Do Language Models Think Ahead? Probing Look-Ahead Signals and Multi-Token Prediction

AUTHOR
Harley T. Gribble

AFFILIATION

Abstract

Decoder-only transformers are trained myopically—optimizing only for the immediate next token. Yet they often generate coherent long-range text, suggesting hidden structure that encodes look-ahead information. We investigate this question along two fronts: (1) Wu et al.'s *myopic probing* methodology, which tests whether hidden states contain signals about future tokens beyond training objectives, and (2) Meta AI's *multi-token prediction* models, which add multiple output heads to explicitly train for future tokens and speed up inference.

Our contributions so far are:
- Synthesizing the literature into a unified framework of "myopia" vs "look-ahead."
- Implementing an initial experiment: frozen GPT-2 with auxiliary heads predicting (t{+}1), (t{+}2), and (t{+}3) from the same hidden state.
- Early findings: frozen heads underperform the vanilla baseline, supporting the idea that myopic training limits linearly-decodable look-ahead.
- Outlining future work: probes at intermediate layers, fine-tuning, and self-speculative decoding.

---

1. Introduction

Core question: Do large language models "think ahead"?

Transformer decoders are trained only to predict the **next token**: [= *-{t=1}^{T} p(x*{t+1}x_{t}).] This **myopic** loss does not explicitly reward carrying information useful for later tokens ((x_{t+2}, x_{t+3},)).

And yet, LLMs generate text with long-range coherence. How much of that comes from incidental correlations vs. explicit forward-looking signals? To answer this, we combine insights from two research directions:

- **Wu et al. (2024):** Diagnostic probing to test whether hidden states encode future token info despite myopic training.
- **Meta AI (2024):** Architectural changes (multi-token heads + speculative decoding) to make forward-looking prediction explicit and efficient.

Our project explores both angles, starting with small-scale reproducible experiments on GPT-2.

---

2. Background

2.1 Transformer decoder overview

- Input sequence (x_{1:T}) → tokenized and embedded.
- Passes through stacked layers: masked self-attention + feed-forward blocks.
- Produces hidden states (H^{(L)} ^{Td}).
- Final token state (h_T^{(L)}) → linear projection (W_{}) → logits → softmax.

Attention (per head): [(X) = !()V, Q=XW_Q,;K=XW_K,;V=XW_V.]

Cross-entropy loss: [L=-_v q_v p_v,q_v=[v=x_{t+1}].]

2.2 What is "myopia"?

- **Myopic model:** trained only on immediate next-token loss.
- **Implication:** model may *incidentally* encode future-relevant info, but is not rewarded for doing so.
- Wu et al. created *myopic baselines* by blocking gradient flow to earlier timesteps, ensuring predictions only used local info.

2.3 Multi-token prediction

- Add (K) parallel output heads to predict tokens (t{+}1,,t{+}K).
- Each head has its own (W^{(k)}, b^{(k)}).
  [= *{k=1}^K !((h_T*^{(L)}W*{(k)}+b^{(k)}),; x*{T+k}).]
- Training all heads jointly provides a non-myopic signal.
- At inference, Meta used **self-speculative decoding**: draft multiple tokens, then verify them efficiently.

---

3. Related Work

Wu et al. (2024): *Do LMs Think Ahead?*

- **Setup:**
  - Trained "myopic" models where gradient updates stop at current timestep.
  - Probes (linear regression) on hidden states → predict future tokens.
  - Per-neuron correlations to check whether any single dimension aligned with targets.
- **Findings:**
  - Even myopic models encode some future signal.
  - Larger models show stronger non-myopic behavior.
  - Evidence that LLMs develop forward-looking internal representations.

Meta AI (2024): *Scaling Multi-Token Prediction*

- **Setup:**
  - Augmented LLMs with multiple prediction heads.
  - Sequential backprop per head to save memory; summed loss.
  - Benchmarked on code and text tasks.
- **Findings:**

- Clear speedups with speculative decoding.
- Accuracy gains for code; mixed results for general text.
- Hypothesis: explicit multi-token training may improve robustness at decision points.

---

4. Our Research Goals

- **Replicate** probing ideas and multi-token setups in a controlled, smaller-scale environment.
- **Question 1:** How much future token info is linearly decodable from frozen GPT-2 hidden states?
- **Question 2:** Do auxiliary heads trained without unfreezing the trunk succeed at (+2,+3) prediction?
- **Question 3:** Could fine-tuning or speculative decoding make multi-token prediction competitive?

---

5. Methods

5.1 Data

- WikiText-2, tokenized with GPT-2 tokenizer.
- Max sequence length = 256.
- Standard train/val split.

5.2 Models

- **Baseline:** GPT-2 (pretrained, Hugging Face). Teacher-forced evaluation at (+k).
- **Auxiliary-head model:** GPT-2 trunk frozen. Three linear heads, each predicting offset (k{1,2,3}) from (h_t^{(L)}).

5.3 Training

- Optimizer: AdamW, 2 epochs.
- Loss: sum of cross-entropies across heads.
- Only head parameters updated.
- Batch size = 8.

5.4 Evaluation

- Metrics: token-level acc@1 and loss.
- Compare vanilla vs aux-heads at (+1,+2,+3).

---

6. Experiment 1 — Frozen GPT-2 with Aux Heads

Results

Offset	Vanilla acc@1	Vanilla loss	Aux-head acc@1	Aux-head loss
+1	0.327	3.893	0.305	5.702
+2	0.330	3.847	0.155	7.240
+3	0.332	3.815	0.095	7.810

Interpretation

- Vanilla consistently outperforms frozen aux-heads.
- Aux-heads degrade sharply with distance (+2,+3).
- Suggests look-ahead information is not linearly decodable at final layer without adapting the trunk.
- Aligns with Wu et al.'s claim: myopic training leaves forward signals weak, though larger/fine-tuned models may differ.

7. Discussion

- **Myopia in practice:** Teacher forcing lets vanilla succeed at (+2,+3), while frozen aux-heads fail.
- **Implication:** Without explicit training signals, look-ahead is not well-represented in the final hidden state.
- **Next steps:** probe earlier layers; allow partial trunk fine-tuning.

8. Future Work

1. Probe intermediate layers (Wu et al. style).
2. Train aux-heads with trunk partially unfrozen.
3. Joint multi-head training (Meta-style).
4. Self-speculative decoding evaluation.
5. Extend experiments to code datasets (Meta showed stronger gains there).

9. Conclusion

Our initial experiments support the hypothesis that standard next-token training is **myopic**: the final hidden state alone does not linearly expose future token info. Multi-token approaches may help—but require training signals that go beyond frozen readouts. This work sets the foundation for deeper exploration of look-ahead in LLMs.

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.