

A chatbot from OpenAI

ChatGPT

- AIGC: AI-Generated Content
 - Al painting
 - Al chat

• AGI: Artificial general intelligence

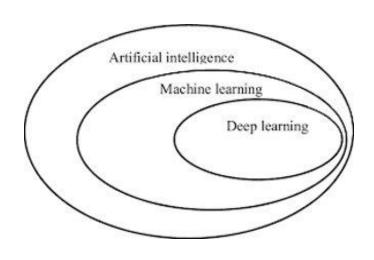
```
_______ modifier_ob.
 mirror object to mirror
mirror_mod.mirror_object
Peration == "MIRROR_X":
irror_mod.use_x = True
irror_mod.use_y = False
### Irror_mod.use_z = False
 _operation == "MIRROR_Y"
irror_mod.use_x = False
 lrror_mod.use_y = True
  lrror_mod.use_z = False
  _operation == "MIRROR_Z"
  rror_mod.use_x = False
  lrror_mod.use_y = False
  rror_mod.use_z = True
  melection at the end -add
   ob.select= 1
   er ob.select=1
   ntext.scene.objects.active
   "Selected" + str(modifier
    rror ob.select = 0
  bpy.context.selected_obj
   ata.objects[one.name].se
  int("please select exaction
  OPERATOR CLASSES ----
     pes.Operator):
     X mirror to the selected
    ject.mirror_mirror_x"
 ontext):
ext.active_object is not
```

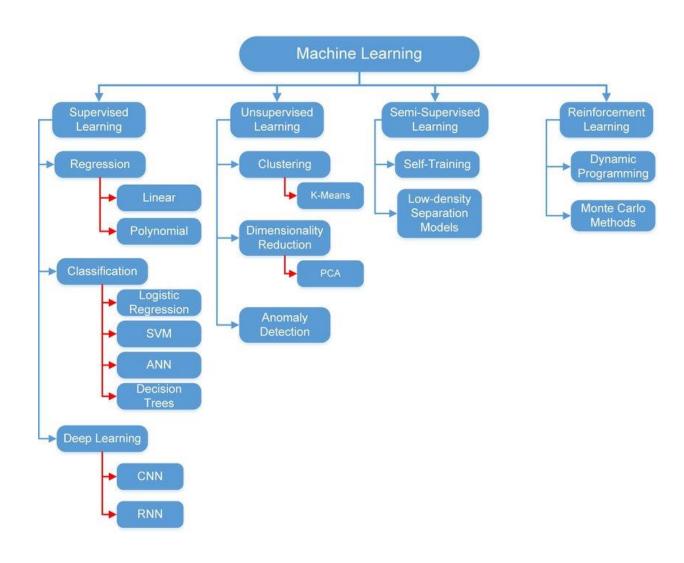
ChatGPT

Chat => Language Model
 NLP(Nature Language Processing)

• GPT => GPT-3.5

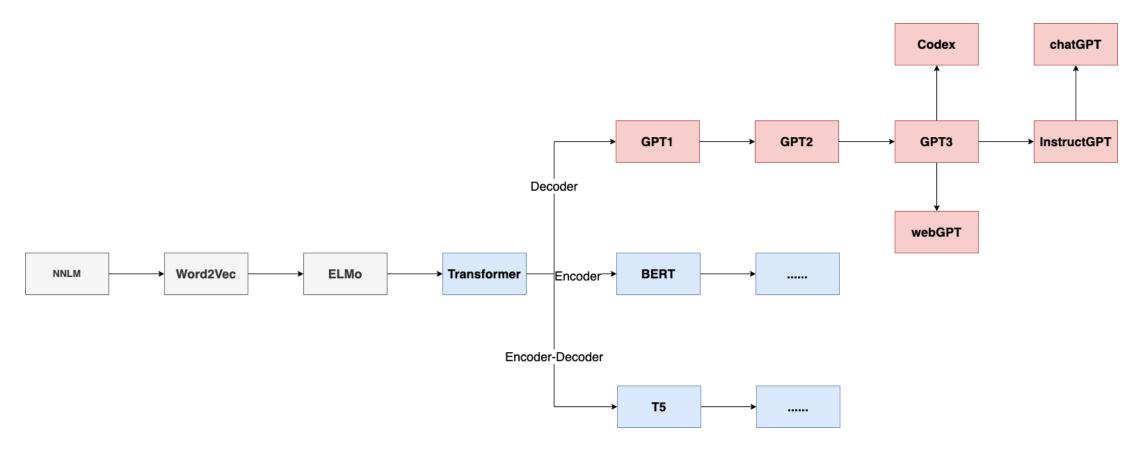
Artificial Intelligence

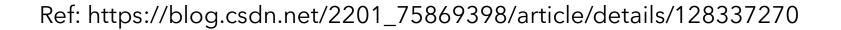




Ref: E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics

Language Model





ChatGPT

- Tune pre-trained GPT 3.5 model
- Train a reward model
- Reinforcement Learning with reward model

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

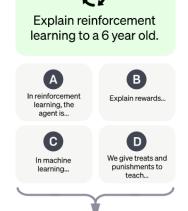
This data is used to fine-tune GPT-3.5 with supervised learning



Step 2

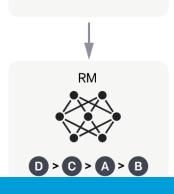
Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



D > C > A > B

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

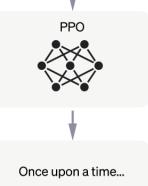
The PPO model is initialized from the supervised policy.

The policy generates an output.

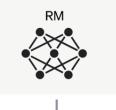
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO











Q & A

Thanks!