📘 **Hierarchical Tokens: Structuring Transformers for AGI**

Author: Rogério Figurelli

Date: April 25, 2025

Abstract

This white paper introduces the concept of Hierarchical Tokens, a novel architectural direction for transformer-based models. Instead of limiting language generation to token-level prediction (word by word), this approach expands the predictive structure to higher-level semantic units such as sentences, paragraphs, sections, chapters, and even domains of knowledge—each treated as composable macro-tokens.

By applying the same mechanisms of attention and probability distribution across multiple scales of abstraction [1][4], this method proposes a path toward Artificial General Intelligence (AGI) that is simpler, more natural, and more human-like. This approach aligns machine generation with how humans plan, organize, and express thought.

---

1. Introduction

Large Language Models (LLMs) such as GPT have revolutionized natural language processing by predicting the next word in a sequence [2][3]. However, they generate language as linear sequences of tokens, relying solely on local context windows and autoregressive mechanisms.

While powerful, this method lacks a true understanding of structure, coherence, or long-term planning. In contrast, humans think hierarchically: we plan entire ideas, organize them into sections, and only then write individual words. Our cognition is top-down, not just reactive [5].

---

2. The Proposal: Hierarchical Tokenization

This paper proposes a simple but radical shift: instead of only predicting the next token, the model should be capable of predicting the next sentence, the next paragraph, and beyond.

These higher-level units can be treated as macro-tokens—large, semantically meaningful structures represented as embeddings, and predicted using attention and probability mechanisms similar to word-level modeling [1][4].

2.1 Hierarchical Layers

Level 0: Word token prediction (standard)
Level 1: Sentence prediction
Level 2: Paragraph prediction

Level 3: Section or chapter-level prediction
Level 4+: Entire documents, collections, or knowledge domains

Each level is aware of and conditioned by the levels above and below it.

---

3. Practical Simulation

Let's simulate how a hierarchical token predictor might behave with a simple example.
Context:

> "The cat sat on the mat and then it purred quietly as the sun warmed its fur."
> "It blinked slowly and curled up into a soft ball of fur."

Task: Predict the next sentence (F_3), using sentence-level prediction with sampling.

The system selects from a distribution of whole-sentence candidates, maintaining semantic and tonal consistency with prior context. This process can scale upward for paragraph prediction, thematic flow, or even document-level organization [6].

---

4. Architecture and Design Implications

This approach does not discard transformers—it extends and empowers them [1].
Hierarchical tokens can be embedded using:

- Sentence encoders
- Hierarchical attention layers
- Cross-level feedback (top-down and bottom-up)

Each level in the hierarchy operates with the same foundational principles:

- Contextual embeddings
- Attention over prior units
- Predictive sampling with uncertainty

Yet the unit of prediction scales upward:
From tokens → to sentences → to paragraphs → to sections
And further: to chapters, entire books, cross-domain themes, or even knowledge disciplines

This architecture is modular and composable, meaning it can operate:

- On top of traditional LLMs (e.g., GPT-style) [2][3]
- Within functional or symbolic reasoning systems
- Alongside chain-of-thought prompting [6], as a scaffolding framework

Thus, hierarchical prediction becomes a meta-layer, able to empower multiple paradigms of cognition, not just one.

---

5. A Cognitive Pathway Toward General Intelligence

Rather than scaling up prediction depth (more layers, more memory, more tokens), this proposal shifts focus to scaling up abstraction.
Each unit in the hierarchy is a thought container:

- A sentence is a contained action or statement
- A paragraph is a block of meaning or argument
- A section organizes a topic or sub-theme
- A chapter structures an arc of reasoning or narrative
- A book represents a closed knowledge expression
- A collection of books or articles forms a domain of understanding

This means we can envision models capable of:

- Predicting not just how a sentence ends, but how an idea continues
- Navigating domains like law, medicine, or literature with structural awareness [5][7]
- Planning across chapters, across sources, or across disciplines

Moreover, this proposal is compatible with current models. It does not require a complete redesign of LLMs:

- GPT-like models can be wrapped with a hierarchical selector
- Chain-of-thought reasoning trees can be integrated as semantic branches [6]
- Symbolic inference systems can be enhanced with hierarchical context planners

In this light, Hierarchical Tokens is not a replacement for existing architectures—it is a cognitive scaffolding that allows models to think in a more structured, purposeful, and scalable way.

---

6. Conclusion

> "It is not more memory that leads to general intelligence—but more meaningful structure."

Hierarchical Tokens is a simple yet powerful idea: treat language the way humans treat it—not as a stream of disconnected words, but as a fractal of thoughts, layered with purpose, logic, and narrative intent.
This proposal invites researchers and developers to rethink the unit of intelligence not as the token, but as the intention behind it—and to explore how structure, not just scale, may be the missing ingredient in the path to AGI.

References

[1] Vaswani et al. (2017). *Attention Is All You Need*. https://arxiv.org/abs/1706.03762
[2] Brown et al. (2020). *Language Models are Few-Shot Learners*. https://arxiv.org/abs/2005.14165
[3] Radford et al. (2019). *Language Models are Unsupervised Multitask Learners*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[4] Raffel et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. https://arxiv.org/abs/1910.10683
[5] Chowdhery et al. (2022). *PaLM: Scaling Language Modeling with Pathways*. https://arxiv.org/abs/2204.02311
[6] Wei et al. (2022). *Chain of Thought Prompting Elicits Reasoning in Large Language Models*. https://arxiv.org/abs/2201.11903
[7] Bai et al. (2022). *Training a Helpful and Harmless Assistant with RLHF*. https://arxiv.org/abs/2204.05862