The Role of Generative AI in Academic Research

Julia Li

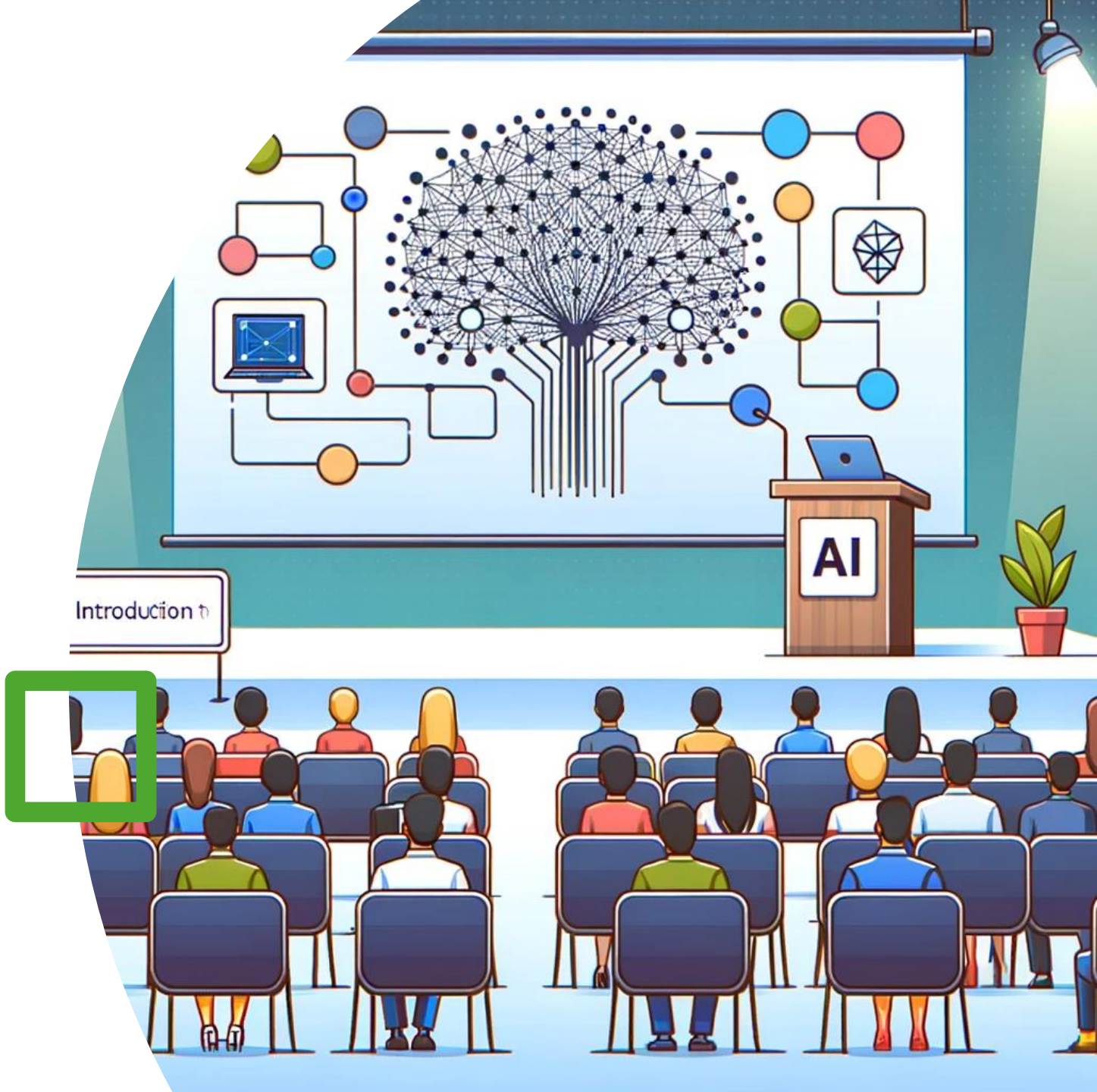
Principal Algorithm Engineer

Microsoft Software Technology Center Asia (STCA)

Agenda

- Introduction to Generative AI
- Generative AI in Research
- Collaborative between Academia and Industry
- Responsible AI for Research

Introduction to Generative AI



What is Generative AI?

Algorithms and models capable of generating new data that resembles a given dataset.

To create diverse, realistic, and high-quality data

Generated Data type

Text

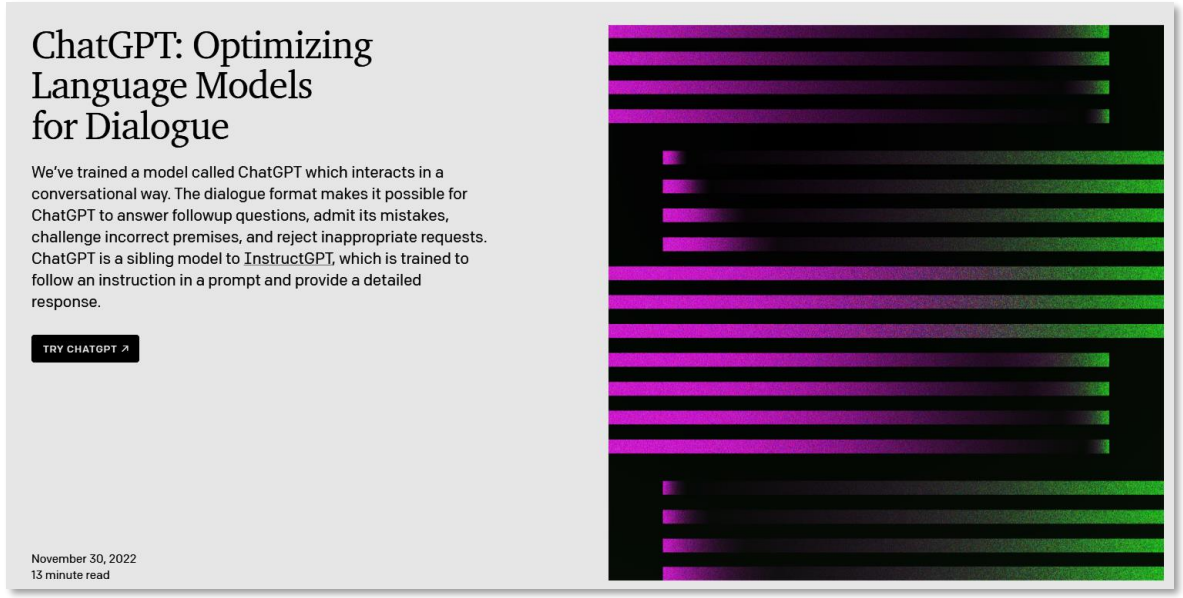
Image

Audio

Video

AI Text Generation： ChatGPT

- Based on GPT (Generative Pre-trained Transformer) 3.5/4 Model
  - OpenAI
  - Released at 2022.11.30



https://openai.com/blog/chatgpt

AI Image Generation

Stable-diffusion

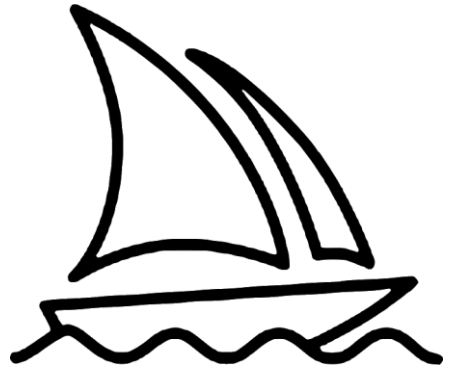
https://beta.dreamstudio.ai/dream

Dall-E

https://labs.openai.com/

Midjourney

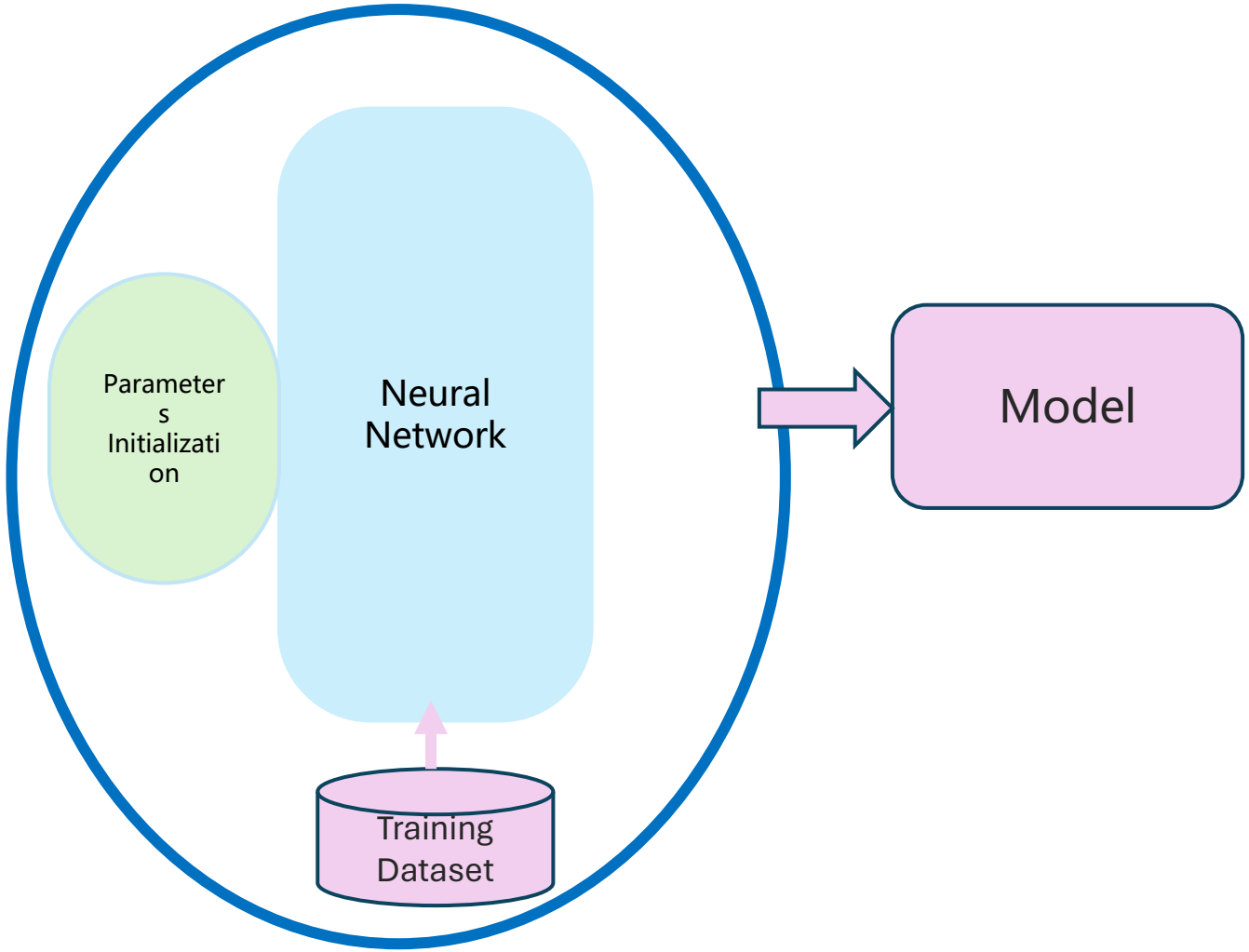
https://www.midjourney.com/



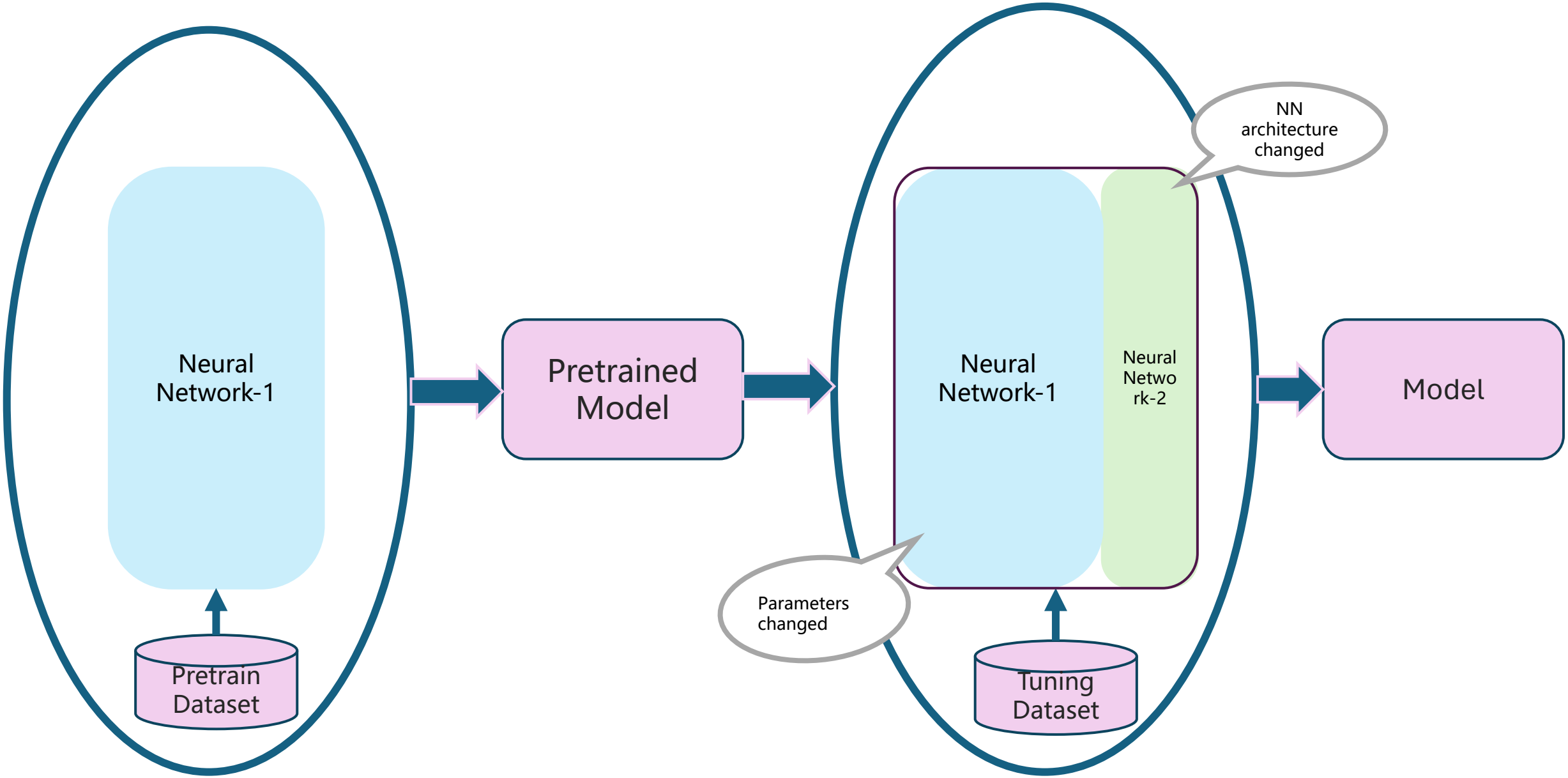
The Evolution of AI Paradigms

- The Core of AI Technology
  - Models: *mathematical frameworks or computational systems that are trained on data to perform tasks such as making predictions, identifying patterns, or making decisions*
  - Model = $f_{algorithm}(Data)$
- AI Paradigm
  - #1: Training from scratch
  - #2: Pretrained model + fine tuning
  - #3: Pretrained large model + prompt engineering

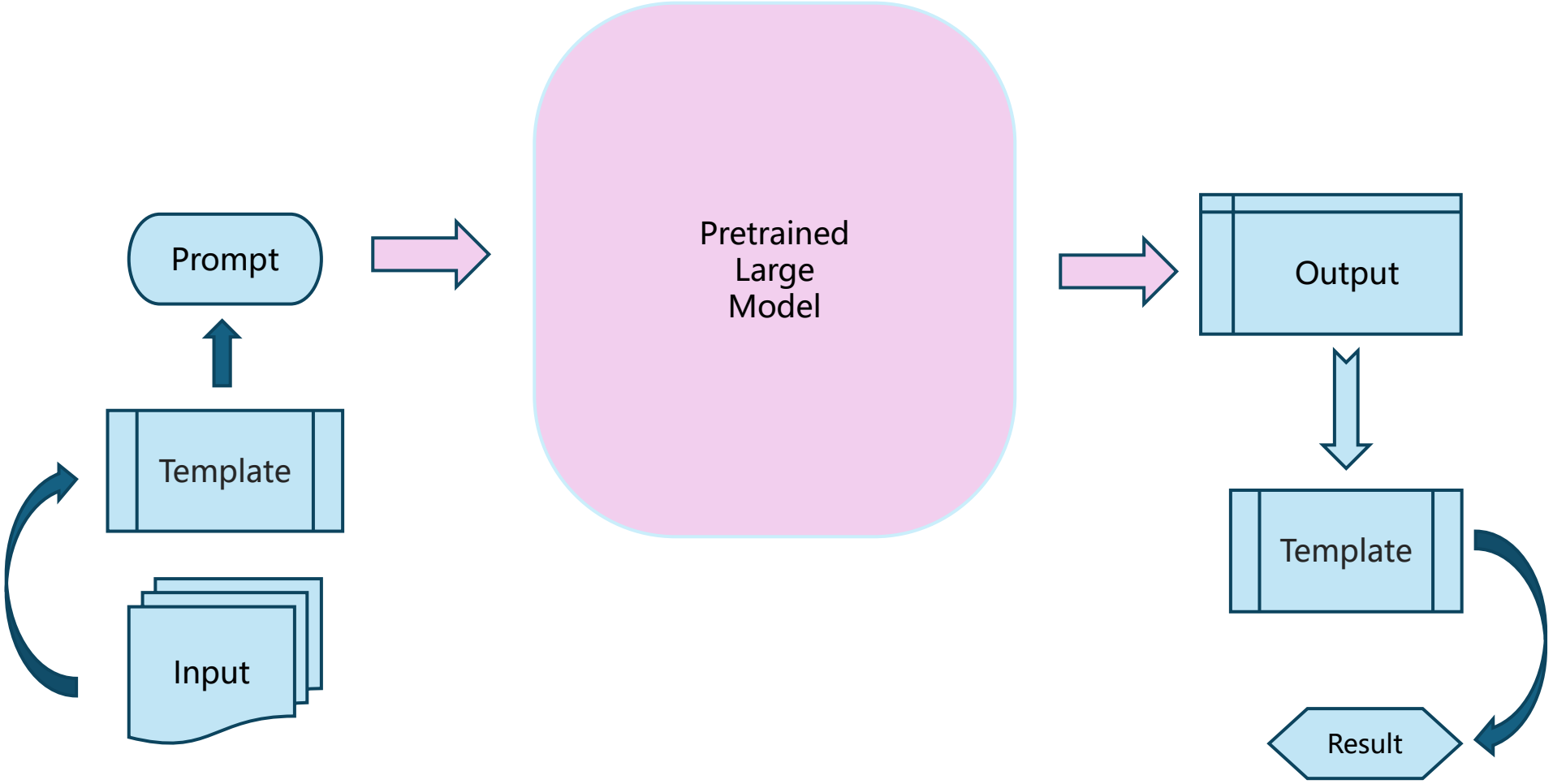
Training From Scratch



Pretrained Model + Fine tuning



Pretrained Large Model + Prompt Engineering



Examples of Generative Models

Generative Pre-trained Transformer (GPT)

Architecture: Transformer-based model for natural language processing.

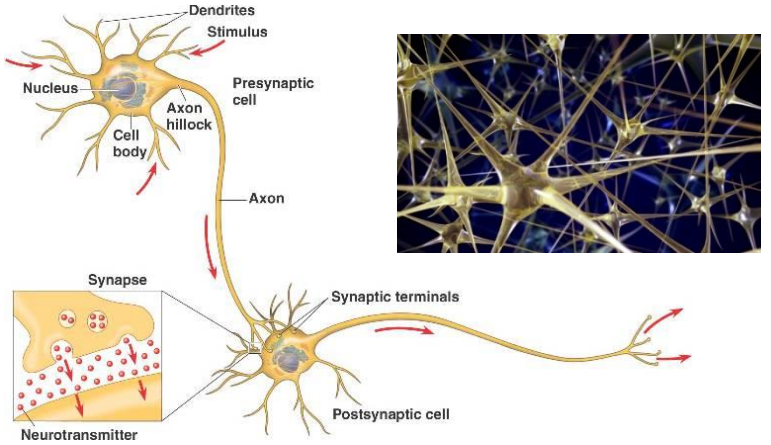
Real-world application examples

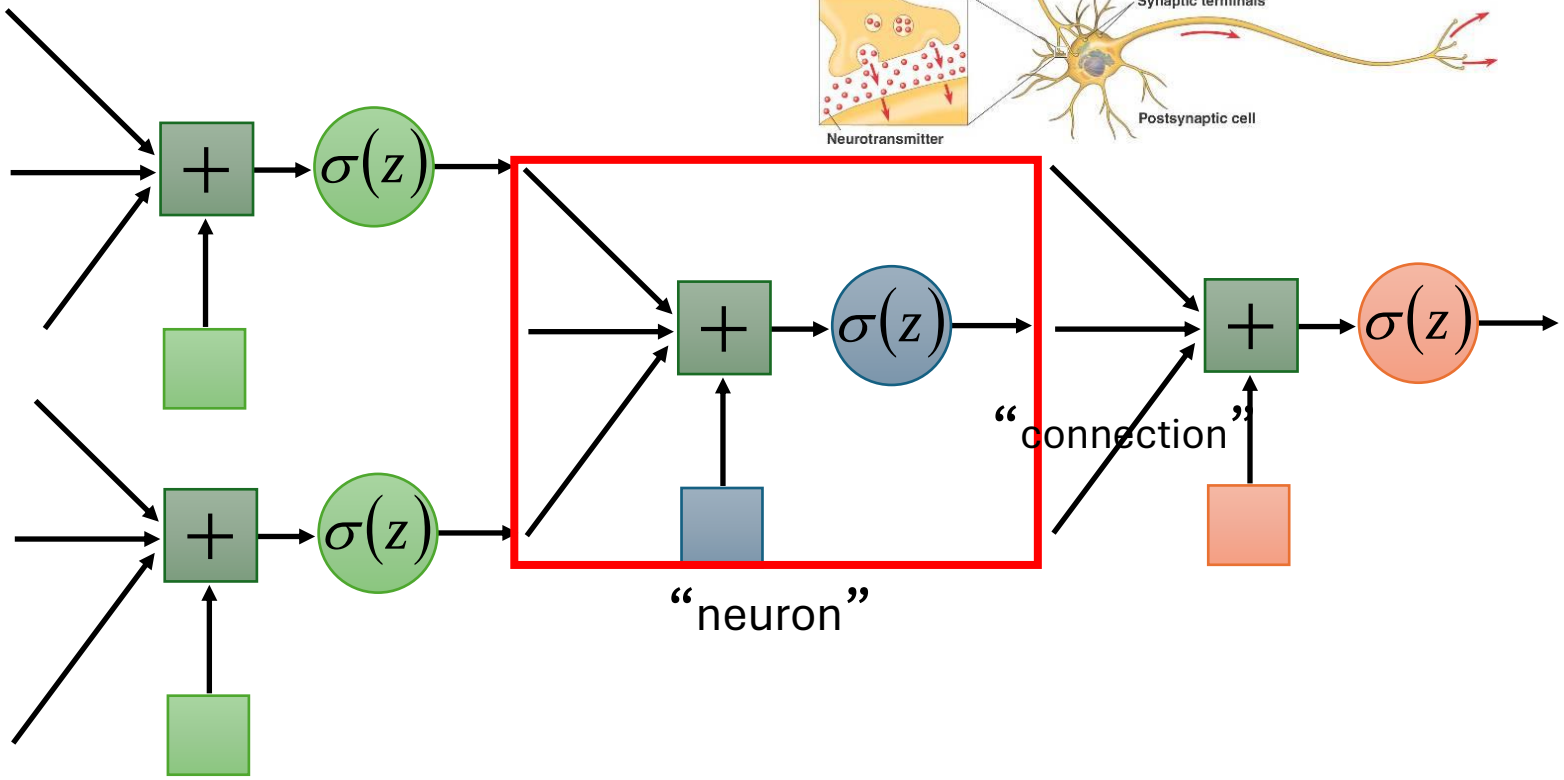
Diffusion Models

Probabilistic generative model for data generation.

Real-world application examples

Neural Network

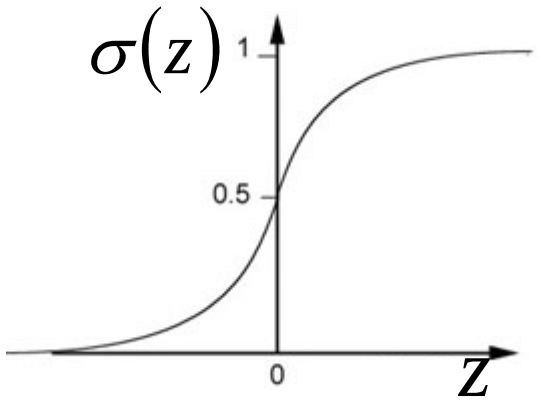




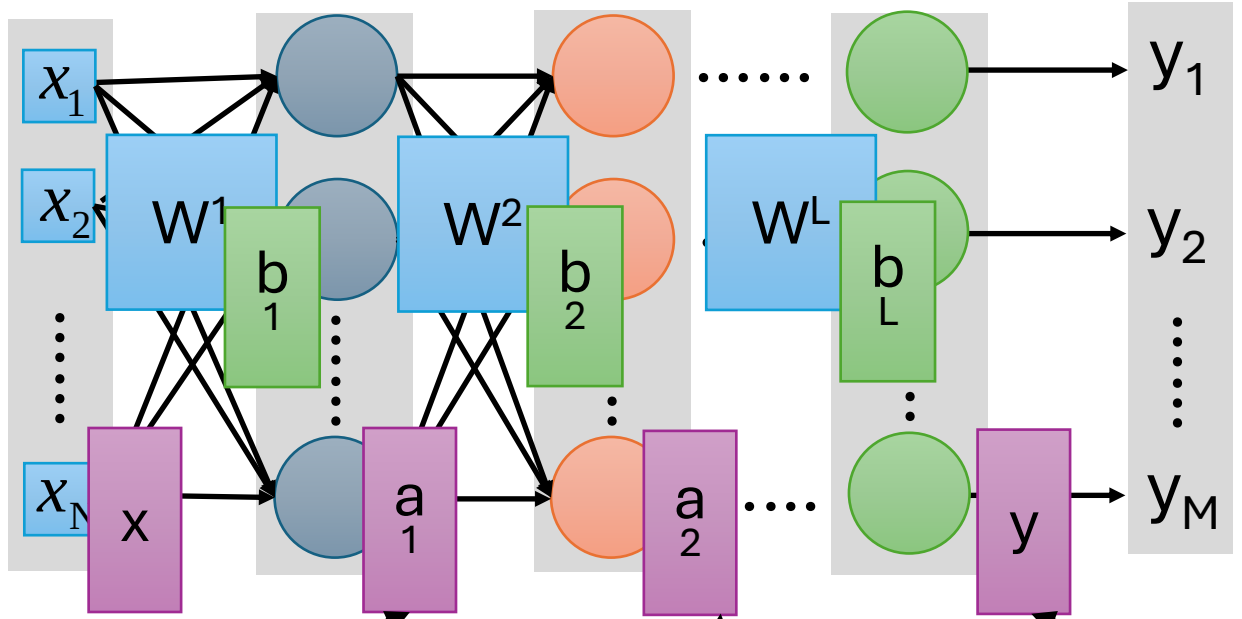
Fully connected neural network (FCNN)

Sigmoid Function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



- hidden layer
  - weighted sum of input data
  - activation function
- layers



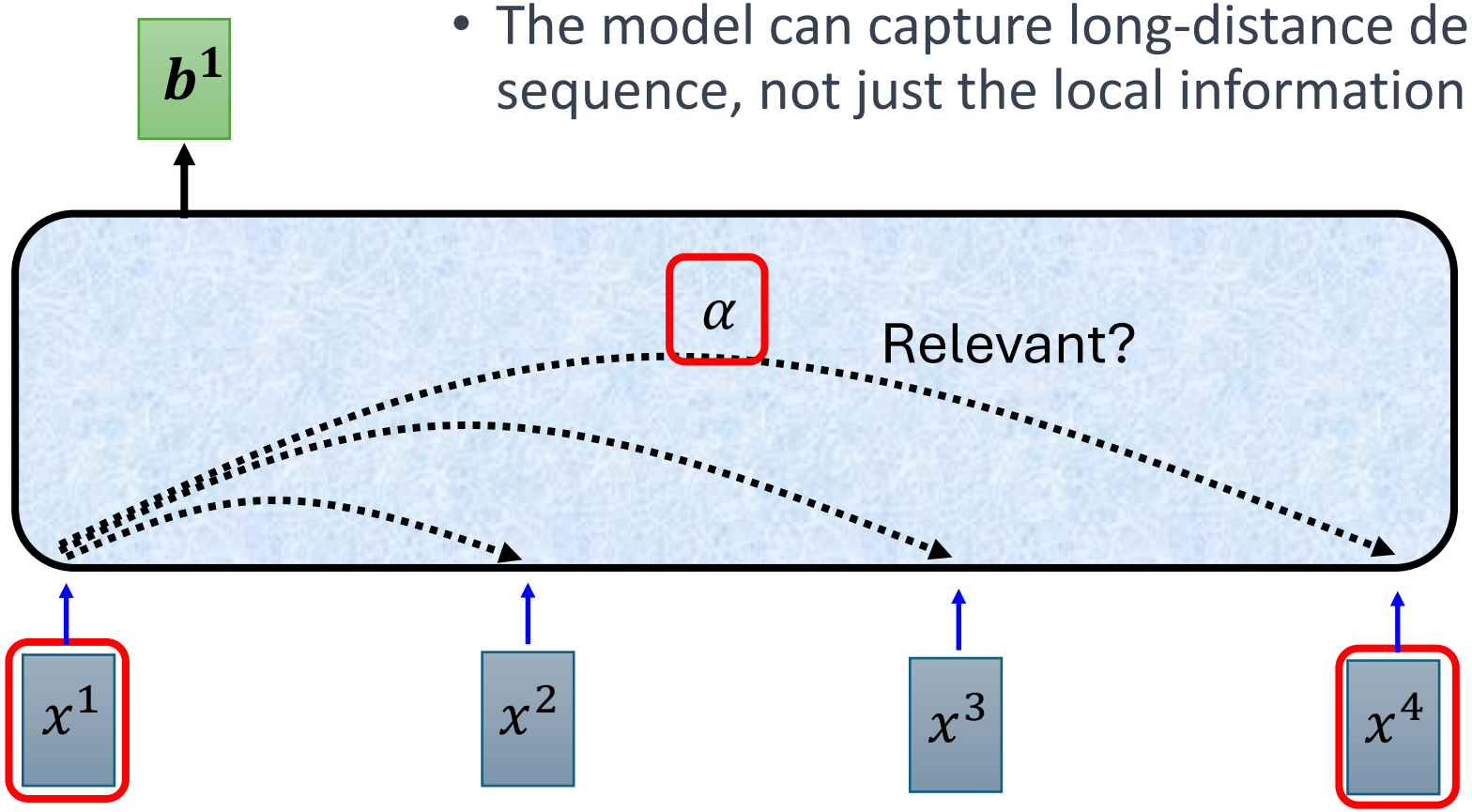
$$\sigma(W^1 x + b_1)$$

$$\sigma(W^2 a_1 + b_2)$$

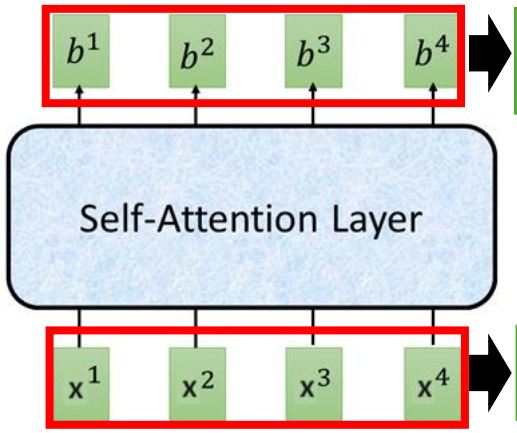
$$\sigma(W^L a^{L-1} + b_L)$$

Self-Attention

- Computing the similarity between each element and other elements in the input sequence as weights
- The model can capture long-distance dependencies in the sequence, not just the local information



Matrix Operations of Self-attention



$$O = Q = W^q I$$

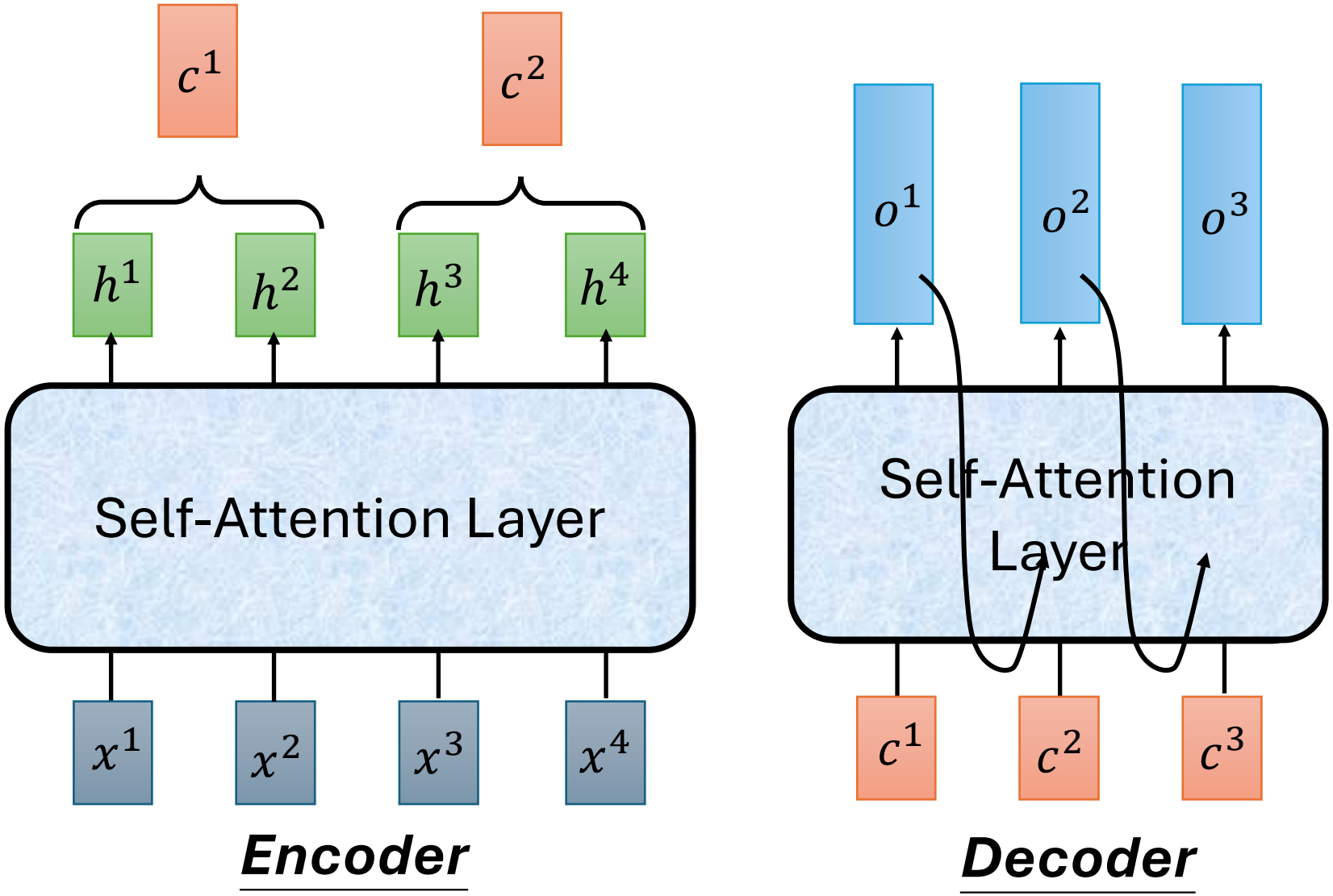
$$K = W^k I$$

$$V = W^v I$$

$$\hat{X} \leftarrow X = K^T Q$$

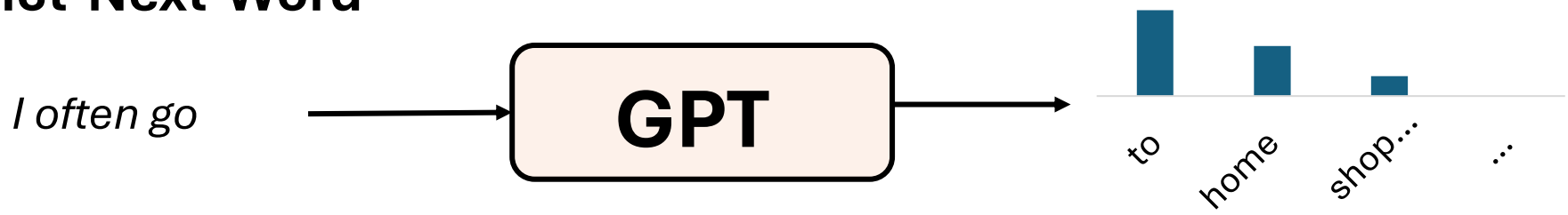
$$O = V \hat{X}$$

Transformer: Seq2seq model with Self-Attention



Text Generation

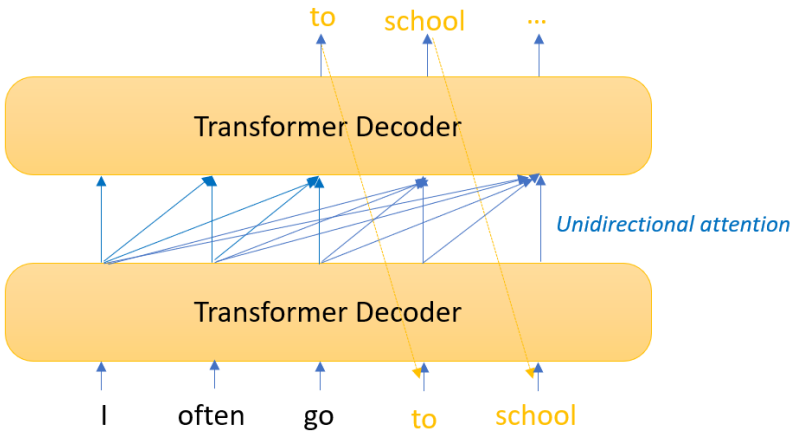
Predict-Next-Word

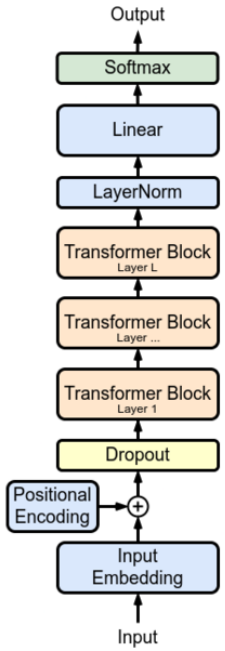


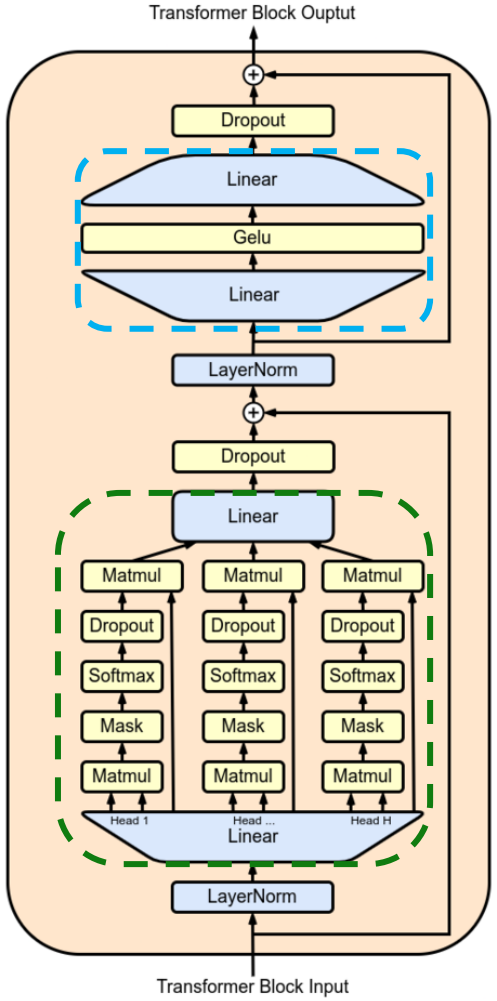
- Input： token sequence
- Output： the next token (word) to the sequence

GPT (Generative Pre-trained Transformer)

Attention-based deep transformers



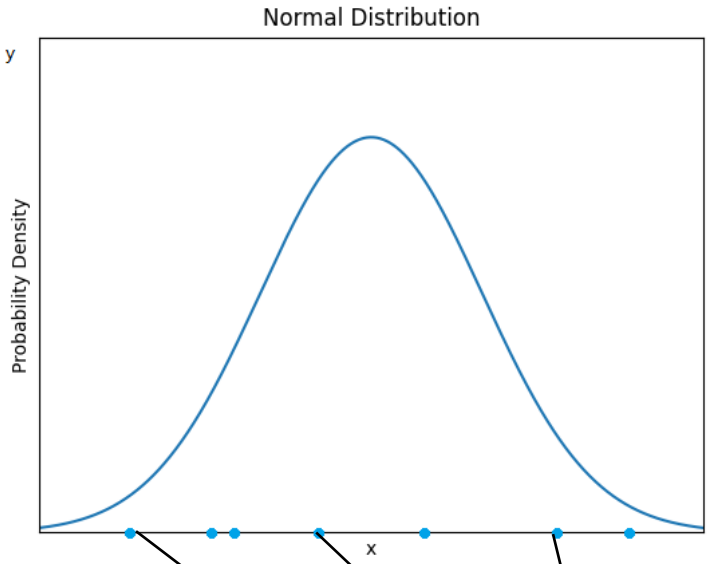


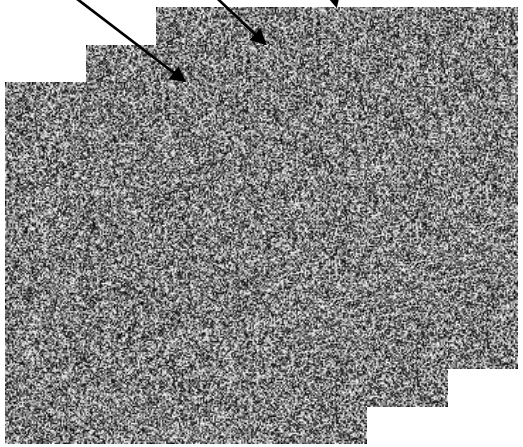


Feedforward Network

Attention

Image Generation





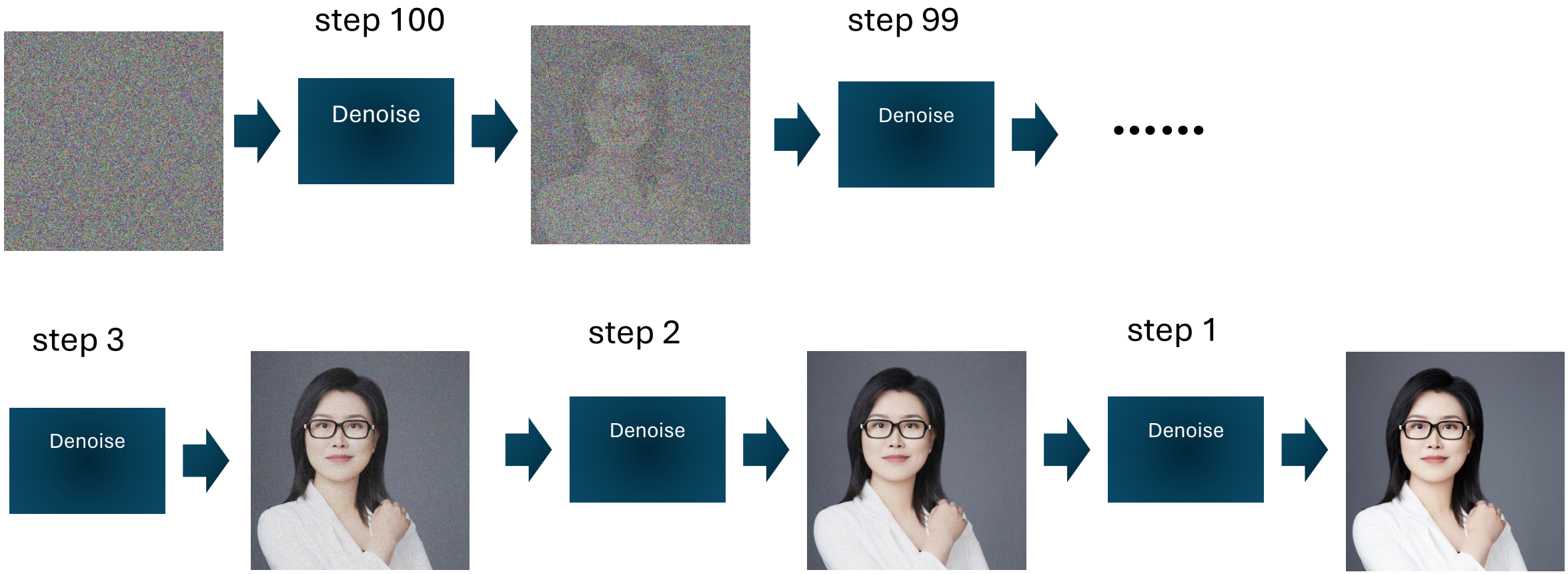
Text

condition

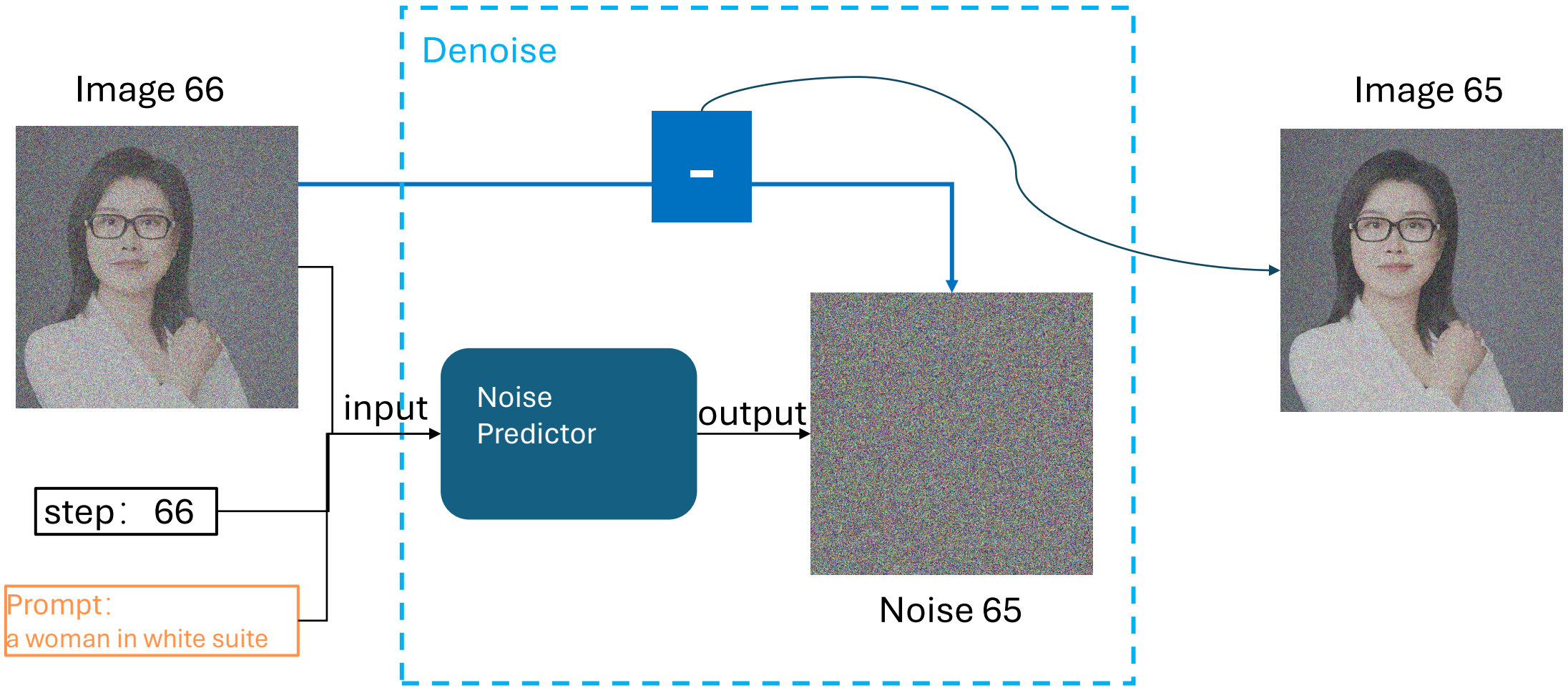
Image Generation Model (Neural Network)

paired data

Diffusion Model: Generate Images from Noise



Denoise => Predict noise



Math principals of diffusion model

Algorithm 1 Training

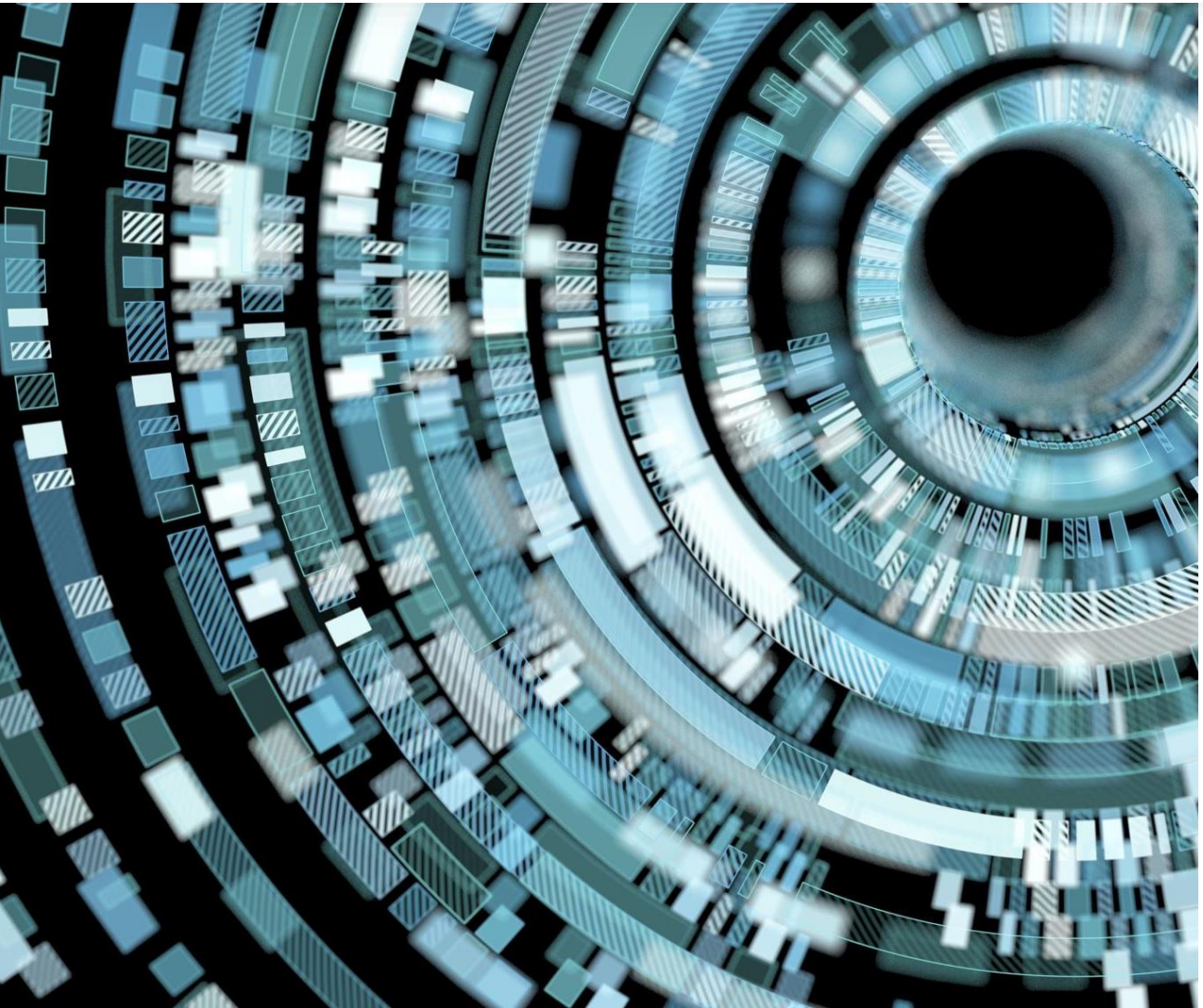
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

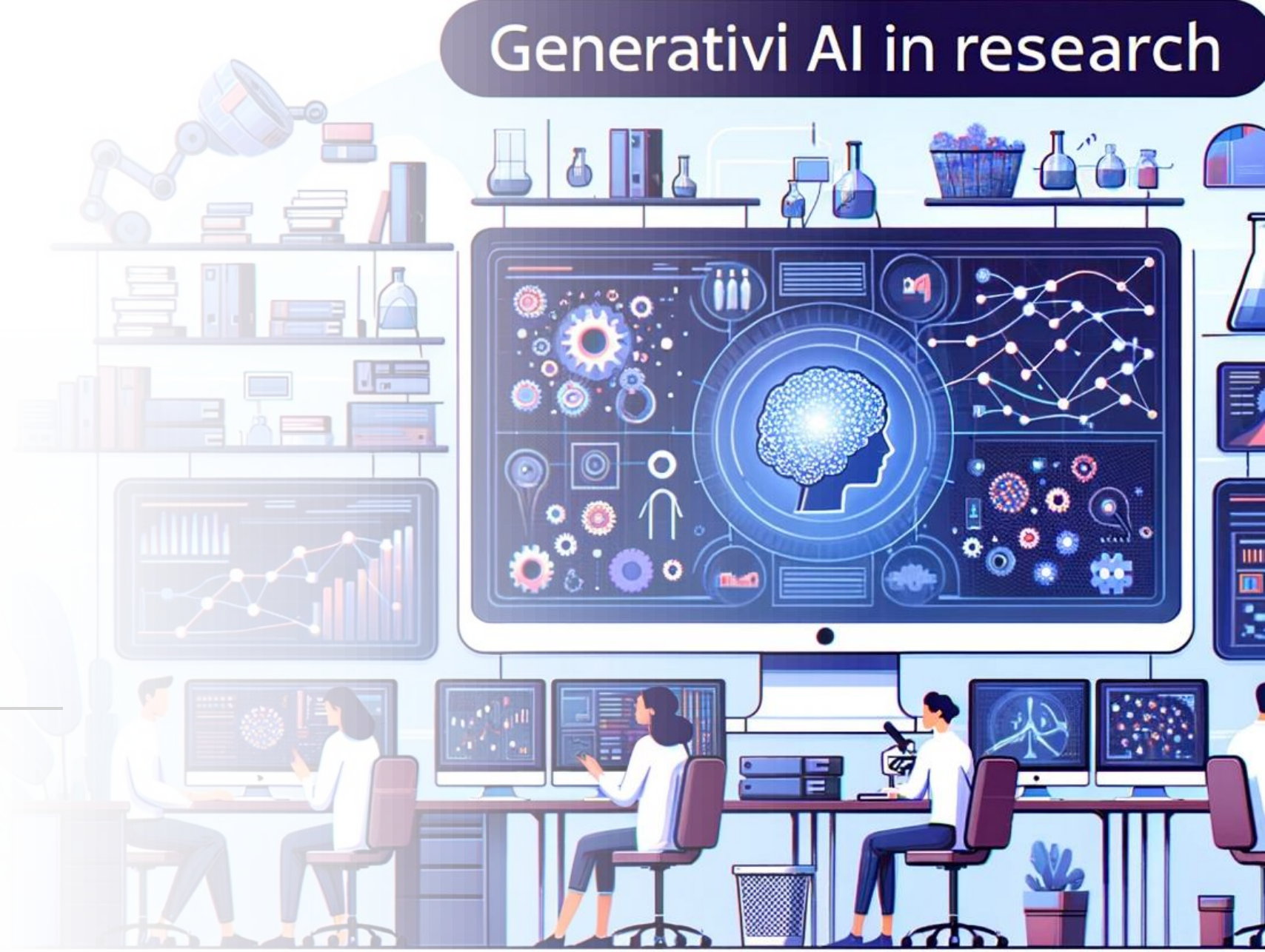
Denoising Diffusion Probabilistic Models
Submitted on 19 Jun 2020, last revised 16 Dec 2020
[2006.11239] Denoising Diffusion Probabilistic Models (arxiv.org)

The Spread of Prompt Engineering



- Prompt Engineering
  - a method to enhancing model performance by crafting effective prompts
  - the process of carefully crafting and refining input prompts to guide machine learning models, particularly language models, towards producing desired outputs
- The Revolution of AI adoption
  - Input with natural language instead of code
  - AI models are accessible for unprofessional users
  - Prompts are description of user’s requirements

Generative AI in Research



Generative AI aids in academic reading

- Automated Abstraction
- Automated summarization
- Knowledge extraction

- The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.
- Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

An academic assistant aids in paper reading

- Basic Information
- Paper Structure
- Key points
- Related information

Structure of “Attention is all your need”

1. Introduction
2. Background
3. Model Architecture

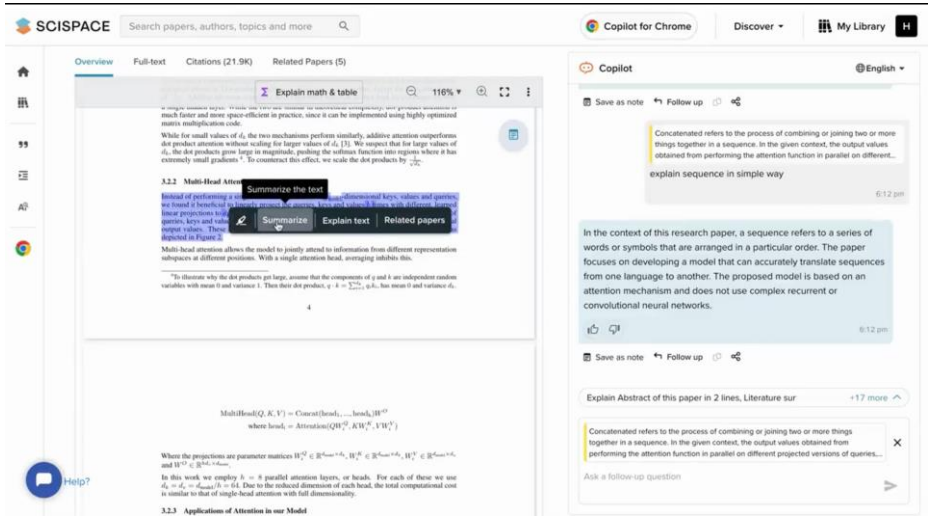
   3.1 Transformer Architecture

   3.2 Attention

   ……
4. Why Self-Attention
5. Training
6. Results
7. Conclusion

Automated Literature Review

- Utilizing AI to streamline the literature review process
- AI-powered tools for literature review
  - Time-saving
  - comprehensive analysis
  - bias reduction



Automated Hypothesis Generation

- Leveraging AI to formulate research hypotheses
  - AI-driven tools analyze vast literature using NLP
  - Identify connections across research areas
- Generate innovative hypotheses
  - Data analytics to discover new patterns or relations
  - Beyond traditional methods

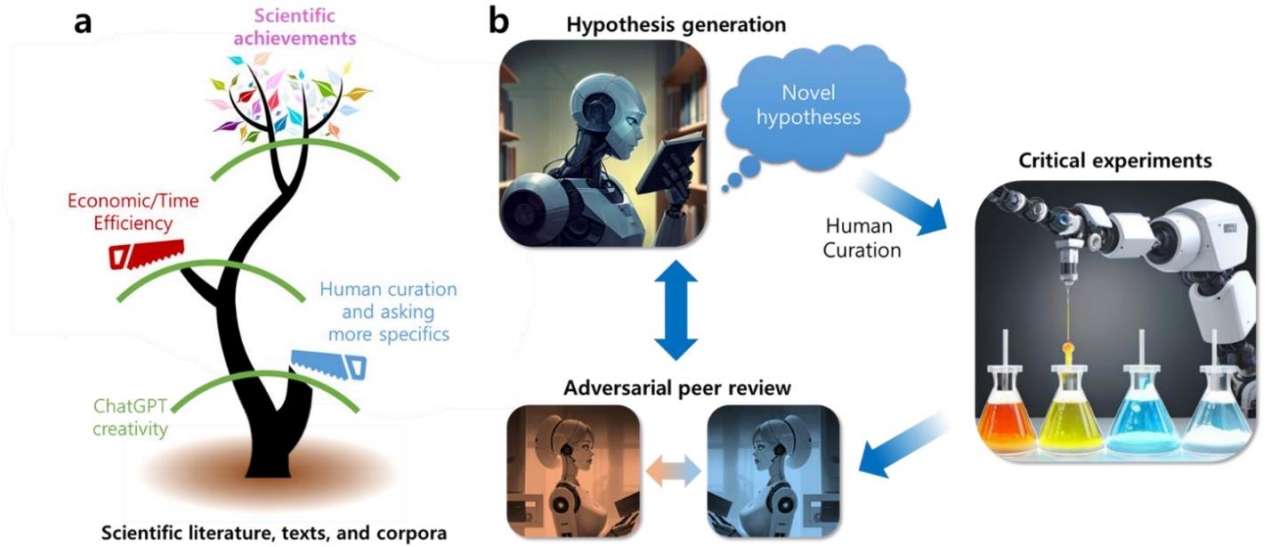


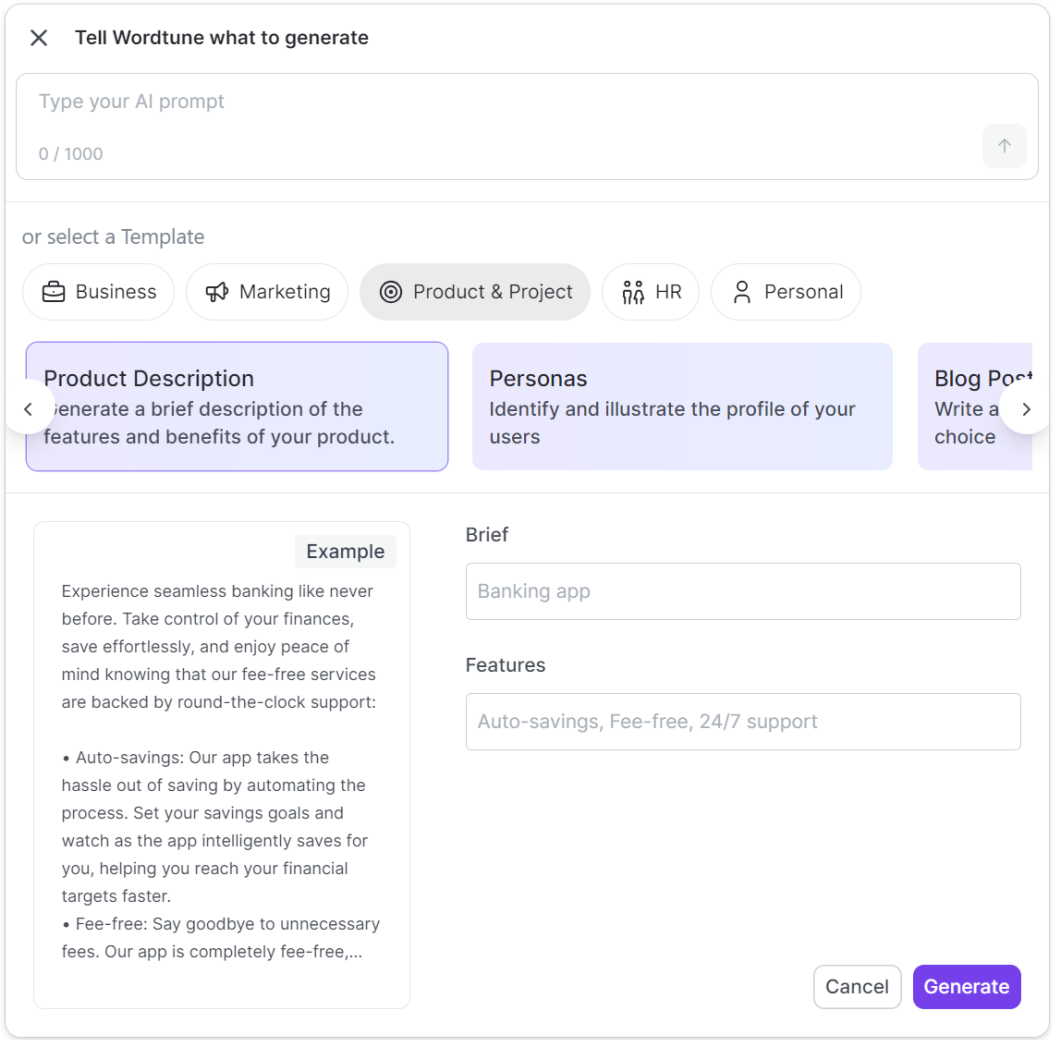
Figure 1. a) A systematic diagram of using ChatGPT as a hypothesis machine to achieve scientific innovations. ChatGPT's creativity is used to derive new hypotheses from existing literature. Then generated hypotheses are curated, tested and accepted. b) The scenario of AI-generated hypothesis generation, human curating, automated experimentation, and adversarial peer reviews. It may become a multi-agent swarm effort with built-in diversity of hyperparameters in an evolutionary setting.

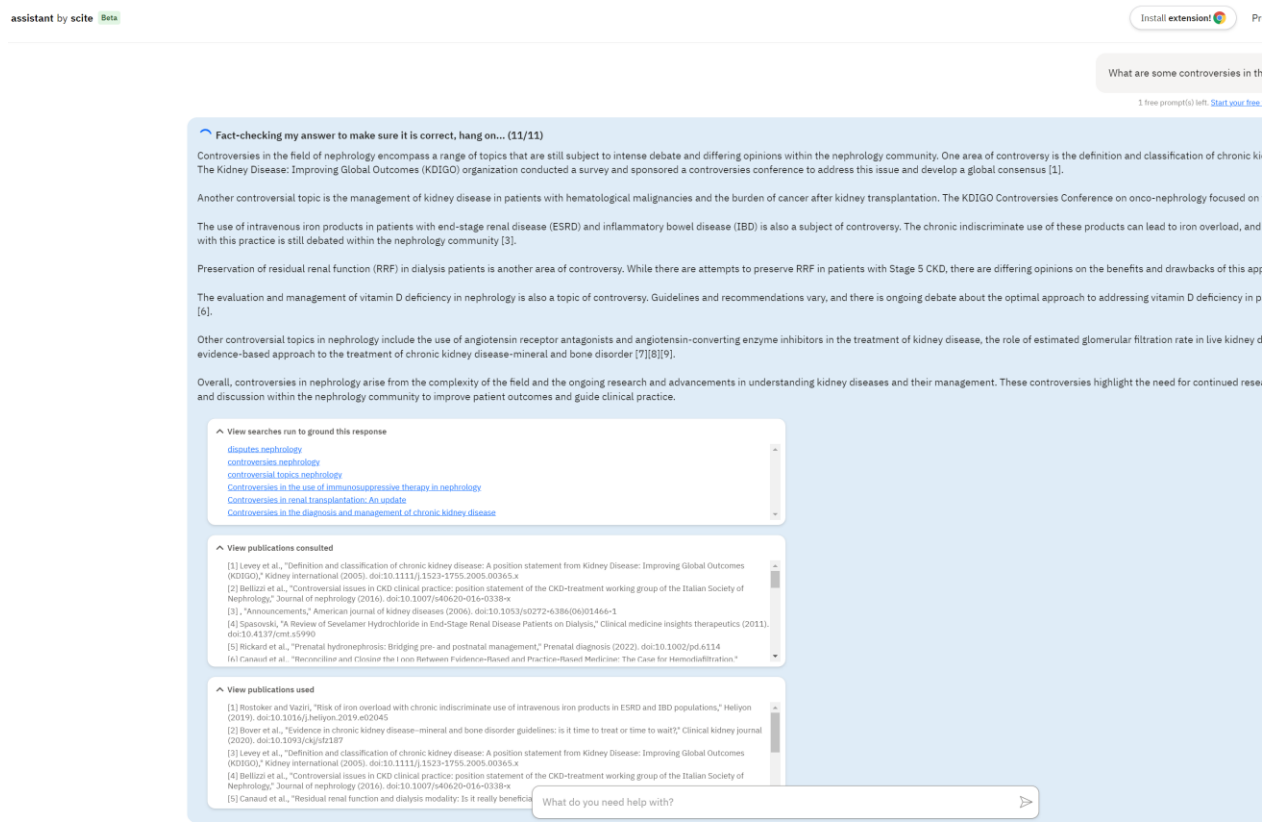
Data Augmentation & Synthetic Data Generation

- Data Augmentation
  - Transformations (rotation, flipping, cropping) to diversify dataset
  - Common in image tasks.
- Synthetic Data Generation
  - Create new data points via algorithms
  - Addresses data scarcity and aids when labeling is challenging.
- Expanding datasets through synthetic data generation
- Enhance training data quality & quantity

Prewrite & Draft Academic Papers

- Prewriting
  - Expand on your research questions with an AI brainstorming session
  - Turn your ideas into an outline with an AI outline generator
- Drafting
  - Streamline the writing process with an AI-powered content generator by offering personalized content that reflects your varied sources and ideas





Junior Researchers Coaching

- Utilizing AI to mentor and guide junior researchers, especially through chat-based interactions
- Benefits
  - Enhanced learning
  - Personalized feedback
  - Accelerated research progress

Collaborative Opportunities between Industry and Academia

Benefits of Collaboration

- Knowledge Exchange
  - Sharing expertise and resources for mutual growth.
- Talent Development
  - Training the next generation of AI researchers and practitioners.
- Problem-Solving
  - Tackling real-world challenges with combined efforts.

Areas for Collaboration (1)

- Joint Research Projects
  - Initiatives in the realm of Generative AI
  - Driven by requirements from industry
- Product Development
  - Academic guidance from academia (e.g. algorithms, prompts)
  - Collect feedback from customers directly

Areas for Collaboration (2)

- Educational Programs
  - Co-developed courses, workshops, and training programs
  - Public lectures and knowledge sharing
- Community Engagements
  - Hackathons
  - Competitions
  - Fostering an ecosystem

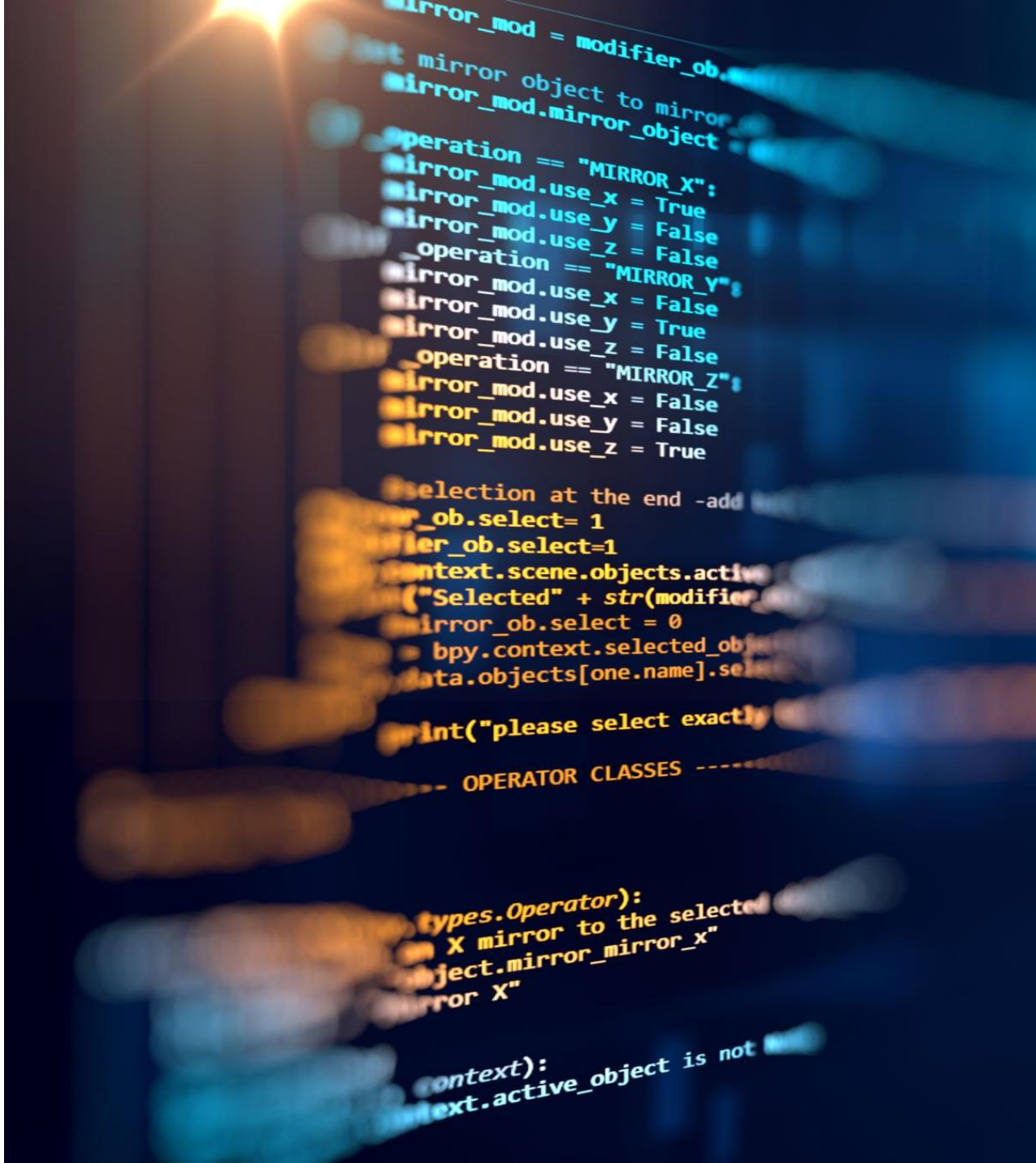
Responsible AI for Research

Bias and Fairness

- Biases in AI models
- Principle: Ensuring fairness and avoiding discrimination.
- Strategies
  - Bias detection
  - fairness-aware algorithms
  - diverse data collection.

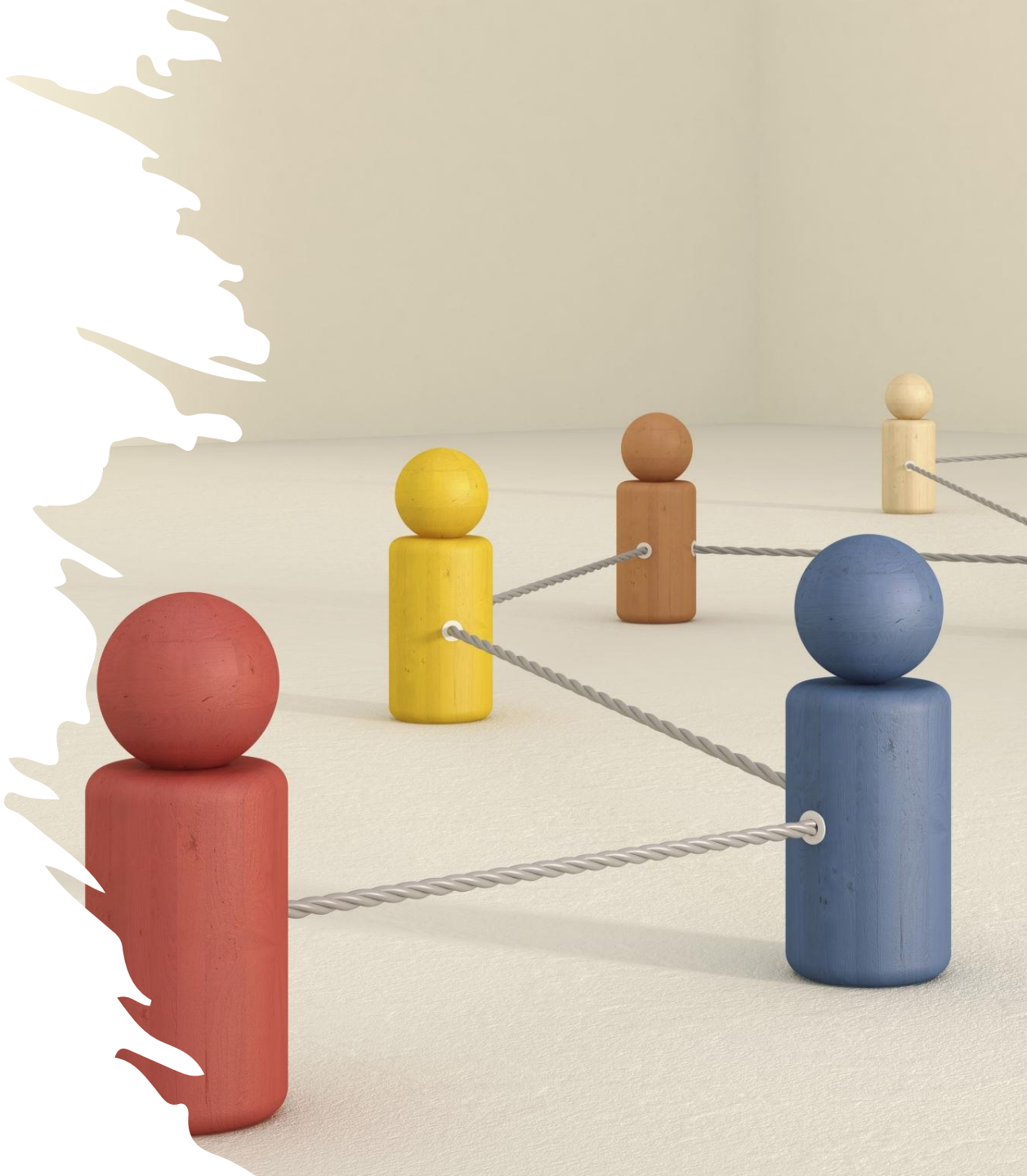
Data Privacy

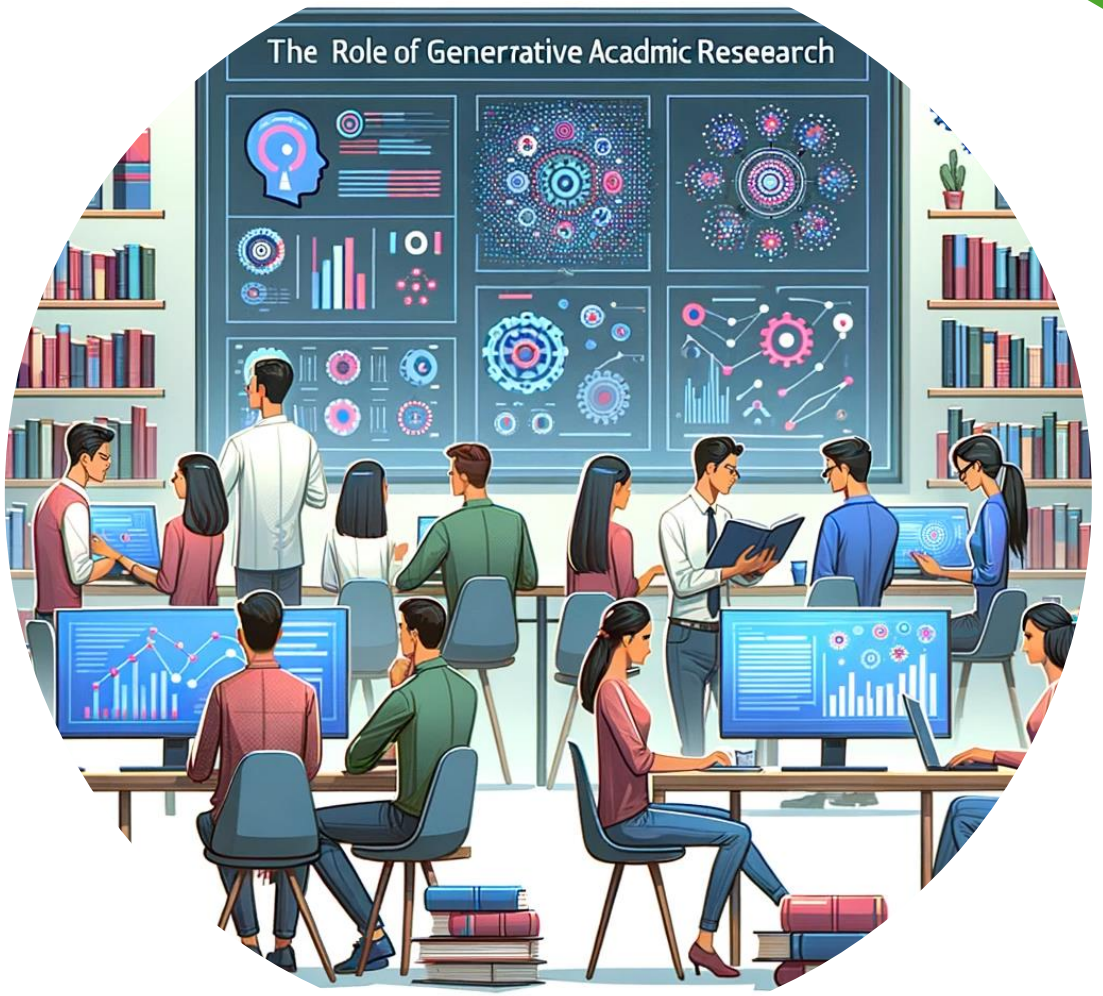
- Ethical handling of data, especially in sensitive research areas to protect individuals' data and ensuring privacy compliance
- Strategies
  - Privacy-preserving data augmentation
  - differential privacy



Accountability and Transparency

- Build trust and understanding among stakeholders by Ensuring clarity and openness in AI systems and research processes.
- Strategies
  - Transparent reporting
  - open-source models
  - explainable AI





Thank you

jull@microsoft.com