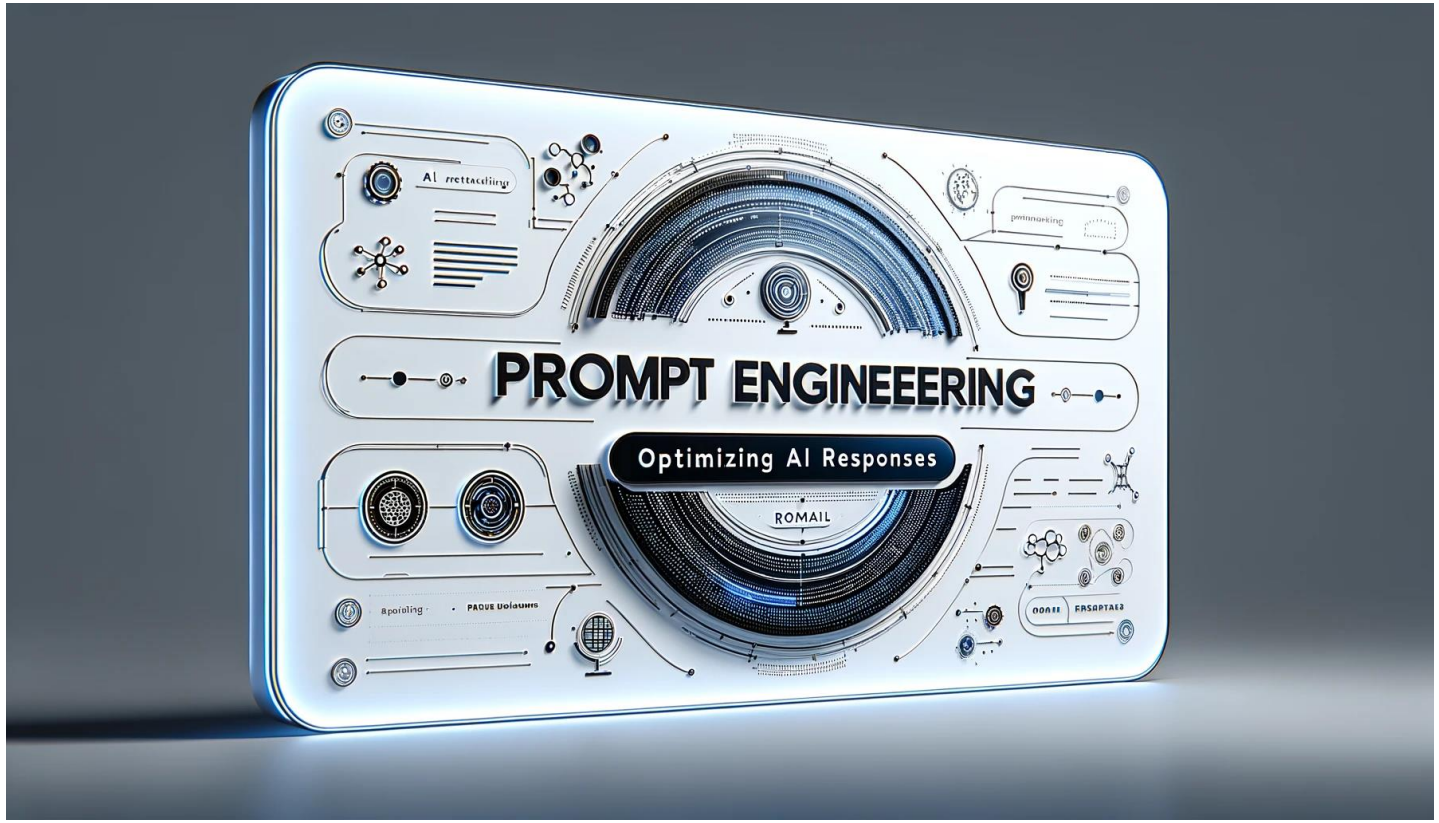
Prompt engineering



ChatGPT (text-to-text)

1. **User-Friendly Interface**: Clean and intuitive design. Includes a simple text input and output area where users can type their queries and receive responses.
2. **Conversational Capabilities**: This interface is optimized for generating conversational AI responses.
3. **Real-Time Interaction**: Responses are generated in real-time, making it suitable for live demonstrations, educational tools, and real-time problem solving.
4. **Accessibility**: Web-based: accessible from any device with internet connectivity and a web browser
5. **Versatile Use Cases**: From casual conversation and troubleshooting to explaining scientific concepts or providing programming help.
6. **Privacy and Security**: OpenAI ensures that interactions with ChatGPT are handled with a focus on privacy and security.
7. **Customizable Settings**: Users can often customize their experience and develop task-specific GPTs.

Generative AI

- Processing in the Neural Network:
  - These embeddings are then fed into a neural network, typically a transformer model
  - The transformer uses layers of attention mechanisms to weigh the importance of different tokens relative to each other
- Contextual Understanding:
  - Embeddings they are updated based on the surrounding context within the sequence
  - Calculate overall meaning and how each token relates to the others
- Token Prediction:
  - The final layer of the transformer model outputs a new vector for each input token
  - A softmax function is applied to convert it into a probability distribution over all possible next tokens
- Generating Output:
  - The token corresponding to the highest probability is typically chosen as the next token in the sequence
  - This process is repeated for generating each subsequent token, using the newly generated tokens as additional context

Prompt engineering tips

- Being Explicit and Detailed:
  - **Clarity is Crucial**: Ensure that the prompt clearly communicates the task or question.
  - **Detail Oriented**: Including specific details can guide the model to generate more accurate and relevant content. Specify the context, desired format, and any particular focus areas.
- Effect of Temperature:
  - **Control Creativity**: Temperature is a parameter that affects the randomness of the response.
  - **Tuning for Task**: Adjusting the temperature can help balance between creativity and accuracy.
- Use of Stop Sequences:
  - **Limit Response Length**: Stop sequences can be used to signal the model when to end a response, preventing overly lengthy or rambling outputs.
  - **Structured Outputs**: They are particularly useful in generating lists, bullet points, or responses with a defined endpoint.

Prompt engineering tips

- Leveraging Examples (Few-Shot Learning):
  - **Contextual Learning**: Providing one or more examples within the prompt can guide the model on the expected style, format, or depth of the response.
  - **Improves Relevance**: Examples act as a direct template for the type of answer or content desired.
- Iterative Refinement:
  - **Prompt Iteration**: Start with a basic prompt and refine it based on the model's responses.
  - **Feedback Loops**: Use feedback from the model's outputs to continuously improve the precision of the prompts, adapting to the strengths and weaknesses observed.
- Ethical Considerations:
  - **Bias and Sensitivity**: Be aware of and actively manage potential biases in AI responses.
  - **Transparency**: Ensure that the use of AI, especially in sensitive contexts, is transparent to users about the involvement of AI in content creation or decision-making processes.