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

Yen-Che Hsiao and Abhishek Dutta

Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA
Email: yen-che.hsiao@uconn.edu

Introduction

Goal:

- Explore how far **small LMs (≤28M parameters)** can go in generating simple, coherent general knowledge.

Dataset:

- Curated **Simple English Wikipedia (64–512 words)**
- Cleaner and more coherent than WikiText-2

Method:

- Hyperparameter tuning on multiple architectures, training parameters, and training strategies

Best Model Found:

- A decoder-only transformer with rotary positional encoding, RMSNorm, GeLU, zero dropout, 8 layers, 8 heads, 512-dim embeddings, and 2048-dim feedforward

Results:

- **Pretraining:**
  - The model with training loss ≈ 0.1 can generate text with basic knowledge in clear and simple English
  - Limited generalization
- **Fine-tuning:**
  - Rephrased-context dataset used for fine-tuning
  - The fine-tuned model showed limited generalization ability.

Model Architecture

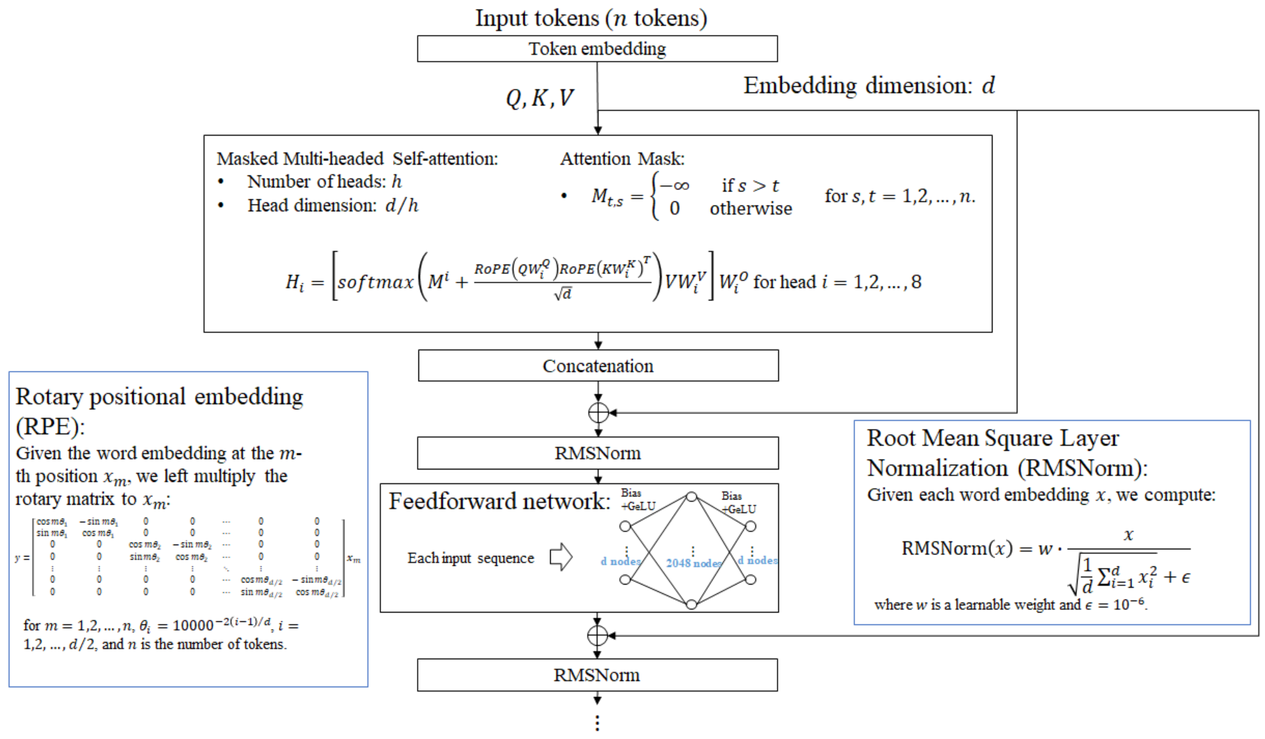
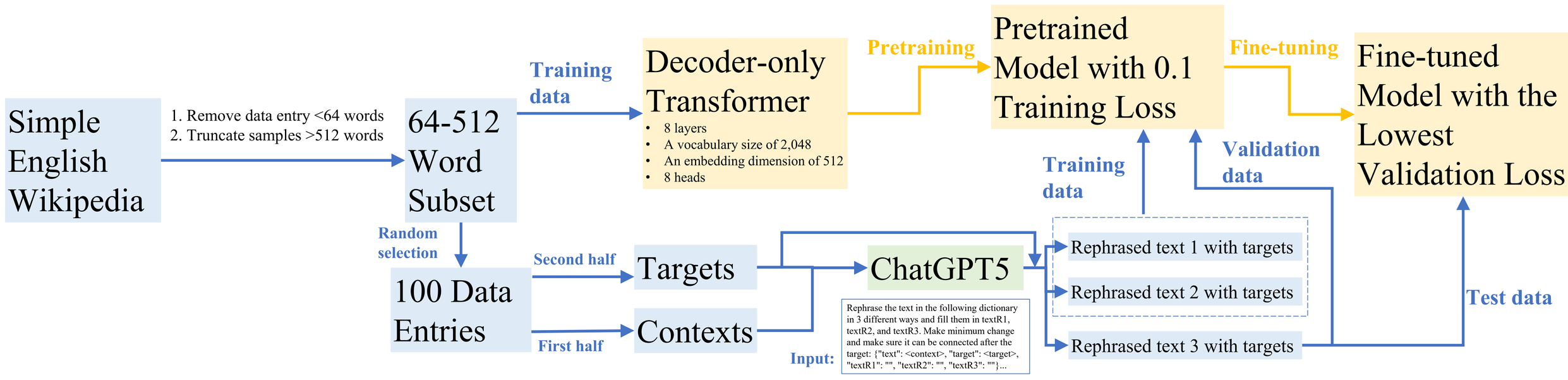
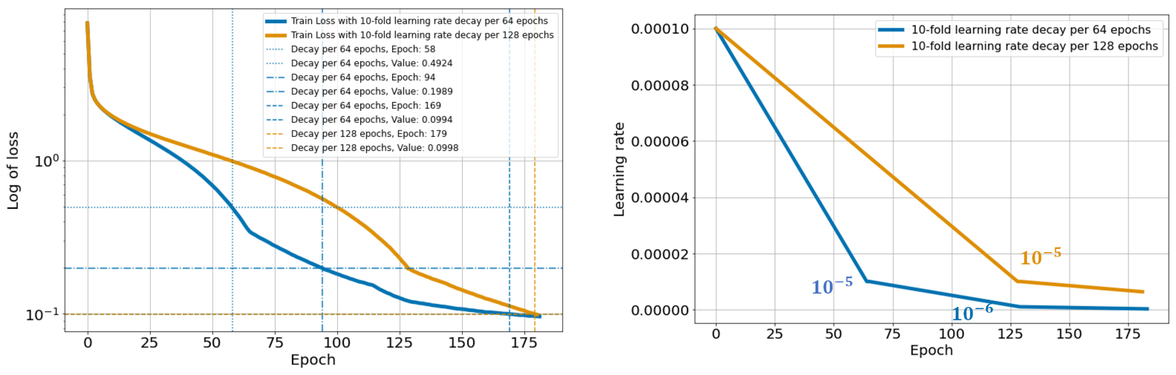


Illustration of the architecture of the initial part of the transformer-based language model used in this study, including the first decoder block. For this work, we found that the model with **8 layers, a vocabulary size of 2,048, an embedding dimension of 512, and 8 heads** performs the best on decreasing the training loss.

Enabling **Generalization Ability** through Fine-tuning on Rephrased Text

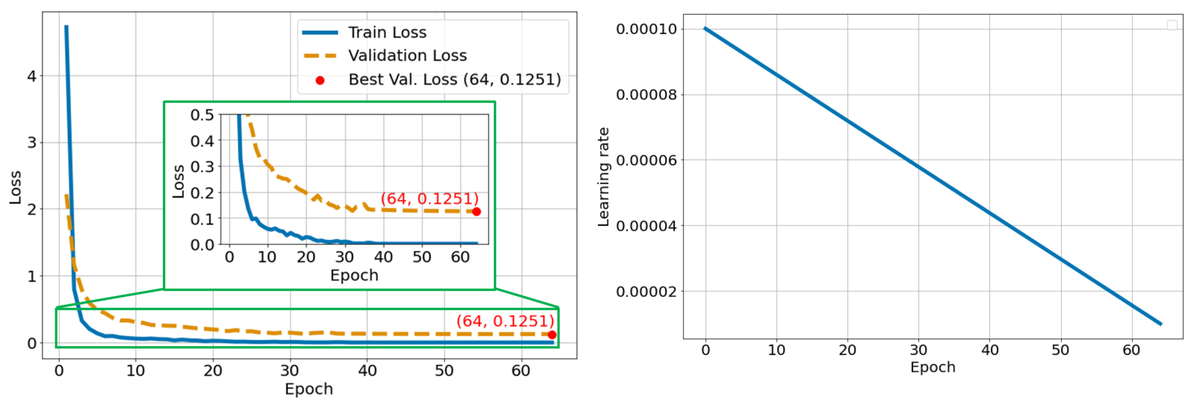


Training Loss and Learning Rate during **Pre-training**



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. We used the model at roughly 0.1 loss trained with a 10-fold decay every 64 epochs for fine-tuning.

Training Loss, Validation Loss, and the Learning Rate during **Fine-tuning**



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. In the figure on the left, 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 model with the best validation loss is selected for evaluating generalization on the test set

Generated Text from the Fine-tuned Model **Showing Generalization Ability** on Unseen Rephrased Text in the Validation/Test set

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

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.