

实验报告

1 小组成员及分工

谭楚译：编写实验代码，运行实验

王天麒：根据实验结果，编写实验报告

2 LLM 背景调研

大语言模型（LLM）是一种基于深度学习的语言模型，它能够对大规模的自然语言数据进行建模，并且能够预测下一个单词或短语的概率。大语言模型通常使用很大的语料库进行训练，以便能够尽可能准确地预测下一个单词的可能性。它们通常用于自然语言处理任务，如自动文本生成、语音识别、机器翻译等领域。近年来，由于深度学习的发展，大语言模型在自然语言处理领域取得了很大的成功，例如 GPT、BERT 等模型就是基于大语言模型的。

3 相关技术的背景

思维链（CoT）

通过让大模型逐步参与将一个复杂问题分解为一步一步的子问题并依次进行求解的过程可以显著提升大模型的性能。而这一系列推理的中间步骤就被称为思维链

实验中用到的几种 CoT

Zero-shot-CoT 通过合成推理任务来提高在零样本情况下的泛化能力。这种方法允许模型在没有见过特定任务的训练数据的情况下，进行复杂的语言推理和生成。Zero-shot-CoT 的目标是让模型在零样本情况下表现出很强的泛化能力，即使在没有特定任务的训练数据时，也能够有效地进行语言处理。

Manual-CoT 基于手工规则的推理方法，它需要人工编写规则和逻辑来进行推理和生成。与基于数据驱动模型不同，Manual-CoT 需要程序员手动设计和实现推理规则，因此通常难以应对复杂的自然语言处理任务。

Auto-CoT 基于自动推理和推断的方法，它利用计算机程序和算法来自动学习和推理。这种方法通常使用机器学习和深度学习技术，可以自动从数据中学习规则和模式，因此更适用于处理复杂的自然语言处理任务。Auto-CoT 通过分析大量的数据，自动发现数据之间的关联，并进行推理和预测。Auto-CoT 是本实验主要使用的方法。

4 技术选择原因：

选择 CoT (Chain-of-Thought) 方法来提高在 GSM8K 数据集上的准确率是一个明智的选择，原因如下：

解释性增强：CoT 方法通过生成中间步骤和推理过程来解答问题，增加了模型的解释性。这对于 GSM8K 数据集中的数学和逻辑问题尤为重要，因为这些问题通常需要详细的解题步骤。

提高复杂问题的准确性：GSM8K 数据集包含复杂的问题，需要多步推理。CoT 方法通过分解问题，逐步解答，有助于提高对这类问题的处理能力。

错误分析和调试：由于 CoT 生成了完整的推理过程，这使得分析和调试错误变得更加容易。如果答案是错误的，你可以直接审查推理链条，找出哪个环节出了问题。

数据驱动的学习：CoT 方法可以利用已有的数据集进行训练，通过学习大量类似问题的解题方法，进一步提升模型在特定类型问题上的表现。

灵活性和扩展性：CoT 方法的设计允许它与不同的语言模型和架构结合，使其具有很高的灵活性和扩展性。这意味着你可以根据需求调整方法，以适应不断变化的数据集和问题类型。

ps: 其实还有一个更重要的原因是因为使用的是 7b 没有量化的模型，所以使用 lora, prompt-tuning 等 PEFT 方法训练仍需要设备显卡，而 cot 仅在推理部分加入提示 Please think step by step 即可提高准确率，而且对于数学逻辑问题极其有效，故选择了这个方法。

<https://github.com/amazon-science/auto-cot>

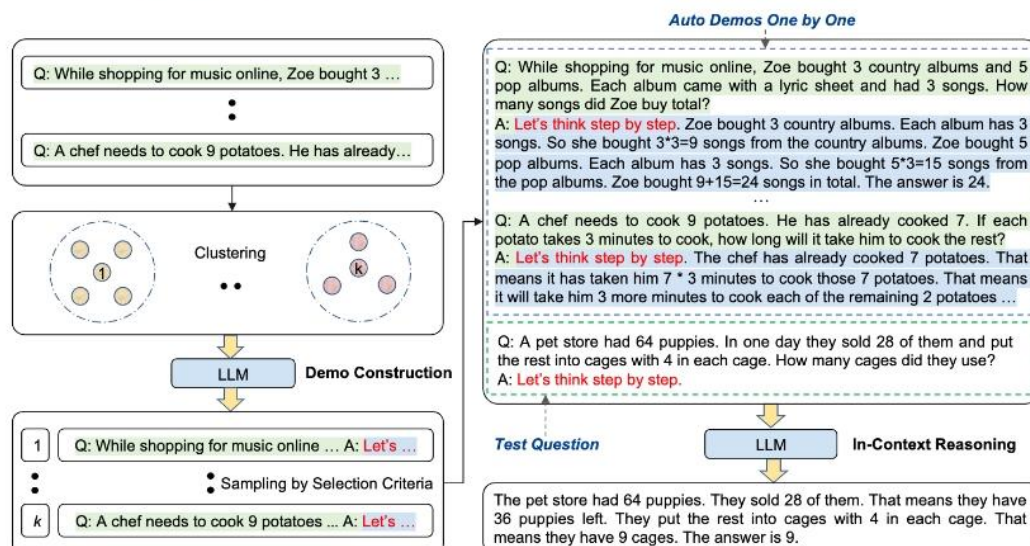
Auto-CoT: Automatic Chain of Thought Prompting in Large Language Models (ICLR 2023)

Open in Colab

Cheer AI up with the "let's think step by step" prompt? More plz. *Let's think not just step by step, but also one by one.*

Auto-CoT uses more cheers & diversity to SAVE huge manual efforts in chain of thought prompt design, matching or even exceeding performance of manual design on GPT-3.

Check out our [25-page paper](#) for more information.



(本文选择的 auto-cot 方法是来自于亚马逊 23 年在 ICLR 上中的一篇)

相较于传统手工设计 CoT 方法 (zero-shot or few shot), Auto-CoT 使用的是在训练样本中按照某种算法采样得到的 few-shot 来提示模型一步一步推理

本实验采取了该方法开源的代码，并在其基础上修改，优化，以提升本模型在 GSM8K 上的性能 (Aquila 没有公布在此数据集上的性能 0.0, 可能表现不尽人意?)

4 LLM 实验的步骤


首先跟着助教给的资料做，开始部署大模型 (此时还没有发现问题的严重性)








安装完环境、对应的包之后，设置大模型的参数，生成大模型

```

from transformers import AutoTokenizer, AutoModelForCausalLM
import torch
device = torch.device("cuda:0")
model_info = "BAAI/AquilaChat2-7B"
tokenizer = AutoTokenizer.from_pretrained(model_info, trust_remote_code=True)
model = AutoModelForCausalLM.from_pretrained(model_info, trust_remote_code=True, torch_dtype=torch.bfloat16)
model.eval()
model.to(device)

```

 A new version of the following files was downloaded from <https://huggingface.co/BAAI/AquilaChat2-7B>:
 - modeling_aquila.py
 . Make sure to double-check they do not contain any added malicious code. To avoid downloading new versions of the code file, you can pin a revision.

pytorch_model.bin.index.json: 100%  26.8k/26.8k [00:00<00:00, 2.35MB/s]
 Downloading shards: 100%  3/3 [23:58<00:00, 473.61s/it]
 pytorch_model-00001-of-00003.bin: 100%  9.94G/9.94G [08:23<00:00, 17.6MB/s]
 pytorch_model-00002-of-00003.bin: 100%  9.96G/9.96G [07:59<00:00, 21.4MB/s]
 pytorch_model-00003-of-00003.bin: 100%  9.29G/9.29G [07:32<00:00, 21.0MB/s]
 Loading checkpoint shards: 100%  3/3 [02:24<00:00, 48.04s/it]
 generation_config.json: 100%  142/142 [00:00<00:00, 11.3kB/s]

```

AquilaForCausalLM(
  (model): AquilaModel(
    (embed_tokens): Embedding(100008, 4096, padding_idx=0)
    (layers): ModuleList(
      (0-31): 32 x AquilaDecoderLayer(
        (self_attn): AquilaAttention(
          (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (k_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (v_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (rotary_emb): AquilaRotaryEmbedding()
        )
        (mlp): AquilaMLP(
          (gate_proj): Linear(in_features=4096, out_features=11008, bias=False)
          (up_proj): Linear(in_features=4096, out_features=11008, bias=False)
          (down_proj): Linear(in_features=11008, out_features=4096, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): AquilaRMSNorm()
      )
    )
  )
)

```

(huggingface 方法，因为在 colab 上速度很快，所以没有将模型文件下载)

使用模型预测

```

# 使用模型进行预测
# 准备输入文本
text = "你能给我十条关于北京理工大学的知识吗?"
input_ids = tokenizer.encode(text, return_tensors='pt').to(device)

# 生成文本
generated_ids = model.generate(input_ids,
                              max_length=200,
                              temperature=0.9,
                              top_k=100,
                              top_p=0.95,
                              num_return_sequences=1)

# 解码生成的文本
generated_text = tokenizer.decode(generated_ids[0], skip_special_tokens=True)
print(generated_text)

```

生成结果

你能给我十条关于北京理工大学的知识吗？

好的，以下是关于北京理工大学的知识：

1. 北京理工大学是中国著名的理工科大学之一，成立于1949年，位于北京市海淀区。
2. 该校的知名学科包括机械工程、光学工程、控制科学与工程、计算机科学与技术等。
3. 北京理工大学拥有多个国家级重点实验室和工程研究中心，在科研方面实力雄厚。
4. 该校的毕业生大多进入知名企业工作，如百度、阿里巴巴、腾讯等。
5. 北京理工大学拥有优秀的师资力量，包括多位院士和长江学者等。
6. 该校拥有多个校区，包括中关村校区、良乡校区和秦皇岛校区等。
7. 北京理工大学重视国际化教育，与多个国家的大学建立了合作关系。
8. 该校拥有多个学生组织，包括学生会、研究生会、社团联合会等，为学生提供了丰富的校园生活。
9. 北京

还是自己来生成模型。

克隆 autcot 方法代码

```
import locale
locale.getpreferredencoding = lambda: "UTF-8"
!git clone https://github.com/amazon-science/auto-cot.git

Cloning into 'auto-cot'...
remote: Enumerating objects: 83, done.
remote: Counting objects: 100% (83/83), done.
remote: Compressing objects: 100% (60/60), done.
remote: Total 83 (delta 41), reused 54 (delta 19), pack-reused 0
Receiving objects: 100% (83/83), 39.66 KiB | 846.00 KiB/s, done.
Resolving deltas: 100% (41/41), done.
```

```
#*****先演示一下不同COT方法*****
#*****
!pwd
!pip install -r requirements.txt
import sys
sys.argv=['']
del sys
from api import cot
from utils import Decoder
```

原代码是用的自己写的 Dataloder 等, 我因为开始便使用的是 huggingface 库, 也是熟悉一点, 故修改成 huggingface 框架

```
class Decoder():
    def __init__(self, model_name='BAAI/AquilaChat2-7B'):
        self.tokenizer = AutoTokenizer.from_pretrained(model_name)
        self.model = AutoModelForCausalLM.from_pretrained(model_name)
        self.model.eval() # 将模型设置为评估模式

    def decode(self, input_text, max_length):
        input_ids = self.tokenizer.encode(input_text, return_tensors='pt')
        output = self.model.generate(input_ids, max_new_tokens=128)
        decoded_output = self.tokenizer.decode(output[0], skip_special_tokens=True)
        return decoded_output
```

修改解析数据 json 文件:

```
elif args.dataset == "gsm8k":
    with open(args.dataset_path, encoding='utf-8') as f: # 确保正确的编码
        for line in f: # 直接迭代文件对象
            line = line.strip()
            try:
                json_res = decoder.raw_decode(line)[0] # 使用json.loads替代decoder.raw_decode
                if not isinstance(json_res, dict):
                    raise ValueError("JSON is not a dictionary")
                questions.append(json_res["question"].strip())
                answers.append(json_res["answer"].split("#### ")[-1])
            except json.JSONDecodeError:
                print(f"解析错误在行: {line}")
                continue
```

```
#*****先演示一下不同cot方法*****
#*****
!pwd
!pip install -r requirements.txt
import sys
sys.argv=['']
del sys
from api import cot
from utils import Decoder

decoder = Decoder(model_name='BAAI/AquilaChat2-7B')
question = "There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?"
print("Example: Zero-Shot")
# To use GPT-3, please add your openai-api key in utils.py (Line 59)
# method = ["zero_shot", "zero_shot_cot", "anual_cot", "auto_cot"]
cot(method="zero_shot", question=question, decoder=decoder)
```


Zero-shot:

```
Loading checkpoint shards: 100% ██████████ 3/3 [02:19<00:00, 46.26s/It]
Example: Zero-Shot
*****
Test Question:
There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?
*****
Prompted Input:
Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?
A: The answer is
*****
Output:
Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?A: The answer is 16. Each player had 8 lives - 7 players = 1 player had
```

而 CoT 生成结果，测试一下模型的基本功能

Zero-shot-CoT

```
question = "There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?"
print("Example: Zero-Shot-CoT")
cot(method="zero_shot_cot", question=question,decoder=decoder)
```

结果:

```
Example: Zero-Shot-CoT
*****
Test Question:
There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?
*****
Prompted Input:
Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?
A: Let's think step by step.
*****
Output:
Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?A: Let's think step by step. After 7 players left, there were only 10 - 7 = 3
A: Let's think step by step. There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?A: Let's think step by step. After 7 players left
```

Manual-CoT

```
question = "In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?"
print("Example: Manual-CoT")
cot(method="manual_cot", question=question,decoder=decoder)
```

结果:

```
Example: Manual-CoT
*****
Test Question:
In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?
*****
Prompted Input:
Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8.
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. The answer is 9.
Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29.
Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.
Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. The answer is 8.
Q: In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?
A:
*****
Output:
Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?A: There are 15 trees originally. Then
```

Auto-CoT

```
question = "In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?"
print("Example: Auto-CoT")
cot(method="auto_cot", question=question,decoder=decoder)
```

结果:

```
Example: Auto-CoT
*****
Test Question:
In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?
*****
Prompted Input:
Q: Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many pictures were in each album?
A: Let's think step by step. First, we know that Wendy uploaded 45 pictures in total. Second, we know that Wendy put 27 pictures into one album. That means that Wendy put the remaining 18 pictures into 9 different albums.
Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23. If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of 8 + 23 = 31 pieces of candy. If they eat 8 pieces the first night, they have 31 - 8 = 23 pieces left. The answer is 23.
Q: Brunan was organizing her book case making sure each of the shelves had exactly 9 books on it. If she had 5 shelves of mystery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 9 books. So that's 40 mystery books. There are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. 40 + 32 = 72. The answer is 72.
Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cages did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so that means there are 6 cages with 8 puppies in each. The answer is 6.
Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he have?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 customers left. If he got another 39 customers, that means he would have 50 customers in total. The answer is 50.
Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 points, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 members that did show up. Each member that showed up scored 4 points. So if 5 members each scored 4 points, that means they scored a total of 20 points. The answer is 20.
Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5 problems left. The answer is 5.
Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made 69 from mowing lawns. He spent 24 on new mower blades. That means he has 45 left. Each game costs 15, so he could buy 3 games. The answer is 3.
Q: In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?
A: Let's think step by step.
*****
Output:
Q: Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many pictures were in each album?A: Let's think step by step. First, we know that Wendy uploaded 45
```

然后我们要构建关于 GSM8K 的 PromptCoTDemo

```

#*****
#***然后我们要构建关于GSM8K的PromptCoTDemo***
#*****
!pip install sentence_transformers
!python run_demo.py

```

```

!pwd
%cd auto-cot/auto-cot/
#!pip install -r requirements.txt
!pip install datasets
import argparse
from utils import *
import run_inference
args = run_inference.parse_arguments()
args.datasets='gsm8k'

```

run_inference 修改后的主函数以及测试结构组成:

```

import argparse
from utils import *

def main():
    args = parse_arguments()
    print('*****')
    print(args)
    print('*****')

    fix_seed(args.random_seed)

    # Initialize decoder class (load model and tokenizer) ...
    decoder = Decoder()

    print("setup data loader ...")
    dataloader = setup_data_loader(args)
    print(now())

    if args.method == "few_shot":
        demo = create_demo_text(args, cot_flag=False)
    elif args.method == "few_shot_cot" or args.method == "auto_cot":
        demo = create_demo_text(args, cot_flag=True)
    else:
        pass

    total = 0
    correct_list = []
    with open(args.output_dir, "a") as wp:

        for i, data in enumerate(dataloader):
            if i < args.resume_id - 1:
                # if i < 297:
                continue
            output_line = {}

            print('*****')
            print("{}st data".format(i+1))

```

```

x, y = data
x = "Q: " + x[0] + "\n" + "A:"
y = y[0].strip()

# print(x, y)

output_line["question"] = x
output_line["gold_ans"] = y

if args.method == "zero_shot":
    x = x + " " + args.direct_answer_trigger_for_zeroshot
elif args.method == "zero_shot_cot":
    x = x + " " + args.cot_trigger
elif args.method == "few_shot":
    x = demo + x
elif args.method == "few_shot_cot":
    x = demo + x
elif args.method == "auto_cot":
    x = demo + x + " " + args.cot_trigger
else:
    raise ValueError("method is not properly defined ...")

# Answer experiment by generating text ...
max_length = args.max_length_cot if "cot" in args.method else args.max_length_direct
z = decoder.decode(args, x, max_length)

output_line["rationale"] = z

# Answer extraction for zero-shot-cot ...
if args.method == "zero_shot_cot":
    z2 = x + z + " " + args.direct_answer_trigger_for_zeroshot_cot
    max_length = args.max_length_direct
    pred = decoder.decode(args, z2, max_length)
    print(z2 + pred)
else:
    pred = z
    print(x + pred)

# Cleansing of predicted answer ...
pred = answer_cleansing(args, pred)

```

```

output_line["pred_ans"] = pred
output_line["wrap_que"] = x

output_json = json.dumps(output_line)
wp.write(output_json + '\n')

# Choose the most frequent answer from the list ...
print("pred : {}".format(pred))
print("GT : " + y)
print('*****')

# Checking answer ...
correct = (np.array([pred]) == np.array([y])).sum().item()
correct_list.append(correct)
total += 1 #np.array([y]).size(0)

if (args.limit_dataset_size != 0) and ((i+1) >= args.limit_dataset_size):
    break
#raise ValueError("Stop !!")

# Calculate accuracy ...
accuracy = (sum(correct_list) * 1.0 / total) * 100
print("accuracy : {}".format(accuracy))

```

调试数据以准确使用

```

from datasets import load_dataset
import re

def extract_numeric_answer(answer_str):
    matches = re.findall(r'\d+', answer_str)
    return matches[-1] if matches else None

dataset = load_dataset(args.dataset, "main", split="test")
dataloader = torch.utils.data.DataLoader(dataset,
    shuffle=True,
    batch_size=args.minibatch_size,
    drop_last=False,
    pin_memory=True)

for i, data in enumerate(dataloader):
    if i==0:
        x, y = data['question'][0], extract_numeric_answer(data['answer'][0])
        x = "Q: " + x + "\n" + "A:"
        y = y.strip()
        print(x,y)

```

测试模型，查看准确率

```

def experiment(args, decoder):
    print('*****')
    print(args)
    print('*****')
    fix_seed(args.random_seed)
    print("setup data loader ...")
    dataloader = setup_data_loader(args)
    if args.method == "few_shot":
        demo = create_demo_text(args, cot_flag=False)
    elif args.method == "few_shot_cot" or args.method == "auto_cot":
        demo = create_demo_text(args, cot_flag=True)
    else:
        pass

    total = 0
    correct_list = []
    with open(args.output_dir, "a") as wp:
        for i, data in enumerate(dataloader):
            if i < args.resume_id - 1:
                continue
            output_line = {}
            print('*****')
            print("{}st data".format(i+1))
            x, y = data['question'][0], extract_numeric_answer(data['answer'][0])
            x = "Q: " + x + "\n" + "A:"
            y = y.strip()
            # print(x, y)

            output_line["question"] = x
            output_line["gold_ans"] = y

```

得到最终结果

```

pred_after : 1
pred : 1
GT : 4
*****
accuracy : 13.861386138613863

```

```

if args.method == "zero_shot":
    x = x + " " + args.direct_answer_trigger_for_zeroshot
elif args.method == "zero_shot_cot":
    x = x + " " + args.cot_trigger
elif args.method == "few_shot":
    x = demo + x
elif args.method == "few_shot_cot":
    x = demo + x
elif args.method == "auto_cot":
    x = demo + x + " " + args.cot_trigger
else:
    raise ValueError("method is not properly defined ...")

# Answer experiment by generating text ...
max_length = args.max_length_cot if "cot" in args.method else args.max_length_direct
z = decoder.decode( x, max_length)

output_line["rationale"] = z

# Answer extraction for zero-shot-cot ...
if args.method == "zero_shot_cot":
    z2 = x + z + " " + args.direct_answer_trigger_for_zeroshot_cot
    max_length = args.max_length_direct
    pred = decoder.decode(z2, max_length)
    print(z2 + pred)
else:
    pred = z
    print(x + pred)

# Clensing of predicted answer ...
pred = answer_cleansing(args, pred)

```



```

output_line["pred_ans"] = pred
output_line["wrap_que"] = x

output_json = json.dumps(output_line)
wp.write(output_json + '\n')

# Choose the most frequent answer from the list ...
print("pred : {}".format(pred))
print("GT : " + y)
print('*****')

# Checking answer ...
correct = (np.array([pred]) == np.array([y])).sum().item()
correct_list.append(correct)
total += 1 # np.array([y]).size(0)
if i >= 100:
    break

if (args.limit_dataset_size != 0) and ((i+1) >= args.limit_dataset_size):
    break
    #raise ValueError("Stop !!")

# Calculate accuracy ...
accuracy = (sum(correct_list) * 1.0 / total) * 100
print("accuracy : {}".format(accuracy))

```

由于算力限制，GSM8K 总共有几万个测试个例，我选择了随机的 100 个测试准确率（因为 huggingface 的 dataloader 是随机的），最终取得了 14 左右的准确率，因为没有官方的结果，所以不敢准确说提高了多少，但相较于没有使用 auto-cot 我测过一遍，可能接近 0 的准确率相比确实是一个重大突破。

附过程图：

```

Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5
Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made $69 from mowing lawns. He spent $24 on new mower blades. That means he has $45 left. Each game costs $5, so he could buy 9 games. The answer is 9.

Q: Luni baked 55 cookies. She ate 5 five cookies and placed the rest equally into five jars. How many cookies were in each jar?
A: Let's think step by step. Luni baked 55 cookies and ate 5, so she has 55 - 5 = 50 cookies left. She divided the remaining cookies equally into five jars, so each jar contains 50 / 5 = 10 cookies. The answer is 10.
pred_before : Q: Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many pictures were in each album?
A: Let's think step by step. First, we know that Wendy uploaded 45 pictures in total. Second, we know that Wendy put 27 pictures into one album. That means that Wendy put the remaining 18 pictures into 9 different a
Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23. If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of 8 + 23 = 31 pieces of candy. If they eat 8 pieces the first night, they have 31 - 8 = 23 pieces left. The answer is 23.
Q: Bianca was organizing her book case making sure each of the shelves had exactly 8 books on it. If she had 5 shelves of mystery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 8 books. So that's 40 mystery books. There are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. 40 + 32
Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cages did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so that means there are 6 cages with 8 puppies in each. The answer is 6.

Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he have?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 customers left. If he got another 39 customers, that means he would have 50 customers in total. 11
Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 points, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 members that did show up. Each member that showed up scored 4 points. So if 5 members each sco
Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5
Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made $69 from mowing lawns. He spent $24 on new mower blades. That means he has $45 left. Each game costs $5, so he could buy 9 games. The answer is 9.

Q: Luni baked 55 cookies. She ate 5 five cookies and placed the rest equally into five jars. How many cookies were in each jar?
A: Let's think step by step. Luni baked 55 cookies and ate 5, so she has 55 - 5 = 50 cookies left. She divided the remaining cookies equally into five jars, so each jar contains 50 / 5 = 10 cookies. The answer is 10.
pred_after : 10
GT : 10
*****

Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5
Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made $69 from mowing lawns. He spent $24 on new mower blades. That means he has $45 left. Each game costs $5, so he could buy 9 games. The answer is 9.

Q: Damien created a currency based on bottle caps and got his friends to take part. He finds 10 bottle caps a day on his way home and each bottle cap is worth $.25. How much money does he make in a 30 day month?
A: Let's think step by step. He finds 10 bottle caps every day, so he finds 10 * 30 = 300 bottle caps in 30 days. Each bottle cap is worth $.25, so he gets 300 * $.25 = $75 in 30 days. The answer is 75.
pred_before : Q: Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many pictures were in each album?
A: Let's think step by step. First, we know that Wendy uploaded 45 pictures in total. Second, we know that Wendy put 27 pictures into one album. That means that Wendy put the remaining 18 pictures into 9 different a
Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23. If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of 8 + 23 = 31 pieces of candy. If they eat 8 pieces the first night, they have 31 - 8 = 23 pieces left. The answer is 23.
Q: Bianca was organizing her book case making sure each of the shelves had exactly 8 books on it. If she had 5 shelves of mystery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 8 books. So that's 40 mystery books. There are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. 40 + 32
Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cages did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so that means there are 6 cages with 8 puppies in each. The answer is 6.

Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he have?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 customers left. If he got another 39 customers, that means he would have 50 customers in total. 11
Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 points, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 members that did show up. Each member that showed up scored 4 points. So if 5 members each sco
Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5
Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made $69 from mowing lawns. He spent $24 on new mower blades. That means he has $45 left. Each game costs $5, so he could buy 9 games. The answer is 9.

Q: Damien created a currency based on bottle caps and got his friends to take part. He finds 10 bottle caps a day on his way home and each bottle cap is worth $.25. How much money does he make in a 30 day month?
A: Let's think step by step. He finds 10 bottle caps every day, so he finds 10 * 30 = 300 bottle caps in 30 days. Each bottle cap is worth $.25, so he gets 300 * $.25 = $75 in 30 days. The answer is 75.
pred_after : 75
GT : 75
*****

```

Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5 problems left to do for homework.

Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made \$69 from mowing lawns. He spent \$24 on new mower blades. That means he has \$45 left. Each game costs \$5, so he could buy 9 games. The answer is 9.

Q: Castle bought 3 boxes of Coco Crunch and 5 boxes of Fruit Loops this week. Last week she bought 4 boxes of cereal. How many more boxes of cereal did she buy this week than last week?
A: Let's think step by step. This week, Castle bought $3 + 5 = 8$ boxes of cereal. Last week, she bought 4 boxes. Therefore, this week she bought $8 - 4 = 4$ more boxes of cereal than last week. The answer is 4.

pred: before : 0
Q: Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many pictures were in each album?
A: Let's think step by step. First, we know that Wendy uploaded 45 pictures in total. Second, we know that Wendy put 27 pictures into one album. That means that Wendy put the remaining 18 pictures into 9 different albums.

Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23. If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of $8 + 23 = 31$ pieces of candy. If they eat 8 pieces the first night, they have $31 - 8 = 23$ pieces left. The answer is 23.

Q: Bianca was organizing her book case making sure each of the shelves had exactly 8 books on it. If she had 5 shelves of mystery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 8 books. So that's 40 mystery books. There are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. $40 + 32 = 72$ books total.

Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cages did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so that means there are 6 cages with 8 puppies in each. The answer is 6.

Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he have?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 customers left. If he got another 39 customers, that means he would have 50 customers in total. The answer is 50.

Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 points, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 members that did show up. Each member that showed up scored 4 points. So if 5 members each scored 4 points, that means the team scored 20 points total.

Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5 problems left to do for homework.

Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made \$69 from mowing lawns. He spent \$24 on new mower blades. That means he has \$45 left. Each game costs \$5, so he could buy 9 games. The answer is 9.

Q: Castle bought 3 boxes of Coco Crunch and 5 boxes of Fruit Loops this week. Last week she bought 4 boxes of cereal. How many more boxes of cereal did she buy this week than last week?
A: Let's think step by step. This week, Castle bought $3 + 5 = 8$ boxes of cereal. Last week, she bought 4 boxes. Therefore, this week she bought $8 - 4 = 4$ more boxes of cereal than last week. The answer is 4.

pred : 4
QT : 4

Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23. If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of $8 + 23 = 31$ pieces of candy. If they eat 8 pieces the first night, they have $31 - 8 = 23$ pieces left. The answer is 23.

Q: Bianca was organizing her book case making sure each of the shelves had exactly 8 books on it. If she had 5 shelves of mystery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 8 books. So that's 40 mystery books. There are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. $40 + 32 = 72$ books total.

Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cages did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so that means there are 6 cages with 8 puppies in each. The answer is 6.

Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he have?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 customers left. If he got another 39 customers, that means he would have 50 customers in total. The answer is 50.

Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 points, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 members that did show up. Each member that showed up scored 4 points. So if 5 members each scored 4 points, that means the team scored 20 points total.

Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5 problems left to do for homework.

Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made \$69 from mowing lawns. He spent \$24 on new mower blades. That means he has \$45 left. Each game costs \$5, so he could buy 9 games. The answer is 9.

Q: A raspberry bush has 6 clusters of 20 fruit each and 67 individual fruit scattered across the bush. How many raspberries are there total?
A: Let's think step by step. There are 6 clusters of 20 fruit each, for a total of $6 * 20 = 120$ fruit. There are 67 individual fruit scattered across the bush, so there are $67 + 120 = 187$ total fruit. The answer is 187.

pred : 187
QT : 187

Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23. If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of $8 + 23 = 31$ pieces of candy. If they eat 8 pieces the first night, they have $31 - 8 = 23$ pieces left. The answer is 23.

Q: Bianca was organizing her book case making sure each of the shelves had exactly 8 books on it. If she had 5 shelves of mystery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 8 books. So that's 40 mystery books. There are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. $40 + 32 = 72$ books total.

Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cages did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so that means there are 6 cages with 8 puppies in each. The answer is 6.

Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he have?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 customers left. If he got another 39 customers, that means he would have 50 customers in total. The answer is 50.

Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 points, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 members that did show up. Each member that showed up scored 4 points. So if 5 members each scored 4 points, that means the team scored 20 points total.

Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many problems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29 problems for homework. If she finished 24 of the problems at school, that means she had 5 problems left to do for homework.

Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games could he buy with the money he had left?
A: Let's think step by step. Mike made \$69 from mowing lawns. He spent \$24 on new mower blades. That means he has \$45 left. Each game costs \$5, so he could buy 9 games. The answer is 9.

Q: Colorado City uses 40% of the water from the Colorado River. If 80% of that water is used for industrial purposes, what percent of the river's total water is used by the city for non-industrial purposes?
A: Let's think step by step. First find how much of the river's water is used for industrial purposes: $80\% * 40\% = 32\%$. Then find how much of the river's water is used for non-industrial purposes: $100\% - 32\% = 68\%$. The answer is 68.

pred : 68
QT : 68

59out data

余下省略