Human-AI Interaction, Lecture 4

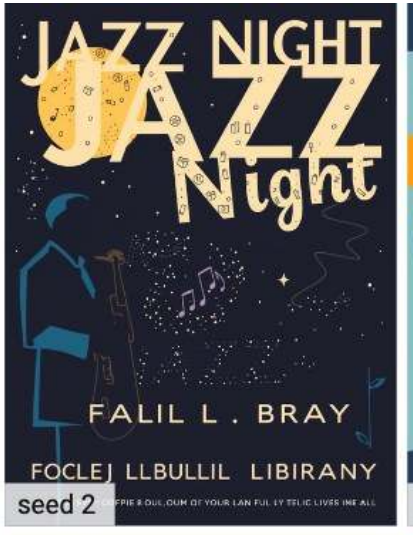
Generative Models

“a poster for a jazz night at a small public library”

sent to an image model, eight times

Eight runs of one prompt

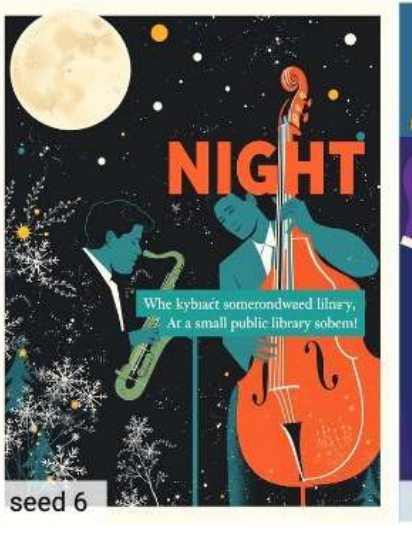


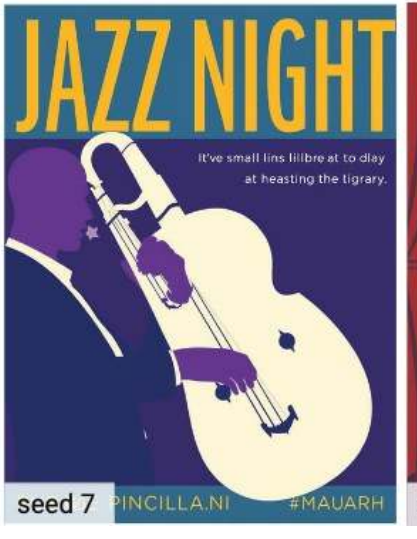


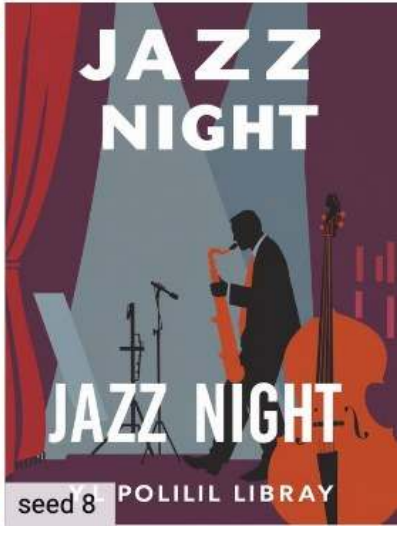












FLUX.1 [schnell]

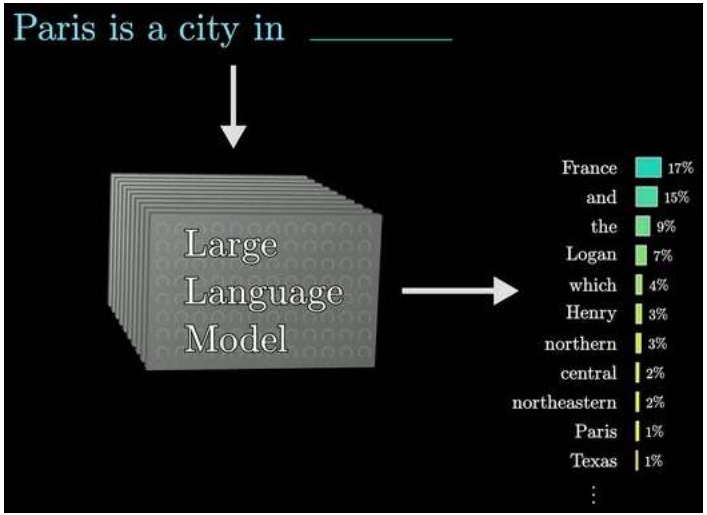
Each of the eight runs used a different **random seed**.

How does a model make text and images, and what does that mean for design?

Today

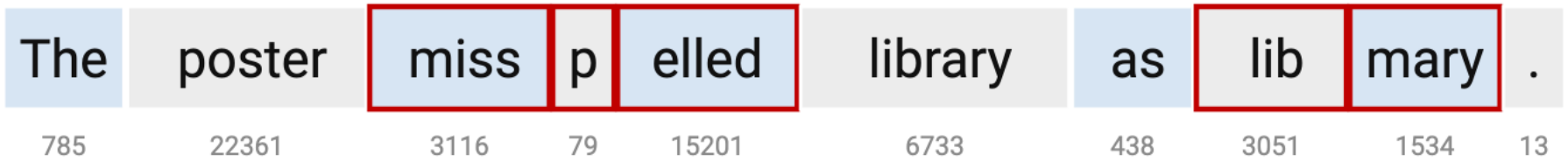
1. How language models write
2. How chat models are trained
3. How image models make images
4. Activity: seeds and temperature
5. Discussion
6. What this means for design

A large language model is “a sophisticated mathematical function that predicts **what word comes next** for any piece of text”



3Blue1Brown, “Large Language Models explained briefly,” 2024, at 0:49

Tokens

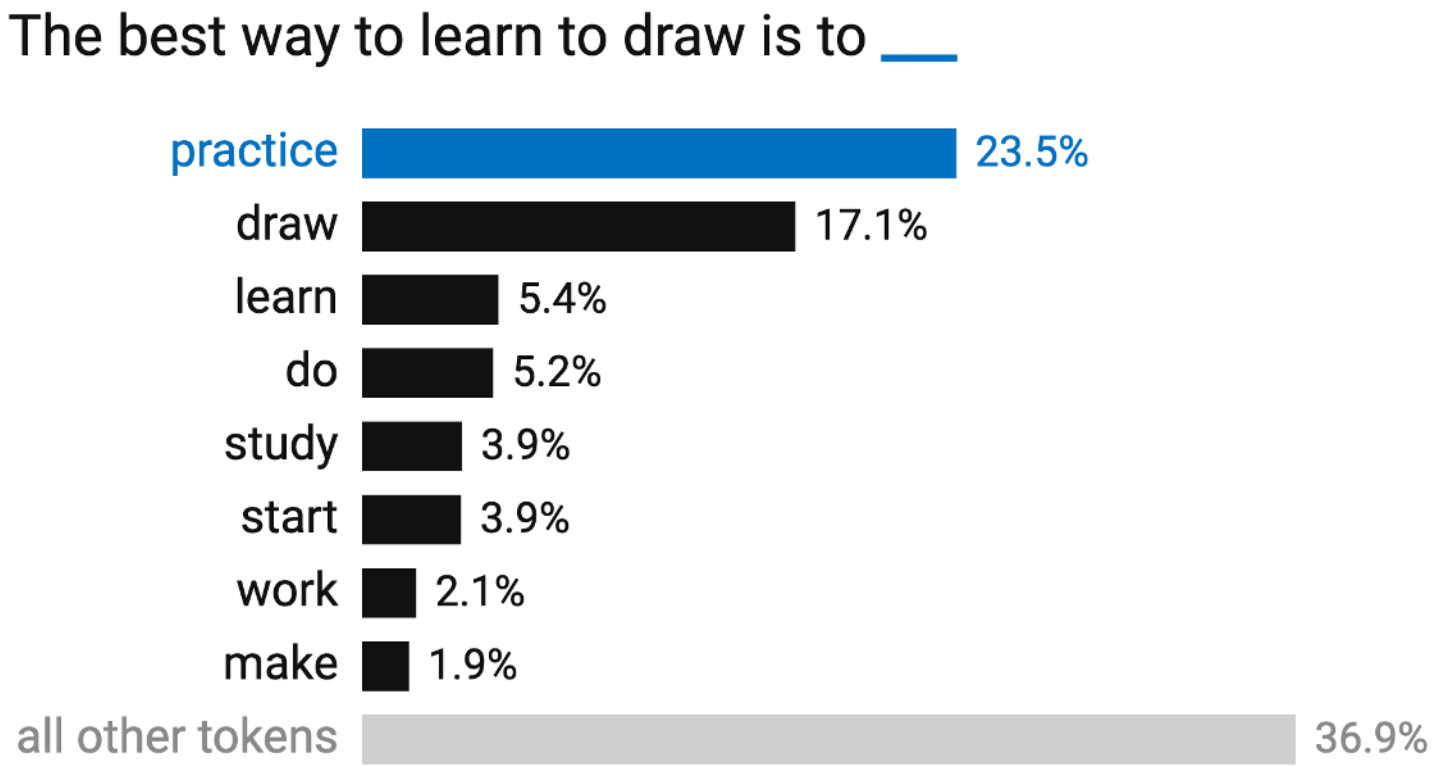


numbers under each token: the token's ID, which is all the model receives

Rare words are split into several tokens.

Qwen3 tokenizer

Next-token probabilities



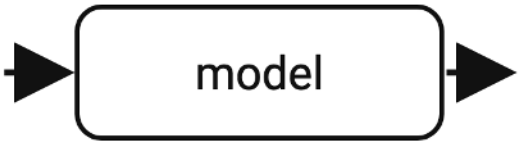
Qwen3-0.6B-Base

Generating text one token at a time

The best way to learn to draw is to



...to learn to draw is to **practice**



...to draw is to practice**.**



each round adds the most likely token (shown in blue), then runs the model again on the longer text

Qwen3-0.6B-Base

The chat template

```
<|im_start|>user
Suggest a name for a jazz night at a small public library.<|im_end|>
<|im_start|>assistant
<think>

</think>

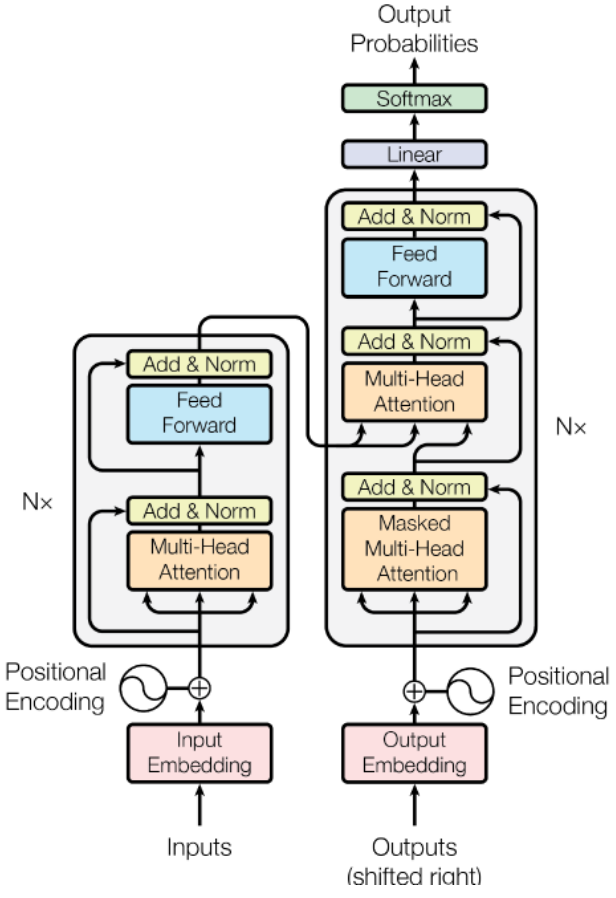
A great name for a jazz night at a small public library could be
**"Jazz at the Grove"**. It evokes a cozy, intimate atmosphere and
ties the event to a library setting. Let me know if you'd like a more
specific or creative name!
```

What Qwen3-0.6B receives (grey and black) and what it wrote back (blue)

Transformer

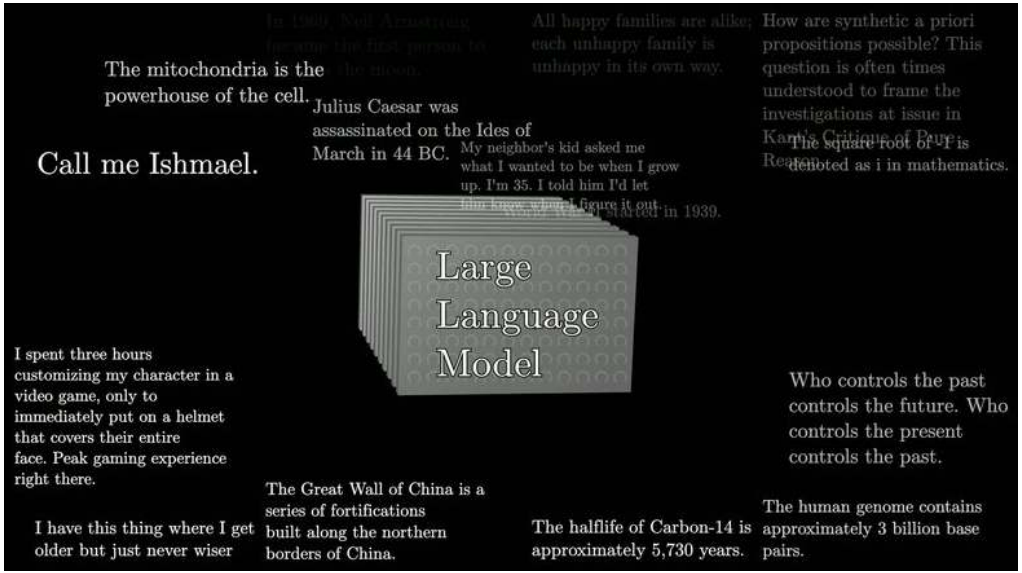
Attention

Feed-forward



Vaswani et al., NeurIPS 2017, Fig. 1

Pretraining: predict the next token

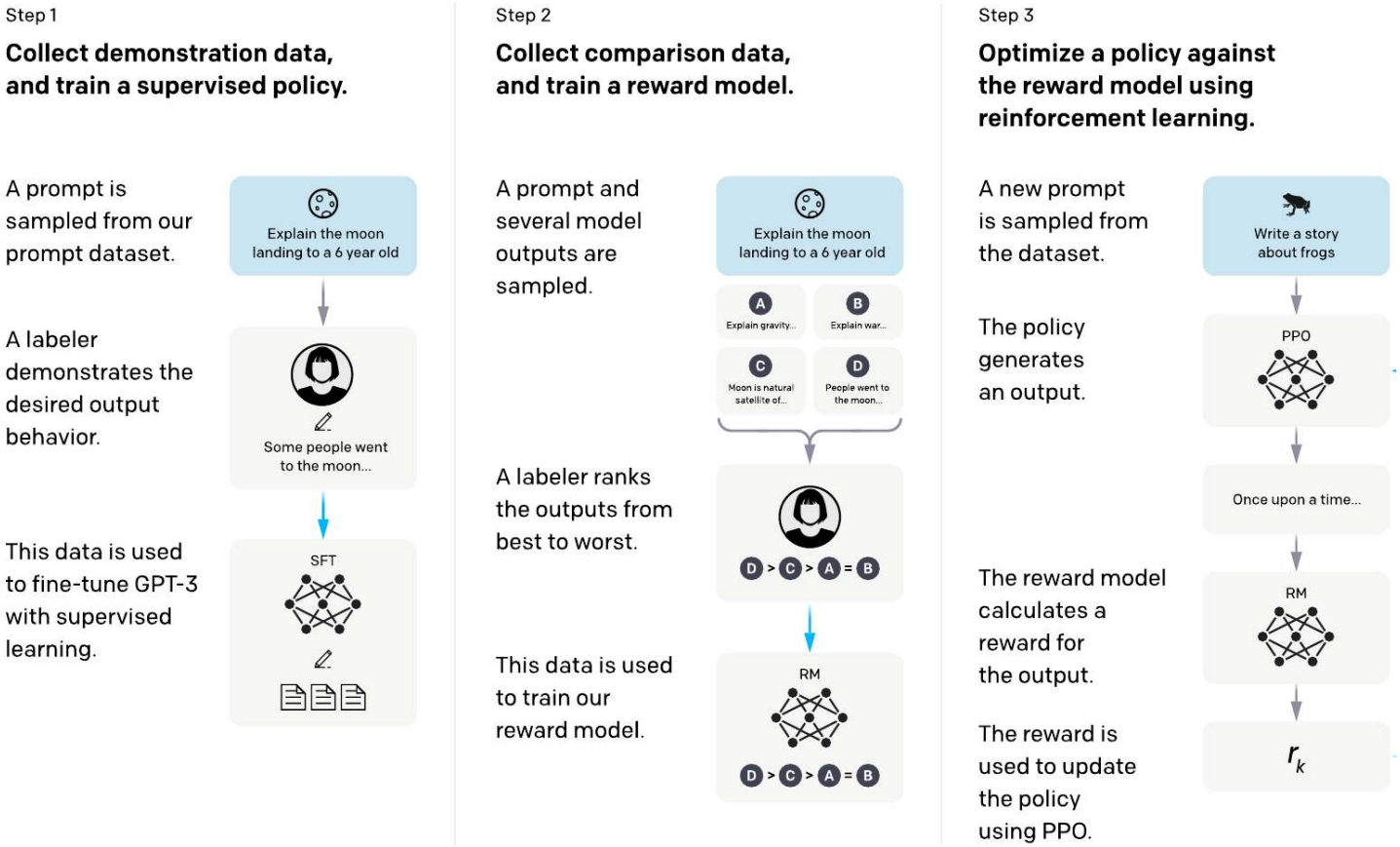


3Blue1Brown 2024, at 1:29

No labels: the next token is the answer.

GPT-3's training text: **over 2,600 years** of nonstop reading

Instruction tuning and human feedback



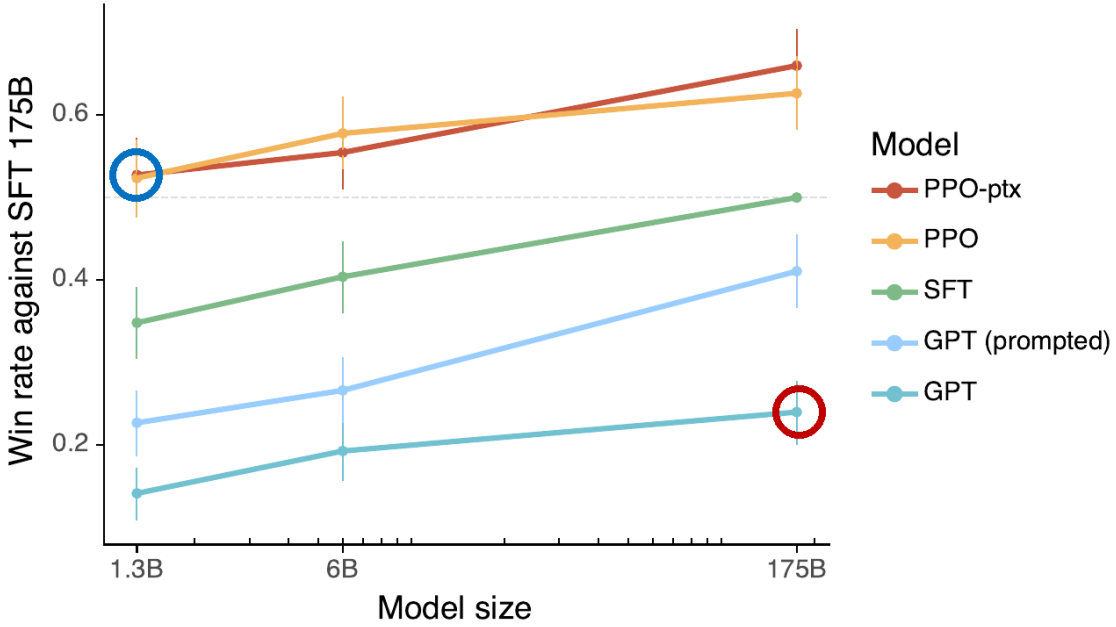
1. Copy answers people wrote

2. A reward model learns people's rankings

3. Train toward the reward model

Ouyang et al., NeurIPS 2022, Fig. 2

InstructGPT results



Ouyang et al., NeurIPS 2022, Fig. 1

People preferred InstructGPT at 1.3 billion parameters (blue) to GPT-3 at 175 billion (red).

Reasoning models

How many times does the letter r appear in the word "library"?

thinking

Okay, so I need to figure out how many times the letter 'r' appears in the word "library". Let me start by writing down the word to visualize it better:

L, I, B, R, U, R, Y. [...]

4. Fourth: R – yes, that's one 'r'.
5. Fifth: U – no.
6. Sixth: R – another 'r' here. [...]

So that's two 'r's in total. [...] Yep, that seems right.

The word "library" consists of the letters **L, I, B, R, U, R, Y**. The letter 'r' appears at the fourth and sixth positions. Therefore, the letter 'r' appears 2 times in the word.

Its steps spell "library" **with a U**, and it still counts two r's.

Qwen3-0.6B; OpenAI, "Learning to reason with LLMs," 2024

3. How image models make images

Today's reading: Lee et al., “Diffusion Explainer,” IEEE VIS 2024, and its live tool



Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion

Seongmin Lee* Benjamin Hoover *,† Hendrik Strobelt† Zijie J. Wang* ShengYun Peng*
Austin Wright* Kevin Li* Haekyu Park* Haoyang Yang* Duen Horng (Polo) Chau*

DIFFUSION EXPLAINER
Learn how Stable Diffusion transforms your text prompt into image!

Figure 1: With Diffusion Explainer, users can visually examine how (A) *text prompt*, e.g., *"a cute and adorable bunny... pixar character"*, is encoded by (B) the *Text Representation Generator* into vectors to guide (C) the *Image Representation Refiner* to iteratively refine the vector representation of the image being generated. (D) The *Timestep Controller* enables users to review the incremental improvements in image quality and adherence to the prompt over timesteps. (E) The final image representation is upscaled to a high-resolution image. Diffusion Explainer tightly integrates a visual overview of Stable Diffusion's complex components with detailed explanations of their underlying operations, enabling users to fluidly transition between multiple levels of abstraction through animations and interactive elements (Fig. 2, Fig. 3).

arXiv:2305.03509v3 [cs.CL] 31 Aug 2024

Abstract

Diffusion-based generative models' impressive ability to create convincing images has garnered global attention. However, their complex structures and operations often pose challenges for non-experts to grasp. We present Diffusion Explainer, the first interactive visualization tool that explains how Stable Diffusion transforms text prompts into images. Diffusion Explainer tightly integrates a visual overview of Stable Diffusion's complex structure with explanations of the underlying operations. By comparing image generation of prompt variants, users can discover the impact of keyword changes on image generation. A 56-participant user study demonstrates that Diffusion Explainer offers substantial learning benefits to non-experts. Our tool has been used by over 10,300 users from 124 countries at https://poloclub.github.io/diffusion-explainer/.

1 Introduction

Diffusion-based generative models [36, 43, 51] like Stable Diffusion [43] and DALL-E [31] have captured global attention for their impressive image creation abilities, from AI developers, designers, to policymakers. However, the popularity and progress of generative AI models have sparked social concerns [44, 9, 11, 12], such as accusations of artistic style theft by developers of AI image generators [11, 12]. Policymakers are also discussing ways to combat malicious data generation and revise copyright policies [14, 15, 38, 1].

There is an urgent need for individuals from many different fields to understand how generative AI models function and communicate effectively with AI researchers and developers [12, 17].

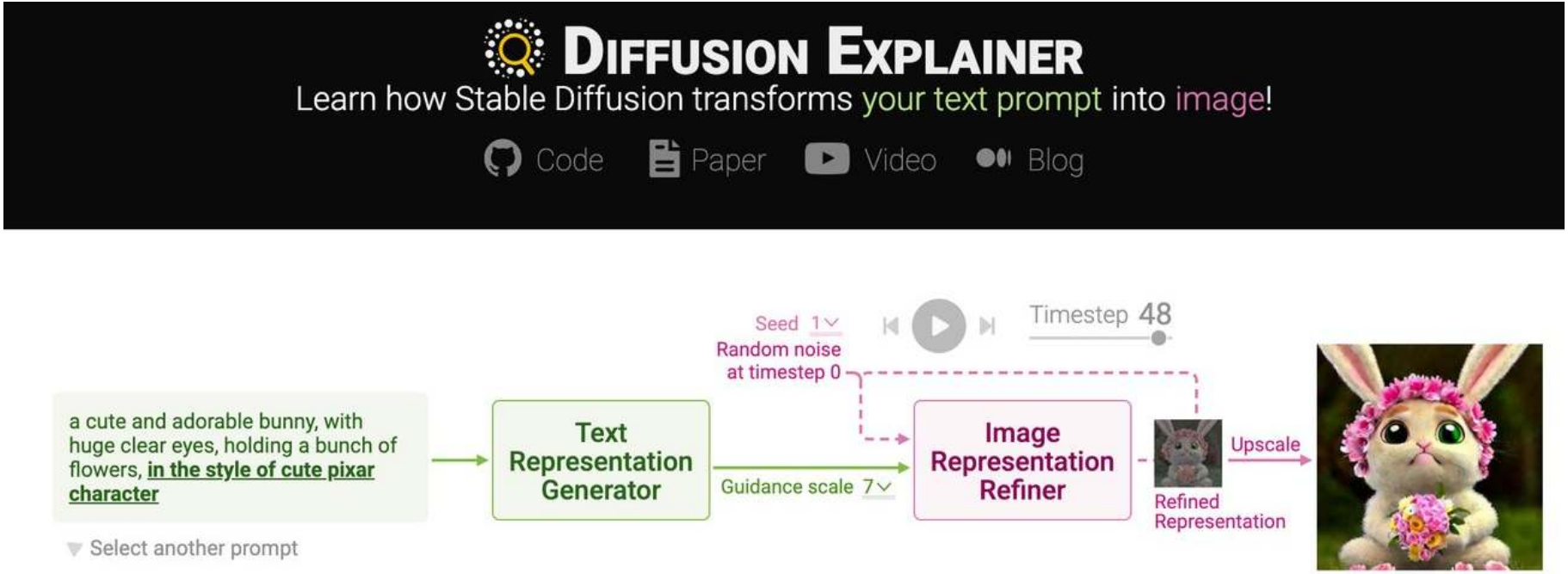
Key challenges in designing learning tools for Stable Diffusion. At the high level, Stable Diffusion iteratively refines noise into a vector representation of a high-resolution image, guided by a text prompt. Internally, the prompt is tokenized and encoded into vector representations by *CLIP Text Encoder* [35]. With the text representations' guidance, Stable Diffusion improves the image quality and adherence to the prompt by progressively denoising the image's vector representation using the *UNet* neural network [37]. The final image representation is upscaled to a high-resolution image [24]. The crux of learning about Stable Diffusion stems from the complex interplay between multiple subcomponents, their intricate operations, and the iterative nature of refinements, which are challenging even for experts to grasp [47]. While some articles [7] and videos [19, 4] explain Stable Diffusion, they often assume notable knowledge of machine learning and focus on mathematical details.

In this work, we contribute:

- **Diffusion Explainer, the first interactive visualization tool designed for non-experts** to explain how Stable Diffusion transforms a text prompt into a high-resolution image, overcoming key design challenges in developing interactive learning tools for Stable Diffusion (Fig. 1). It tightly integrates a visual overview of Stable Diffusion's complex structure with detailed explanations of their underlying operations via animations and interactive elements (Fig. 2, Fig. 3, 4-A,3). It also provides a new way to visualize the impact of keyword changes on the complex image generation process by comparing image generation of prompt variants (Fig. 4, 4-A,3).
- **Reflection and design lessons derived from human evaluation with 56 non-experts** showcase Diffusion Explainer's substantial

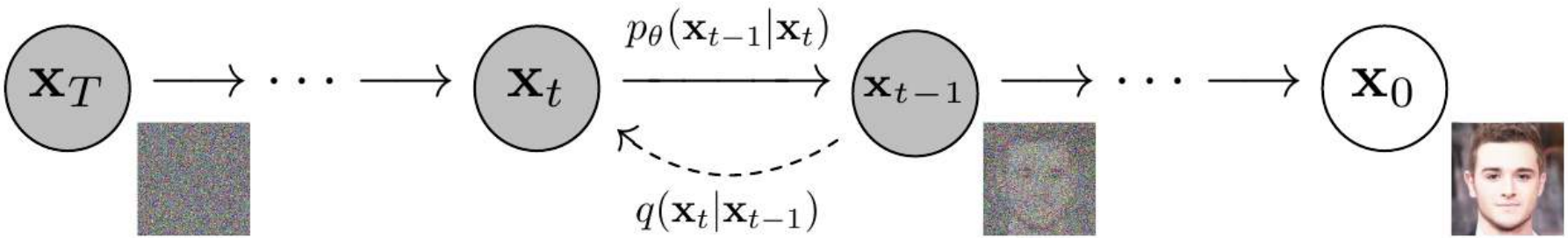
*Georgia Tech. {seongmin@gatech.edu, {jayw,bpeng69,jjgao,sigfried,kevin.li,haekyu}@gatech.edu, {haoyang,polo}@gatech.edu
†IBM Research. hendrik.strobelt@ibm.com

Diffusion Explainer



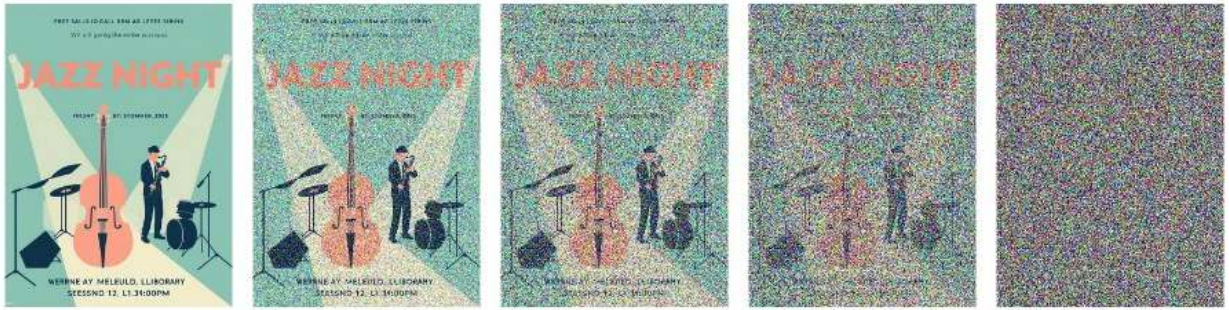
Diffusion Explainer (Lee et al., IEEE VIS 2024), poloclub.github.io/diffusion-explainer

Training a diffusion model



Ho, Jain & Abbeel, NeurIPS 2020, Fig. 2

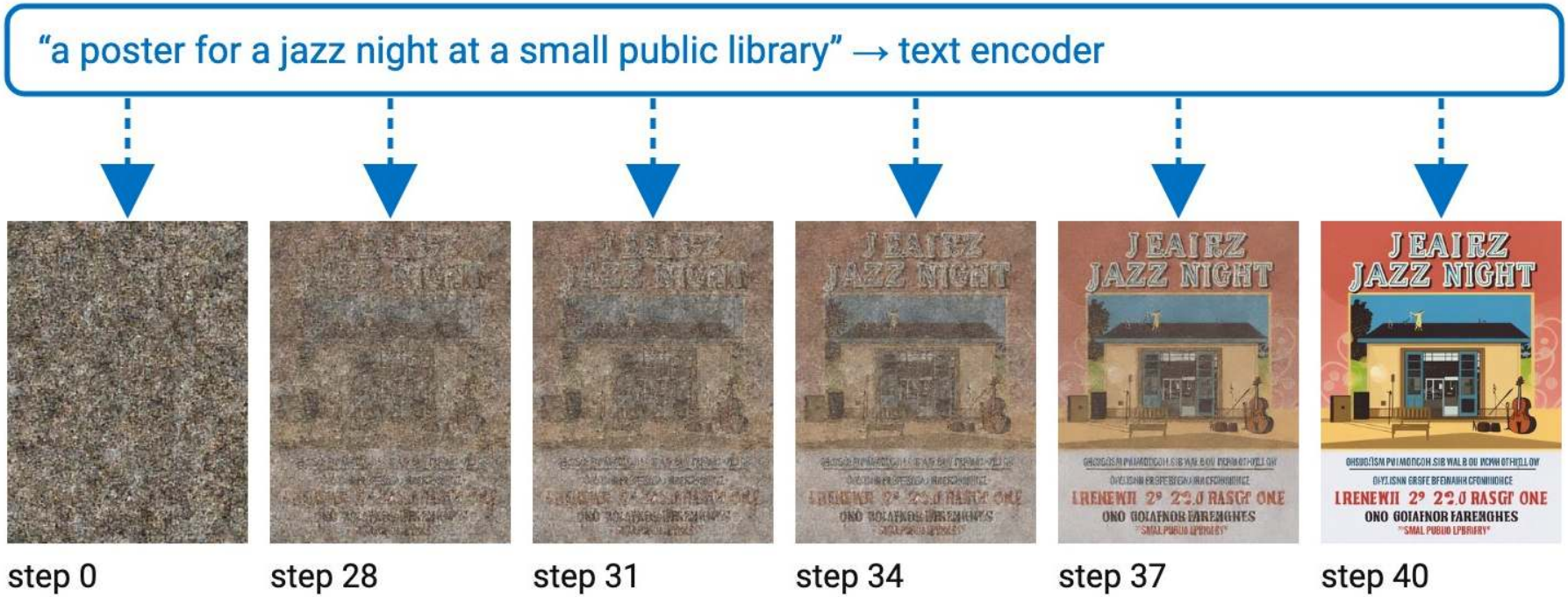
add a little more noise at each step



the network learns to guess
the noise that was added

Our seed 4 poster, with more and more noise

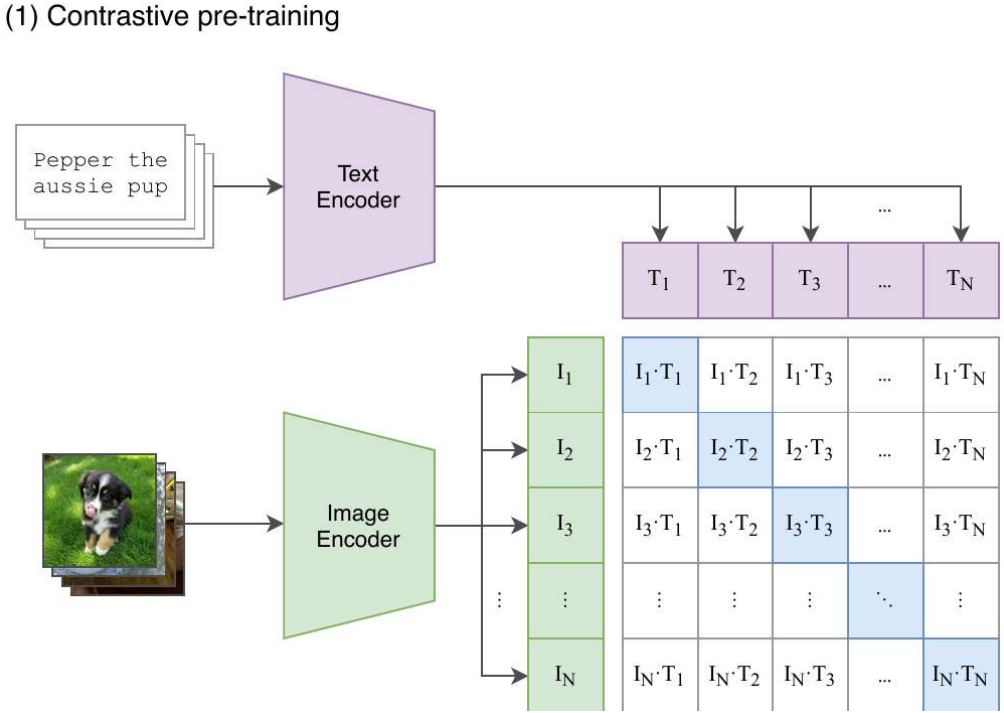
Generating an image



A **text encoder** turns the prompt into numbers.

FLUX.2 [klein] 4B Base; Rombach et al., CVPR 2022 (cross-attention)

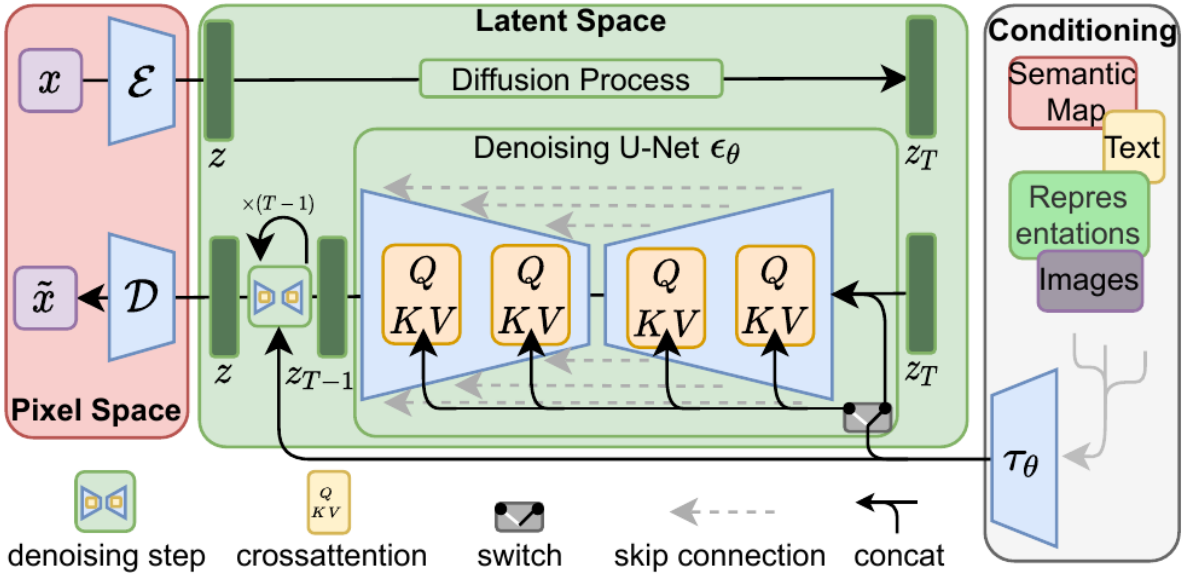
CLIP



Trained on **400 million** (image, text) pairs from the internet.

Radford et al., ICML 2021, Fig. 1

Latent diffusion



Stable Diffusion v1:
512 × 512 × 3 = 786,432
64 × 64 × 4 = 16,384

Rombach et al., CVPR 2022, Fig. 3

Running the same seed again



seed 4

=



seed 4, run again later

FLUX.1 [schnell]

Newer image models



2. Trained using latent adversarial diffusion distillation, `FLUX.1 [schnell]` can generate high-quality images in only 1 to 4 steps.

Black Forest Labs, FLUX.1 [schnell] model card, Hugging Face

Unlike DALL·E, which operates as a diffusion model, 4o image generation is an autoregressive model natively embedded within ChatGPT. This fundamental difference introduces several new

OpenAI, Addendum to GPT-4o System Card: Native image generation, 2025, p. 1

Activity: seeds and temperature

In pairs, 15 minutes

1. Diffusion Explainer: change only the seed
2. Then change only the guidance scale
3. A chat tool: run one prompt three times
4. If you can, try temperature 0 and 1
5. Note what changed and what stayed the same

poloclub.github.io/diffusion-explainer

What changed in our eight posters

Changed

colors

layout

lettering

instruments

Stayed the same

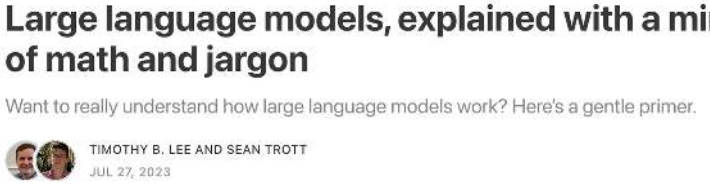
flat illustration

big “Night” lettering

no library anywhere



Discussion: the three readings



Large language models, explained with a minimum of math and jargon

Want to really understand how large language models work? Here's a gentle primer.

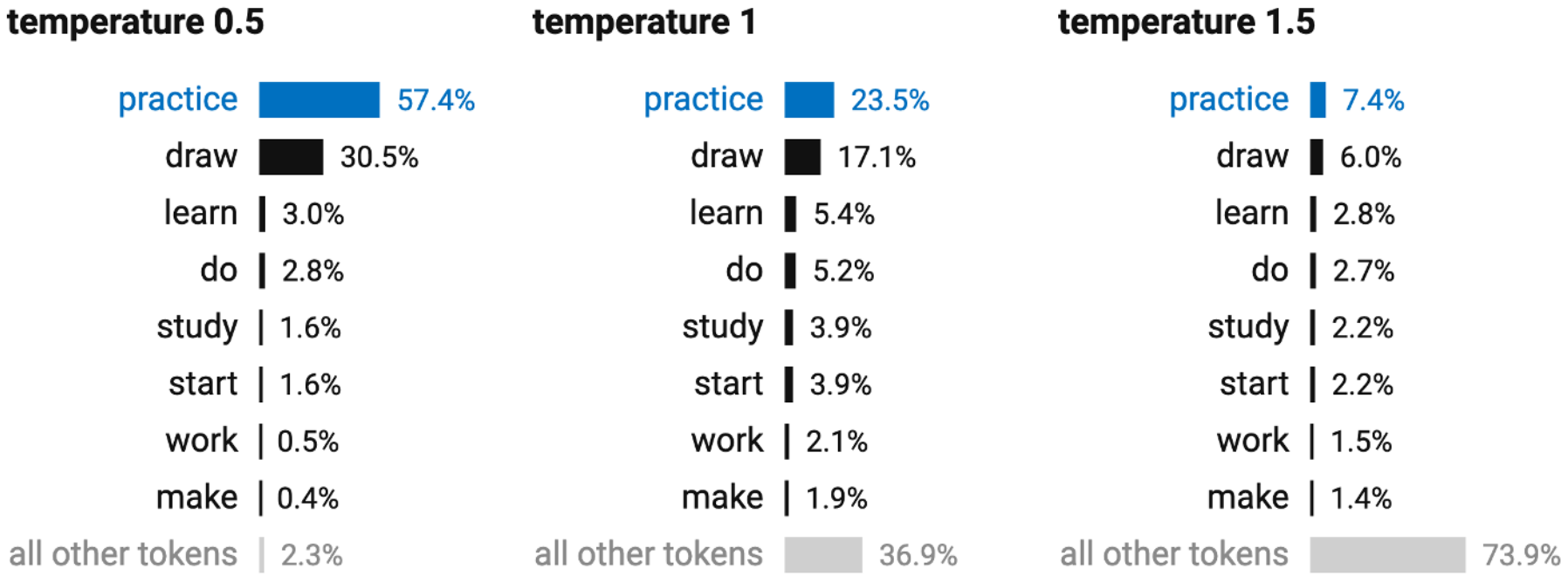
TIMOTHY B. LEE AND SEAN TROTT
JUL 27, 2023

1. After today, what could you explain to a friend about LLMs, and what not?

2. Which changed the image more, the seed or the guidance scale?

3. What in the readings is already out of date?

1. Sampling



Temperature reshapes the probabilities before the tool picks one.

Qwen3-0.6B-Base

Six runs of one prompt

The best way to learn to draw is to …

temperature 0, run 1

…practice. The more you practice, the better you will become. The

temperature 0, run 2

…practice. The more you practice, the better you will become. The

temperature 1, seed 1

…: ↵ A. make real drawings with your hands ↵ B. study drawings

temperature 1, seed 2

…practice a lot. And not just practice on paper, but practice forever

temperature 1, seed 3

…practice. Use your own resources, but don't worry about how you

temperature 1.5, seed 1

…: ↵ running how its limits ↵ discussion allowed the fi mple fakes

Qwen3-0.6B-Base

Designing for sampling



Show several at once

Midjourney documentation, Discord Quick Start



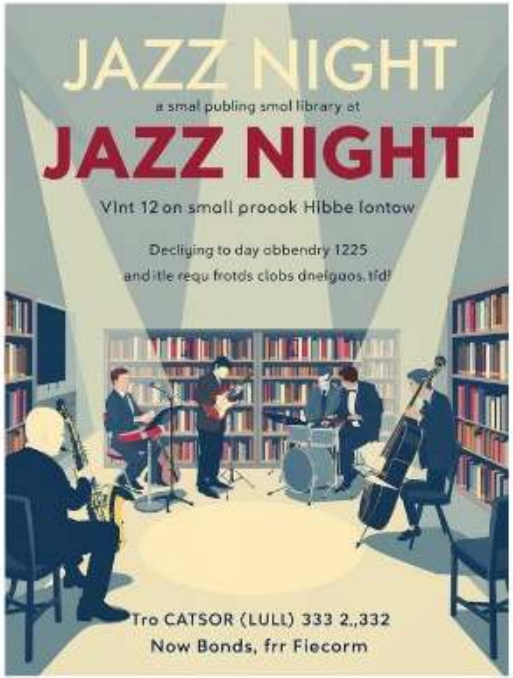
Let people set the seed

Diffusion Explainer

Adding words with the same seed



seed 4: "...small public library"



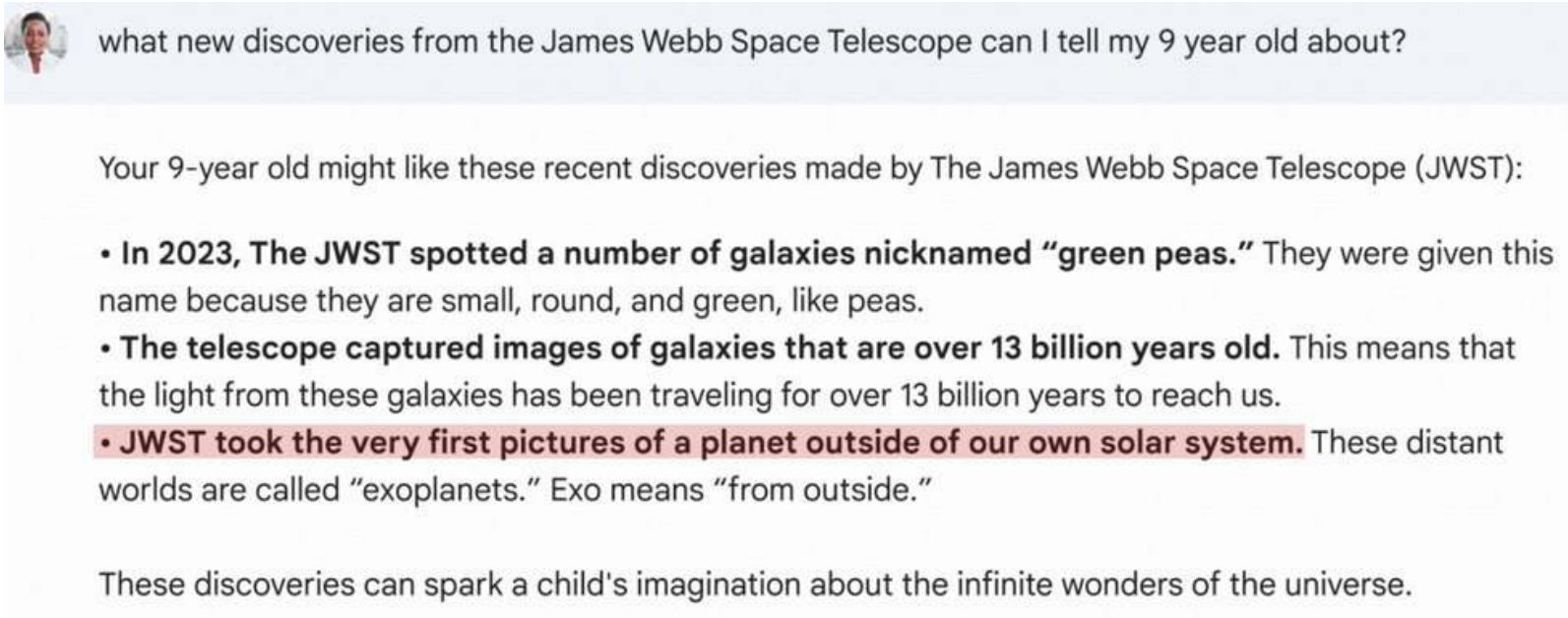
seed 4: "...small public library, **with bookshelves behind the band**"

"While seeds can help with consistency, they have the least impact on the final image."

Midjourney documentation, "Seeds"

FLUX.1 [schnell]

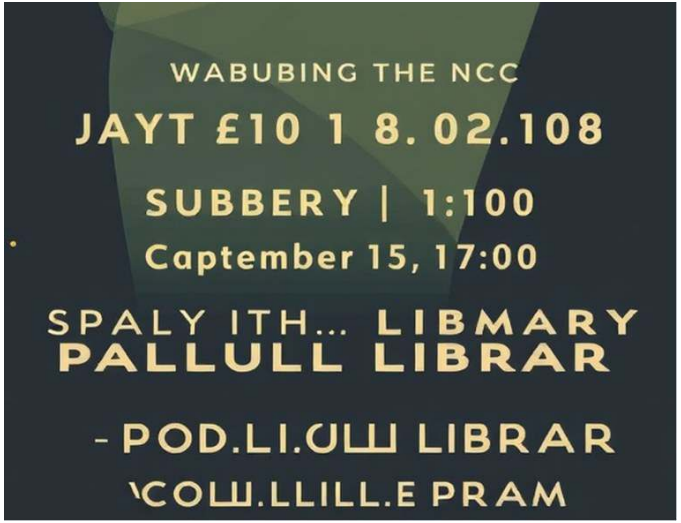
2. Fluent errors



The first picture of a planet outside our solar system was taken in **2004**.

Google's video introducing Bard, in S. Pichai, "An important next step on our AI journey," Google, 2023, at 0:11; Engadget and Reuters, 2023

A language model is “haphazardly stitching together sequences of linguistic forms … **without any reference to meaning**”



Bender, Gebru, McMillan-Major & Shmitchell, FAccT 2021, p. 616; poster: FLUX.1 [schnell]

3. Imprecise prompts

The prompt said nothing about color,
layout, lettering or instruments.

It did ask for a library, and no poster shows one.

Milestone 1: Watch people use AI

Due Thursday of week 3

Next: Lecture 5

Mental Models and Theory of Mind

Read before class:

- Nass, Steuer & Tauber, “Computers are social actors,” CHI '94
- Eslami et al., “I always assumed that I wasn't really that close to [her],” CHI '15