

From Pretraining to Alignment: Development and Evaluation of a Modern Small Language Model

Ruiyang Yu
University of Michigan
ruiyangy@umich.edu
GitHub Repository

Abstract—Large Language Models (LLMs) have revolutionized natural language processing, yet their massive size often obscures the underlying training dynamics. This project aims to build and train a decoder-only transformer, by extending the nanoGPT framework. We incorporate state-of-the-art architectural primitives, including Rotary Positional Embeddings (RoPE) and SwiGLU activations, to enhance performance at a small scale. In the first phase, we pretrained the base model on 1B tokens from the FineWeb-Edu dataset. In the second phase, we implemented a Supervised Fine-Tuning (SFT) pipeline using the SmolTalk corpus to align the model’s output with human intent. Our results demonstrate that modern architectural efficiency combined with high-quality instructional data allows a small-scale model to transition from chaotic text generation to a coherent conversational assistant.

I. INTRODUCTION

Large Language Models (LLMs) have revolutionized natural language processing, yet their massive size often obscures the underlying training dynamics. This project focuses on demystifying these dynamics by constructing a modern, decoder-only transformer. The capability of these models to follow instructions and maintain coherent conversations is not an inherent property of the architecture but a result of specific alignment techniques. While pretraining teaches a model the statistical structure of language, it does not inherently teach it to be helpful or logical; therefore, we employed Supervised Fine-Tuning (SFT) to align the model’s behavior with human intent.

The motivation for this project was two-fold. First, instead of relying on pre-made black-box APIs, we sought to engineer the model’s core components—such as position encoding and activation functions—to understand their mechanical impact. Second, we aimed to investigate the critical transition from “pretraining” (knowledge acquisition) to “alignment” (behavioral shaping).

Deviating from the standard GPT-2 architecture, this project implements state-of-the-art architectural improvements found in models like LLaMA, including Rotary Positional Embeddings (RoPE) and SwiGLU activations. In the preliminary phase, we successfully pretrained a base model on the FineWeb-Edu dataset (1B tokens). The subsequent work focused on the alignment phase: implementing an SFT pipeline using the SmolTalk corpus to transition the model from a text generator into a conversational assistant.

II. RELATED WORK

The foundation of this project is the Transformer architecture introduced by Vaswani et al. [1]. However, specific implementation details diverge from the original design and the GPT-2 standard [2] to align with recent advancements seen in the LLaMA family of models [3].

A. Architectural Primitives

Unlike absolute positional embeddings used in BERT or GPT-2, RoPE [4] encodes position by rotating the query and key vectors in a high-dimensional space. This allows the model to capture relative positions more effectively and generalizes better to sequence lengths longer than those seen during training. Additionally, we replace the standard GELU activation with SwiGLU [5]. Empirical studies suggest SwiGLU offers better training stability and performance for compute-optimal models. We also utilize RMSNorm [6] for pre-normalization stability.

B. Data-Centric AI

Recent work, such as “Textbooks Are All You Need” [7], posits that data quality is more important than quantity. We utilize *FineWeb-Edu* [8], a dataset filtered for educational value, to test this hypothesis. For alignment, we build upon the *SmolTalk* dataset [9], which provides synthetic conversations designed to teach small models instruction-following capabilities.

III. METHODOLOGY

The codebase is built in PyTorch, extending the nanoGPT framework [10] to support the architectural changes described below.

A. Base Model Architecture (Phase 1)

Our model builds upon the foundational decoder-only architecture but integrates architectural primitives used in modern foundation models.

- **RoPE:** We implemented Rotary Positional Embeddings to encode positional information by mathematically rotating the Query and Key vectors.
- **SwiGLU:** We enhanced the feed-forward layers by replacing GELU with SwiGLU, a gated linear unit that uses the Swish function.

- **RMSNorm:** We utilized Root Mean Square Normalization in a pre-norm configuration to reduce computational overhead and improve stability.

B. SFT Data Pipeline (Phase 2)

The critical contribution of the second phase is the SFT pipeline designed to ingest the *SmolTalk* corpus.

1) *Chat Formatting:* We adapted the tokenizer to handle ChatML-style formatting. Raw conversations are serialized into a single sequence using special tokens: `<|user|>`, `<|assistant|>`, and `<|end|>`.

2) *Loss Masking:* A naive approach to fine-tuning would calculate loss on all tokens, effectively training the model to predict the user’s questions as well as its own answers. To prevent this, we implemented a selective masking strategy. As shown in the logic below, we assign a label of ‘-100’ (ignored by PyTorch’s CrossEntropyLoss) to all user and system tokens, ensuring the model is optimized solely on the assistant’s responses.

```

1 if role == "user":
2     # MASK user tokens
3     ids.append(self.SID["user"])
4     labels.append(-100) # Ignored index
5     labels.extend([-100] * len(tokens))
6
7 elif role == "assistant":
8     # TRAIN on assistant tokens
9     ids.append(self.SID["asst"])
10    labels.append(self.SID["asst"])
11    # Train on content
12    labels.extend(tokens)

```

Listing 1. Selective Masking Logic in Dataset

3) *Training Optimization:* Training utilized mixed-precision (BF16) to reduce memory footprint and gradient accumulation to simulate larger effective batch sizes (256) on GPU resources.

IV. RESULTS

A. Pretraining Stability

The model was pretrained on a sample of the FineWeb-Edu dataset (approx. 1B tokens). As shown in Fig. 1, the training loss demonstrates a healthy convergence curve, indicating that the integration of RoPE and SwiGLU did not introduce instability despite the small scale.

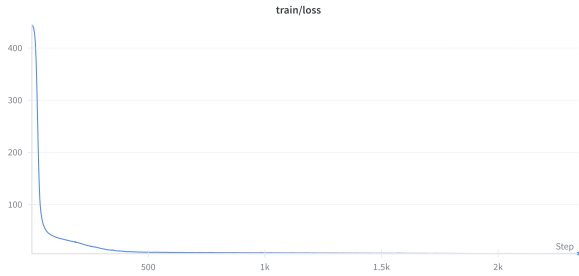


Fig. 1. Training loss curve from Weights & Biases during the pretraining phase on FineWeb-Edu.

B. Qualitative Alignment Analysis

To evaluate the impact of Supervised Fine-Tuning, we compared the generation capabilities of the Base Model against the SFT Model using standard general knowledge prompts.

As shown in Table I, the Base Model exhibits typical “completion behavior”—it attempts to continue the text as if it were writing a story or an article, often hallucinating context or losing logic. In contrast, the SFT Model demonstrates “chat behavior.” It recognizes the conversational intent, provides a direct answer, and terminates generation appropriately using the `<|end|>` token.

TABLE I
QUALITATIVE COMPARISON: BASE VS. SFT MODEL

Prompt	Base Model Output	SFT Model Output
Who is the president of the United States?	...and the government is located in Washington. The president is the head of state and the commander in chief...	The current president of the United States is Joe Biden.
What is the capital of France?	...in the region of Europe. It is a beautiful city with the Eiffel Tower and the Louvre...	The capital of France is Paris.
Once upon a time	”...such a last year, it will be contempt to, and it will be a good thing that is the idea that nasal, I will be, and in the list of the or or a one will take to the same....”	In a kingdom far away, there lived a knight who protect the realm.

This shift in behavior validates the efficacy of the masking strategy implemented in the methodology. By constraining the loss function to specific tokens, we successfully reshaped the model’s probability distribution to favor helpful, aligned responses.

V. CONCLUSION

This project successfully demonstrated the end-to-end development of a modern Small Language Model. By upgrading the standard GPT-2 architecture with RoPE and SwiGLU, we achieved stable pretraining on high-quality educational data. Furthermore, the implementation of a custom SFT pipeline with token masking successfully transitioned the model from a generic text completer to an aligned conversational assistant. Future work can focus on scaling the model parameters and incorporating Direct Preference Optimization (DPO) for further alignment.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [2] A. Radford, J. Wu, R. Child, et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [3] H. Touvron, T. Lavril, G. Izacard, et al., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [4] J. Su, M. Ahmed, Y. Lu, et al., “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, vol. 568, p. 127063, 2024.

- [5] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.
- [6] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [7] S. Gunasekar, Y. Zhang, J. Jyothi, et al., "Textbooks are all you need," *arXiv preprint arXiv:2306.11644*, 2023.
- [8] A. Lozhkov, L. Ben Allal, L. von Werra, and T. Wolf, "FineWeb-Edu: The Finest Collection of Educational Content," 2024. [Online]. Available: <https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>
- [9] L. Ben Allal, A. Lozhkov, E. Bakouch, et al., "SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model," *arXiv preprint arXiv:2502.02737*, 2025.
- [10] A. Karpathy, "nanoGPT," GitHub repository, 2023. [Online]. Available: <https://github.com/karpathy/nanoGPT>