# Token Trajectory Analysis for Principled Transformer Acceleration

Tristan Peat and Prof. Kartik Goyal

December 2024

## 1 Introduction

The sequential nature of autoregressive sampling in large language models (LLMs) presents a fundamental bottleneck for inference speed. While methods like speculative decoding [9, 3] partially address this by using smaller models to draft token sequences, they rely on empirical heuristics rather than theoretical guarantees. Recent work has revealed that Transformer architectures [13] exhibit systematic clustering behaviors [7, 6, 8], suggesting their token representations converge to predictable configurations before reaching final layers.

We propose unifying three frameworks to improve inference efficiency: deep equilibrium networks (DEQ) [2], which directly solve for convergence points rather than computing through multiple layers; optimal transport (OT), which provides tools for analyzing probability measure evolution; and speculative decoding, which accelerates inference through parallel token verification. Our key insight is that understanding the geometry of token distributions through transformer layers can lead to principled methods for predicting final configurations without full forward passes.

## 2 Related Works

### 2.1 Speculative Decoding

Speculative decoding [9, 3] accelerates inference in LLMs while maintaining exact decoding accuracy by using a more efficient draft model to predict likely outputs while verifying them with the main model in parallel. The process involves a target model $M_p$ that produces distribution $p(x_t|x_{<t})$ and an approximation model $M_q$ producing $q(x_t|x_{<t})$. Recent improvements include self-speculative decoding [15], which eliminates the need for a separate draft model, and multi-draft approaches [12] that generate token trees for parallel verification. For probability distributions $p(x)$ from target model $M_p$ and $q(x)$ from approximation model $M_q$, the core sampling algorithm accepts samples $x \sim q(x)$

when $q(x) \leq p(x)$ and rejects with probability $1 - \frac{p(x)}{q(x)}$ otherwise, drawing new samples from adjusted distribution $p'(x) = norm(max(0, p(x) - q(x)))$.

## 2.2 Deep Equilibrium Networks

Deep equilibrium networks (DEQ) [2] directly solve for fixed points in sequence modeling rather than computing through multiple layers. For a deep feedforward sequence model defined as $z_{1:T}^{i+1} = f_\theta^i(z_{1:T}^i; x_{1:T})$, DEQ finds the equilibrium point $z_{1:T}^\star$ satisfying $f_\theta(z_{1:T}^\star; x_{1:T}) = z_{1:T}^\star$ using root-finding methods. The approach enables efficient forward passes through quasi-Newton methods and backward passes through implicit differentiation, achieving constant memory consumption during training.

## 2.3 LayerSkip and Sparsity

LayerSkip [5] speeds up inference by combining layer dropout (higher rates in later layers) with early exit capabilities, while maintaining accuracy through self-speculative verification. This approach complements work on transformer sparsity [4, 11], where techniques like DejaVu [10] exploit input-dependent sets of attention heads and MLP parameters that approximate full model outputs. These methods demonstrate that Transformers often contain efficient computational paths that can be leveraged without compromising performance [16].

## 2.4 Attention Patterns and Token Dynamics

Recent work has revealed that transformer token representations follow predictable trajectories governed by attention mechanisms. The emergence of attention sinks [14] and clustered patterns [1] suggests an underlying geometric structure that can be formalized through optimal transport theory [6, 7].

In this framework, token embeddings behave as particles whose evolution is determined by attention-weighted interactions. For autoregressive models, [8] shows that the evolution of the k-th token embedding $x_k(t)$ follows a gradient flow in probability space:

$$\dot{x}_k(t) = P_{x_k(t)} \left( \frac{1}{Z_k(t)} \sum_{j=1}^{k} e^{\beta \langle Q(t)x_k(t), K(t)x_j(t) \rangle} V(t) x_j(t) \right) \tag{1}$$

where $P_{x_k(t)}$ projects onto the tangent space, $Z_k(t)$ is the partition function, and $Q(t)$, $K(t)$, $V(t)$ are the query, key, and value matrices. This formulation shows that tokens follow least-action paths in a Wasserstein metric space induced by attention.

This optimal transport perspective provides theoretical justification for combining deep equilibrium methods with early exiting. By analyzing token distribution convergence in Wasserstein space, we can predict final configurations without full forward passes and identify natural stopping points. The transport cost between intermediate and final distributions offers a principled metric for early exit decisions, replacing empirical heuristics with geometric guarantees.

# 3 Methods

We hypothesize that understanding the geometry of learned token representations in Transformer architectures—specifically how attention parameters (K, Q, V) influence token trajectories—can lead to principled improvements in inference efficiency. While prior work has demonstrated various methods for accelerating inference through speculative decoding or early exiting, these approaches often rely on empirical heuristics rather than theoretical guarantees.

Our work aims to bridge this gap by developing a theoretical framework that unifies deep equilibrium networks, optimal transport, and speculative decoding. Specifically, we seek to:

1. Characterize the geometry of token distributions across transformer layers, focusing on how attention mechanisms influence convergence patterns

2. Develop methods to predict or solve for equilibrium points in token configurations without requiring full forward passes

3. Leverage these theoretical insights to improve inference efficiency through more informed speculative decoding

The practical focus of our work will be on inference optimization in pretrained models, avoiding the additional complexity of training-time modifications. We propose to combine deep equilibrium methods with optimal transport to develop more principled approaches for speculative decoding and early exiting. This framework should allow us to: (1) Identify when token representations have effectively converged, (2) Make theoretically-grounded predictions about likely next tokens, (3) Determine optimal exit points that maintain model faithfulness.

The core innovation of our approach lies in replacing empirical heuristics with theoretical guarantees derived from analyzing the geometry of Transformer computations. While maintaining practical applicability, we aim to develop a deeper understanding of how token distributions evolve through transformer layers and how this knowledge can be exploited for more efficient inference.

## 3.1 Evaluation

We will evaluate our approach on two primary metrics:

**Faithfulness**: The method should maintain output quality within bounded error compared to full inference. We will assess this across tasks of varying complexity, such as base language modeling (perplexity), reasoning and comprehension, and code generation.

**Efficiency**: Following [15], we will measure speedup as the acceleration of average inference time per token compared to autoregressive baselines. Our theoretical framework should provide guarantees about the trade-off between computational savings and output quality.

### 3.2 Semester Timeline

| Phase | Timeline | Key Deliverables |
|---|---|---|
| Literature Review | Dec 2024 - Jan 2025 | Implementation of speculative decoding, DEQ networks, and LayerSkip baseline |
| Method Design | Feb - Mar 2025 | Algorithm combining DEQ and speculative decoding; theoretical proofs |
| Experimentation | Mar - Apr 2025 | Comprehensive benchmarks and ablation studies |
| Writing | Apr - May 2025 | Compile manuscript for submission |

Table 1: Research Timeline

## 4 Main Deliverable

The primary deliverable will be a research paper introducing a novel theoretical framework that unifies deep equilibrium networks with speculative decoding through optimal transport. We aim to submit this work to a top-tier machine learning or natural language processing venue, such as NeurIPS 2026. The paper will present both theoretical contributions (proofs of convergence and optimality) and empirical validation on standard language modeling benchmarks.

## References

[1] Saurabh Agarwal, Bilge Acun, Basil Hosmer, Mostafa Elhoushi, Yejin Lee, Shivaram Venkataraman, Dimitris Papailiopoulos, and Carole-Jean Wu. Chai: Clustered head attention for efficient llm inference, 2024.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models, 2019.

[3] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023.

[4] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023.

[5] Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. Layerskip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 12622–12642. Association for Computational Linguistics, 2024.

[6] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics, 2024.

[7] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024.

[8] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking, 2024.

[9] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding, 2023.

[10] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, and Beidi Chen. Deja vu: Contextual sparsity for efficient llms at inference time, 2023.

[11] Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss, 2021.

[12] Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. Spectr: Fast speculative decoding via optimal transport, 2024.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[14] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024.

[15] Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. Draft & verify: Lossless large language model acceleration via self-speculative decoding, 2024.

[16] Haizhong Zheng, Xiaoyan Bai, Xueshen Liu, Z. Morley Mao, Beidi Chen, Fan Lai, and Atul Prakash. Learn to be efficient: Build structured sparsity in large language models, 2024.