Tokens and Context Length in Large Language Models

I. Introduction

Tokens and context length are fundamental concepts in modern large language models (LLMs). They determine how textual input is represented, processed, and retained during both training and inference.

II. Definition of Tokens

A *token* is the smallest unit of text processed by a language model. Tokens do not necessarily correspond to complete words and may represent subwords, punctuation symbols, numerical values, or special markers.

Language models convert raw text into tokens using **subword tokenization** techniques.

III. Tokenization Examples

A. Word-Level Example

Given the sentence:

```
The bank approved the loan
```

A possible tokenization is:

```
["the", "bank", "approved", "the", "loan"]
```

B. Subword-Level Example

The word:

```
unbelievable
```

may be tokenized as:

```
["un", "##believ", "##able"]
```

Thus, a single word may correspond to multiple tokens.

IV. Types of Tokens

Language models process several categories of tokens:

- Complete words
- Subword units
- Punctuation symbols
- Numerical values
- Special tokens (e.g., `[CLS]`, `[SEP]`, `<EOS>`)

V. Rationale for Token-Based Processing

Tokens are used instead of full words because token-based processing:

- Enables handling of unknown or rare words
- Reduces vocabulary size
- Supports multilingual language modeling
- Improves generalization across linguistic variations

For these reasons, modern LLMs rely on subword tokenization.

VI. Definition of Context Length

Context length refers to the maximum number of tokens that a language model can process simultaneously.

Formally, context length defines the effective memory window of the model.

VII. Context Length Example

If a model has a context length of 512 tokens:

- Only the most recent 512 tokens are considered
- Tokens beyond this limit are truncated or ignored

VIII. Typical Context Lengths

Model Type	**Context Length (Tokens)**
BERT	512
GPT-2	1,024
GPT-3	2,048
Modern LLMs	8k, 16k, 32k+

Exact limits vary depending on the specific model architecture.

IX. Importance of Context Length

Context length directly influences:

- Long-form conversations
- Document comprehension
- Retrieval-Augmented Generation (RAG) systems
- Retention of earlier information

A. Illustrative Scenario

```
User: My name is Arshad.
[long conversation]
User: What is my name?
```

If the initial statement lies outside the context window, the model cannot answer correctly.

X. Relationship Between Tokens and Context Length

- Context length is measured strictly in tokens
- It is independent of word or character count
- Longer words may generate more tokens
- Large documents may exceed the context limit

XI. Token Budget Constraint

The total number of tokens processed must satisfy:

$$\text{Input Tokens} + \text{Output Tokens} \leq \text{Context Length}$$

A. Example

- Context length: 4096 tokens
- Input size: 3500 tokens
- Maximum output size: approximately 596 tokens

XII. Reasons for Context Length Limitations

Context length is constrained due to:

- Quadratic computational complexity of attention mechanisms
- Increased memory requirements
- Higher computational cost

As a result, strategies such as chunking and selective retrieval are widely used.

XIII. Tokens and Context Length in RAG Systems

In Retrieval-Augmented Generation systems:

- Documents are divided into smaller chunks
- Only relevant chunks are retrieved
- Input size is kept within the context window

XIV. Common Errors

Common mistakes include:

- Assuming one word corresponds to one token
- Ignoring output token consumption
- Providing entire documents without chunking

XV. Key Observations

- Tokens are the fundamental units processed by language models
- Context length defines the memory capacity of the model
- All limits and costs are measured in tokens
- Larger context windows improve long-text reasoning at higher computational cost

XVI. Concise Examination Definition

Tokens are subword units processed by language models, and context length denotes the maximum number of tokens a model can consider simultaneously.

XVII. Memory Aid

Tokens define what the model reads, and context length defines how much the model can remember.