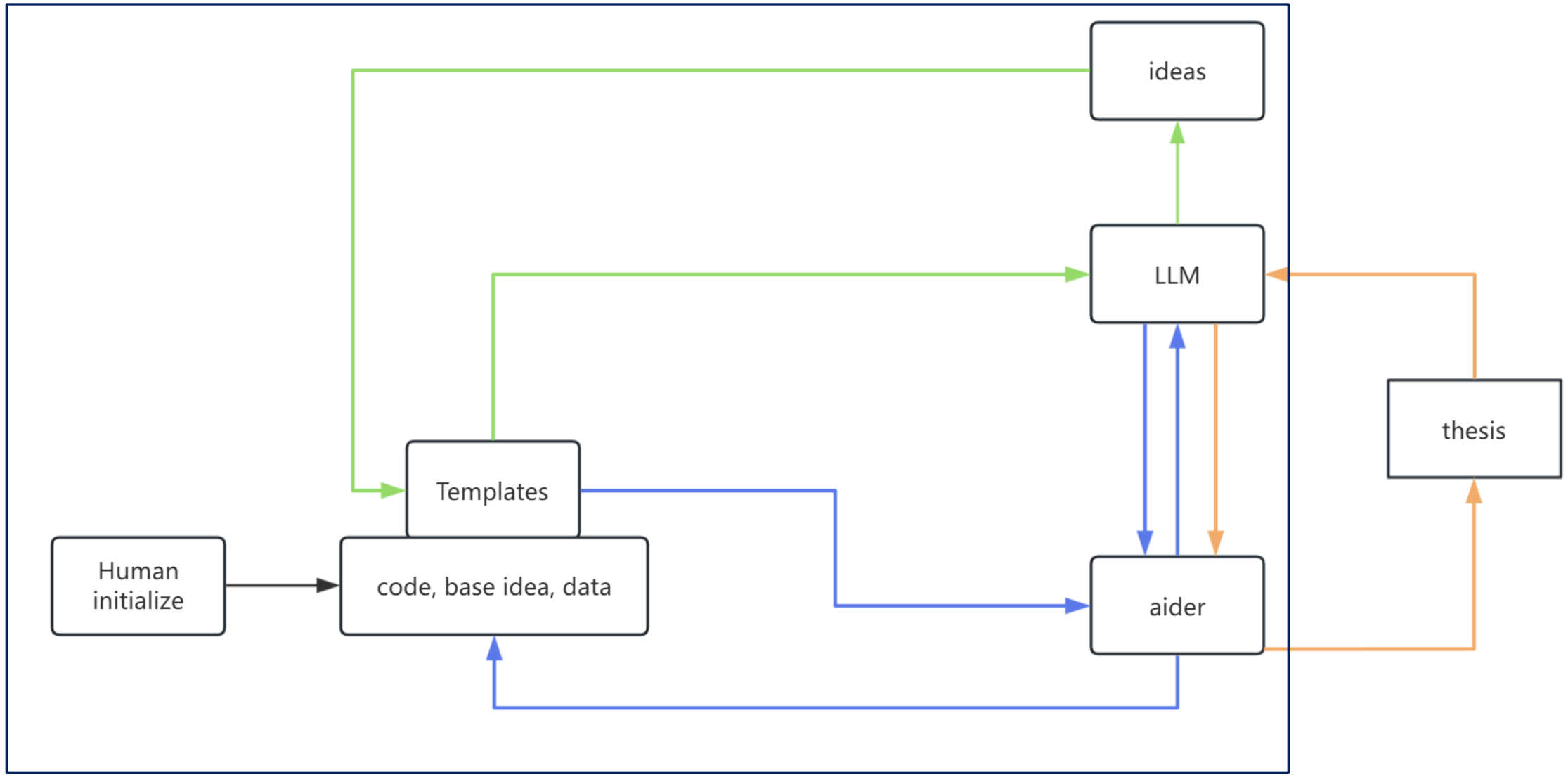
# 实验报告

罗培骜 2021200235

# 整体介绍

- 自动想法生成、想法查新、自动实验、实验结果深入分析、文章撰写、文章质量评判

- 实际实现：
- 自动想法生成、想法查新、自动实验、实验结果分析
- api为chatgpt-4o
- 整个自动化实验的功能完整地实现，但是效果很差，以下主要对效果进行报告


ideas
LLM
thesis
Templates
Human initialize
code, base idea, data
aider

# 实验内容——TEMPLATES

- CNN daily mail summarization
  - templates/transformer 路径下 experiment.py
  - cnn新闻数据集的summarization任务，用调用现成transformer模型结构实现（有自行搭建，但是结构的复杂程度让大模型没有办法对其进行有效更改，故采用已有结构简化模型结构部分）
- Macro prediction
  - templates/macro_pred 路径下 experiment.py
  - 出口总额预测任务，采用简单的encoder only transformer结构，decoder用线性层实现

# 实验内容——IDEA GENERATION

单个idea示例


```
"Name": "dynamic_batch_size",
"Title": "Dynamic Batch Size Adjustment: Optimizing Memory Usage During Training",
"Experiment": "Implement a dynamic batch size adjustment mechanism that monitors GPU memory usage dur
"Interestingness": 7,
"Feasibility": 5,
"Novelty": 6,
"ISNOVEL": false,
"Reference": [
    "Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu. (2020). Dynamically Adjusting Transforme
    "Qing Ye, Yuhao Zhou, Mingjia Shi, Yanan Sun, Jiancheng Lv. (2020). DBS: Dynamic Batch Size For D
]
```


- summarization：
  templates/transformer 路径下 ideas.json
  均是已有的非新颖且简单的idea（比如dynamic batchsize等），均通不过novelty检验
- macro pred：
  templates/macro_pred 路径下 ideas.json
  基本新颖

- 更迭template以及template对应的领域发现：

参考资料越少的领域，会给出越新颖的idea；template的灵活程度会影响大模型给出的想法（直接用pretrained model或者是用pytorch实现一个类，大模型的想法会因为模型结构具体可调而给出更具体的针对模型结构的想法，因为模型是预训练就避开改动模型结构的想法）

# 实验内容——IDEA NOVELTY

单个idea示例


```
"Name": "dynamic_batch_size",
"Title": "Dynamic Batch Size Adjustment: Optimizing Memory Usage During Training",
"Experiment": "Implement a dynamic batch size adjustment mechanism that monitors GPU memory usage dur
"Interestingness": 7,
"Feasibility": 5,
"Novelty": 6,
"ISNOVEL": false,
"Reference": [
    "Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu. (2020). Dynamically Adjusting Transforme
    "Qing Ye, Yuhao Zhou, Mingjia Shi, Yanan Sun, Jiancheng Lv. (2020). DBS: Dynamic Batch Size For D
]
```


大模型对novelty的初判断几乎无效，基本全部为novelty打分到5分以上（10分制）

且该分数与提供了文献数据库查询结果后的novelty判断也难以理解，参考文献存在的多少和年限对novelty判定好像没有什么影响；偶尔，即便没有搜索到相关文献，大模型仍然判定为not novel

实验idea与对应任务比对可以发现，idea基本上完全由experiment.py里面采用了什么模型决定，而没有很好参考实验描述，对于两个不同任务（都采用transformer），给出了很类同的想法，关注点在于transformer模型而非具体的时间序列预测、或者是summarization任务

# 实验内容——EXPERIMENT

- 对于单个想法，代码实现困难
- 对于一个想法的一轮尝试，平均要迭代三次才能成功运行出结果
  报错的原因常是：类型错误、输出输出形状错误、无中生有调用方法
  对于类型错误，反复的迭代仍然不能使其成功运行，会不断地修改无关部分，导致实验失败
- 对于轮次之间的迭代，没有太多参考前次迭代的结果来进行更新，即便实验结果不理想，也会常常再第二三轮提前结束该次想法实现

# 实验内容——EXPERIMENT & RESULT

```
# Title: Enhancing Transformer Model Efficiency through Attention-Based Optimization
# Experiment description: Modify the TEncoder class to enable attention-based pruning or re-weighting of model com

## Run 0: Baseline
Results: ['next period prediction: 21.99 ; fit direction ratio: 0.8889 ; test direction ratio: 0.5']
Description: The baseline run was conducted to establish a reference point for evaluating the impact of subsequent

## Run 1: Attention-Based Pruning
Results: ['next period prediction: 20.97 ; fit direction ratio: 0.9000 ; test direction ratio: 0.5']
Description: In this run, attention-based pruning was implemented in the TEncoder class. The goal was to enhance m

## Analysis of Runs
The initial runs of this experiment provide valuable insights into the potential for optimizing Transformer models
```

idea的实验结果示例

最后的整体分析中，对于效果不理想，且提前结束的原因，的全部笼统的归于模型不适合或者想法不合适，没有结合其修改的代码或者具体任务分析；此外，对于尝试了多轮的实验，总结时经常会丢失很多轮次的结果分析，只基于片段

经常出现的情况是大模型对代码进行了有效的修改，但disable了它，或者是修改了模型结构需要新的输出输入，但再train函数中没有改动，导致结果和原来完全一致

# 总结

- 整个自动化实验功能是完整可实现的
- 但是其效果至少在采用本实验的template，prompt，chatgpt-4o为AI Scientist的情况下较差。受制于成本原因，没有对实验进行大批量的测试，但在已有的10次左右运行的结果是：
  - 实验的运行成功率很低，需要多次迭代修改；
  - 实验很可能因为多次修改仍然运行失败直接失败没有结果；
  - 运行成功的实验倾向于在实验结果没有显著变化的情况下提前结束，且没有给出太多有意义的结论，且会丢失过程中的很多次迭代的结果的分析，最后的总体分析只是基于片面的实验片段。
- 与AI Scientist项目的最大差异在于，template的任务以及其中的experiment代码，推测表现差异也一部分来源任务本身的选择，以及experiment代码的构架与大模型进行交互的合适性