Автоматическое формирование ответов в чате техподдержки с применением методов выравнивания языковых моделей.

Выполнил: Черных С.Д
Руководитель: Алексейчук А. С.

Обзор Литературы

1. Борис Шапошников, Новые методы алаймента языковых моделей, Тинькофф, 2023
2. Никита Драгунов, Alignment языковых моделей. Prompt engineering & supervised fine-tuning // Practical ML Conf 2023, Яндекс Поиск, 2023
3. Ирина Степанюк, Интенты в саппорте на основе sentence embeddings и при чем тут LLM // Tinkoff.AI NLP Monolog Meetup #2, Тинькофф, 2023
4. Анатолий Потапов, Как собрать свой датасет для предобучения LLM // Tinkoff.AI NLP Monolog Meetup #2, Тинькофф, 2023
5. Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, Deyi Xiong. Large Language Model Alignment: A Survey, College of Intelligence and Computing, Tianjin University, 2023

Обзор известных методов решения поставленной задачи

1. Sentence embeddings (векторных представлений предложений) для определения интентов пользователя в системах поддержки.
2. Prompt engineering: разработка подсказок, которые направляют генерацию текста в желаемом направлении.
3. Fine-tuning: выравнивание модели на небольшом наборе данных с целевой задачей.
4. Supervised fine-tuning: выравнивание модели на наборе данных с метками, указывающими на желаемые характеристики текста.
5. Reinforcement learning: использование обучения с подкреплением для обучения модели генерировать текст, соответствующий желаемым характеристикам.
6. Сбор данных, построение архитектуры модели, обучение модели.

Выравнивание языковых моделей: ключ к безопасному и надежному ИИ

Почему важно решать эту задачу?

1. Снижение нагрузки на операторов чата
2. Повышение скорости и качества обслуживания клиентов
3. Круглосуточная доступность поддержки
4. Снижение затрат на обслуживание

Актуальность и новизна

Актуальность:

1. Высокая нагрузка на операторов чат-поддержки: чат-боты могут взять на себя часть рутинных задач, освободив время операторов для решения более сложных проблем.
2. Низкая скорость и качество обслуживания клиентов: чат-боты могут обеспечить круглосуточную доступность поддержки и более быстрый ответ на запросы клиентов.
3. Высокие затраты на обслуживание: чат-боты могут помочь снизить расходы на обслуживание клиентов.

Применение методов выравнивания языковых моделей:

1. Позволяет повысить точность и релевантность ответов чат-ботов.
2. Снижает риск предвзятости, дискриминации и других деструктивных действий.
3. Позволяет создавать чат-ботов, способных вести более сложные диалоги с клиентами.

Постановка задачи

Разработать модель для автоматического формирования ответов на сообщения пользователей в чате технической поддержки сервиса. Модель будет реализована на языке Python с использованием библиотеки машинного обучения PyTorch, включая использование предобученных моделей. В работе будет использованы методы выравнивания языковых моделей для улучшения качества ответов. Будут построены несколько моделей, основанных на различных методах выравнивания языковых моделей, и их эффективность будет оценена, чтобы сделать выводы о наилучшем подходе к формированию ответов в чате технической поддержки.

Описание алгоритма

Я еще не написал, почему не буду рассматривать тот или иной метод, поэтому и тут ничего не напишу

Планируемые результаты

1. Разработан метод автоматического формирования ответов в чате техподдержки с применением методов выравнивания языковых моделей.
2. Реализована модель, использующая данный метод.
3. Проведена оценка эффективности разработанного метода и модели.
4. Сформулированы рекомендации по внедрению разработанного метода и модели в реальных условиях.

Предполагаемый научный результат

1. Разработан новый метод автоматического формирования ответов в чате техподдержки, основанный на методах выравнивания языковых моделей.
2. Определены ограничения разработанного метода и пути его дальнейшего совершенствования.
3. Сформулированы рекомендации по внедрению разработанного метода в реальных условиях.

Выводы

Будет готовая работа, будет и вывод, пока не буду ничего выдумывать