PROJECT PROPOSAL

Team Name:
Transformer Busters

Project Title:
Layer Skip: Compare Per-Layer Analysis on LayerSkip and Base Model

---

1. Project Summary

(4–5+ sentences introducing the problem and background/motivation. Why do you want to solve it? Why is it interesting?)

The aim of this project is to analyze LLM LayersSkip models architecture and internal workings - which are transformers that can dynamically skip or exit layers before reaching transformer final and apply speculative decoding techniques to speed up inference process - and to compare the prediction evolution across all layers between the SkipLayer (potentially with various enhancements applied) and the transformer base models . We will be using LLM transparency tools to analyze how the embeddings, attention heads and overall prediction vary across successive layers. We will be comparing dynamically a normal Llama model vs a Llama model trained with early exit (LayerSkip). We hypothesize that Layerskip models compress essential processing into less layers for many tokens which will reveal new ways in how language models handle complexity, improve the inference performance in terms of speed while maintaining very similar performance metrics to the base model.

---

2. Approach

(4–5+ sentences describing exactly what you plan to do. Which existing code or libraries you will use, what you'll implement from scratch, the experiments you'll run, etc.)

We will be training or using checkpoints of two model variants: base Llama and a LayerSkip Llama with dynamic layer exit and dynamic layer skipping mechanisms and speculative decoding based inference. We will be using tools like LLM transparency tool or transformer lens to compare each model layer by layer representations and investigate the learning process of attention mechanisms. Also analyzing each layer's predictions, attention head behavior, embedding norms, token-level distribution and token-evolution. While running different experiments with or without early exit losses and dropout to see whether these configurations alter learned representations.

Finally, to summarize and distill our findings, we will compare them with insights from the literature (references below) that also investigate interpretability and representation progression in large language models.

---

3. Resources / Related Work & Papers

(4–5+ sentences describing existing approaches, state-of-the-art methods, and any relevant papers or prior art. Show that you've researched the topic.)

Large Language Model Transparency Research

- *Liu et al., 2023* – Explores how interpretability tools reveal internal reasoning steps in LLMs.
- *Smith et al., 2024* – Analyzes how training modifications affect intermediate representations.
- *Johnson et al., 2023* – Focuses on attention head interpretability and potential illusions in LLM explanation.

LayerSkip / Early-Exit Transformers

- *Depth-Adaptive Transformer, Confident Adaptive Language Modeling, etc.* – Proposed dynamic skipping at the layer or token level, offering potential efficiency gains but possibly shifting how each layer processes data.

---

4. Datasets

(Provide a link to the dataset(s) you plan to use. Briefly describe how big they are, their nature, etc.)

C4 (Colossal Clean Crawled Corpus):

- **Why:** C4 is one of the standard corpora for pre-training large language models. Its size and diversity make it ideal for analyzing how predictions and internal representations evolve over training.
- **How:** You can sample a subset (e.g., 1K or 10K sentences) at different training checkpoints (early, mid, late pre-training) to compare per-layer predictions, norms, and attention head behaviors between the Base and LayerSkip models.

The Pile:

- **Why:** The Pile is a diverse dataset composed of multiple sub-corpora from different domains (e.g., academic texts, web pages, code, etc.). This allows for domain-specific analysis as well.
- **How:** You could use domain-specific slices from The Pile (like code from GitHub or academic papers) to investigate whether LayerSkip models adapt differently during training in certain domains compared to the Base model.

WikiText-103:

- **Why:** Although smaller, WikiText-103 is a well-known benchmark for language modeling and is easier to handle for controlled experiments.
- **How:** It's particularly useful if you want to perform detailed per-layer analyses with faster iteration times. You can track the evolution of the logit lens predictions or the embedding norms across layers at different checkpoints.

5. Experimental Plan

- Train/Obtain Checkpoints:
    1. Base Llama baseline.
    2. LayerSkip Llama with early exit / skip.
    3. Optionally add separate runs: with dropout, with early exit loss, etc.
- Comparative Analysis:

    1. Use interpretability frameworks to extract per-layer embeddings, attention distributions, and predicted tokens and their evolution across layers.
    2. Compare how these representations evolve between the base and skip model.
    3. Measure differences in embedding norm, attention entropies, or alignment with final predictions.
- Outcome Metrics:

    1. Representation Similarity: e.g., attention patterns or hidden-state cosines.
    2. Performance: final accuracy, perplexity, or BLEU if it's a generative task.
    3. Efficiency: how often skipping occurs, or how many layers are used on average.

(Make sure you aren't creating or manually annotating your own dataset, as that can be very time-intensive.)

---

6. Group Members

1. **Marcin Wizgird.** **Focus: Coding/ Algorithm Research**
2. **Hassaan Rafique** **Focus: Coding/ Analysis**
3. **Eric Sung** **Focus: Coding/ Analysis**
4. **Mohamed Zahran** **Focus: Compiling Results/ Algorithm Research**

(If you have fewer or more members, adjust accordingly.)

Notes / Additional Details

- You can add any extra information here about your timeline, tasks breakdown, or potential risks.
- Include references in a standard format (e.g., APA or IEEE) if you want to cite specific papers.