547 people lived there.
			100	o Oriental. According to the 2020 census, 147,547 people lived there.
3	Nuclear force is the force between nucleons. It is the force that pulls protons and neutrons into atoms. It is very hard to break the bond,	or tie, between protons and neutrons in an atom, because nuclear force holds them together. When the bond is broken, this is called nuclear fission.	10	and is very difficult to measure.
			50	or tie, between protons and neutrons in an atom, because nuclear force holds them together. When the bond is broken, this is called nuclear fission.
			100	or tie, between protons and neutrons in an atom, because nuclear force holds them together. When the bond is broken, this is called nuclear fission.
4	Hubert Miles Gladwyn Jebb, 1st Baron Gladwyn, GCMG, GCVO, CB, known as Gladwyn Jebb (25 April 1900 – 24 October 1996), was a prominent British civil servant, diplomat and politician as well as the first Acting Secretary-General of the United Nations. Acting UN Secretary-General. After World War II, he served as Executive Secretary of the Preparatory Commission of the United Nations in August 1945. He was appointed Acting United Nations Secretary-General from October 1945 to February 1946 until the appointment of the first Secretary-General Trygve Lie. Ambassador. Returning to London, Jebb was Deputy to the Foreign Secretary Ernest Bevin at the Conference of Foreign Ministers before serving as the Foreign Office's United Nations Adviser (1946-47). He represented the United Kingdom at the Brussels Treaty Permanent Commission with personal rank of Ambassador. He became the United Kingdom's Ambassador to the United Nations from 1950-1954 and to Paris from 1954-1960. Political career. In 1960 Jebb was made a heredit	ary peer and as Baron Gladwyn became involved in politics as a member of the Liberal Party. He was Deputy Leader of the Liberals in the House of Lords 1965-1988 and spokesman on foreign affairs and defence. An supporter of the European Union, he served as a Member of the European Parliament 1973-1976 where he was also the Vice-President of the Parliament's Political Committee. He tried to be elected to the European Parliament in 1979. When asked why he had joined the Liberal party in the early 1960s, he replied that the Liberals were a party without a general and that he was a general without a party. Like many Liberals, he passionately believed that education was the key to social reform. Death. He died in 1996, and is buried at St. Andrew's, Bramfield in the county of Suffolk. Lady Gladwyn. Jebb's wife, Cynthia, Lady Gladwyn, was a noted diarist of their times in Paris and a hostess of Liberal and London politics. She was the great-grand daughter of Isambard Kingdom Brunel. Publications and papers. Publications by Baron Gladwyn include:	10	ary peer and was appointed as the ambassador to the United Nations in 1962. He was appointed as the United Nations Secretary-General in 1962. He was appointed as the United Nations Secretary-General in 1963. He was the Secretary-General of the United Nations from 1976 to 1979. Death. On 24 October 1996, Jebb died of heart failure in London. He was 79 years old. He was buried at the Battle of London....
			50	ary peer and in 1973 he was a Member of Parliament (MP) for the Lim administration of British Union President General George B. Miles from September 1961 to July 1974. Death. Jebb died in October 1996 at the age of 88 in Nice, France of natural causes. He died in Nice on 24 October 1996 while in prison, he was buried at St Mary's Cemetery in Nice....
			100	ary peer and made a candidate a member of the Liberal Party. He was elected party leader to the Liberal Party. In 1961 Jebb was elected party leader of the Liberal Party. He served as the Secretary of State for National Unity (South Street) for two years. Personal life. Jebb married Rachel Lewis (1913-1990) in an alliance during the Secretary-General era. They had two children. Death. Gladwyn died in Lincolnshire on 24 October 1996 at the age of 74. He died from a heart attack, on 24 October 1996 in Charleroi, Mauritius. He was buried at Lincoln Cemetery in Rome, Italy. According to James Bond, Jebb was the most recent child of Jebbin and Lois Regnall. He was the last surviving person to have been head of the (Lord Mayor) House of Settlements since 1970....

1.Mati is a city in the Philippines. It is the capital of the province of Davao Oriental. According to the 2020 census, 147,547 people lived there.
2.Malita is a municipality in the Philippines. It is the capital of the province of Davao Occidental. According to the 2020 census, 118,197 people lived there.
3.Kabugao is a municipality in the Philippines. It is the capital of the province of Apayao. According to the 2020 census, 16,215 people lived there.
4.Conner is a municipality in the province of Apayao, Philippines. According to the 2020 census, 27,552 people lived there.
5.Crossodactylus cyclospinus is a frog. It lives in Minas Gerais, Brazil. People have seen it in exactly two places, both on the Jequitinhonha River.
6.Carate Urio is a \"comune\" in the Province of Como in the Lombardy region in Italy.
7.Gera Lario is a \"comune\" in the Province of Como in the Lombardy region in Italy.
8.Mariveles is a municipality in the province of Bataan, Philippines. According to the 2020 census, 149,879 people lived there.
9.Oroquieta is a city in the Philippines. It is the capital of the province of Misamis Occidental. According to the 2020 census, 72,301 people lived there.
10.Le Coudray-Montceaux is a commune. It is in \u00cele-de-France in the Essonne department in north France.
11.Old Occitan is a Old Romance language which is an early form of the Occitan language.
12.Mariana Avitia Mart\u00ednez (born September 18, 1993) is a Mexican archer. Avitia competed at the 2008 Summer Olympics and the 2012 Summer Olympics.
13.Ernesto Horacio Boardman L\u00f3pez (born 23 February 1993) is a Mexican archer. Boardman competed at the 2016 Summer Olympics.
14.The anthropenic shrub frog (\"Pseudophilautus hoipolloi\") is a frog. It lives in southwestern Sri Lanka. People have seen it between 15 and 684 meters above sea level.
15.Talkeetna (Dena'ina: \"K'dalkitnu\") is a census-designated place (CDP) in Matanuska-Susitna Borough, Alaska, United States. It began as a district headquarters of the Alaska Railroad in 1916.3

Figure 9: Part of the data entries of the Simple English Wikipedia dataset with word count between 16 and 31.

During the training, each short-context sample is padded or truncated to a maximum of 128 tokens, and each long-context sample is padded or truncated to a maximum of 1,024 tokens.

Separate tokenizers are constructed for the short- and long-context subsets using the GPT-2 style BPE (Radford et al., 2019) trained on each subset. Both tokenizers include the special tokens `[PAD]`, `[UNK]`, and `[EOS]`, with a vocabulary size of 2,048.

C.1 Results for the 2-stage training

To address the issue of incorrect predictions on long-context samples, we use a 2-stage training method. In the first stage, the model is trained on the short-context subset for 10 epochs; in the second stage, it is subsequently trained on the long-context subset for another 10 epochs. Figure 16 shows the training losses of 8-layer models trained with different batch sizes and an initial learning rate of 10^{-4} for the cosine scheduler as shown in the blue and green solid line in Figure 13(b). The best performance is achieved with a batch size of 2, yielding a minimum loss of approximately 0.5699 in the first stage and 1.5696 in the second stage. We further extend the training with a batch size of 2 to 50 epochs for each stage with the same initial learning rate of 10^{-4}, as shown in the orange and red dotted line in Figure 13(b). The results in Figure 17 show that the model trained for 50 epochs achieved a lower training loss of 0.1501 on the short-context subset and a training loss of 0.4359 on the long-context subset.

We inspect the generated text from the model trained with a batch size of 2 for 10 and 50 epochs on the long-context subset (Table 5). Both models fail to generate accurate text for either short- or long-context samples. However, the 50-epoch model produces a longer initial match, starting with " , he served as Executive Secretary of the", compared to the 10-epoch model. While these results indicate that the model can produce coherent sentences, the 2-stage training method does not resolve the long-context prediction problem and also introduces the issue of forgetting short-context samples.

C.2 Results for Interleaved Training

To address the problem of forgetting short-context samples, we train the model by interleaving datasets and context lengths, alternating each epoch between the short-context subset with a maximum sequence length of 128 tokens and the long-context subset with a maximum sequence length of 1024 tokens. The learning rate decays linearly from 10^{-4} to 10^{-5}, as shown in Figure 13(c). As shown in Figure 18, the lowest training loss is comparable to that of the 2-stage training model. The generated text in Table 6 shows that more words from the short-context subset (labels 1–3) are correctly

Table 5: Generated text from the 8-layer model (vocabulary size 2,048, embedding dimension 512, 8 heads) obtained from 2-stage training with a cosine scheduler and an initial learning rate of 10^{-4}. Common words between the target and the generated text are highlighted in yellow.

Label	*Input*	*Target*	Trained epochs	*Output*
1	Champmotteux is a commune. It is in	Ile-de-France in the Essonne department in north France.	20	the Auvergne-Rhone-Alpes region in the Ain department in the east of France. In 2009, 1,353 people lived in the commune. The inhabitants are the "Champm-lik"...
			100	the Allancesse of France in south-western France. It was the 1910s-25 km of the area for a long time...
2	Mati is a city in the Philippines. It is the capital of the province of	Davao Oriental. According to the 2020 census, 147,547 people lived there.	20	Mato. It is the largest city in the Philippines. It is on the banks of the Mboro Metro...
			100	Metroio, as of Augusto. By Augusto, it is the 12th largest city in the region with an urban population of 253,591, even though the largest city in Matiban is in the Fort Kanoa region. More than 80,000 people live in the city...
3	Nuclear force is the force between nucleons. It is the force that pulls protons and neutrons into atoms. It is very hard to break the bond,	or tie, between protons and neutrons in an atom, because nuclear force holds them together. When the bond is broken, this is called nuclear fission.	20	and is very hard to tell if it is in a massive nucleus. Nuclear force is used to make electricity, which is also used to make electricity. Nuclear force is used to electricity in the electricity.
			100	because all force hold fewer forces. There are some defects forces that are common enclosed to detect using the nucleuses, allowing explosion of the nuclear weapons...
4	Hubert Miles Gladwyn Jebb, 1st Baron Gladwyn, GCMG, GCVO, CB, known as Gladwyn Jebb (25 April 1900 – 24 October 1996), was a prominent British civil servant, diplomat and politician as well as the first Acting Secretary-General of the United Nations. Acting UN Secretary-General. After World War II	, he served as Executive Secretary of the Preparatory Commission of the United Nations in August 1945. He was appointed Acting United Nations Secretary-General from October 1945 to February 1946 until the appointment of the first Secretary-General Trygve Lie. Ambassador. Returning to London, Jebb was Deputy to the Foreign Secretary Ernest Bevin at the Conference of Foreign Ministers before serving as the Foreign Office's United Nations Adviser (1946-47). He represented the United Kingdom at the Brussels Treaty Permanent Commission with personal rank of Ambassador. He became the United Kingdom's Ambassador to the United Nations from 1950-1954 and to Paris from 1954-1960. Political career. In 1960 Jebb was made a hereditary peer and as Baron Gladwyn became involved in politics as a member of the Liberal Party. He was Deputy Leader of the Liberals in the House of Lords 1965-1988 and spokesman on foreign affairs and defence. An supporter of the European Union, he served as a Member of the European Parliament 1973-1976 where he was also the Vice-President of the Parliament's Political Committee. He tried to be elected to the European Parliament in 1979. When asked why he had joined the Liberal party in the early 1960s, he replied that the Liberals were a party without a general and that he was a general without a party. Like many Liberals, he passionately believed that education was the key to social reform. Death. He died in 1996, and is buried at St. Andrew's, Bramfield in the county of Suffolk. Lady Gladwyn. Jebb's wife, Cynthia, Lady Gladwyn, was a noted diarist of their times in Paris and a hostess of Liberal and London politics. She was the great-grand daughter of Isambard Kingdom Brunel. Publications and papers. Publications by Baron Gladwyn include:	20	, he was the Secretary-General of the United Nations, and was the first Acting Secretary-General of the United Nations. He was also the first Acting Secretary-General of the United Nations. Gladwyn was a member of the United Nations and was a member of the United Nations. He was also a member of the United Nations General Assembly from 1957 to 1961. He was the first Acting Secretary-General of the United Nations General Assembly from 1961 to 1962. He was the first Acting Secretary-General of the United Nations General Assembly from 1962 to 1963. He was the first Acting Secretary-General of the United Nations General Assembly from 1963 to 1964. Gladwyn served as the Acting Secretary-General of the United Nations General Assembly from 1964 to 1965. Gladwyn died on 24 October 1996 in Washington, D.C., aged 77. His wife, Jebb, was the first Acting Secretary-General of the United Nations. He was the first Acting Secretary-General of the United Nations General Assembly. He was the first Acting Secretary-General of the United Nations General Assembly from 1957 to 1961. Gladwyn died on 24 October 1996 in Washington, D.C. from a stroke, aged 87. ... His daughter, Prince Edward, was the second son of the United Nations General Assemblyman James Calvin. His daughter, Prince Edward, was the first Acting Secretary-General of the United Nations General Assembly. Honors. ...
			100	, he served as Executive Secretary of the Presbyterian Church between 1954 and 1956, and possibly the oldest living United Nations High Commissioner of the Presbyterian ministry at Harvard-War, Leeds, Plymouth, Plymouth and Edinburgh, and the then 221 specializing in making special wides within just one monthly. He served on a short-lived voter for that post until the Secretary General's death on 19 December 2017, who was the last vote owner of the United Nations, serving as UN Secretary General. He was also the last Vatgenian legal secretary, to be the last Governor-General of North South America. ... On 11 December 1963, he wasaring a life peer during a flight as an MCC (MAC-North Part links). The MCC-North Part links were signed by the UN as an independently handled burglaries, with the then 645 million voluns and 1.07 million voluns and in 2002 more was merged with a monthly. The United Kingdom government closed the network and made an effort to remove certain riots. The UN forces split in two pieces were to create the MCC-North Part links: the UN forces spread it for more than a million week and the UN forces spread on Strathcle, but Glidarling of the United Kingdom spread it for more than a million volun participating in the military. Galli prodignantly hit the UN's ship caused the bullet to spend on his tour in the UK. He died on 24 October 1996 at the age of 94. Executive Disputed Shropshire Declaration. Personal Lectures. ... He is best known for his forces about the Corruption Act of 1965. He is also known for his net worth of $3.5 million, which was also US$3.2 million; for a total of $3.5 million, he had to pay a $2.2 million loan and $2.4 million from the U.S. Majority Whip during his time as the U.S. Minister of Defence. He was also one of the only African American ladies and the third African-American to ever lunar weapon tried to fix the disintegration of the lady, Janey Evans. United States. ...

Table 6: Generated text from the 8-layer model (vocabulary size 2,048, embedding dimension 512, 8 heads) obtained from interleaved training with a linear scheduler and an initial learning rate of 10^{-4}. Common words between the target and the generated text are highlighted in yellow.

Label	***Input***	***Target***	**Trained epochs**	***Output***
1	Champmotteux is a commune. It is in	Ile-de-France in the Essonne department in north France.	20	Bourgogne-Franche-Comte in the Doubs department in east France. The commune is in the southern suburbs of Paris, from the center...
2	Mati is a city in the Philippines. It is the capital of the province of	Davao Oriental. According to the 2020 census, 147,547 people lived there.	20	Davao Oriental. According to the 2020 census, 106,409 people lived there. Mati is known for its merchants and casteous mouths,...
3	Nuclear force is the force between nucleons. It is the force that pulls protons and neutrons into atoms. It is very hard to break the bond,	or tie, between protons and neutrons in an atom, because nuclear force holds them together. When the bond is broken, this is called nuclear fission.	20	and is not forced to stabilize. It is the part of nuclear weapons. It is a part of the bond that is not far away. Nuclear forces are also very strong. Nuclear forces are not fully heated. Nuclear forces are not fully heated. Nuclear forces are not fully heated...
4	Hubert Miles Gladwyn Jebb, 1st Baron Gladwyn, GCMG, GCVO, CB, known as Gladwyn Jebb (25 April 1900 – 24 October 1996), was a prominent British civil servant, diplomat and politician as well as the first Acting Secretary-General of the United Nations. Acting UN Secretary-General. After World War II	, he served as Executive Secretary of the Preparatory Commission of the United Nations in August 1945. He was appointed Acting United Nations Secretary-General from October 1945 to February 1946 until the appointment of the first Secretary-General Trygve Lie. Ambassador. Returning to London, Jebb was Deputy to the Foreign Secretary Ernest Bevin at the Conference of Foreign Ministers before serving as the Foreign Office's United Nations Adviser (1946-47). He represented the United Kingdom at the Brussels Treaty Permanent Commission with personal rank of Ambassador. He became the United Kingdom's Ambassador to the United Nations from 1950-1954 and to Paris from 1954-1960. Political career. In 1960 Jebb was made a hereditary peer and as Baron Gladwyn became involved in politics as a member of the Liberal Party. He was Deputy Leader of the Liberals in the House of Lords 1965-1988 and spokesman on foreign affairs and defence. An supporter of the European Union, he served as a Member of the European Parliament 1973-1976 where he was also the Vice-President of the Parliament's Political Committee. He tried to be elected to the European Parliament in 1979. When asked why he had joined the Liberal party in the early 1960s, he replied that the Liberals were a party without a general and that he was a general without a party. Like many Liberals, he passionately believed that education was the key to social reform. Death. He died in 1996, and is buried at St. Andrew's, Bramfield in the county of Suffolk. Lady Gladwyn. Jebb's wife, Cynthia, Lady Gladwyn, was a noted diarist of their times in Paris and a hostess of Liberal and London politics. She was the great-grand daughter of Isambard Kingdom Brunel. Publications and papers. Publications by Baron Gladwyn include:	20	, he was a Lieutenant Colonel in the Royal Navy, and served as a Lieutenant colonel in the British Army. He was also a member of the British Royal Family. He was also a member of the British Royal Family. He was a member of the House of Lords. He was a cousin of former President of the United Nations High Commissioner for Human Rights. He was also a member of the House of Lords. He was a member of the House of Lords. He was a cousin of former Prime Minister Henry G. Harding, and he was a member of the House of Lords. He was a member of the Labour Party. He was a member of the House of Lords from 1947 until his death in 1996. He was a member of the House of Lords. He was a cousin of former Prime Minister Lord Human Rights leader Mike Walsh. He was a member of the Lieutenant Governorship and was a member of the Lieutenant Governorship. He was a cousin of former Prime Minister Lord Human Rights leader Mike Walsh. He was a cousin of former Prime Minister Lord Human Rights leader Mike Walsh. He was a member of the Lieutenant Governorship and was a member of the Lieutenant Governorship. He was a cousin of former Prime Minister Lord Human Rights leader Mike Walsh. He was a cousin on the Lieutenant Governorship. He was a cousin of former Prime Minister Lord Human Rights leader Mike Walsh. He was a cousin on the Lieutenant Governorship. He was a cousin on the Lieutenant Governorship. He was a cousin on the Lieutenant Governorship...

1.Consiglio di Rumo was a \"comune\" in the Province of Como in the Lombardy region in Italy. On 11 February 2011, the former municipalities of Consiglio di Rumo, Germasino and Gravedona merged to form the new municipality of Gravedona ed Uniti.
2.Gravedona was a \"comune\" in the Province of Como in the Lombardy region in Italy. On 11 February 2011, the former municipalities of Consiglio di Rumo, Germasino and Gravedona merged to form the new municipality of Gravedona ed Uniti.
3.Germasino was a \"comune\" in the Province of Como in the Lombardy region in Italy. On 11 February 2011, the former municipalities of Consiglio di Rumo, Germasino and Gravedona merged to form the new municipality of Gravedona ed Uniti.
4.Gravedona ed Uniti is a \"comune\" in the Province of Como in the Lombardy region in Italy. It was created on 11 February 2011 from the former municipalities of Consiglio di Rumo, Germasino and Gravedona.
5.In English, the word \"free\" has two meanings, which are very different from each other:\nRichard Stallman summarised the difference in a slogan: \"Think free as in free speech, not free beer.\"
6.Manly is a suburb of Auckland and is located in the Whangaparaoa Peninsula. It has a primary school, an established shopping center and residential areas of Big Manly Beach in the north and Little Manly Beach in the south. It was once a seaside holiday location until it became within commuting distance of Auckland City.
7.Johann Neumann was an Austrian footballer. He played for Wiener AC.\nInternational.\nHe played in eight matches for the Austria national football team from 1911 to 1923, scoring two goals. His first match was on 10 September 1911 in a 2-1 away win versus Germany. His last match was on 10 June 1923 in al away loss versus Sweden.
8.Russell Eric Wilson Mawhinney was born 28 March, 1960 in Ranfurly. He was a New Zealand cricketer who played for Northern Districts, Griqualand West and Otago in first-class cricket. He is currently married to TVNZ presenter Matty McLean.

Figure 10: Part of the data entries of the Simple English Wikipedia dataset with word count between 32 and 63.

1.Dolby Laboratories, Inc. (often known simply as Dolby) is a company specializing in audio noise reduction, audio encoding/compression, spatial audio, and High-dynamic-range television imaging. Dolby licenses its technologies to consumer electronics manufacturers. \nIt was founded by Ray Dolby (1933\u20132013) in London, England, in 1965. He moved the company headquarters to San Francisco in 1976. \nThe first movie with Dolby sound was \"A Clockwork Orange\" in 1971.
2.Salko Hamzic (born 17 September 2006) is a Bosnian football goalkeeper. He plays for Austrian 2nd league club FC Liefering.\nCareer.\nHamzi\u0107 began his career at UFC Siezenheim. In December 2015 he moved to SV Austria Salzburg. In February 2019 he moved to FC Red Bull Salzburg's youth team. He then went through all age levels in the academy from the 2020/21 season.\nIn May 2023 the goalkeeper was in the squad of the second-class farm team FC Liefering for the first time. For the 2023/24 season he moved into the Liefering squad. He made his debut in the 2nd league on 15 September 2023 when he was in the starting line-up on matchday seven of that season against SV Stripfing.
3.Muhammad Tawhidi, known online as the Imam of Peace is a Shiite Imam and Influencer. He was born in Qom, Iran in between the time period of 1982-1983. As of January 2022, Tawhidi has served as the Vice President for the Global Imams Council in Najaf, Iraq.\nViews on Islam.\nHis views on Islam are that Islam needs to be reformed to survive. He believes that terrorism are forbidden in the Quran, and made a speech denouncing the Islamic State of Iraq and Syria along with it's affilates such as Boko Haram.
4.The Castrovirreyna Province is one of seven provinces in the Huancavelica Region of Peru. The capital of this province is the city of Castrovirreyna.\nGeography.\nThe Chunta mountain range traverses the province. Some of the highest peaks of the province are listed below:\nPolitical divisions.\nThe province is divided into thirteen districts, which are:\nEthnic groups.\nIndigenous citizens of Quechua descent live in this place. Spanish is the language which the majority of the population (77.20%) learnt to speak in childhood. 22.30% of the residents started speaking using the Quechua language (2007 Peru Census).

Figure 11: Part of the data entries of the Simple English Wikipedia dataset with word count between 64 and 127.

predicted compared to the 2-stage training results in Table 5, indicating that interleaved training can mitigate the forgetting problem for short-context samples and the model can produce coherent sentences for both short and long contexts. However, the model still fails to produce correct predictions for samples in the long-context subset.

D Generated Text from the Model Trained on the 64–512 Word Subset

Table 7: Generated text from the model with 8 layers, a vocabulary size of 2,048, an embedding dimension of 512, and 8 heads, trained on the 64–512 word subset of Simple English Wikipedia using a learning rate schedule with 10-fold linear decay every 64 epochs and an initial learning rate of 10^{-4}. Input text labeled (seen) corresponds to training samples, while text labeled (rephrased) corresponds to rephrased versions of the seen inputs generated by ChatGPT-5. Outputs are generated using greedy decoding. Common words between the target and generated text are highlighted in yellow.

Label	***Input***	***Target***	**Loss Threshold**	***Output***
1 (seen)	"Blue Moon" is a 1934 song recorded by Richard Rodgers and Lorenz Hart and has become a standard jazz ballad. It was hit single in 1935, 1949, Elvis Presley released his version of "Blue Moon" for his 1956 album "Elvis Presley". It became a huge hit for The Marcels in 1961 as an international number one hit single.	It has been covered by numerous artists over the years.	0.1	It has been covered by numerous artists over the years. It was certified diamond by the Recording of the song earned more than 30 million reviews than praised it.. . .
			0.2	It has been covered by numerous artists over the years. It was covered by the then-latch hit "Clash of Lions" in 1964.. . .
			0.5	It has been covered by numerous artists over the years. It was covered by the 1964 movie "Wonder Years".. . .
1 (rephrased)	"Blue Moon" is a 1934 song by Richard Rodgers and Lorenz Hart that has become a standard jazz ballad. It was a hit single in 1935 and 1949. Elvis Presley released his version on his 1956 album "Elvis Presley". The song became an international number one hit for The Marcels in 1961.		0.1	It also reached the sops from 12 studio albums, 1961ing their version of "The Marshall Show".. . .
			0.2	It also reached the top five in her the U.S. It ended in February 1979.. . .
			0.5	It also reached the top five in her the U.S. The song is a semi-fatale American standard. It is played by John Steve.. . .
2 (seen)	Claude Kahn (9 November 1935 – 17 November 2023) was a French classical pianist, internationally known for his interpretations of especially the music of Chopin, but also of French music (Debussy, Fauré, Ravel) as soloist or accompanied by great orchestras in the world. He founded and directed a piano competition in 1970, to become international as the Concours International de Piano Claude Kahn. He founded the conservatoire of Antibes in 1971 and directed it until 1991.	He died in 17 November 2023, eight days after his 88th birthday.	0.1	He died in 17 November 2023, eight days after his 88th birthday. Kahn was appointed chairman of the Department of Modern Artists and in charge of three and most influential American Theatre studies, one of the best modern of Romanian musical theatre works in history. . . .
			0.2	He died in 17 November 2023, eight days after his 88th birthday. Kahn was appointed Commander of the Ordre national Group in Paris on 8 November 2023.
			0.5	He died in 17 November 2023, eight days after his 88th birthday. Kahn was appointed chairman of the Department of Modern Artists and in charge of three and most influential American Theatre studies, one of the best modern of Romanian musical theatre works in history. . . .
2 (rephrased)	Claude Kahn (9 November 1935 – 17 November 2023) was a French classical pianist, internationally recognized for his interpretations of Chopin and French composers such as Debussy, Fauré, and Ravel, both as a soloist and with major orchestras. In 1970, he founded and directed a piano competition that later became the Concours International de Piano Claude Kahn. He also established the conservatoire of Antibes in 1971 and directed it until 1991.		0.1	In 1988, he was elected public broadcast mette, internationally with 13 other public interests, on behalf of Frederic Norman.Kahn died on 17 November 2023, at the age of 87.. . .
			0.2	In 1997, he was elected "C member of the National Orchestes", representing the particularly the fifth national anthem of Chopin and Faure.\n Kahn died on 17 November 2023 in Nimes, France at the age of 87. His cause of death was plouguin. . . .
			0.5	Kahn died on 17 November 2023, at the age of 88. the group was formed out of More than two, than a, Cesa 19, which after its creation, by the medical research.. . .
3 (seen)	Wael Ghonim born 23 December 1980) is an Internet activist and computer engineer with an interest in social entrepreneurship. Awards.\n Ghonim topped "Time" magazine's yearly list of the world's 100 most influential people. On 26 April, he arrived in New York to be honored at the 2011 Time 100 Gala ceremony where he began his speech with a moment of silence to mark those killed in protests around the Arab world.	On 3 May, World Press Freedom Day, Wael Ghonim was awarded with the Press Freedom prize from the Swedish division of Reporters Without Borders.\n Ghonim also received the JFK Profile in Courage Award. On 23 May, Caroline Kennedy, daughter of President John F. Kennedy, presented the awards to Elizabeth Redenbaugh and Wael Ghonim, who was named a recipient on behalf of "the people of Egypt". Kennedy said she could think of no better recipients.	0.1	On 3 May, World Press Freedom Day, Wael Ghonim was awarded with the Press Freedom prize from the Swedish division of Reporters Without Borders.\n Ghonim also received the JFK Profile in Courage Award. On 23 May, Caroline Kennedy, daughter of President John F. Kennedy, presented the awards to Elizabeth Redenbaugh and Wael Ghonim, who was named a recipient on behalf of "the people of Egypt". Kennedy said she could think of no better recipients. . . .
			0.2	On 3 May, World Press Freedom Day, Wael Ghonim was awarded with the Press Freedom prize from the Swedish division of Reporters Without Borders.\n Ghonim also received the JFK Profile in Courage Award. On 23 May, his JFK Press and world-choice are awarded for the best Reporters Woman and Winnipeg Jesender.\n On 7 June, Head Atlantais signed a agreement with his successor Eisenhower "Despri des India" . . .
			0.5	His other work has also been worldwide. It won the Silver Spring, Favorite daily aesthetics competition in 2014 for aesthetics.\n In 2015 Ghonim was listed as the numberGoma and Asafus into the top ten of the Safari are listed as critical of Srbish and Internet Force, the US, and Malmora. . .
3 (rephrased)	Wael Ghonim (born 23 December 1980) is an Internet activist and computer engineer with an interest in social entrepreneurship. He topped Time magazine's list of the world's 100 most influential people. On 26 April 2011, he arrived in New York to be honored at the Time 100 Gala, where he began his speech with a moment of silence for those killed in protests across the Arab world.		0.1	One of the books was founded at Saint Paul and ghosts' 1984 hit single-seat conspiracy alphabet before joiningrad and SARC PIN. Most important books in His work include "The Psychic Coll: The New Cast, Noble System, and the SAS".. . .
			0.2	His first book, "What I Home My Inten Home", was published inight monthly by Fortune 5 and all report in connection with the CIST on Week Whizote.\n Background.\n Gala joined BBC in 2007. He stepped down for his preservation in Fortune 5 before releasing it. The BBC then stepped down from BBC when his brother, Ronnie and Billie all duplicate about it. Gala currently started BBC Workshop. . .
			0.5	His first book, "What I Song, African Anders" was published. A daily news about IQos was shown. It has been described as having a hard time going to be hit. The book describes the effects of computers and services.. . .

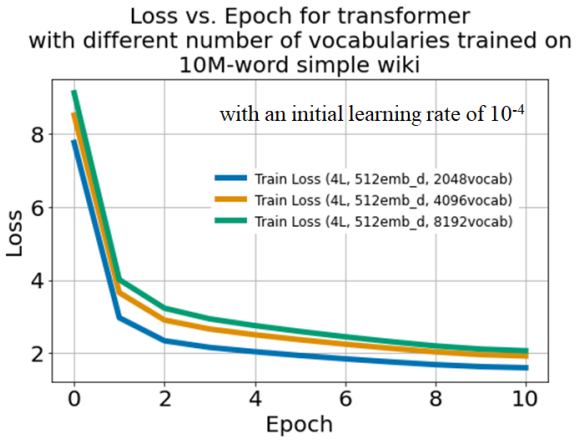


Figure 12: Training loss of three models trained on the 10M-word subset with vocabulary sizes of 2,048, 4,096, and 8,192, using the same hyper-parameters: embedding dimension 512, 4 layers, 8 heads, and an initial learning rate of 10^{-4} with a cosine scheduler. The lowest training loss decreases as the vocabulary size decreases, with the 2,048-token model achieving the lowest loss of 1.5978.

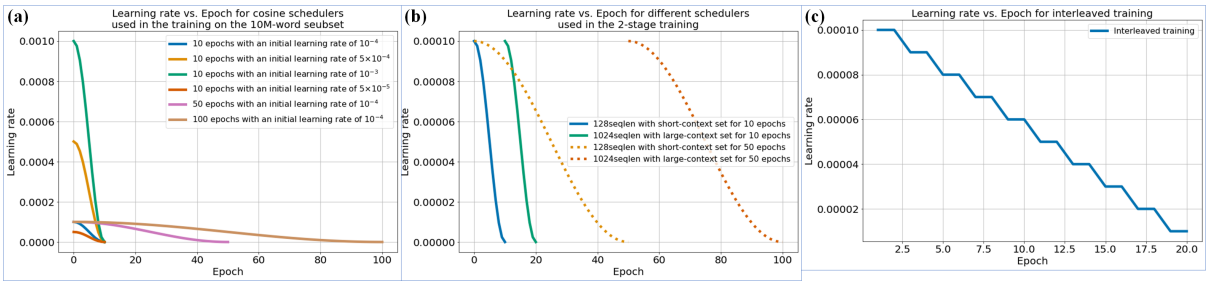


Figure 13: (a) Learning rate schedules from the cosine scheduler for different epochs and initial learning rates. (b) Learning rate schedule used in 2-stage training. (c) Learning rate schedule used in interleaved training.

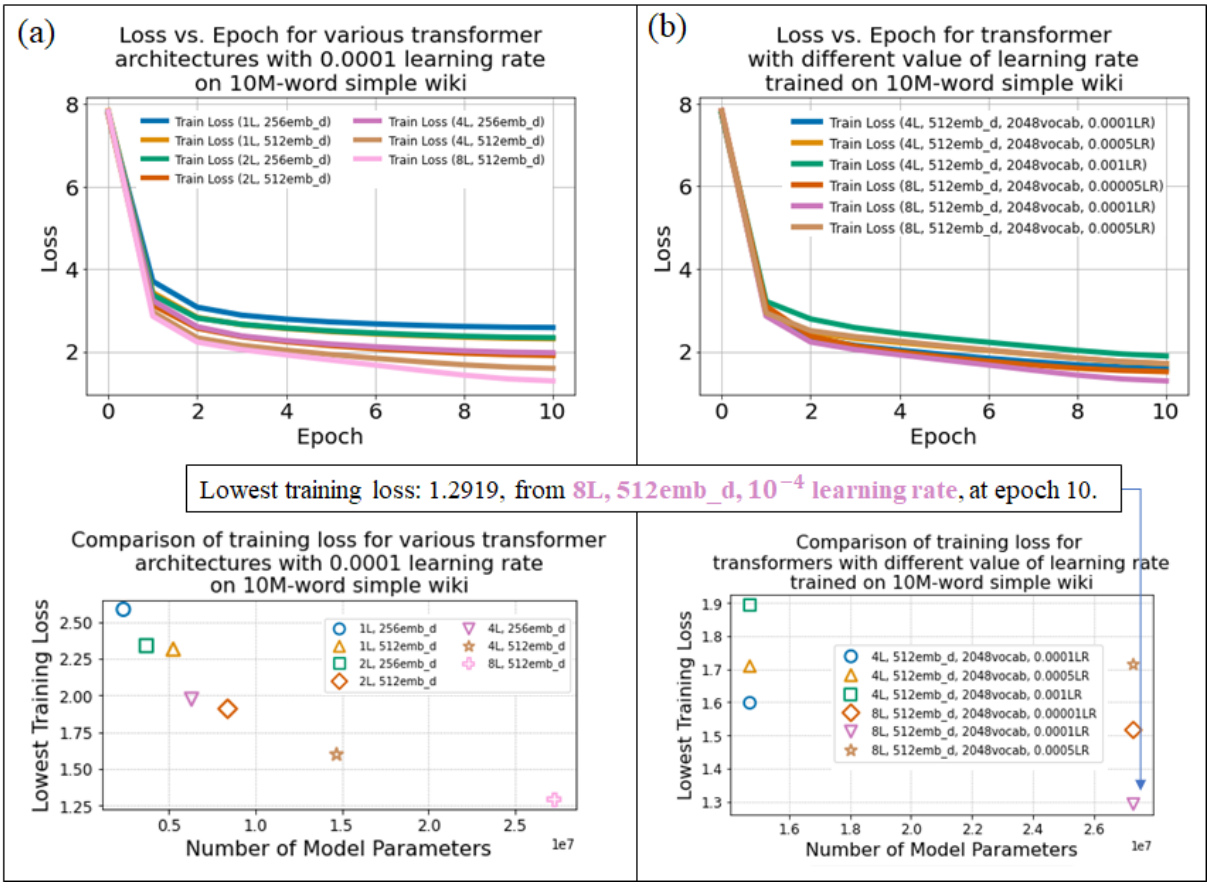
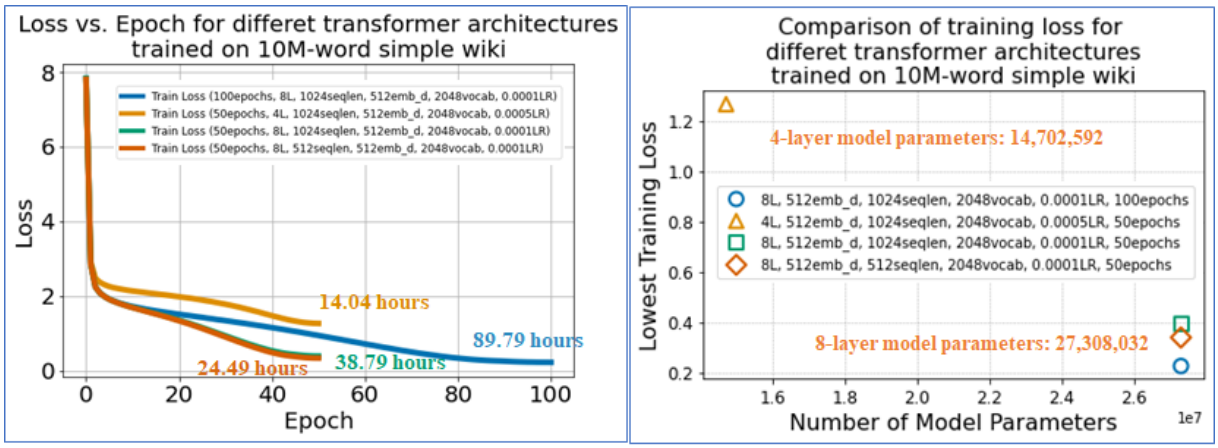


Figure 14: (a) Top: Training loss of models trained on the 10M-word subset with embedding dimensions of 256 or 512, layer counts of 1, 2, 4, or 8, and identical hyper-parameters: vocabulary size 2,048, 8 heads, and an initial learning rate of 10^{-4} with a cosine scheduler. Bottom: Best training loss from the top plot versus number of model parameters, showing that increasing the number of layers and embedding dimension reduces training loss at the cost of more parameters. (b) Top: Training loss of models trained on the 10M-word subset with initial learning rates of 1×10^{-3}, 5×10^{-4}, 1×10^{-4}, or 5×10^{-5}, layer counts of 4 or 8, and identical hyper-parameters: vocabulary size 2,048, embedding dimension 512, and 8 heads. Bottom: Best training loss from the top plot versus number of model parameters, showing that the 8-layer model with an initial learning rate of 1×10^{-4} achieves the lowest loss (1.2919).



Lowest training loss: 0.2259, from 8L, 512emb_d, 10^{-4} learning rate, at epoch 100.

Figure 15: (Left) Training loss of models trained on the 10M-word subset with embedding dimensions of 256 or 512, 4 or 8 layers, and varying numbers of epochs, using identical hyper-parameters: vocabulary size 2,048, 8 heads, and an initial learning rate of 10^{-4} with a cosine scheduler. (Right) Best training loss from the left plot versus number of model parameters. The 8-layer model trained for 100 epochs achieves the lowest loss of 0.2259.

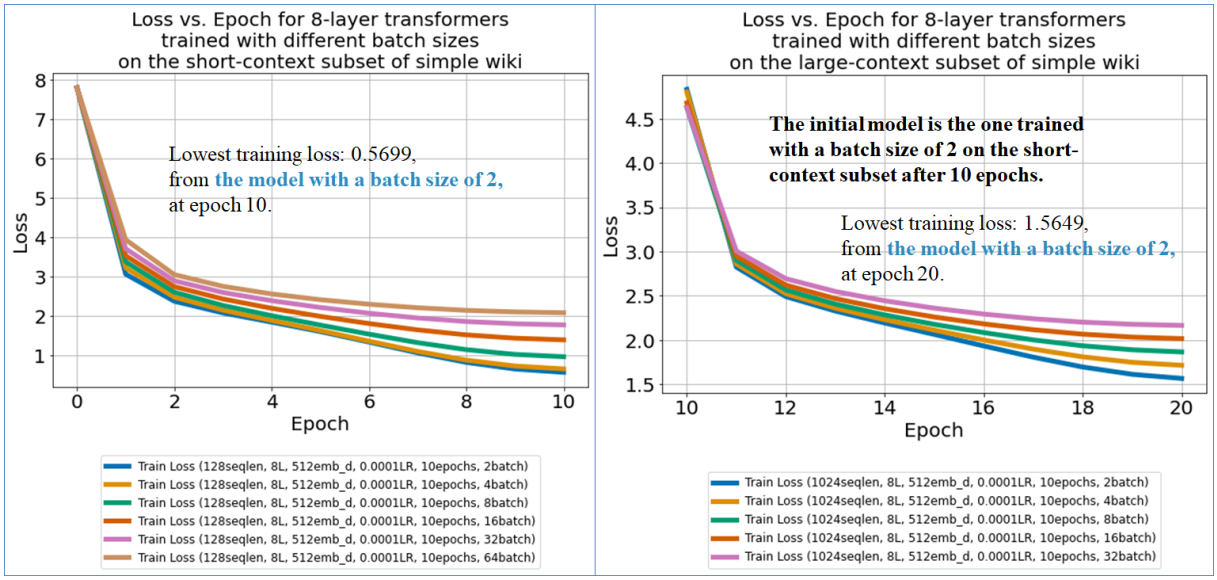


Figure 16: (Left) Training loss of the same model trained on the short-context subset with batch sizes ranging from 2 to 64. (Right) Training loss of the same models evaluated on the long-context subset after training on the short-context subset with different batch sizes. The model trained with a batch size of 2 achieves the lowest loss of 0.5699 on the short-context subset and 1.5649 on the long-context subset. All models share the same architecture: embedding dimension 512, 8 layers, vocabulary size 2,048, and 8 heads, trained with an initial learning rate of 10^{-4} using a cosine scheduler.

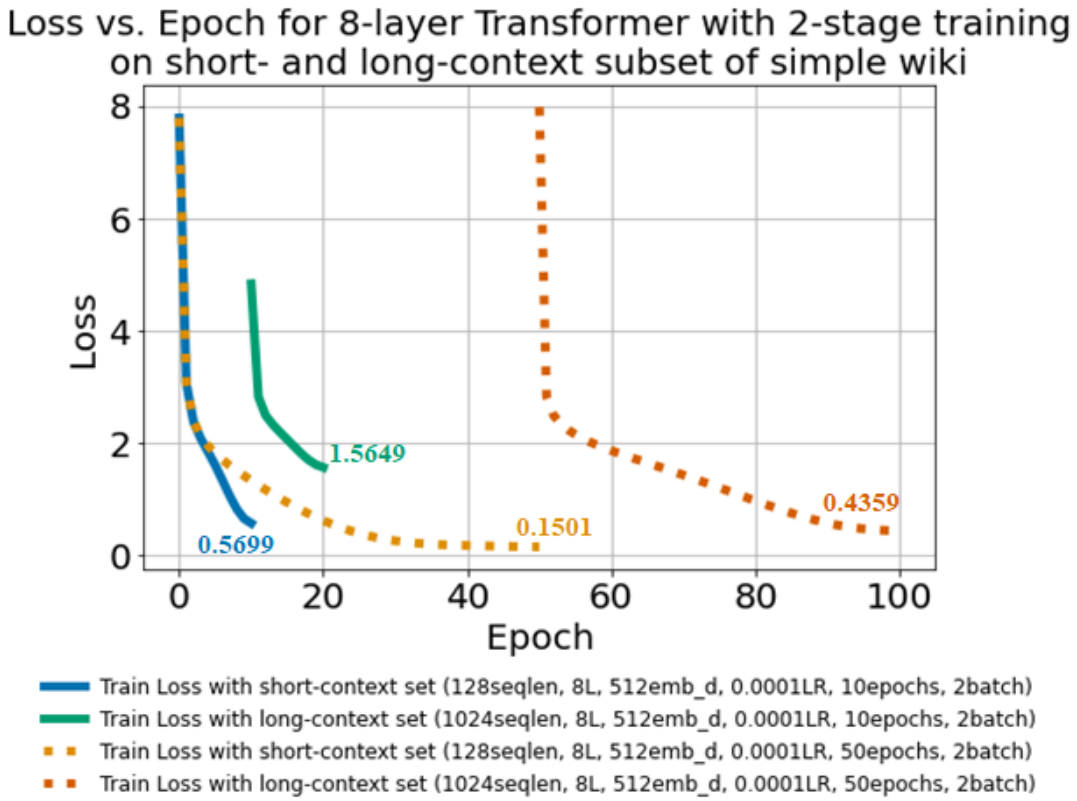


Figure 17: Training loss of 8-layer transformer models from 2-stage training with different numbers of epochs. Solid lines indicate training for 10 epochs on the short- and long-context subsets, and dotted lines indicate training for 50 epochs. The 50-epoch model achieves lower losses (0.1501 on the short-context subset and 0.4359 on the long-context subset) compared to the 10-epoch model, indicating that increasing the number of training epochs improves performance in 2-stage training.

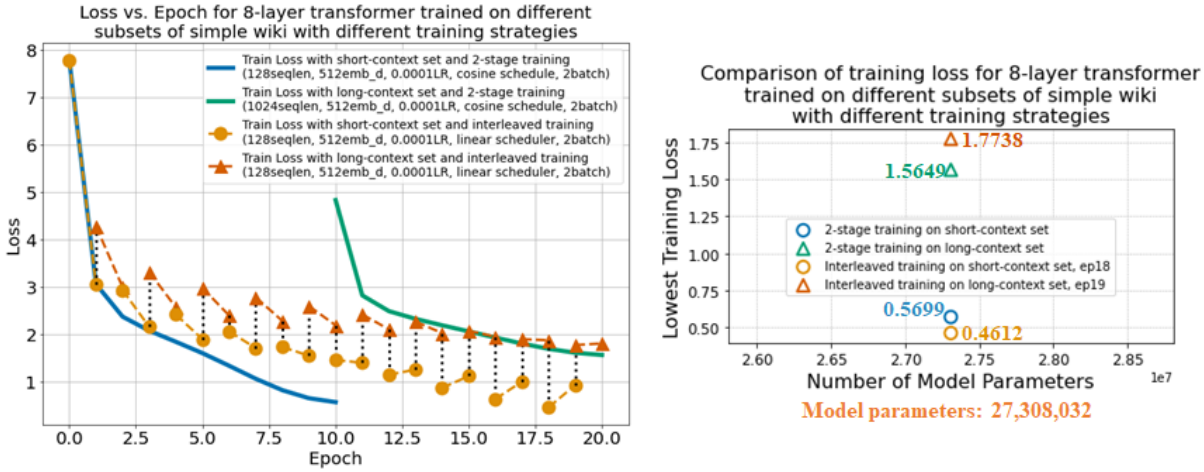


Figure 18: (Left) Training loss of 8-layer transformers with 2-stage (solid) and interleaved (dashed) training. (Right) Best training loss versus parameter count. Interleaved training achieves a lower training loss on the short-context subset (0.4612 compared to 0.5699 for 2-stage training) but a higher training loss on the long-context subset (1.7738 compared to 1.5649 for 2-stage training).