# One-Step Prompting: End-to-End Technical Lifecycle

Input: 'Explain AI in one simple sentence.'

Output: 'AI is a technology that enables machines to learn like humans.'

## 1. Text Normalization & Tokenization

The input string is cleaned and split into smaller units called tokens.   
Real LLMs use Byte Pair Encoding (BPE) or WordPiece tokenization.  
Example tokens: ["Explain","AI","in","one","simple","sentence","."]

## 2. Vocabulary Mapping

Tokens are mapped to unique numeric IDs using vocabulary lookup tables (stoi/itos).   
This enables the model to process numeric data instead of text.

## 3. Token Embeddings + Positional Encoding

Each token ID is mapped to a dense vector (embedding) that represents its semantic meaning.   
A positional encoding vector is added to preserve word order information.

## 4. Transformer Self-Attention

Self-attention calculates relationships between tokens to understand context.  
Q = X·Wq, K = X·Wk, V = X·Wv  
Attention = softmax((Q·Kᵀ)/√d\_k)  
Output = Attention·V

## 5. Feed Forward Network

Each token’s contextualized representation passes through a feed-forward neural network for non-linear transformation.  
H = ReLU(Z·W1 + b1)·W2 + b2

## 6. Output Projection and Softmax

The model projects hidden states to vocabulary logits.   
Softmax converts these logits into probabilities representing the likelihood of the next token.

## 7. Decoding (Generation)

Using the token probabilities, the model generates text sequentially — token by token — until the <EOS> token appears.  
Common decoding methods include greedy search, top-k, and nucleus sampling.

## 8. End-to-End Python Example

The accompanying Python script (one\_step\_prompting\_demo.py) demonstrates the entire lifecycle:  
- Tokenization  
- Embedding generation  
- Attention simulation (untrained)  
- Bigram model decoding to generate the final output.

## Output Example

AI is a technology that enables machines to learn like humans.

## Key Takeaways

• BERT → understands the request.  
• GPT → generates the response.  
• T5 → both understands and transforms text.  
• One-step prompting involves a single clear input and one direct output without iterative refinement.