Agenda

- 1) Overview of Machine Learning
- 2) Reinforcement Learning Basics
- 3) Reinforcement Learning: Main Definitions
- 4) Reinforcement Learning Python Example
- 5) Reinforcement Learning in Training Process of ChatGPT

Overview of Machine Learning:

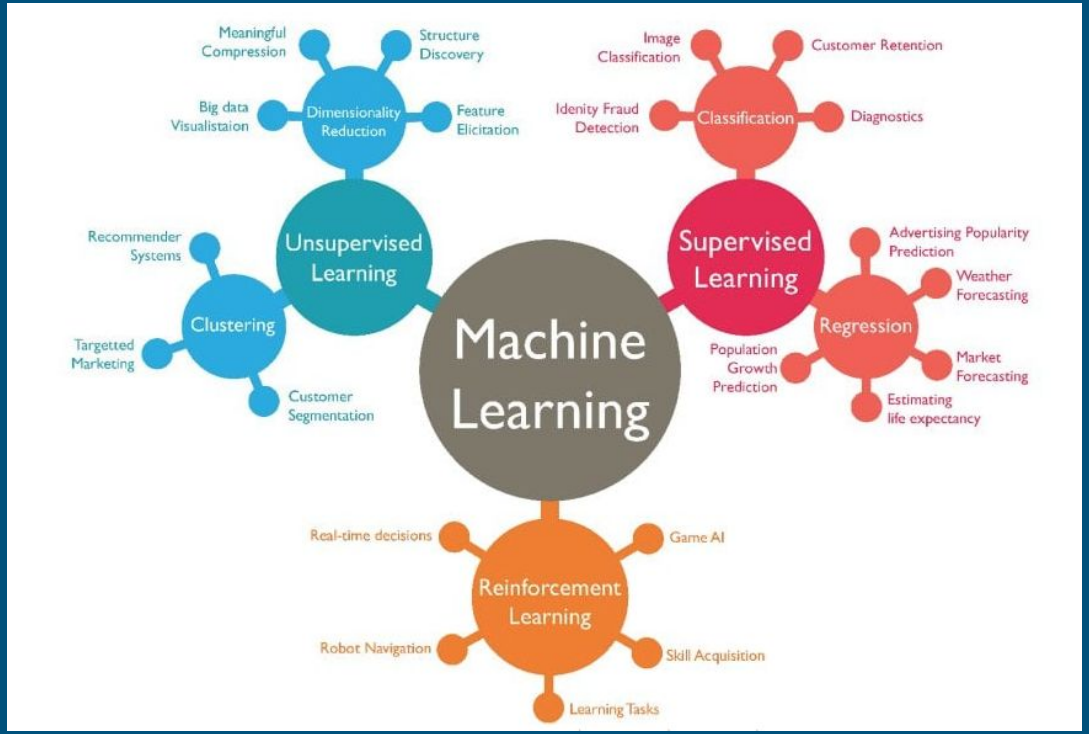
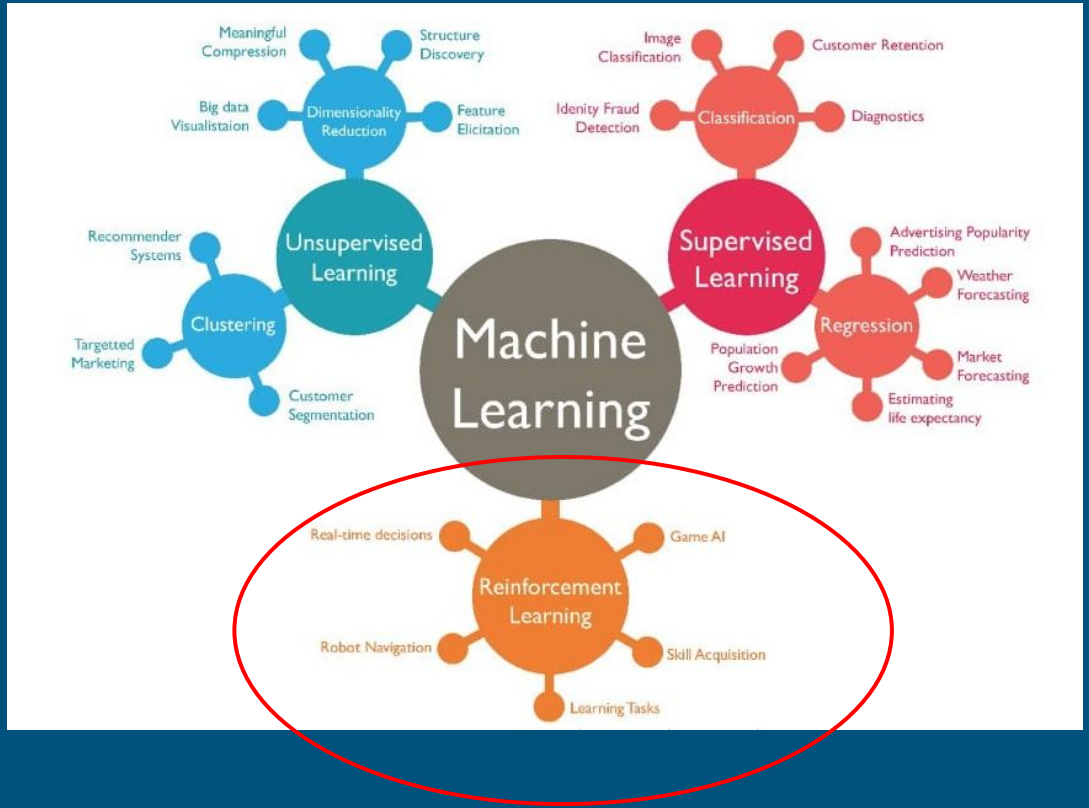


Image source:
https://datasciencedojo.com/blog/machine-learning-101/#

Overview of Machine Learning:

Reinforcement Learning



Reinforcement Learning Basics

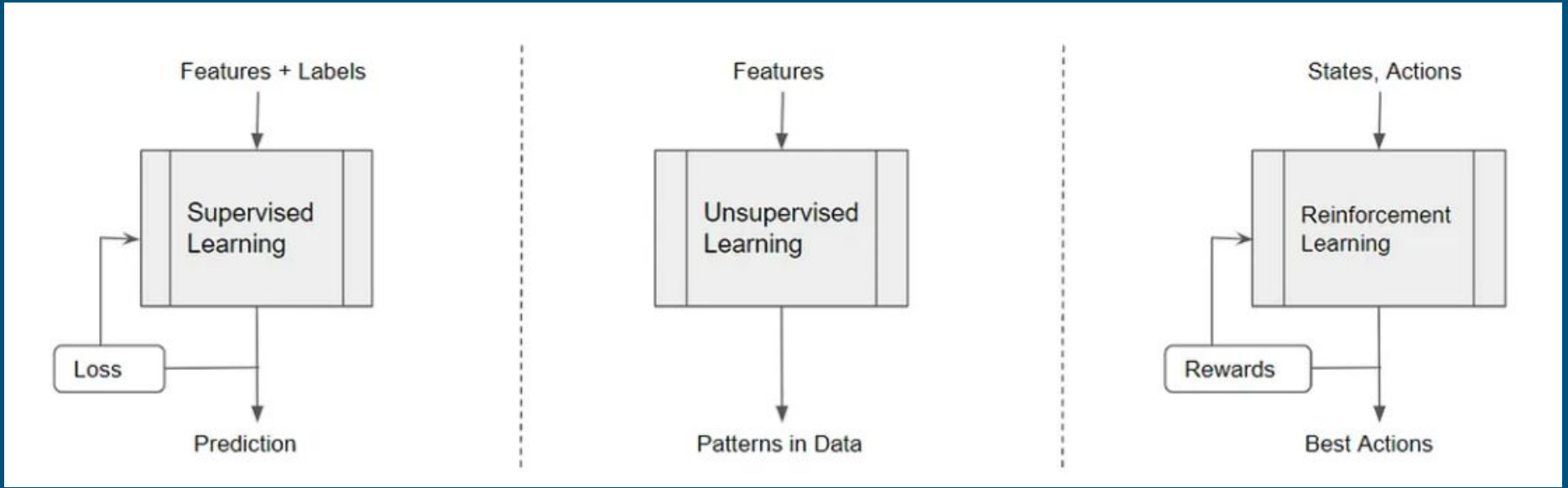
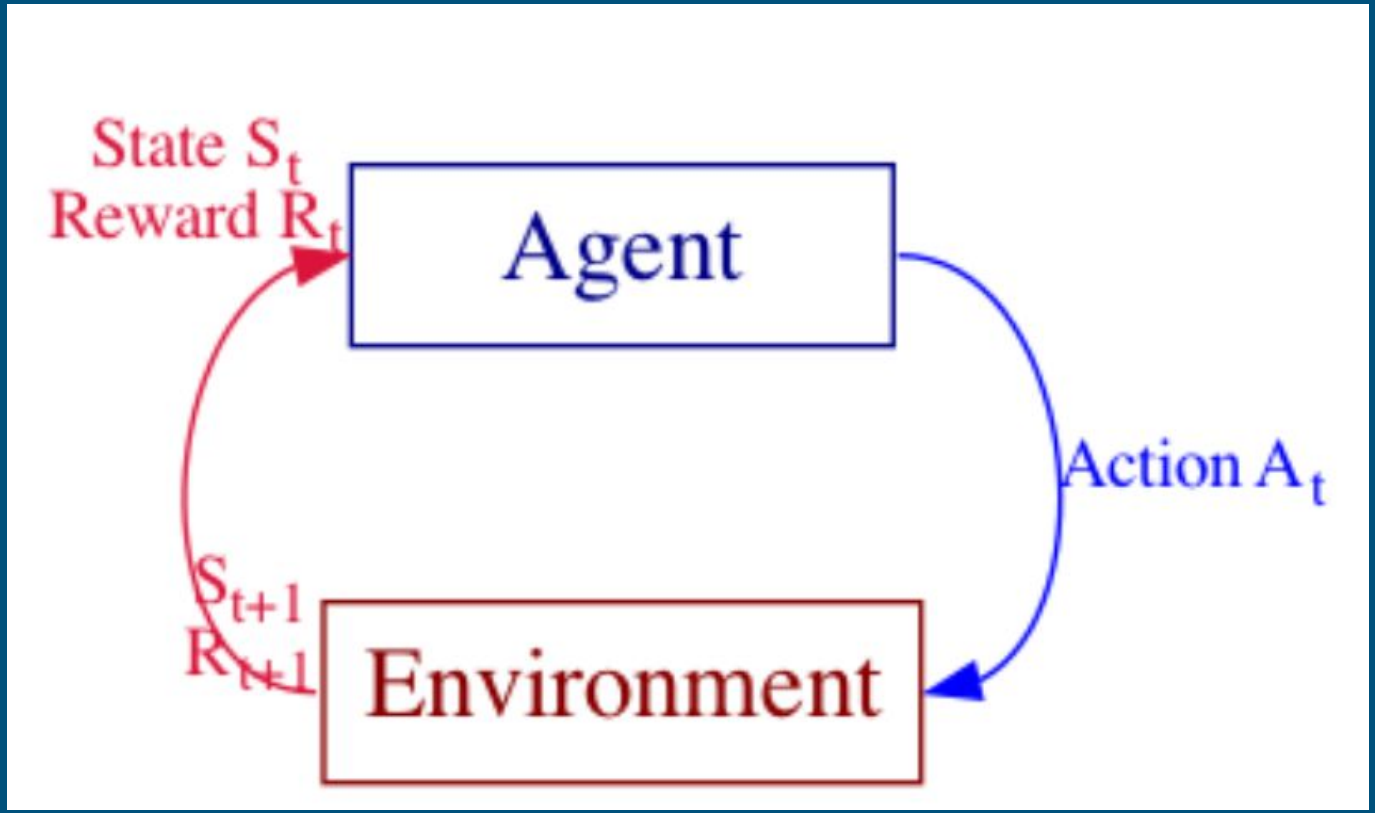
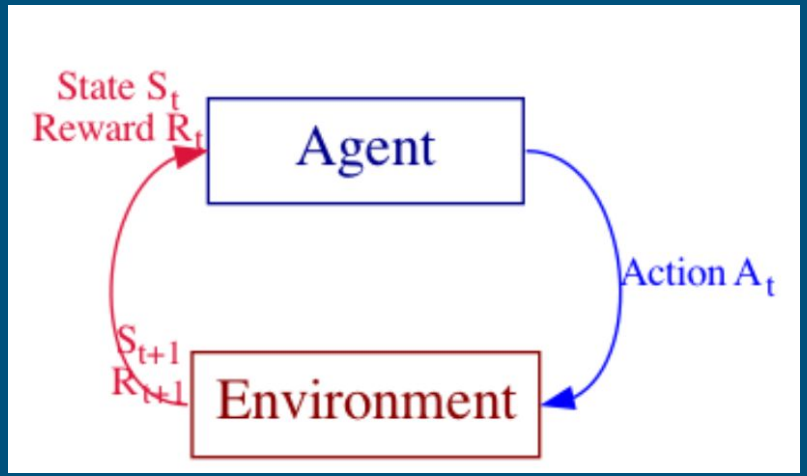


Image source:
https://towardsdatascience.com/reinforcement-learning-made-simple-part-1-intro-to-basic-concepts-and-terminology-1d2a87aa060

Reinforcement Learning Basics



Reinforcement Learning Basics



Environment:Physical world in which the agent operates
State — Current situation of the agent
Reward — Feedback from the environment
Policy — Method to map agent's state to actions
Agent: Decision Maker that take action

Image source:
"Foundation of Reinforcement Learning", Rao Jelvis

Reinforcement Learning Basics

Environment:Physical world in which the agent operates
State — Current situation of the agent
Reward — Feedback from the environment
Policy — Method to map agent's state to actions
Agent: Decision Maker that take action

Image source:
https://towardsdatascience.com/reinforcement-learning-made-simple-part-1-intro-to-basic-concepts-and-terminology-1d2a87aa060

Reinforcement Learning: Main Definitions

Episodes: As Series of Atomic Experiences

State, Actions, Reward, New State, Action…

$$(S_0, A_0, R_1, S_1, A_1, R_2, \cdots, S_{T-1}, A_{T-1}, R_T, S_T)$$

Reinforcement Learning: Main Definitions

Policy: Mapping the State to Action

$$\pi(a|s) = P(A_t = a|S_t = s)$$

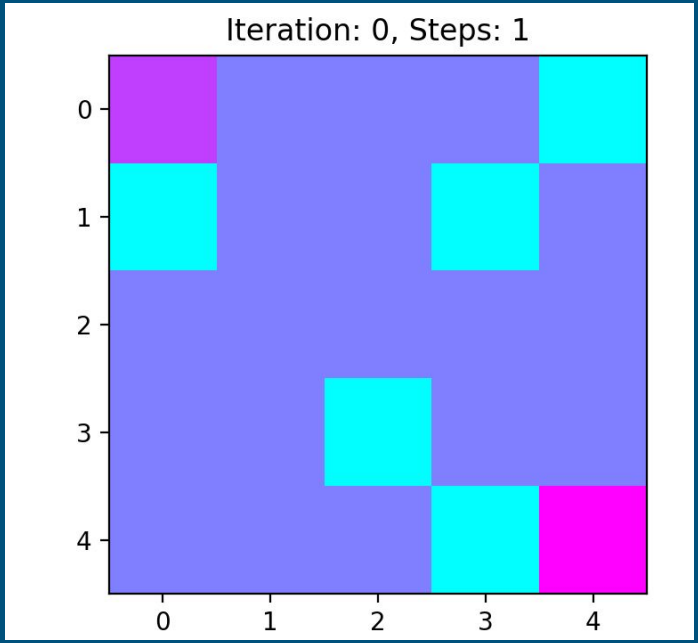
Reinforcement Learning: Main Definitions

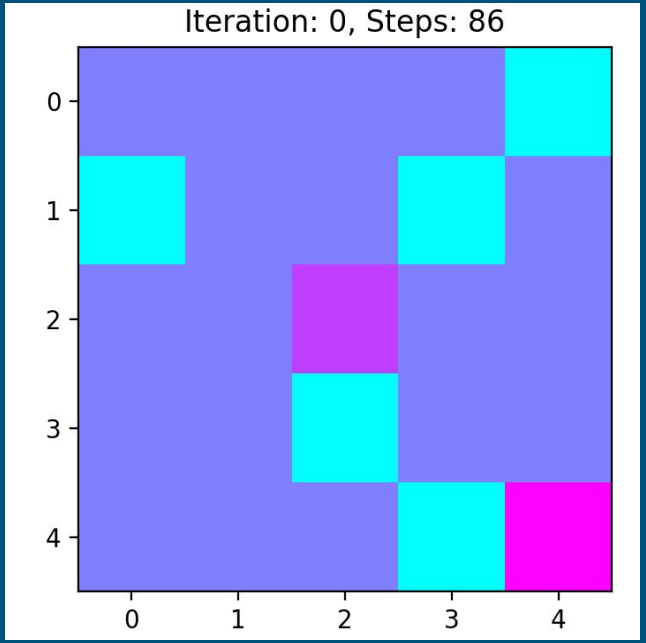
The Learning Eq: Q-learning

$$Q^{new}(S_t, A_t) \leftarrow (1 - \underbrace{\alpha}_{\text{learning rate}}) \cdot \underbrace{Q(S_t, A_t)}_{\text{current value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \Big(\underbrace{R_{t+1}}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(S_{t+1}, a)}_{\text{estimate of optimal future value}} \Big)$$

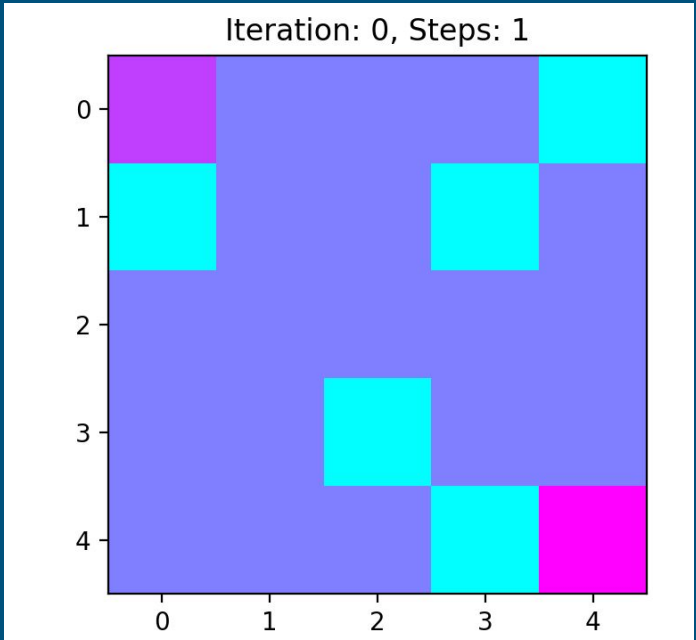
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{new value (temporal difference target)}}$$

RL Example:
Coding Example:





RL Example:
Coding Example:



Reward Structure:

Reward Arriving at goal: **+1**
Rewards At each step: **-0.1**
Reward Hitting Obstacles: **-1**

Available Actions:

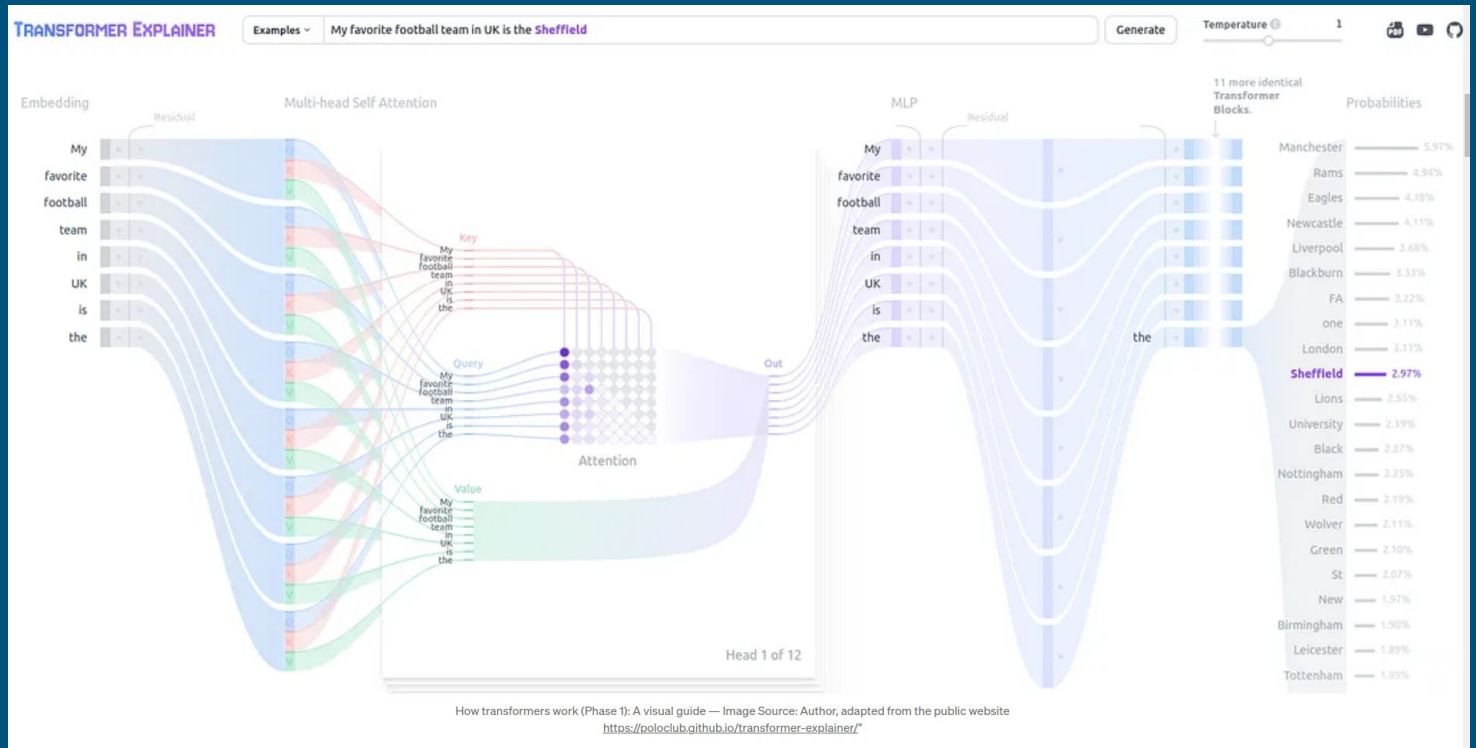
0: Up, 1: Right, 2: Down, 3: Left

Three Phases of ChatGPT Training

1. Phase 1: Pre-training
2. Phase 2: Supervised Fine-tuning (SFT)
3. Phase 3: Reinforcement Learning from Human Feedback (RLHF)

Phase 1: Pre-training

The goal is the “next word prediction”



How transformers work (Phase 1): A visual guide — Image Source: Author, adapted from the public website https://poloclub.github.io/transformer-explainer/"

Phase 2: Supervised Fine-tuning (SFT)

Language is inherently flexible — there are many valid ways to respond to any prompt

For example, if asked “How do I make pasta?”, valid responses could include:

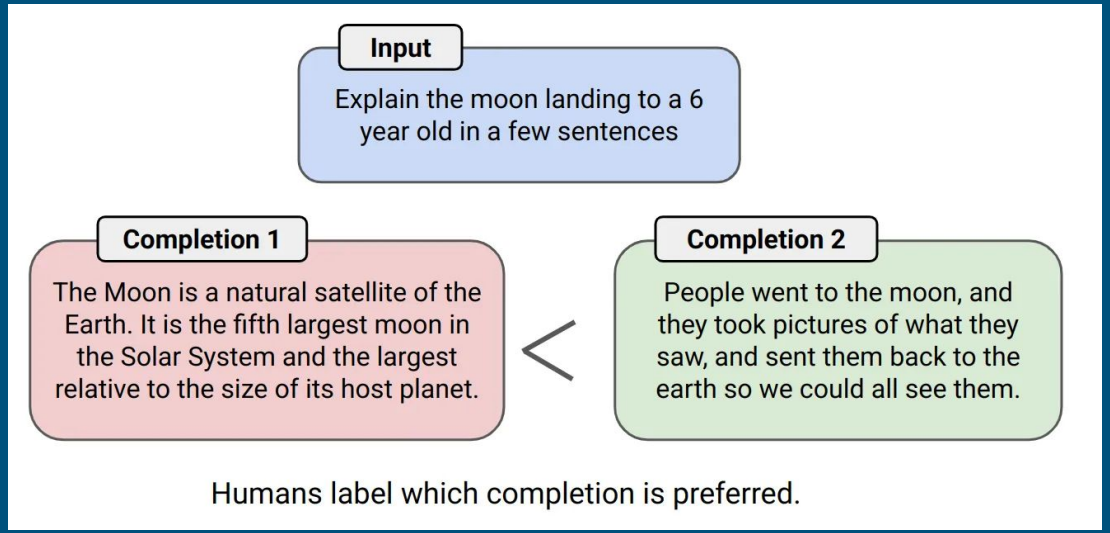
A) You can learn by watching Youtube
B) Giving you a listing of ingredient, explaining cooking steps, or suggesting serving sizes

Phase 2: Supervised Fine-tuning (SFT)

Language is inherently flexible — there are many valid ways to respond to any prompt

Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences?	The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to pump blood faster, stop digesting, and produce adrenaline and cortisol.
Create a shopping list from this recipe: Trim the ends off zucchini. Cut zucchini in half lengthwise; scoop out pulp, leaving 1/2-in. shells. Finely chop pulp. In a skillet, cook beef, zucchini pulp, onion, mushrooms and peppers over medium heat until meat is no longer pink; drain. Remove from the heat. Add 1/2 cup cheese, ketchup, salt and pepper; mix well. Spoon into the zucchini shells. Place in a greased 13x9-in. baking dish. Sprinkle with remaining cheese.	Zucchini, beef, onion, mushroom, peppers, cheese, ketchup, salt, pepper

Phase 3: Reinforcement Learning from Human-Feedback



Phase 3: Reinforcement Learning from Human-Feedback