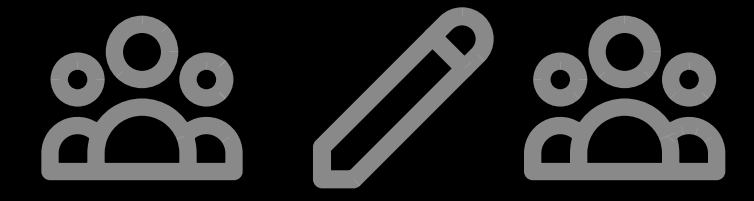


How do AI models process



# multimodal inputs?

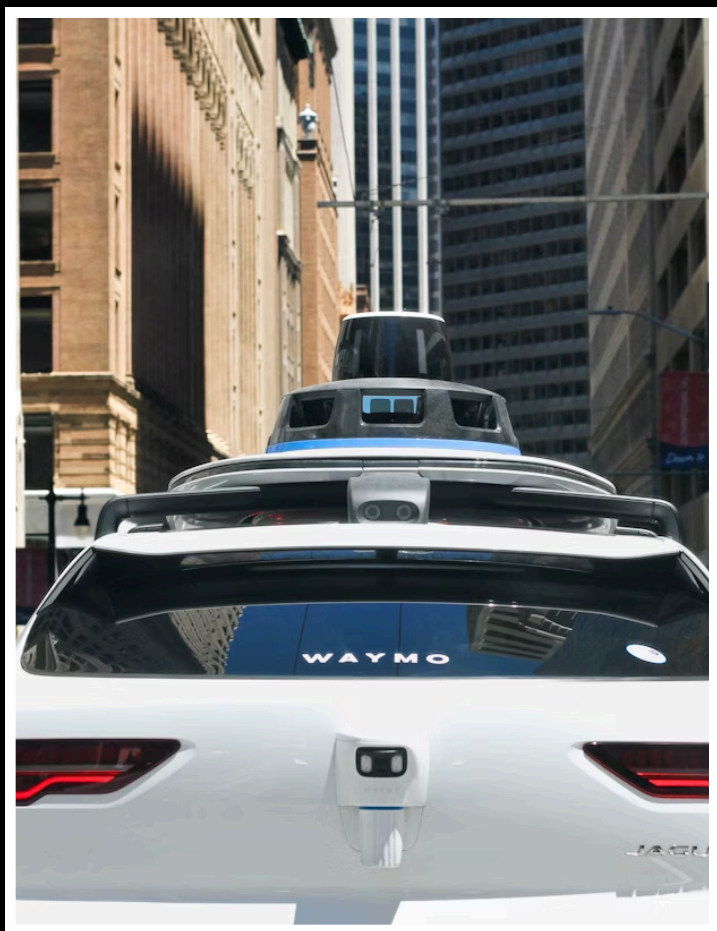
NeurIPS AI Education Resource Showcase



Multimodal AI models are models that take in and process multiple types of data (images, audio, text) all at once.



Why is this relevant?



## Self Driving Cars

Sensory data (LiDAR), camera information, and audio → actions that guide the user safely and accurately.



## Medical Analysis

Patient scans (images) + patient records → fast, accurate diagnostics for individual patients.



# First, how does input arrive to an AI model?

## Text Encoding

To allow machines to understand text, we need some way to convert them into numerical representations. While not commonly used, one basic method is called one-hot encoding, where each word is given a certain index (denoted by a 1 in the matrix).

The  
cat  
leaps.

 $\rightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$ 

## Image Encoding

Similarly, we can encode images by matrices of RGB values, each denoting the color of one pixel within the image. This means each value can range from 0-255.

 $\rightarrow \begin{bmatrix} 231, 37, 59, 187, 109 \\ 187, 110, 20, 32, 53 \\ 112, 78, 23, 90, 135 \\ 255, 18, 209, 215, 11 \\ 68, 99, 156, 101, 234 \end{bmatrix} \begin{bmatrix} 231 \\ 33 \\ 98 \\ 21 \\ 77 \end{bmatrix} \begin{bmatrix} 38 \\ 27 \\ 156 \\ 19 \\ 176 \end{bmatrix}$ 

Note that many other encoding methods exist for different types of input!

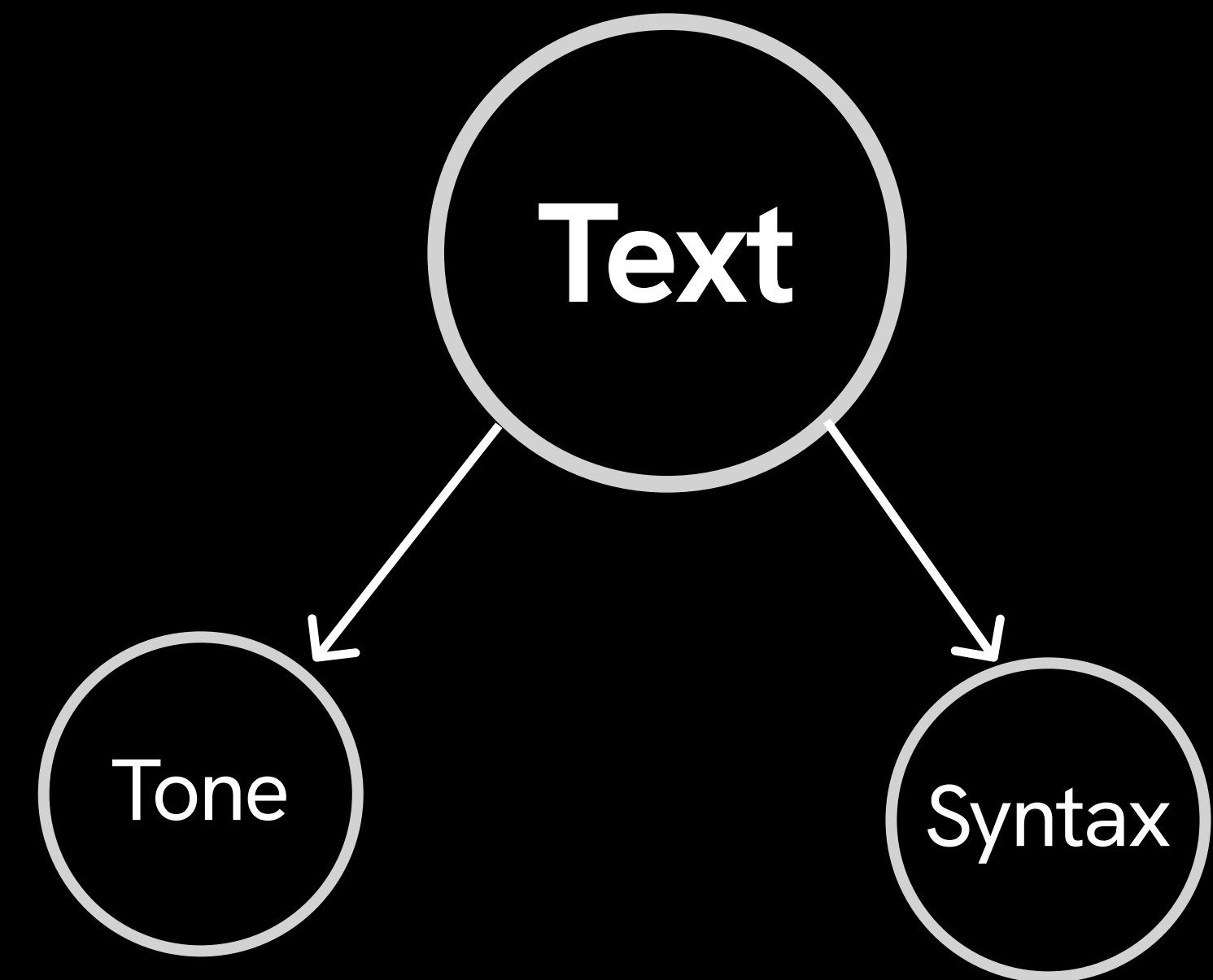
# To learn the patterns within the numbers, we use neural networks!

Oftentimes, there are patterns in these encodings that we can use to extract meaning.

Transformers are types of neural networks that help us parse images, text, etc.! In multimodal models, we use the outputs of these transformers to train our model to draw connections.

## What's a Neural Network?

Modeled after the human brain, a neural network uses layers of "neurons" to minimize errors and improve accuracy. They're particularly useful for complex pattern recognition and prediction!





Let's go through an example with  
images and text!

# OpenAI's CLIP (Contrastive Language-Image Pretraining) model splits the encoding for images and text.

- ① We first encode our image and text pair separately.

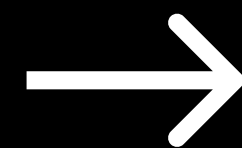
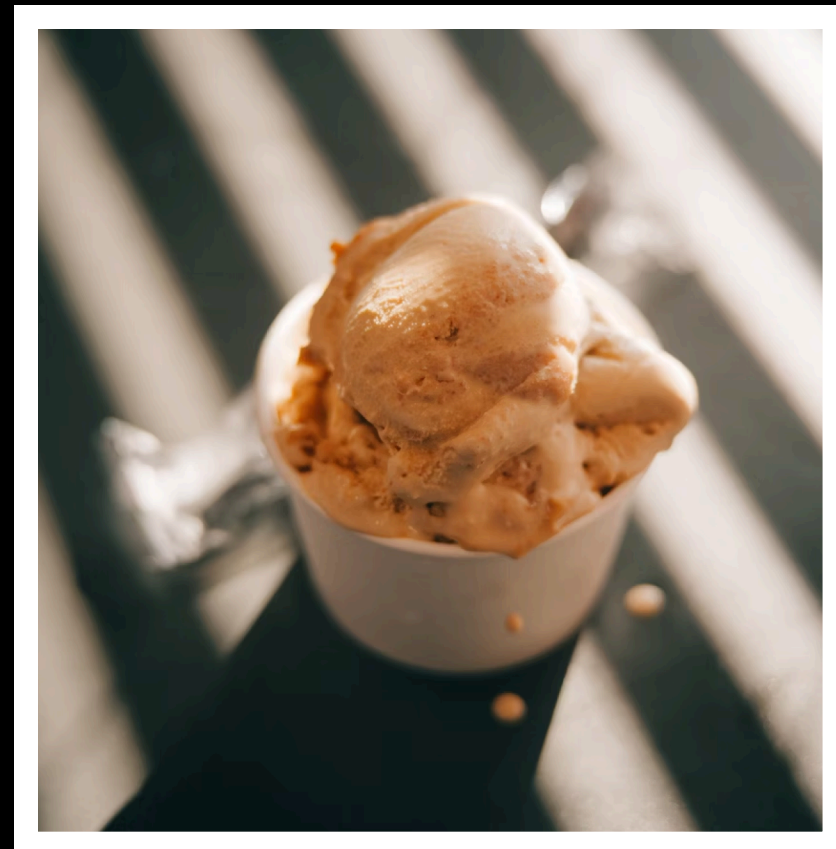
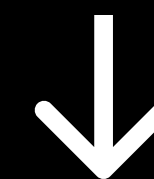


Image Encoder  
(such as Resnet, ViT)

"A cup of orange ice  
cream on the ground  
in the sunlight."



Text Encoder  
(such as BERT)

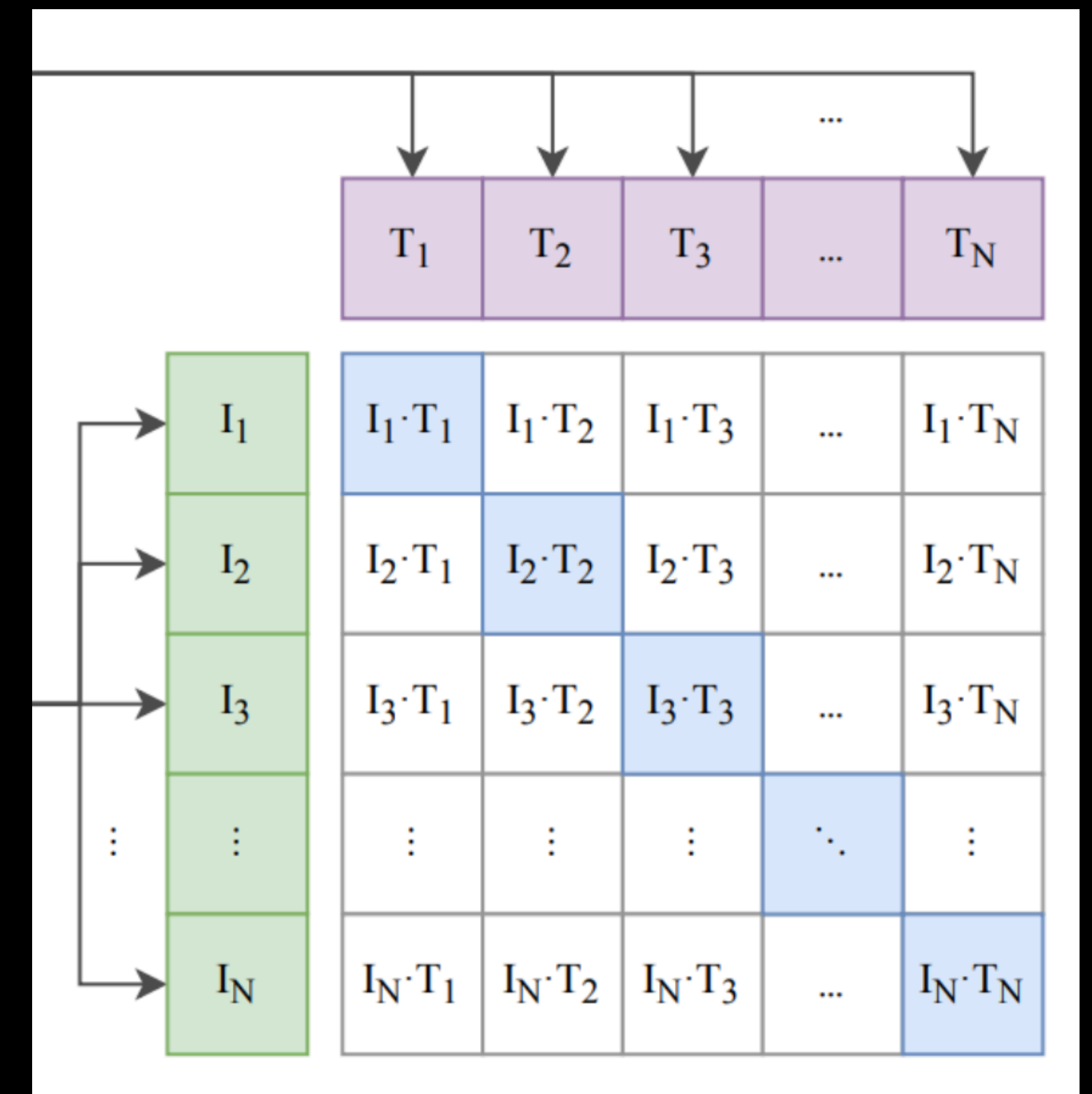


② Now that the image and text are in the same format, we want our model to adjust weights so that the numbers for the image and text are as similar as possible.

In this image, let  $I_i$  represent the encoding data of the  $i$ th image. Similarly, let  $T_i$  represent the data of the  $i$ th caption.

$I_i \cdot T_i$  represents the distance between the two inputs' encoding.

Our model aims to minimize the distances along the diagonal since they're the matching pairs.

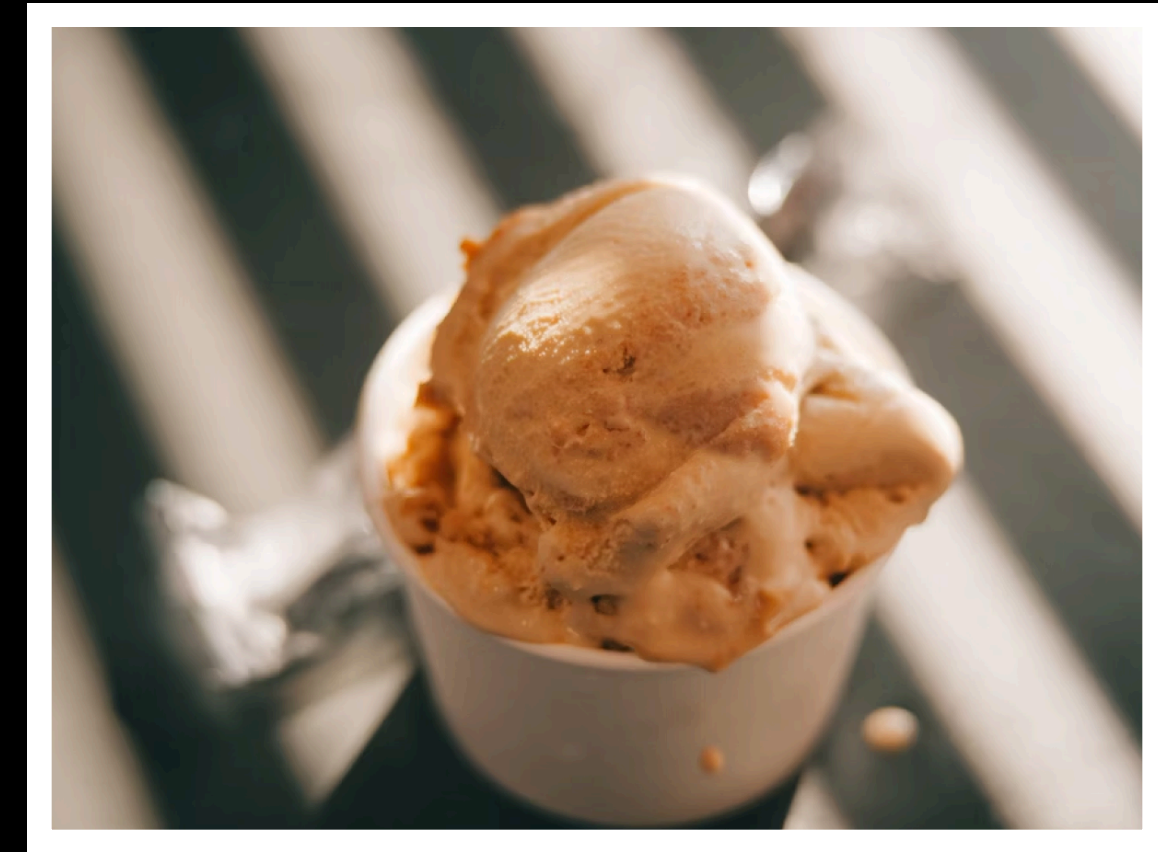




### ③ Now, we can use our CLIP model to classify images given both text and images!

This is especially effective because of CLIP's ability for zero-shot image classification (where the model has not been trained on images it's being prompted on).

CLIP is used in image generators such as DALL-E to translate text descriptions into corresponding images.



"A cup of orange ice cream on the ground in the sunlight."

0.97 ✓

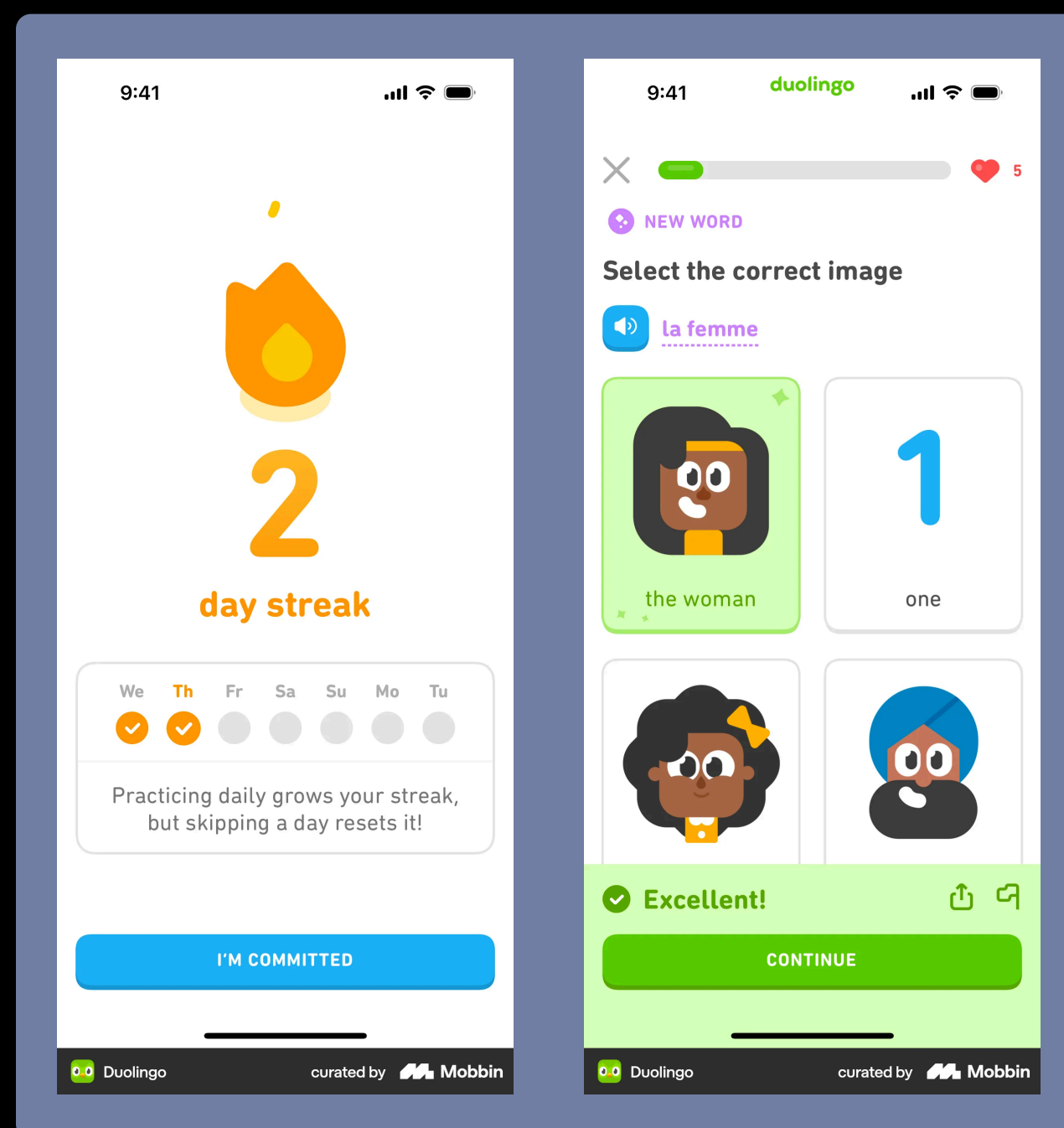
"Vanilla ice-cream cone in a cup on the street."

0.32

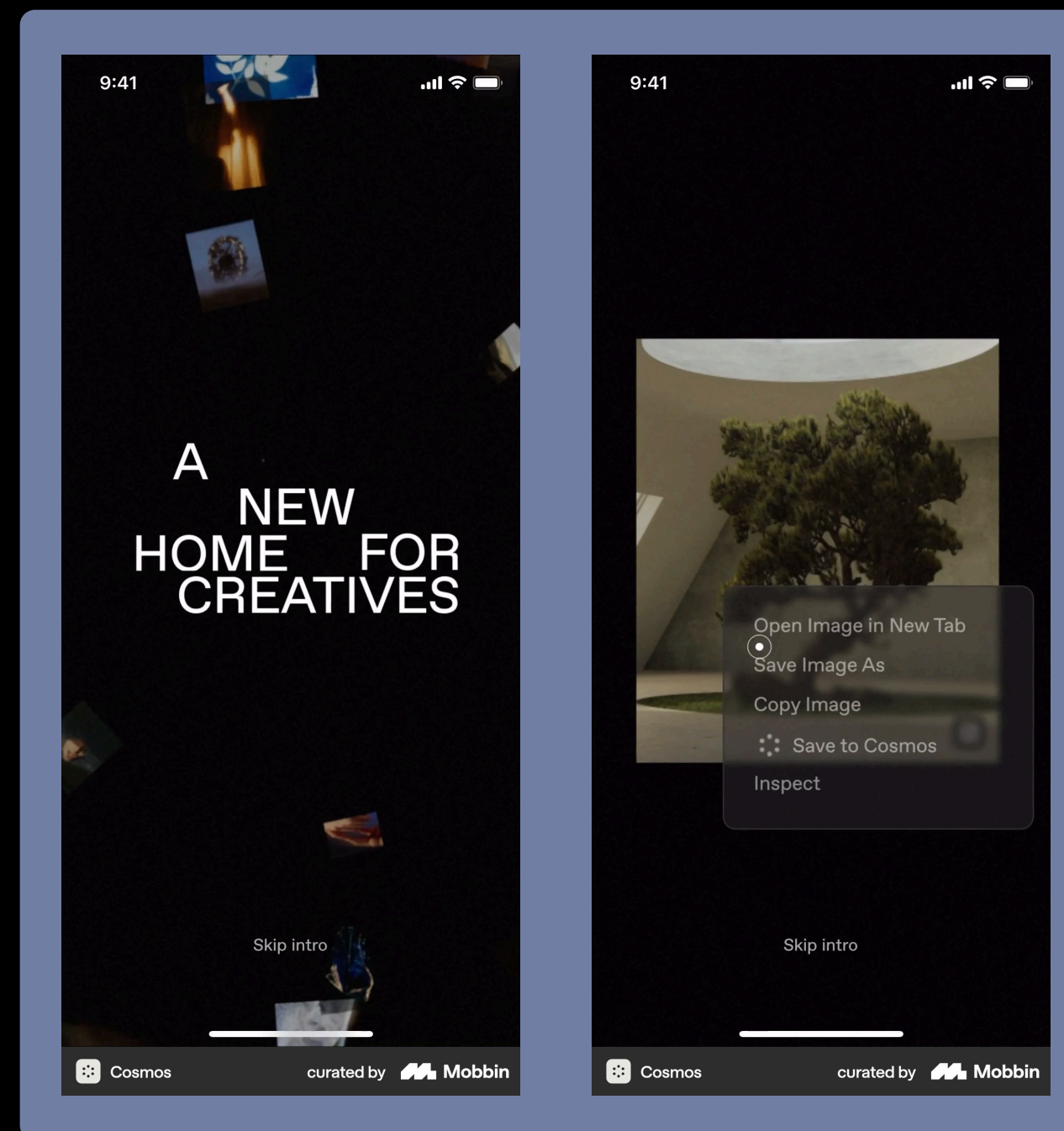


While a general CLIP model trained on 400 million image-text pairs has been released, the architecture has been widely used and applied in unique problems requiring multimodality.

Let's check out one application in UI/UX Design!



*Duolingo, Sourced from Mobbin*



*Cosmos, Sourced from Mobbin*



Is this user interface  
design good?

# What is UI/UX Design?

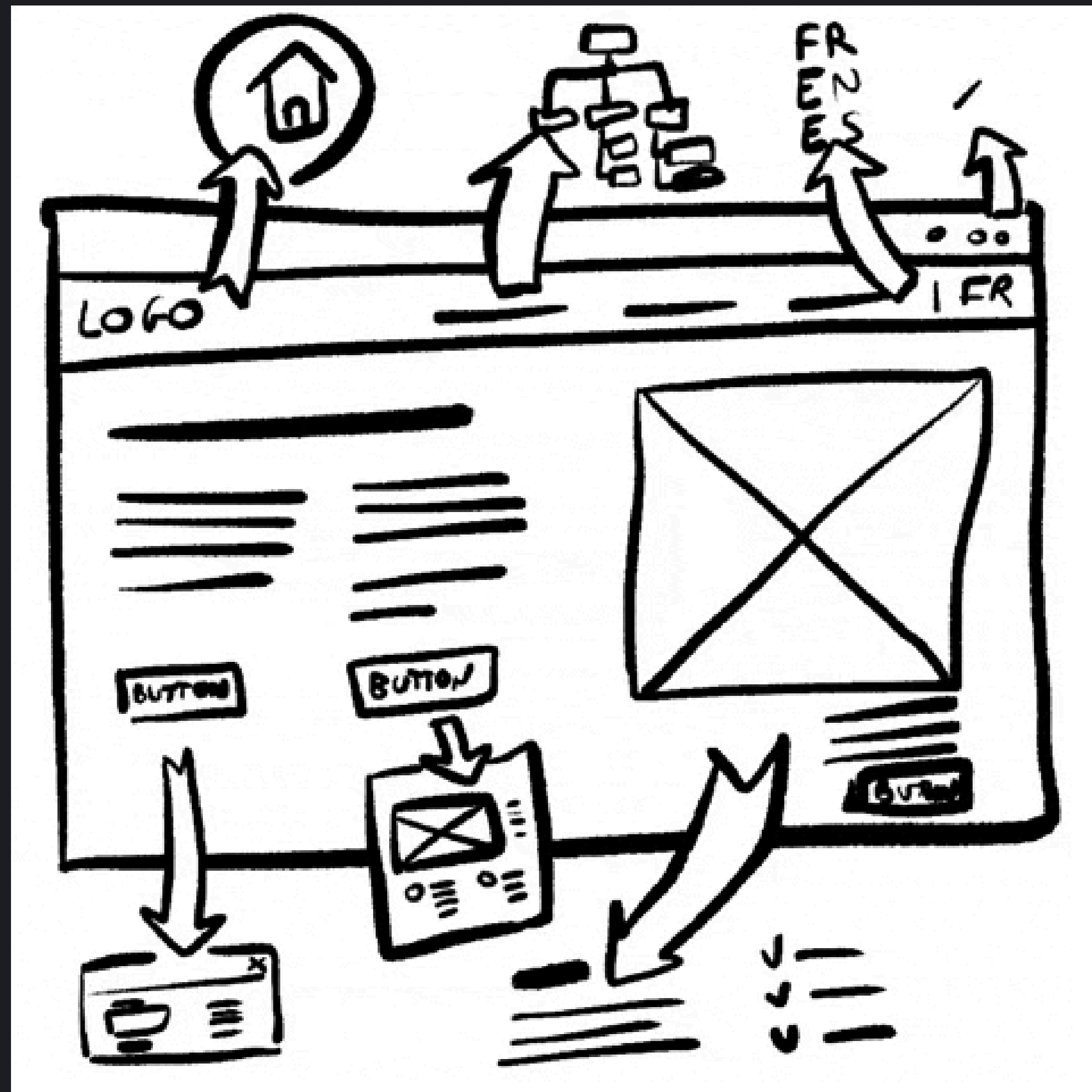
## User Interface

The visual feel and aesthetic of the product. Involves brand identity, aesthetic.

## User Experience

Overall user flow: accessibility, complexity, how intuitive your product is as a whole.

When designing, we're finding the most optimal and intuitive experience for our users.





What makes a user interface  
or experience bad?

confusing sign up flows  
switching costs  
non-descriptive errors  
excessive steps to  
complete easy tasks  
poor mobile  
responsiveness  
inconsistent navigation  
lack of undo options  
too many pop-ups or modals  
unclear button labels  
tiny tap targets  
unexpected logouts  
slow load times  
hidden settings  
forced account  
creation  
overuse of  
animations  
confusing iconography  
poor contrast  
bad  
readability  
lack of feedback on actions  
missing search functionality  
cluttered layout  
no autosave  
inconsistent UI elements  
annoying cookie  
banners  
difficulty returning to previous steps

**A lot of things!**



# Multimodal AI is a good tool for this problem because:

## High Costs for Testing

Because traditional forms of UI testing involve the use of human testers, costs can add up quickly. For aspiring UI designers seeking feedback on their designs, AI could provide a cheap and effective solution to this issue and simulate various user personas.

## No Clear Guidelines

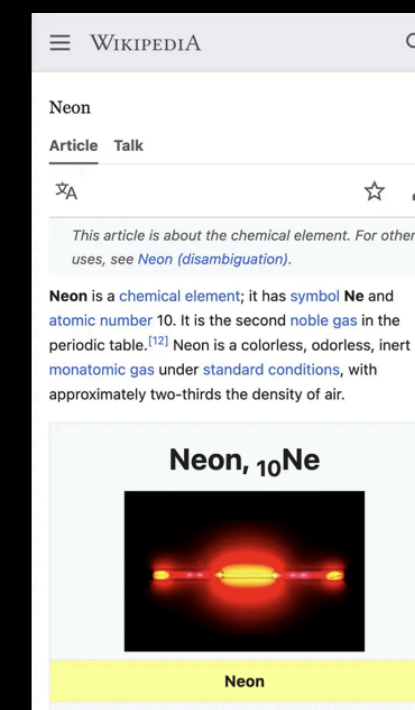
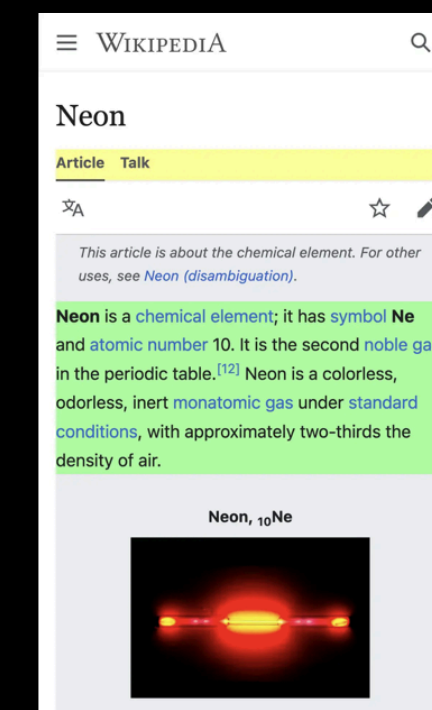
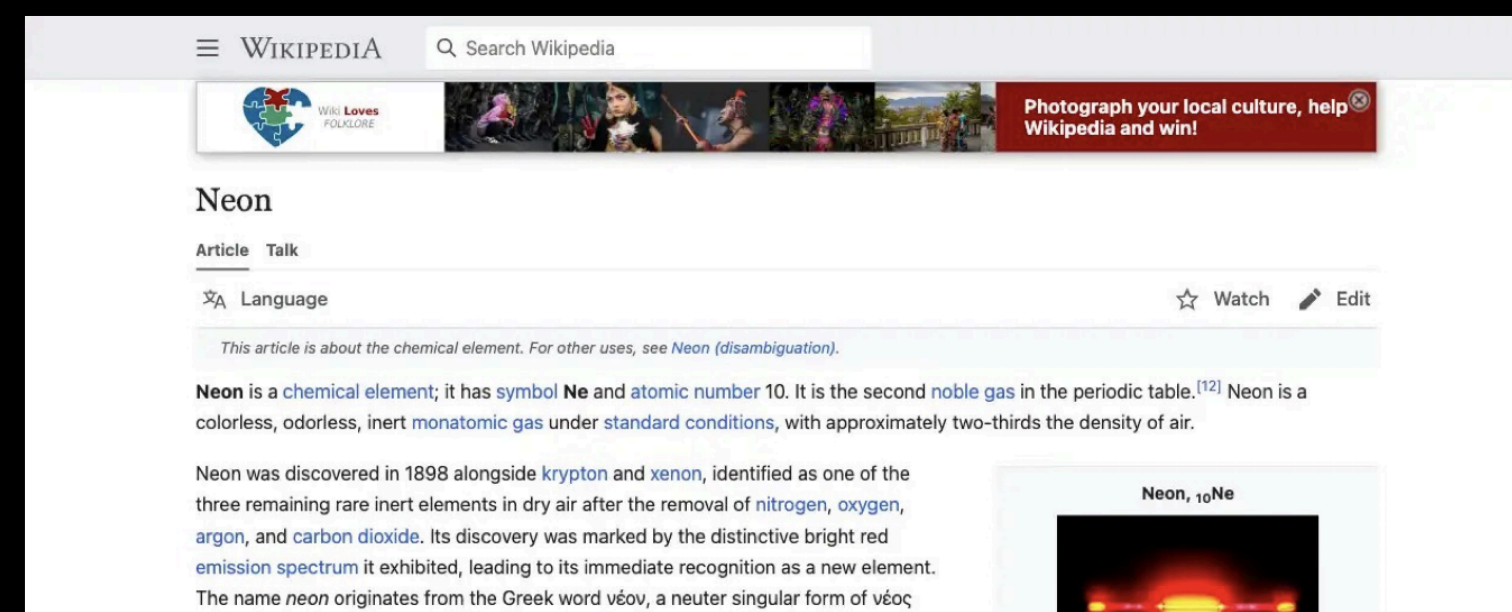
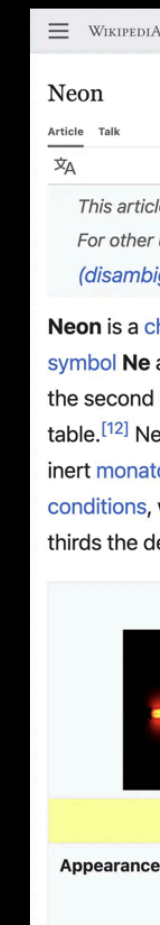
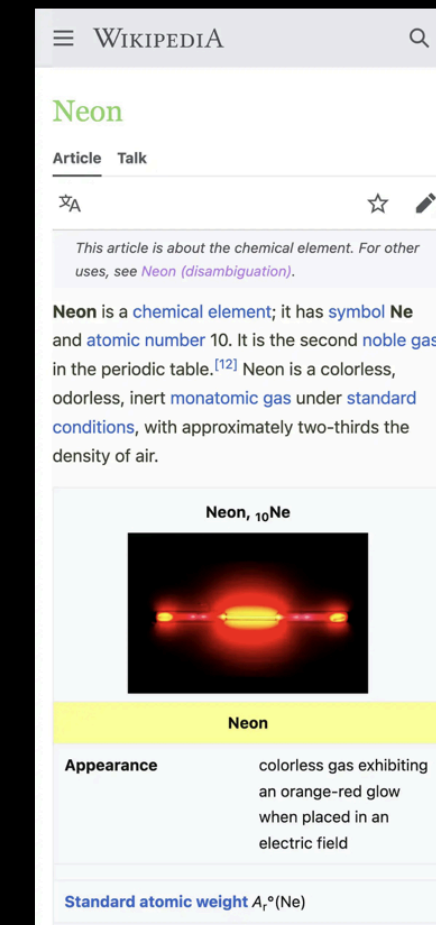
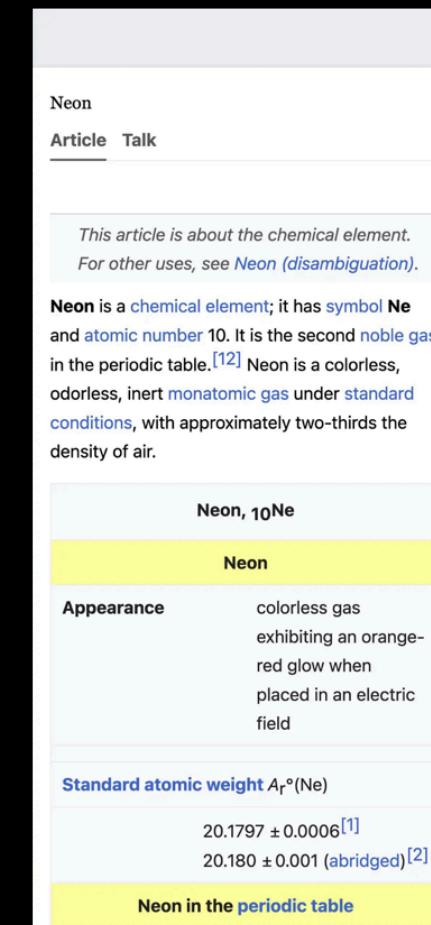
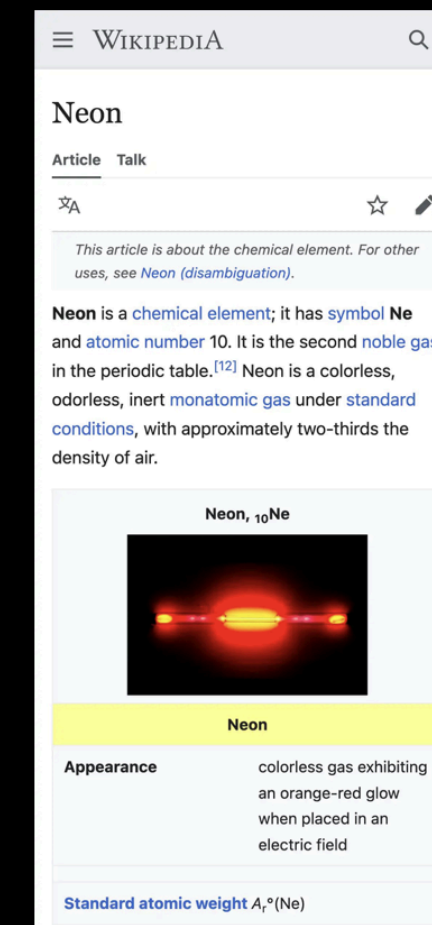
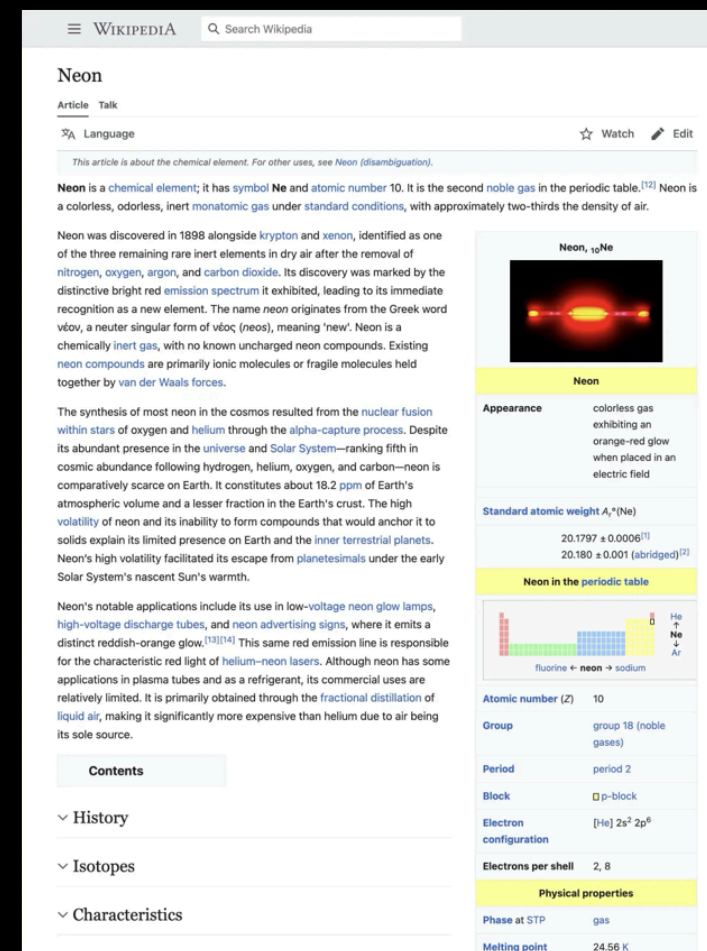
Due to the diversity of interfaces and changing accessibility needs, guidelines for UI design often cannot be generalized to the vast majority of user experiences. AI could provide a nuanced perspective, taking into account context that fixed guidelines can't.



# UIClip

<https://uimodeling.github.io/uiclip/>

- Trains a CLIP model to grade UI. Training data is comprised of screenshots with HTML modifications.
- Associates images of bad UI design with reasons why it's bad in plain text.

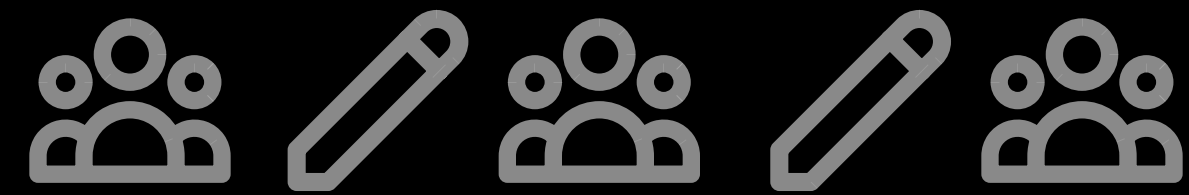




Recap!

# What did we learn?

1. The definition and importance of multimodal AI models.
2. How do AI models understand raw data given to it?
3. How neural networks help us train multimodal AI models.
4. How CLIP models work and their awesome applications!



# Thank you!

Questions? ([ryanchou111@gmail.com](mailto:ryanchou111@gmail.com))