📘 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, 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. 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.

---

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.

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.

Rank	Candidate Sentence (F_3)	Probability (%)
1	A breeze drifted in through the window, rustling the curtains gently.	28.6%
2	The room was silent except for the occasional sound of breathing.	18.4%
3	Outside, a bird chirped and the cat's ear twitched in response.	14.9%
4	Nothing moved for a while as the afternoon light faded slowly.	10.7%
5	Its tail flicked slightly, though it remained perfectly still.	8.3%
6	The warmth of the mat seemed to lull it deeper into sleep.	6.5%
7	A soft creak in the hallway didn't seem to disturb it.	4.9%
8	It breathed rhythmically, lost in some feline dream.	3.4%
9	Dust danced in the light, suspended in golden air.	2.3%
10	The cat seemed unaware of the world around it.	1.0%

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.

---

4. Architecture and Design Implications

This approach does **not discard transformers**—it **extends and empowers them**.

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)
- Within **functional or symbolic reasoning systems**
- Alongside **chain-of-thought prompting**, 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
- 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**
- 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.

Author

Rogério Figurelli
April 25, 2025