

进行的尝试：

1. Chat-GPT：

1. 第一次开会时给我们看的几个问题：3月1日的GPT已经可以解决了
2. 在StepGame数据集上的k的尝试：（chat会学习，所以一次问一堆）  
（需要新的chat进行“训练”，不然就得视之前问了啥来定表现了）

remember what you have seen:

0.Story: X is to the left of K and is on the same horizontal plane.

Question: What is the relation of the agent X to the agent K? left

1.Story: G is at the 6 o'clock position relative to R.

Question: What is the relation of the agent G to the agent R?

below

2.Story: If H is the center of a clock face, B is located between 10

Now is your turn. Answer the following question about spatial reasoning, using and only using the options of left, right, above, below, upper-left, upper-right, lower-left, lower-right, or overlap.

9.Story: F is placed at the upper left of V. V is there and D is at the 2 position of a clock face. E is placed at the lower right of Q.

Question: What is the relation of the agent D to the agent V?

10.Story: F is diagonally right and below B. C is sitting at the top



31. U is to the right of N.
32. C is below and to the left of Y.
33. P is to the left of M.
34. U is above B.
35. P is diagonally below and to the left of F.
36. D is to the right of L.
37. Q is to the right of W.
38. A is to the right and slightly above Q.
39. Z is below T.
40. Z is to the upper right of W.
41. A is above P.
42. I is diagonally above and to the left of M.
43. S is below and to the left of U.
44. V is below Q.
45. V is to the left of T.
46. G is to the left of X.
47. C is to the upper left of Y.
48. S is to the left of D.

## 2. GPT-3.5-turbo:

1. Prompt: 什么样的prompt可以得到想要的回答; 正确率为

[k=1]  $18 / 40 = 0.45$

[k=2]  $12 / 40 = 0.3$

[k=3]  $10 / 40 = 0.25$

[k=4]  $9 / 40 =$

[k=5]  $8 / 40 =$

[k=6]  $10 / 40 =$

[k=7]  $5 / 40 =$

[k=8]  $4 / 40 = 0.1$

[k=9]  $6 / 40$

[k=10]  $7 / 40$

3. ada:

1. prompt: 包括curie等, 用相同的Prompt得不到想要的结果 / 胡言乱语

2. fine-tune (as classification problem): 1000条train [k=1] acc=0.225

1. 是否合理? (相当于在一定程度上提供了rule; 但prompt也)

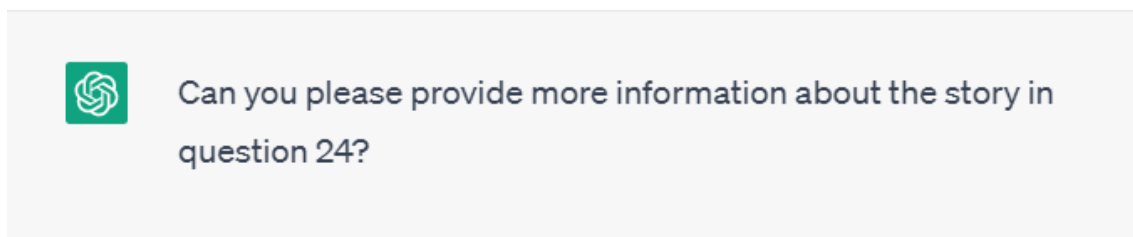
2. 用了数字编码, 可能损失了embedding信息;

3. ada模型本身不够好, 可能需要更多train data (x10倍之类的)

问题? :

1. Chat-GPT的随机性: 会胡言乱语!

23	No relation.	lower-left
----	--------------	------------



2. 通过max-length限制, 不确定是否会向这个参数考虑在内; 但是发现只会直接截断。

9	above and to the left	above
---	-----------------------	-------

3. 判断正确率的时候需要人工操作....请问是否有更好的处理方式?

16	bottom right	lower-right
7	top	above

4. 会回答完全相反的情况: AB主体难以区分?

14	lower-right	upper-left
----	-------------	------------