CHAPTER 04 / FOLLOW THE GENERATION

One token at a time.

The loop behind a common autoregressive response.

Read context. Score tokens. Choose. Repeat.



Simplified autoregressive text generation. Candidate scores shown are hand-picked illustrations, not a live model result.

Context

In this autoregressive example, available tokens and learned parameters shape the next-token scores.

Generation

A decoding rule chooses the next token. Append it to the context and run the loop until a stopping condition.

Verification

Likely text can still be wrong. Retrieval, tools, and external checks are separate system capabilities.

↗ **Fluent is a property of the answer. Verified is a property of the evidence.**

PUT IT TO WORK

For a policy answer, retrieve the approved policy, cite the relevant passage, and verify the claim against it before a user relies on it.

CLOSE THE PAGE. EXPLAIN IT.

Where in your application does an answer get checked against reality?