Session 1 Handout: From Text Prediction to Transformers

FAST-14 How Large Language Models Work · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Why this session matters

Every odd thing an LLM does (a different answer each time, an invented citation, a bill that grows with file size) traces back to one idea: it predicts the next token from its context. Understand that and you can predict its behavior.

Key concepts

1. Next-token prediction

The model reads the context and outputs a probability for every possible next token. It picks one, adds it, and repeats. For "Our shuttle leaves the inn every", "morning" might get a high probability and "day" less. Sampling settings decide how it picks.

2. A short history

Rules (Georgetown-IBM 1954, ELIZA 1966) → statistics (Shannon's n-grams, 1951) → neural language models (Bengio, 2003) → RNNs and LSTMs → attention for translation (2014-15) → the transformer (Vaswani et al., 2017).

3. Tokens

A token is a chunk of text: a common word, part of a word, or punctuation. Rough English rule: 1 token ≈ 0.75 words. Rare names, numbers, and non-English text use more tokens. **You pay per token** (input and output priced separately), and limits are in tokens.

4. Embeddings and model families

An embedding is a list of numbers capturing meaning; similar meanings sit close together. Families: **decoder-only** (chat models; generate text), **encoder-only** (embeddings, search, reranking; still widely used), **encoder-decoder** (translation, summarization).

5. Attention

Each token builds a query, and every earlier token offers a key and a value. Attention weights say how much each earlier token matters. Multi-head attention looks for several kinds of relationships at once. Causal masking stops the model from seeing future tokens. The KV cache stores earlier keys and values so generation is faster.

Worked example (fictional)

Cape Harbor Inn's FAQ: about 1,450 words, which is about 1,950 tokens. Each question sends the FAQ plus 100 tokens and gets 150 back. At $1 / $4 per million input / output tokens: about **$0.00265 a question, $7.95 for 3,000 a month**.

Temperature

Low (0 to 0.2): consistent, best for facts and extraction. Higher (0.7 to 1.0): more variety, good for drafting. Temperature 0 is close to deterministic, not guaranteed identical. Some reasoning models fix the setting.

Checklist

- ☐ I can explain next-token prediction in one sentence.
- ☐ I can estimate tokens and cost for a document.
- ☐ I can say what attention does in plain words.
- ☐ I chose a temperature for the FAQ bot and for marketing copy (W1).