Understanding Tokens in LLMs

The Fundamental Units of Language Models

Created by Daniel Zaldaña https://x.com/ZaldanaDaniel

What is a Token?

A token is the smallest unit of text that an LLM processes. It can be:

- Complete words: "elephant", "the"
- Word pieces: "ing", "pre", "post"
- Characters: "a", "?"
- Special tokens: [START], [END], [PAD]

Mathematical Foundation

Vocabulary Space:

$$\mathcal{V} = \{t_1, t_2, ..., t_{|V|}\}$$

where $|V|$ is vocabulary size (typically 32K-50K)

Token Embedding:

$$E(t_i) = \vec{e_i} \in R^d$$

where d is embedding dimension

Tokenization Process

BPE Algorithm:

1. Start with character vocabulary
2. Iteratively merge most frequent pairs
3. Stop at target vocabulary size

Mathematical Formulation:

$$score(x, y) = \frac{count(xy)}{\sum_{a,b \in V} count(ab)}$$

Probabilistic Framework

Next Token Prediction:

$$P(t_k | t_{1:k-1}) = \frac{\exp(h_k^T W t_k)}{\sum_{j \in V} \exp(h_k^T W t_j)}$$

where:

- h_k is the context vector
- W is the token embedding matrix
- t_k is the candidate token

Information Content

Token Information:

$$I(t_i) = -\log_2 P(t_i)$$

Sequence Entropy:

$$H(T) = -\sum_{i=1}^{n} P(t_i) \log_2 P(t_i)$$

Practical Impact

Context Window Size:

$$C_{tokens} = \text{max\_position} \times \text{batch\_size}$$

Memory Usage:

$$M = C_{tokens} \times d_{model} \times \text{bytes\_per\_parameter}$$