Investigating Multi-Token Prediction and Look-Ahead Behavior in Large Language Models

Harley T. Gribble

Abstract

Large Language Models (LLMs) are typically trained to predict the next token in a sequence, a setup that encourages myopic, one-step-ahead reasoning. However, recent work suggests that hidden states may contain information about tokens beyond the immediate next one, raising questions about whether models implicitly "think ahead." In parallel, Meta AI has proposed explicit multi-token prediction as a way to improve efficiency and accuracy. This study explores these two lines of research — Wu et al. (2024) on probing hidden states for look-ahead information, and Meta AI (2024) on multi-token prediction — while also implementing small-scale GPT-style models to reproduce and study these behaviors.

---

1. Introduction

Language models such as GPT rely on decoder-only transformer architectures that generate tokens sequentially. Their design presumes a next-token-only learning objective. Yet, questions remain about whether models learn strategies that implicitly extend beyond this local scope, and whether training them to predict multiple tokens in parallel could improve performance.

This project investigates both the theoretical and empirical dimensions of this problem by (1) analyzing Wu et al.'s experiments on myopia in transformers, (2) studying Meta AI's multi-token prediction approach, and (3) implementing minimal models to explore these concepts in practice.

---

2. Background

2.1 Neural Networks and Transformers

Neural networks transform inputs into outputs using weighted layers, activation functions, and backpropagation with gradient descent. Transformers extend this by incorporating self-attention layers, allowing each token to contextualize itself with respect to all others. Decoder-only transformers, such as GPT, restrict attention causally so each token can only attend to itself and prior tokens.

2.2 Next-Token Prediction Objective

In standard autoregressive training, the model receives a sequence of tokens and learns to maximize the probability of the next token. Cross-entropy loss measures the gap between predicted probabilities and the true next token, and backpropagation adjusts weights accordingly.

---

3. Related Work

3.1 Wu et al. (2024): Do Language Models Think Ahead?

Wu et al. examined whether LLMs encode information about future tokens. They created "myopic" variants of language models by modifying how gradients were backpropagated, ensuring models only updated weights for the immediate next token. To test whether future information was nonetheless present: - They trained **linear probes** on hidden states to predict future tokens.
- They ran **correlation checks** between neuron activations and target labels.

Findings: Hidden states contained predictive information about tokens beyond the immediate next one. In small models this was incidental, while in larger models evidence suggested explicit encoding of future-related features.

3.2 Meta AI (2024): Multi-Token Prediction

Meta explored adding **multiple prediction heads** to the final hidden layer, each trained to predict a future token (t+n). Training summed the cross-entropy loss across all heads. They paired this with **self-speculative decoding**, where draft tokens from multiple heads were later verified by a standard autoregressive pass.

Findings: Multi-token prediction improved efficiency and was particularly effective for code generation tasks. Results were mixed for large-scale language modeling: sometimes standard models remained more accurate, but multi-token approaches consistently improved inference speed.

---

4. Methodology

To explore these ideas, we implemented simplified GPT-style models from scratch:

- **Numpy implementation:** Step-by-step version of embeddings, attention, feed-forward layers, and softmax to better understand the math.
- **PyTorch implementation:** A practical decoder-only transformer with masking, positional embeddings, and multiple stacked layers.
- **Multi-token extension:** Modified output head to include additional prediction heads in parallel, inspired by Meta's design.

These minimal experiments serve as a sandbox for exploring both the Wu et al. probing approach and Meta's multi-head output strategy.

---

5. Preliminary Insights

- Reproducing GPT-style models clarified how cross-entropy loss and backpropagation flow through embeddings, attention layers, and output heads.
- Adding additional prediction heads is straightforward: each head learns its own weights and contributes its own loss to the global update.
- The main conceptual gap lies in testing myopia: probes must be trained separately on frozen hidden states to reveal latent information.

---

6. Next Steps

- Implement linear probes on hidden states of the PyTorch model, replicating Wu et al.'s analysis.
- Extend multi-token experiments with speculative decoding to test efficiency gains.
- Compare probe accuracy and myopia gap in synthetic datasets to validate findings.
- Prepare results for potential submission to a workshop or conference.

References

- Wu, X., et al. (2024). *Do Language Models Think Ahead?* arXiv:2404.00859.
- Meta AI (2024). *Scaling Multi-Token Prediction with LLMs.* arXiv:2404.19737.