**实验一：重复性实验，考察模型多轮训练后是否具有记忆能力**

全部仅训练标准v2\_add\_change数据

* 使用v2\_add\_change的QA数据\_layers8\_modulesall至0.6左右，提升10%~20%左右。（0728实验）
* 使用400篇全文数据进行训练，形式为Q：能否背出xxx文章，A：全文内容。记忆能力差。
* 使用lora微调标准数据，看多轮后是否能够记住

Rank=8时，800steps （约4epochs）记不住训练集 cmmlu 0.762，ceval 0.7619；1600steps（约8epochs）记不太住训练集 cmmlu 0.7553 ceval 0.7619。

Rank=32时，800steps（约4epochs）记不太住，只有一两个问题能背出来，挺多不行的cmmlu 0.7687，ceval 0.781；1600steps（约8epochs）也记不住 cmmlu 0.7229，ceval 0.7619。

有一个化学物品的QA很容易被记住，分析可能是和开源模型本身的知识/分布接近

* 冻结部分模型参数

freeze\_trainable\_layers=9,freeze\_trainable\_modules=all ，800steps （约4epochs）有一定记忆能力 cmmlu 0.7229，ceval 0.7524；1600 steps（约8epochs），cmmlu 0.7363，ceval 0.7524

冻结embedding

多轮迭代后对于短文本具有一定的记忆能力，对长文本记忆能力较差。

Lora训练通用能力掉的慢，但记忆能力也会差很多

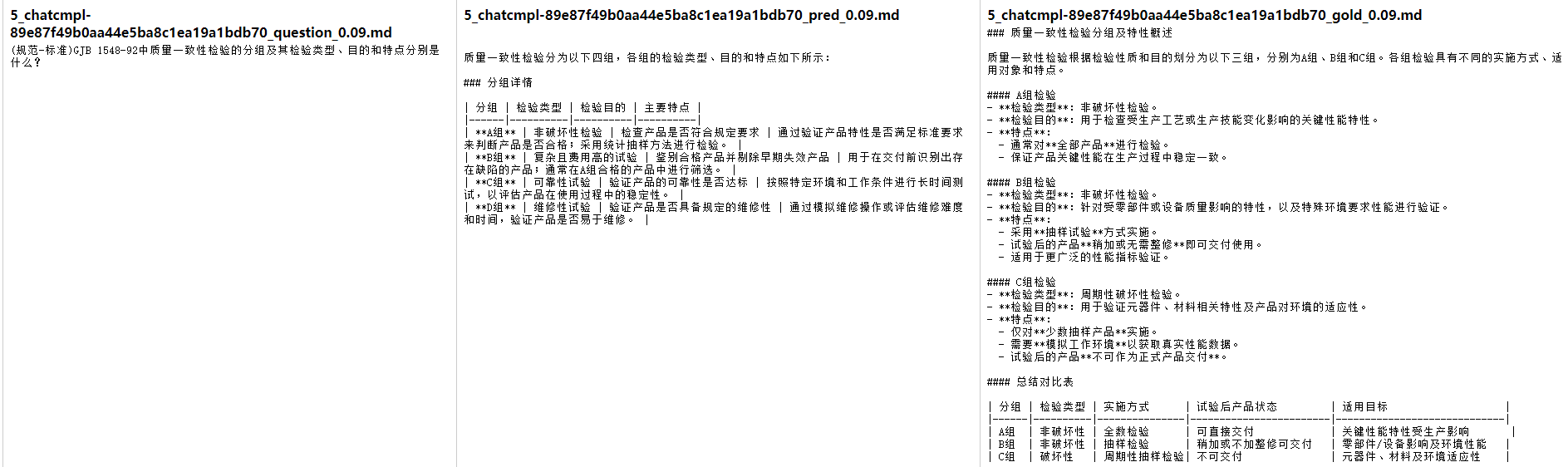
0728实验结果分析：

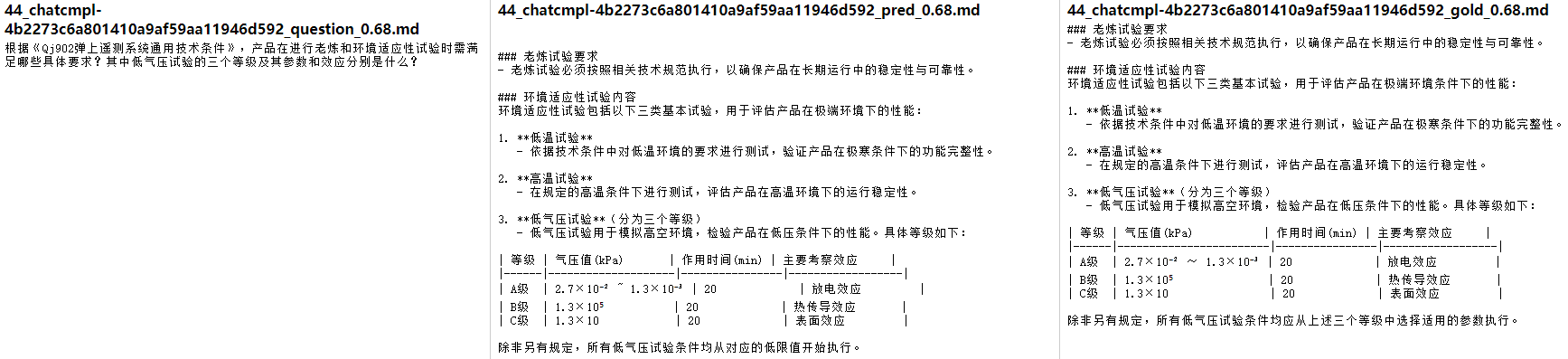
QA质量检验（数据本身以及minHash），prompt修改

（1）chunk分块过程未考虑语义完整性与全文背景，导致对于同一片文章生成了不同的QA，一个关注全文，一个关注局部，最后评测指标不佳



1. prompt的评价指标有待修改，部分完全一样的预测结果评分只有0.68，而且都是0.68，说明目前的prompt在某一固定方面的能力是缺失的。

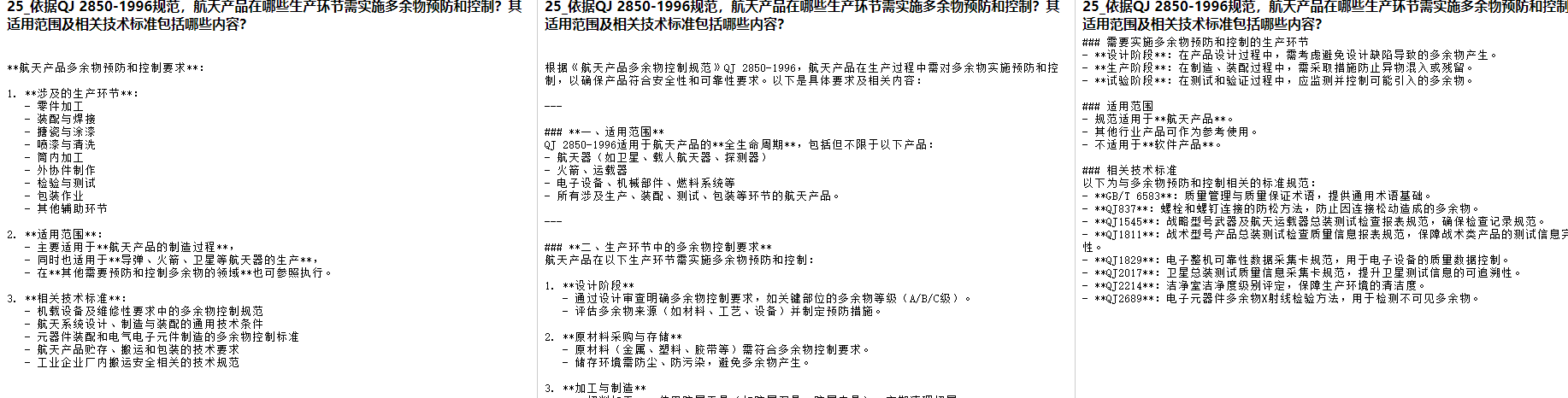




（3）Prediction\_our的结果和GJB 101A-97无关，但和((规范-标准))GJB 150.？A-2009军用装备实验室环境试验方法有关，详细实验项目的内容是军用装备实验环境实验方法中的多个子部分。



1. 输出格式是否需要套到Qwen3的格式上，比如输出的三级目录，加粗加黑等
2. 是否还要加“引用了xxx规范”这样难以记忆的点



（5）think模式是否需要

* 大模型善于记忆point还是较短paragraph的描述

思考：模型对于事实性关键点（比如北京是中国的首都）会有很强的记忆能力，因为这类信息在训练数据中重复率高，表述一致。当关键点以简洁形式在上下文中反复出现，记忆效果最佳。关键点的记忆也符合attention机制的原理，关键点会在向量空间中形成概念聚类。

对于描述性内容，模型更可能记忆核心观点而非具体表述细节，当询问道细节描述时，模型倾向于根据学到的模式重组内容，很有可能会偏离原文。

**实验二：开源大模型对与规章制度是否具有一定能力**

**实验三：开源大模型的通用知识如何保持，主要保持cmmlu和ceval指标**

背景：

Cmmlu是一个综合性的中文评估基准，用于评估语言模型在中文语境下的知识和推理能力，包括67个主题。67个主题包括开发集和测试集，开发集中包含5道问题，测试集中包含100+道问题。

Ceval

目前evalscope中的评测不具有随机性，评测cmmlu和ceval的1000道题目是固定的（鹏钰提供脚本）

Chinese-Qwen3-235B-Thinking-2507-Distill-data-110k-SFT和Chinese-DeepSeek-R1-Distill-data-110k的数据在顺序上是一一对应的

* Qwen3-4B cmmlu 0.7799,ceval 0.7619

测评保存路径：/mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250801\_082853

测试了一下4B在cmmlu，ceval评测集上的输出ouput长度分布：{'1000': 768, '2000': 99, '3000': 43, '4000': 22, '5000': 19, '6000': 11, '7000': 6, '8000': 6, '9000': 2, '10000': 2, '20000': 4, '30000': 18, '40000': 0}

* 使用Chinese-Qwen3-235B-Thinking-2507-Distill-data-110k-SFT下随机sample的10000条数据进行训练

200steps（约1epoch），cmmlu 0.7341, ceval 0.714

测评结果保存路径：/mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250801\_073149

训练数据Output长度分布{'1000': 2372, '2000': 2594, '3000': 2767, '4000': 1453, '5000': 409, '6000': 162, '7000': 66, '8000': 50, '9000': 42, '10000': 14, '20000': 71, '30000': 0, '40000': 0}

* 使用Chinese-DeepSeek-R1-Distill-data-110k下随机 'repo\_limit': {'stem\_zh/chem': 3157, 'EduChat-Math': 3000, 'meta-math/GSM8K\_zh': 2000, 'coig/neo': 1000, 'Haijian/Advanced-Math': 570}的数据分布

200steps（约1.5epoch）cmmlu 0.7721,ceval0.8

测评结果保存目录：/mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250801\_095629

训练数据Ouput长度分布{'1000': 2781, '2000': 5053, '3000': 906, '4000': 392, '5000': 232, '6000': 137, '7000': 67, '8000': 59, '9000': 33, '10000': 21, '20000': 46, '30000': 0, '40000': 0}

Gsm8k:0.934, math\_500:0.94,aime24:0.5667,aime25:0.4667

测评结果保存路径：/mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250801\_114858

Live\_code:0.4615,human\_eval:0.9573(高到离谱。。。)

测评结果保存路径：/mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250801\_132643

* 使用Chinese-DeepSeek-R1-Distill-data-110k全量数据

1700steps（约1epoch）cmmlu 0.7475,ceval 0.781，gsm8k 0.9287，math\_500 0.908, aime24 0.6333, aime25 0.3667,live\_code0.3736 ,human\_eval 0.7683,if\_eval 0.8588 0.8176 0.8071 0.7586,mmlu\_pro 0.6816,mmlu\_redux 0.8188

测评结果保存目录：/mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250802\_021358

mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250802\_024503

mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250802\_032657

mnt/data/Llmei/data/data\_process\_shuyz/evalscope/outputs/20250802\_064025

3400epoch（约2epoch）还没测

目前随机性的调配数据没有方向且没有意义，应该针对cmmlu,ceval指标进行针对性提高

露姐建议：

（1）分析任务本身，看是否有针对性的数据集可以提高

（2）若一部分评测的能力提升了，这部分数据可以固化，再调整别的比例

（3）Ame的数据要关注长度（tokenizer以后的，不是简单的字符计算），不管是数学还是代码，鹏程实验室有文章认为 16k-32k的长度更为合适

下一步工作内容：

1. 测一下code
2. 等比例放大，把数据量扩大
3. multi-turn的数据
4. Function call的数据（要蒸数据）
5. 评测加MT bench， BFCL v3

与鹏钰讨论：

1. less is more，可能简单的数据更适合刷榜msk8k,cmmlu榜单
2. 二轮训练，先训难的知识点，再用简单知识收口
3. 思考Ame数据集为什么效果提不上去

领域内数据筛选标准：

不与领域强相关的问题删去（比如电力xxxx这种比较通用的问题）

引用（比如参加xxx章节，xxx规范）删去

实验四：如何尽可能覆盖chunk所有知识点的QA