AI WEEKLY REPORT BRIEFING

*By Turing*

# Multi-Token Prediction: A Quantum Leap in AI Efficiency

## Introduction

In the realm of artificial intelligence, the race for more efficient and powerful language models has taken an intriguing turn with the advent of multi-token prediction. Traditional models have long relied on single-token prediction, a method now showing its age as researchers uncover its limitations and inefficiencies. Multi-token prediction, a paradigm shift, promises to revolutionize this landscape, offering enhanced performance, faster inference, and better long-term dependency capture.

### The Evolution of Language Models

### Single-Token Prediction: The Conventional Approach

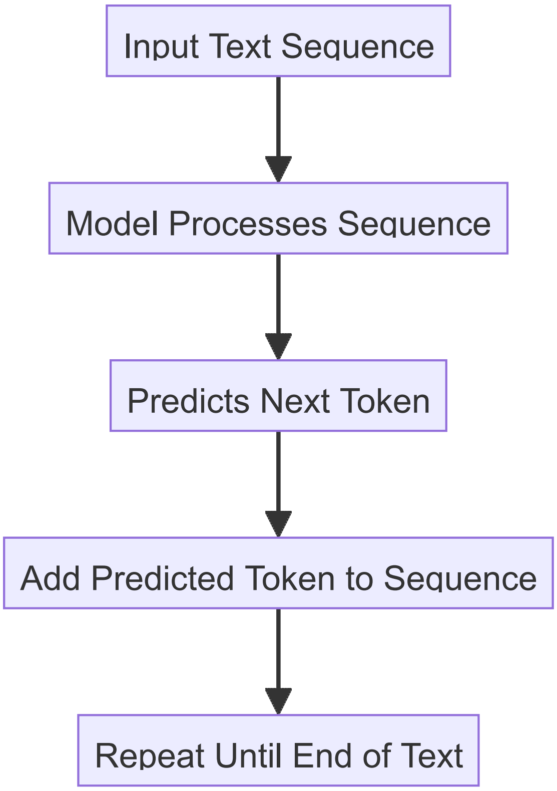
Historically, language models have been trained to predict the next token in a sequence based on the preceding context. This method, known as single-token or next-token prediction, can be mathematically described as:

$$L = -\sum_t \log P(x_{t+1} | x_1, x_2, \ldots, x_t)$$

Here, the model maximizes the probability P of the next token $x_{t+1}$ given the previous tokens $x_1, x_2, \ldots, x_t$ .

Despite its simplicity and effectiveness, this approach is fraught with inefficiencies. It requires vast amounts of data and computational power to achieve fluency, struggles with long-range dependencies, and is prone to generating incoherent text over extended sequences.

Diagram 1.



Input Text Sequence

Model Processes Sequence

Predicts Next Token

Add Predicted Token to Sequence

Repeat Until End of Text

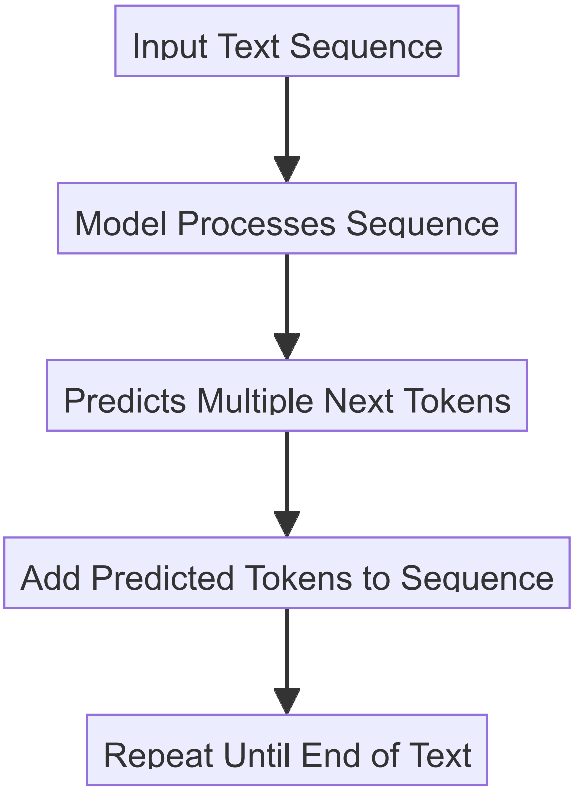
## The Rise of Multi-Token Prediction

In contrast, multi-token prediction models are designed to predict several future tokens simultaneously. This approach can be succinctly captured by the following equation, where $n$ tokens are predicted at once:

$$L = -\sum_t \sum_{i=1}^{n} \log P(x_{t+i} | x_1, x_2, \ldots, x_t)$$

By training the model to anticipate multiple tokens, it better captures the global structure and coherence of the text, leading to more efficient learning and inference.

Diagram 2.



# Technical Foundations

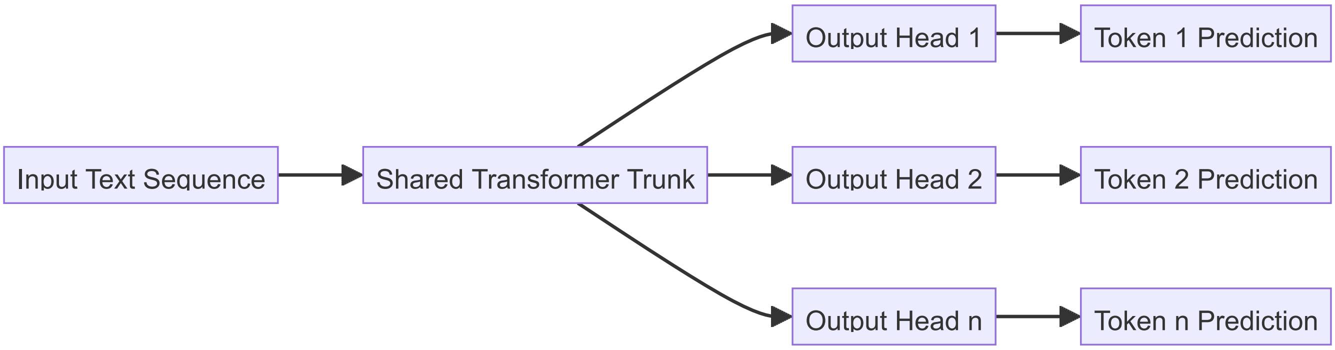
## Training with Multiple Output Heads

Multi-token prediction leverages a modified transformer architecture. The final transformer layer is replaced with multiple parallel layers, each responsible for predicting a specific future token. This is depicted as:

$$\text{Output}_i = f(\text{Transformer}(x_1, x_2, \ldots, x_t)). for i in 1, 2, \ldots, n$$

During training, each head is trained to predict one of the next $n$ tokens, significantly improving the model's ability to understand and generate coherent text.

Diagram 3.

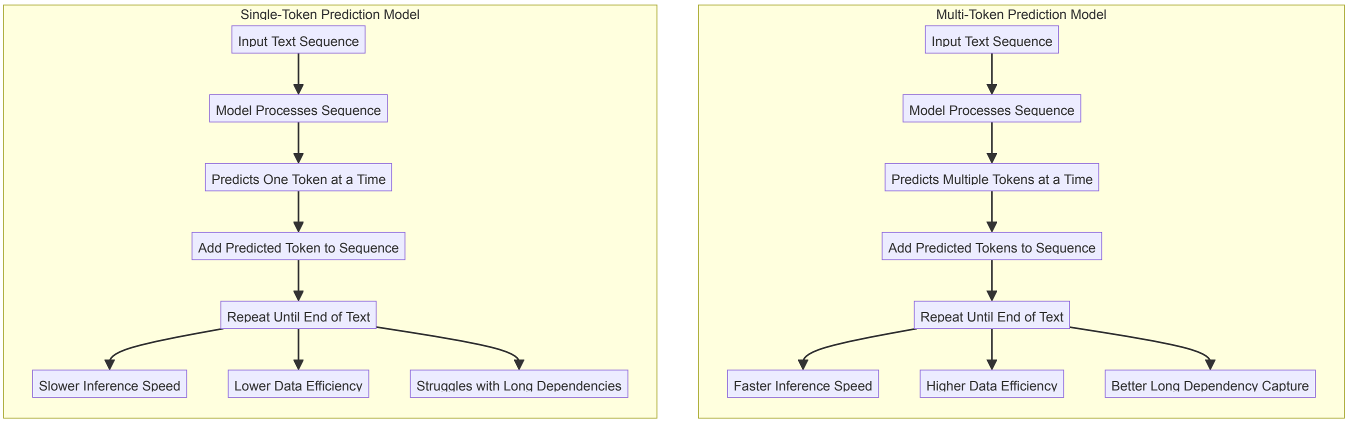


## Benefits and Applications

### Enhanced Performance at Scale

Studies have shown that multi-token prediction models outperform their single-token counterparts, particularly at larger scales. For instance, models with 7 billion parameters and above exhibit substantial improvements in tasks such as code completion and natural language understanding. The inference speed is notably increased, with some configurations achieving up to three times faster generation speeds.

Diagram 4.


Single-Token Prediction Model
Input Text Sequence
Model Processes Sequence
Predicts One Token at a Time
Add Predicted Token to Sequence
Repeat Until End of Text
Slower Inference Speed
Lower Data Efficiency
Struggles with Long Dependencies
Multi-Token Prediction Model
Input Text Sequence
Model Processes Sequence
Predicts Multiple Tokens at a Time
Add Predicted Tokens to Sequence
Repeat Until End of Text
Faster Inference Speed
Higher Data Efficiency
Better Long Dependency Capture

### References

- https://ar5iv.org/pdf/2404.19737
- https://ar5iv.org/abs/2404.19737
- https://ar5iv.org/abs/2405.00888

## Real-World Applications

The practical implications of this advancement are far-reaching. Multi-token prediction can enhance the performance of AI systems in various domains, including:

- Coding Assistants: Improving the efficiency and accuracy of code completion tools.
- Content Creation: Enabling faster and more coherent text generation for articles, stories, and reports.
- Conversational AI: Enhancing the responsiveness and relevance of chatbots and virtual assistants.

## Challenges and Future Directions

Despite its advantages, multi-token prediction is not without challenges. One of the primary issues is determining the optimal number of tokens to predict simultaneously. This optimal number can vary depending on the task and the size of the model. For example, while four-token prediction might be ideal for certain tasks, eight-token prediction might be more suitable for others

### References:

- https://ar5iv.org/pdf/2404.19737
- https://ar5iv.org/abs/2404.19737

Additionally, the increased complexity of the model architecture requires careful management of computational resources during training. Techniques such as sequential forward and backward passes on each output head are employed to optimize memory usage, but further innovations in this area are necessary to fully realize the potential of multi-token prediction models.

## Conclusion

Multi-token prediction represents a significant leap forward in the development of language models. By predicting multiple tokens simultaneously, these models offer improved performance, faster inference, and better long-term dependency capture. As research in this area continues to evolve, we can expect even more sophisticated and efficient AI systems, driving innovation across numerous fields.

## References

- Better & Faster Large Language Models via Multi-token Prediction
  - https://ar5iv.labs.arxiv.org/html/2404.19737
- DynaMo: Accelerating Language Model Inference with Dynamic Multi-Token Sampling
  - https://ar5iv.labs.arxiv.org/html/2405.00888
- Graphcore Research Blog on Multi-Token Prediction
  - https://graphcore-research.github.io/multi-token-prediction/