Prompt Variables and Definitions

- **Goal/Task**
  - What you want to accomplish with the prompt.

- **Prompt**
  - Any type of information (e.g. text, image, question) you give to an AI model that elicits a response.

- **System prompt**
  - Defines guidelines and boundaries for the entire experience.
  - Different than a regular prompt because it's targeted towards the overall experience rather than a specific prompt.

- **AI model type (e.g. GPT-3, ChatGPT, GPT-4, Claude 2, Llama 2)**
  - There are pros and cons to each model. You'll need to experiment with which one works best with your prompt.
    - Here a few resources to help you get started:
    - Voiceflow - You can use Voiceflow to test your prompts and variables in different models.
    - When to use ChatGPT vs. GPT-4 - Voiceflow Pathways Article
    - Llama-2, GPT-4, or Claude-2 - Marktechpost

- **Knowledge base (e.g. text, documents, URLs)**
  - Information given to the models about a specific topic.
  - Expands what the model knows and helps it make decisions.
  - There is room for hallucinations when knowledge bases are involved, so the content and output needs to be optimized and tested thoroughly on a regular basis.

- **Temperature**
  - The level of creativity of the response on scale of 0-1, 0 being the least creative and 1 being the most creative.
  - 0.7 seems to be where many people start.
  - Hallucinations can happen even when the temperature is at 0, so you still need to test the content.

- **Tokens**
  - Tokens are the breakdown of prompts into smaller groups of characters or words.
  - The models process and generate the prompt input/output based on the tokens.
  - The tokens can greatly affect the understanding, output, and behavior of the prompt, which makes this variable a bit of a wild card.
  - Each model has a different token breakdown.

    - The number of tokens varies depending on the model, language, and other factors, so there is no consistent way to quantify the tokens at this point.
    - Tokens affect the cost of the experience.
    - The more tokens you use, the larger the cost.
        - Here's a resource that breaks it down in more detail: What are tokens? - Microsoft

- **Prompt response type (this might be Voiceflow-specific, but wanted to define anyway)**
    - **Prompt:** Response is only based on the information given in the prompt. No context is included from previous turns. Good for very specific responses.
    - **Memory:** This broadens the context of the conversation and allows the LLM to respond only using the context from the previous turns (8 turns with Voiceflow).
    - **Prompt and memory:** The model will generate a response based on both the prompt and memory.
        - Here's more detailed documentation - Conversation memory - Voiceflow documentation

- **Examples of expected behavior and output**
    - Screenshots or copies of the actual responses, taken while testing each prompt. The more examples you have, the better.

- **Behavior and output considerations**
    - Anything that might go wrong with the response or anything within a response that could have a harmful effect on people.

- **What worked well about the prompt**
    - What was successful about the prompt?
    - Was there a specific word that got a better response than others?
    - Was the response descriptive enough?
    - Does the output have enough detail?
    - Does the right tone of voice come across?
    - Are you happy with the length of the response?
    - Is the persona consistent with the rest of the experience?
        - Keep asking yourself these types of questions to find the output you're looking for.

- **What could be improved**
    - If you're not happy with the output, what could be better?