对 ChatGPT 的原理进行理解。

1.ChatGPT 的功能是，给 ChatGPT 一个语言指令，ChatGPT 给出相应的回复。

2.将其功能进行抽象，就是给定 x（一个语言指令），输出 y（回复）。这是属于强化学习（reinforcement learning,RL）的领域。强化学习研究的问题是，对于某个 agent，在情况 x 下，做出最优行动 y，以使得最终的效用最高。比如，某个人打乒乓球，每一时刻，他观察到球的状态，对手的位置，这些记为 x，他要决定自己应该做出什么动作 y，最终的效用是这局球有没有获胜。在 RL，最优的从 x 到 y 的映射，称为最优政策，表示在 x 情况下，执行 y 动作是最优的。可以将 y 表示成 x 的函数，y=f(x)。

3.ChatGPT 要学习 y=f(x)的函数中的参数。（1）由于 x 和 y 都是语言文字，f(x)的函数形式大体上是 NLP 中的神经网络形式，而 OPENAI 公司本身有一个 y=f(x)的模型，其中的参数是通过网络上的数据训练得到的。而网络上的数据，x 是指令且 y 是“应该的（或最优的）回复”的较少，其形式主要是 x 是上句，y 是下句（上句的延续）。所以 OPENAI 雇佣了一些人，这些人人工产生了一些 x 是指令且 y 是“应该的回复”的样本，然后将其已有的 f(x)模型拿过来，用这些新数据精调（fine-tune），即以已有的模型的参数值作为该模型参数的初始值，然后用新数据来训练，轻微修改模型参数值，使得模型能够更好地针对这种任务（给定一个语言指令，输出“应该的回复”）有好的表现。（2）这第（1）步用的是监督学习的方法，因为自变量为 x，因变量为“应该的回复”。而前面说过这一问题也可以看做强化学习。但是要使用强化学习的方法，得有奖励函数（奖励函数是指 agent 做出动作 y，得到的奖励是 z，z 是动作 y 和当前环境状态 x 的函数，该函数就是奖励函数）。为了得到奖励函数，OPENAI 用已有的模型生成一些样本，比如 x 下生成 y1，x 下生成 y2，x 下生成 y3,...（同一个 x 下 f(x)生成的结果不一样，是因为 y 是类型变量，f(x)精确来讲是只在 x 下该模型生成 y1,y2,y3 的概率）。对于上述样本，人工给 y1,y2,y3 分值 z1,z2,z3,...。人工认为该回复越令人满意，则给得分值越高。然后用这些样本训练，得到 z=r(x,y)的函数形式，这就得到奖励函数。（3）将第（1）步得到的政策作为初始政策，进行强化学习训练。大概过程是，随机给定一个初始状态 x，用政策 f(x)生成行动 y，用奖励函数得到该动作 y 的奖励 z，然后用 z 来更新政策 f(x)中的参数,...。之前了解的 RL 中学习算法一般是 Q-learning，这里用得是 Policy Gradient 方法。

4.对 Policy Gradient 方法本身做了理解。

5.在理解 Policy Gradient 方法的时候对强化学习加深了理解。