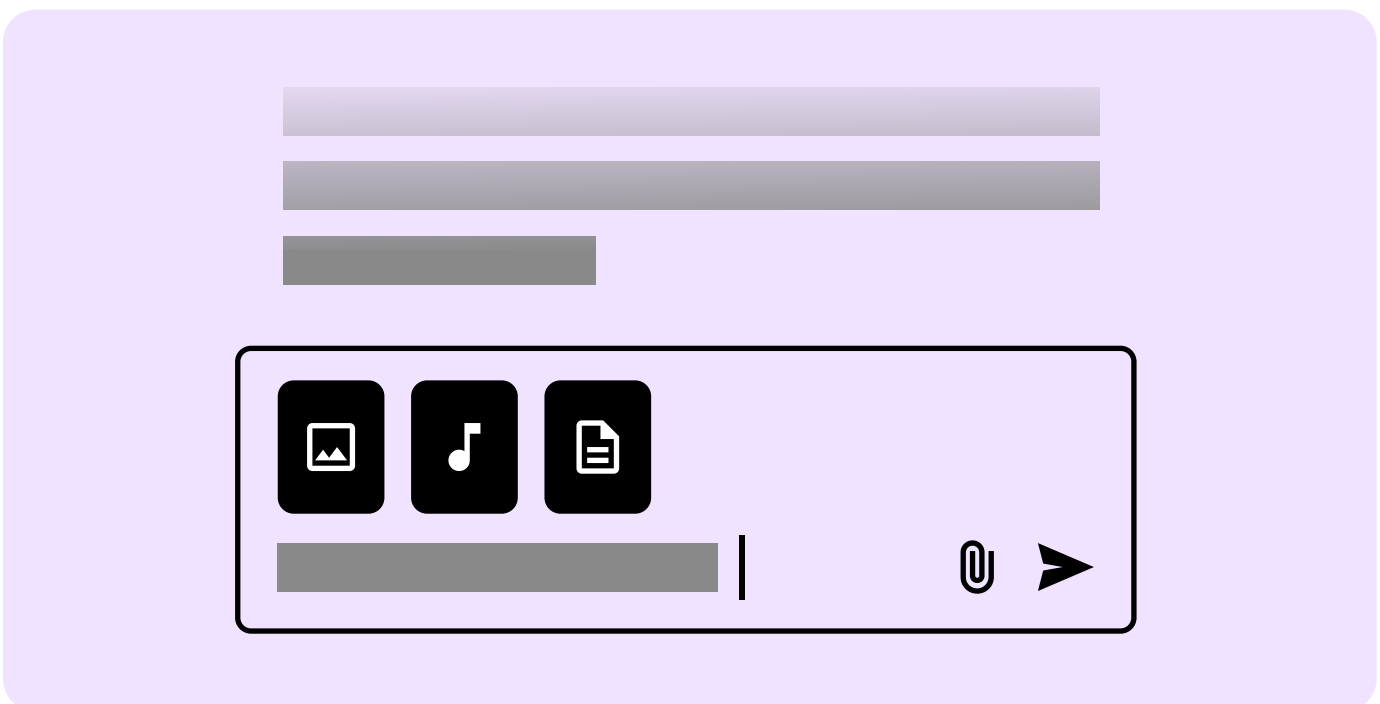
MULTIMODAL PROMPTS

Combine text with images or other media to improve efficiency and accuracy



Overview

While textual prompts have been a foundational method for interacting with LLMs, they are not always the most user-friendly or efficient form of input.

Challenges with Textual Input:

- Descriptive Limitations: Articulating specific requests with words can be cumbersome and imprecise. In many instances, a picture truly is worth a thousand words, enabling users to convey complex ideas more effectively through visual means rather than textual descriptions.
- Format Efficiency: When information is already available in a different format (such as images, audio, or video), converting it into text can be inefficient and may lead to a loss of detail. Providing input in its original format ensures better accuracy and preserves the richness of the information.
- Literacy and Comfort: Literacy levels vary across user bases, and some individuals may feel uncomfortable or find it challenging to express themselves in writing. This can create barriers to effective communication and hinder user engagement.
- Physical Limitations: There are times when users may not have their hands free for typing, such as when they are multitasking or in situations where manual input is impractical.
- Device Constraints: Typing long prompts on small touch devices, like smartphones or tablets, can be particularly challenging and time-consuming, often leading to errors and frustration.

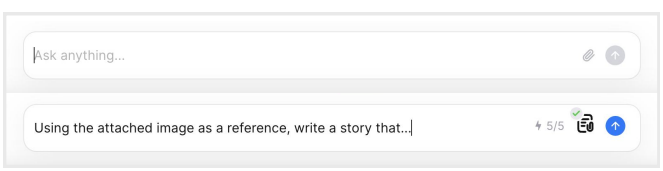
Solution

To address these challenges, incorporating multimodal prompts—input methods that combine text, images, voice, and other formats—can significantly enhance user experience and interaction accuracy. By leveraging multiple modes of input, systems can become more intuitive, inclusive, and effective, catering to a broader range of users and their unique needs.

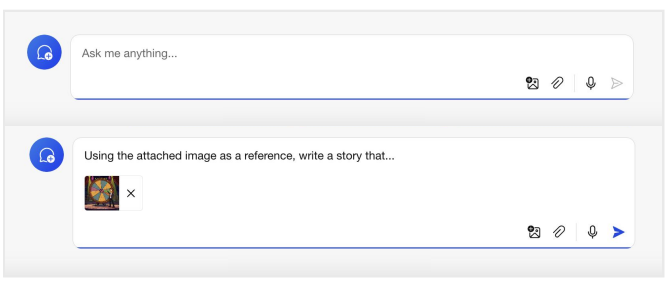
The multimodal prompt input

The most common implementation of this solution is a prompt input component. This component primarily allows users to write or dictate their inputs, and it also supports attaching files in various formats.

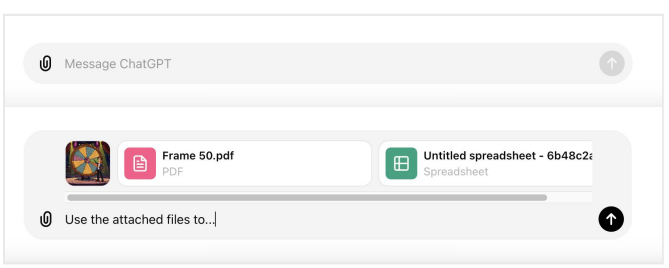
This approach is widely used in general-purpose conversational interfaces, where interacting with the model is mainly done through natural language.



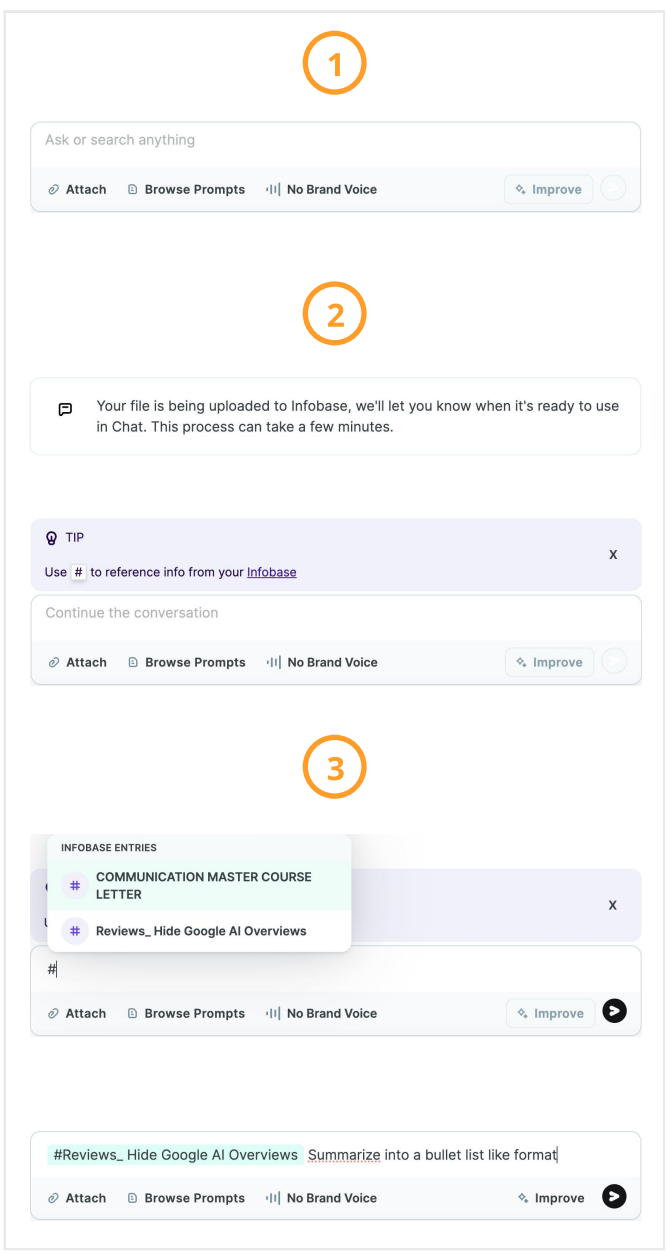
YOU allows a single file attachment and makes this obvious by replacing the button with a static icon.



Bing displays a thumbnail preview of the attached file and allows dictation as an alternative to typing. Even though only a single file can be attached, the interface doesn't make this clear.



ChatGPT takes in multiple files and file formats and displays nice, visual previews. Doing this in the prompt container leads to space limitations, as it can be observed above.

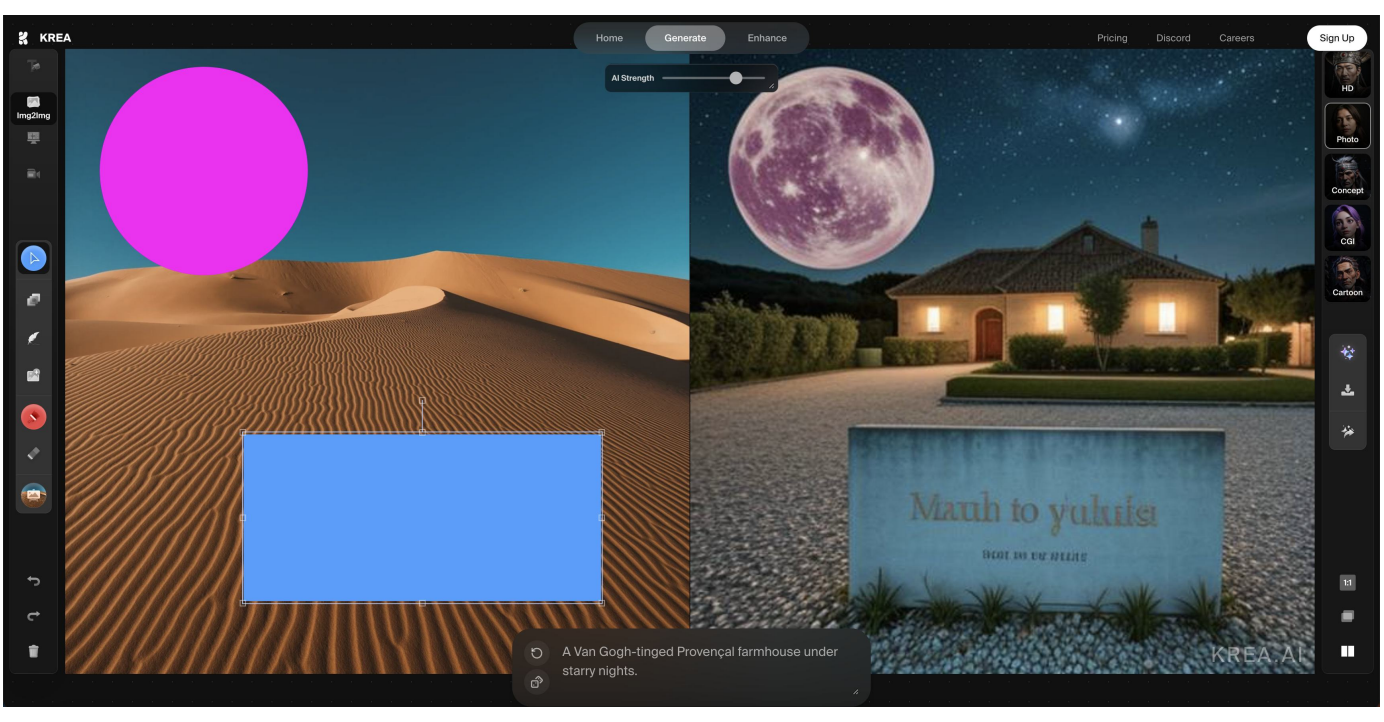


Copy.ai adds the attached files to an infobase first. Users can reference them in prompts by invoking a dropdown with the # key. The reference files can be reused in multiple prompts this way.

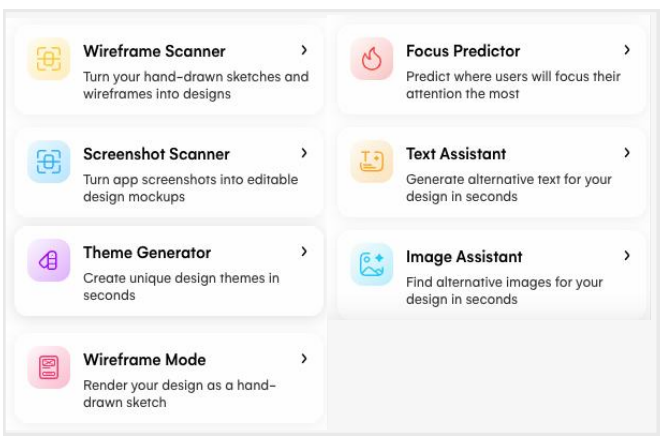
Specialised alternative prompts

While text is a good primary choice for conversational interfaces, specialized products that address specific user needs can benefit from prioritizing other forms of input, leading to more usable interfaces.

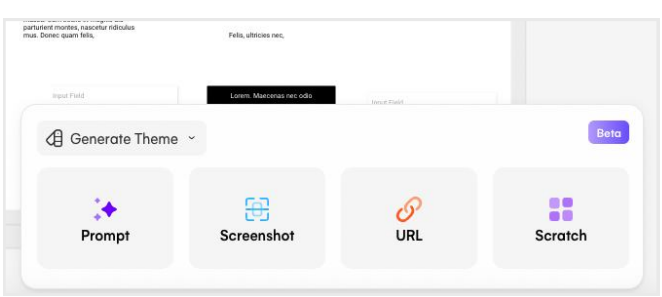
In many of these cases, textual instructions are still permitted, but they are not the primary way of interacting with the models.



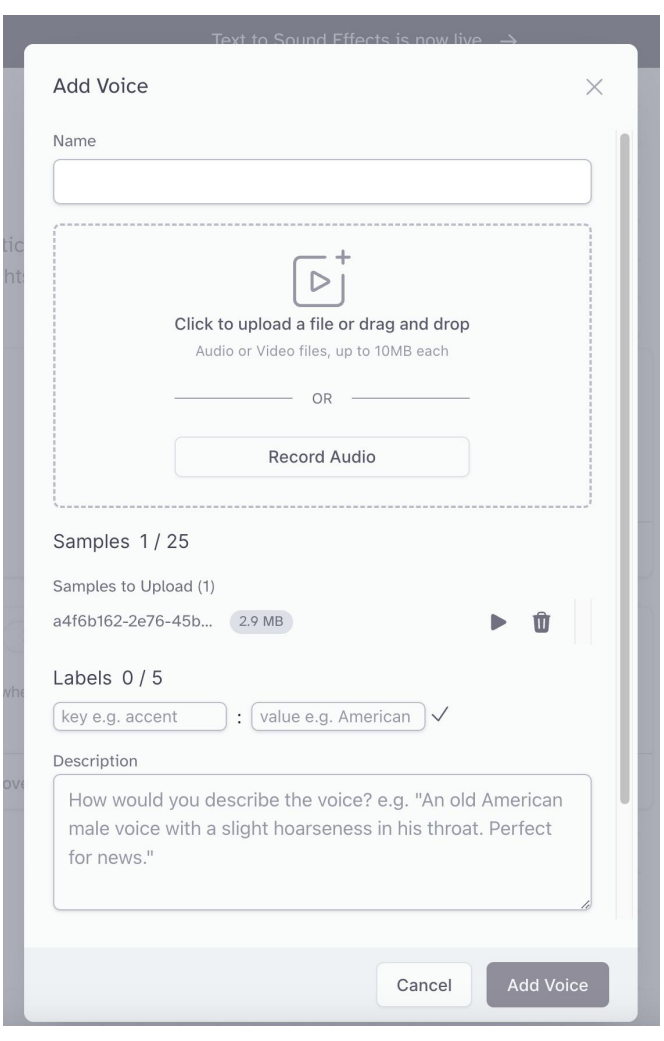
Krea combines reference images with graphic primitives and text in a split screen interface that speeds up image generation and leads to more precise results.



UIzard advertises it's "magic" features in a dedicated menu in the sidebar. These are some AI capabilities that take as input hand-drawn sketches or screenshots, or simply hide the complexities of prompt writing under a button.



For generating a theme, UIzard offers multiple input options. This is definitely one of those scenarios where a picture, or an URL to a website is worth a thousand words.



When creating a synthetic voice, ElevenLabs focuses on sample audio files as primary input. The textual instructions are secondary and optional.