大模型面经之llama3训练如何保证数据质量

瑶光 瓦力算法学研所 2024年10月09日 19:33 安徽

技术专栏

本篇将介绍llama3模型训练的数据质量控制方法

1、训练数据清洗

- **安全性过滤**：对训练数据进行筛选，排除包含个人信息、有害内容和成人内容的文本。
- **文本清洗**：使用HTML解析器提取文本、代码和数学公式，同时去除markdown标签，保留HTML中的alt标签。
- **文本去重**：
  - URL去重：保留每个网页的最新版本URL。
  - Document-level去重：使用全局MinHash算法判定并去除重复文档。
  - Line-level去重：根据每30M文档中出现超过6次的行进行判定和去重。
- **启发式去重**：通过n-gram覆盖比检测重复内容，使用定义的"脏词"过滤成人内容，通过token分布的KL距离检测异常符号。
- **基于模型的低质过滤**：使用多种模型评估文档质量，如Llama2-chat和DistilledRobera。
- **代码和推理数据：**专门定制 HTML parser 从网络文本中抽取出数学推导、理工科里的推理内容以及与文本交织在一起的代码，通过这些数据对 Llama2 进行提示微调，然后使用 Llama2 生成标注数据，交给 DistilledRoberta 从网络文本中分辨出这部分数据。
- **多语言数据。**基于 fasttext 的语言分类模型将所有数据分类成 176 种语言，在每种语言内部执行 document-level 和 line-level 去重，使用每种语言专门的模型和启发式规则过滤低质量样本。

2、不同来源训练数据配比

- 使用知识分类器和规模定律实验来确定不同数据来源在训练集中的占比，最终得到的数据配比如下：
  - 约50%的token与通用知识相关。
  - 25%的数学和推理数据。
  - 17%的代码数据。
  - 8%的多语言数据。

3、后训练数据质量控制

- 后训练阶段使用的训练数据大多是基于已有大模型合成的，因此对数据质量的要求更高。

（1）启发式规则：

- 清洗频繁出现的脏数据，如emoji符号、感叹号和道歉前缀等。

（2）基于模型的方法：

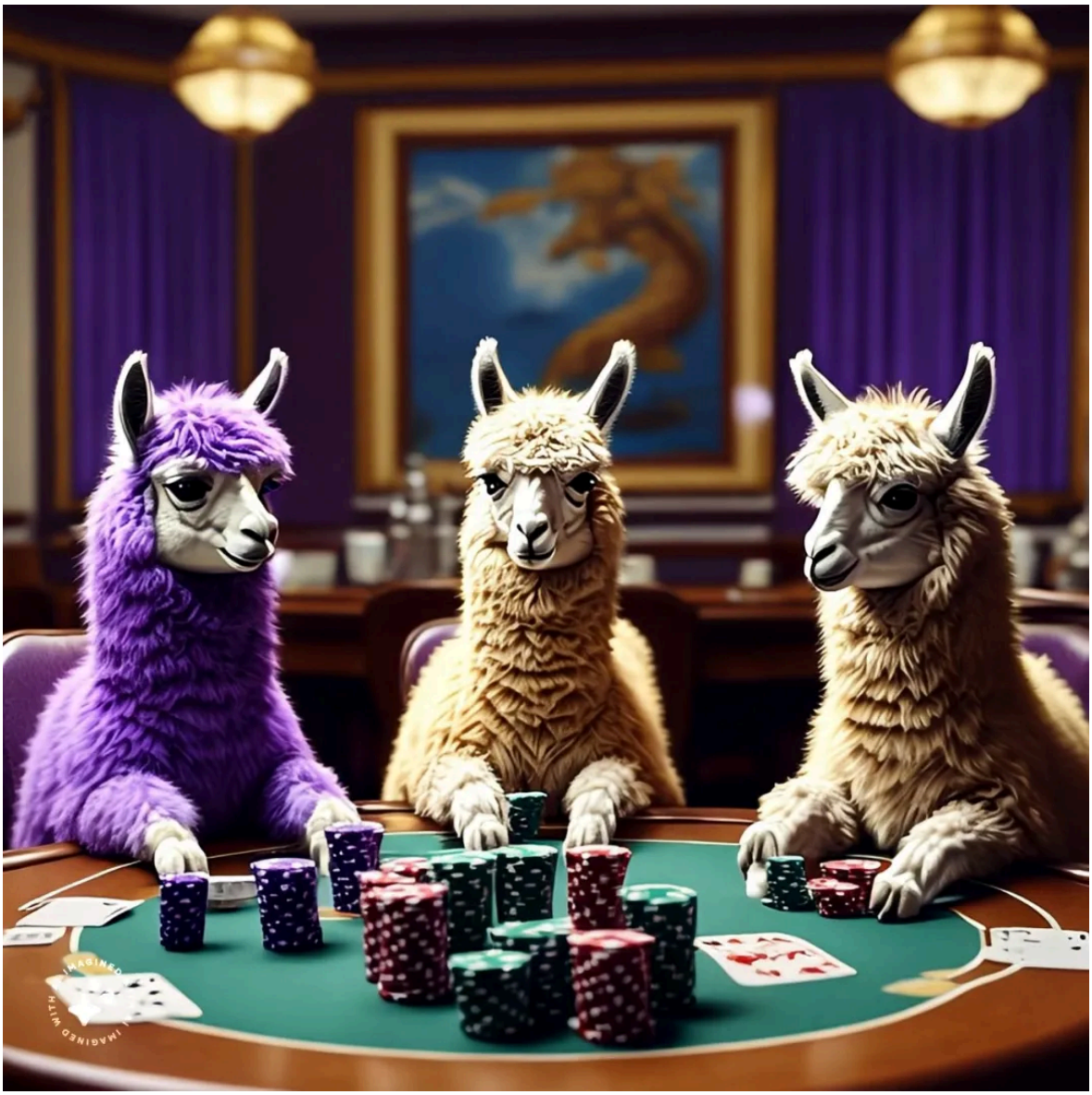
- **主题分类**：使用Llama3 8B微调的主题分类器对数据进行二级主题分类。
- **质量分**：使用奖励模型和Llama-prompt对样本进行质量打分。
- **复杂度**：设计Llama3 70B的prompt，抽取SFT数据中的意图，意图越多说明问题越复杂。
- **语义去重**：使用Roberta对样本聚类，根据质量分和复杂度排序，过滤高度相似的样本。

4、数据生产的新方法

- **执行反馈**：
  - Step 1：随机采样大量代码片段，通过提示让模型总结代码片段中的问题。
  - Step 2：告诉Llama3这些问题、代码以及一些编程语言通用规则，生成这些问题的解决方法。
  - Step 3：抽取解决方法中的源代码，对其进行编译，还可以生成一些单元测试，将编译不过或者单元测试不通过的样本，可以尝试让Llama3来修复，可以修复大约20%的样本。

- **编程语言互译**：
  - 针对冷门语言（如typescript/PHP）数据较少的问题，使用Llama3将热门语言（如Python/C++）翻译成冷门语言的版本，并结合语法分析、编译和跑单测来保证质量。

- **回译**：
  - Step 1：让Llama3为代码生成注释、解释等信息。
  - Step 2：让Llama3根据这些信息生成代码。
  - Step 3：让Llama3判断生成的代码和原始代码是否一致，将不一致的过滤掉来保证质量。



参考资料

https://zhuanlan.zhihu.com/p/712904111

https://link.zhihu.com/?target=https%3A//ai.meta.com/research/publications/the-llama-3-herd-of-models/

想要获取技术资料的同学欢迎关注公众号，进群一起交流~