How LLMs Generate Text - The Complete Process

From Input to Output: Understanding Each Step

1 Input Processing

- User provides prompt or query
- Text is received by the API endpoint
- Context and conversation history are gathered

2 Tokenization

- Text is broken into tokens (subword units)
- Special tokens added (start, end, system)
- Token IDs mapped for model processing

3 Model Processing

- Tokens pass through transformer layers
- Self-attention mechanisms identify relationships
- Neural network computes probability distributions

4 Next-Token Prediction

- Model predicts most likely next token
- Sampling parameters (temperature, top-p) applied
- Token selected based on probability and randomness

5 Output Generation

- Process repeats until stop condition met
- Tokens are decoded back to human text