



Министерство науки и высшего образования Российской Федерации
Федеральное Государственное бюджетное образовательное
учреждение высшего образования
«Московский авиационный институт»
(национальный исследовательский университет)

Институт № 8 «Компьютерные науки и прикладная математика»
Кафедра 805 «Математическая кибернетика»

Выпускная квалификационная работа на тему:

**«Автоматическое формирование ответов в чате
техподдержки с применением методов
выравнивания языковых моделей»**

Студент группы **М80-405Б-20**: Черных Сергей Дмитриевич

Руководитель: к.ф.-м.н., доцент каф. 805

Алексейчук Андрей Сергеевич

Москва 2024



Актуальность и новизна

Актуальность задачи обусловлена следующими проблемами:

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.

Для решения этих проблем предлагается использование языковых моделей, которые позволяют

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.

Обоснование необходимости решения поставленной задачи

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






Постановка задачи

Разработать модель для автоматического формирования ответов на сообщения пользователей в чате технической поддержки сервиса на языке Python, включая использование предобученных моделей.

Использовать методы выравнивания языковых моделей для улучшения качества ответов, для этого построить несколько моделей, основанных на различных методах, и оценить их эффективность.



Обзор известных методов решения поставленной задачи

1. Sentence embeddings (векторных представлений предложений) для определения интенгов (кластер сообщения) пользователя в системах поддержки.
2. Prompt engineering: разработка подсказок, которые направляют генерацию текста в желаемом направлении.
3. Fine-tuning: дообучение модели на небольшом наборе данных с целевой задачей.
4. Supervised fine-tuning: дообучение 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.



Основные этапы построения модели

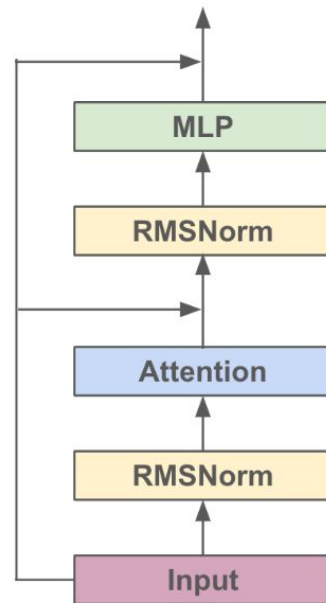
- 1) Выбор стека
- 2) Выбор метода prompt engineering
- 3) Получение результатов на полученных данных
- 4) Оценка результатов



LangChain

Модель для экспериментов

TinyLlama — компактная модель с 1,1 миллиардом параметров. Компактность позволяет обслуживать множество приложений, требующих ограниченного объема вычислений и памяти. Архитектура TinyLlama похожа на LLaма2 (большая языковая модель с 70 миллиардом параметров). К сожалению, доступ к LLaма2 не выдали, поэтому детального обзора и сравнения архитектур предоставить не могу. Модель имеет высокие рейтинги по метрикам на HellaSwag, Obqa, WinoGrande, ARC_c, ARC_e, boolq, piqa, в сравнении с аналогичными моделями с малым количеством параметров.



декодер-блок



Эксперимент 1

- 1) Модель TinyLlama-1.1B
- 2) Фреймворк Pytorch
- 3) Prompt engineering: простая генерация текста с дополнительным сообщением, что необходимо ответить.

Answer user complain: I wanna throw out the pizza you delivered is awful and you should probably eat it yourself.

Результаты:

Hmmm...that sounds like fun! What kind of toppings did they have? And how much was it for a single slice? Or would that be cheaper than buying them individually??

Из-за малого количества ресурсов модель работала очень долго более 5 минут для вычисления ответа. Модель генерировала диалог заказчика и сотрудника службы техподдержки сервиса. (добавить пример)

Вывод:

Эксперимент оказался неудачным, нужны более совершенные методы prompt engineering.



Эксперимент 2

- 1) Модель TinyLlama-1.1B
- 2) Фреймворк Pytorch, Hugging Face
- 3) Prompt engineering: персонализация. Роли разделены на ассистента и пользователя, пользователь принимает роль user, модель в роль assistant.

```
<s>[INST] {{ This is ridiculous! My order has  
been stuck on 'preparing' for over 45 minutes.  
}} [/INST]
```

```
The only thing I can think of is that the  
order is being held up by the server. I'm not  
sure if this is a bug or a feature request,  
but I'm not sure how to fix it
```

Результат:

Улучшилось качество генерации, но значительно упала скорость, на доступных вычислительных мощностях генерация одного ответа модели занимала более 5 минут. (входные выходные данные добавить)

Вывод:

Эксперимент оказался неудачным. Слишком долгая генерация ответа.



Эксперимент 3

- 1) Модель TinyLlama-1.1B
- 2) Фреймворк Hugging Face, Langchain
- 3) Prompt engineering: персонализация и системные сообщения. Для модели подаётся системное сообщение, что она является ассистентом, заказчик ставится в роль user, модель в роль assistant.

```
<s>[INST] <<sys>> Answer user complains  
<</sys>>{{ This is ridiculous! My order has  
been stuck on 'preparing' for over 45 minutes.  
}} [/INST]
```

```
Sure, I'd be happy to help! Here's a possible  
solution: 1. Identify the problem: The first  
step in resolving a customer complaint is to  
identify the problem. In this case, the  
customer is complaining about the quality of  
the product. 2. Collect relevant information:  
Once you've identified the problem, gather as  
much relevant information as possible. This  
could include the customer's name, order  
number, product description, and any relevant  
details about the complaint.
```



Эксперимент 3

Результат:

Улучшилось качество генерации, чистота кода повысилась, простота интеграции в сервисы выросла, но значительно упала скорость, на доступных вычислительных мощностях генерация одного ответа модели занимала около 5 минут.

Вывод:

Эксперимент оказался неудачным. Слишком долгая генерация ответа.



Эксперимент 4

Изменён фреймворк относительно эксперимента 3

- 1) Модель TinyLlama-1.1B
- 2) Фреймворк Langchain, Ollama
- 3) Prompt engineering: персонализация и системные сообщения. Для модели подаётся системное сообщение, что она является ассистентом, заказчик ставится в роль user, модель в роль assistant.

```
<s>[INST] <<sys>> Answer user complains  
<</sys>>{{ This is ridiculous! My order has  
been stuck on 'preparing' for over 45 minutes.  
}} [/INST]
```

```
I understand your frustration. As an  
artificial intelligence (AI) system, it may  
take some time to process and respond to  
certain inputs. However, I would like to  
remind you that this is a temporary state, and  
your order will be prepared shortly. The AI  
assistant is currently processing the request  
for you to await delivery..
```



Эксперимент 4

Результат:

Улучшилось качество генерации, чистота кода повысилась, простота интеграции в сервисы стала наибольшей из всех экспериментов, скорость генерации ответа составила в среднем ~ 13 с на один запрос на доступных вычислительных мощностях

Вывод:

Эксперимент оказался удачным. Лучшая скорость ответа, простота интеграции на наивысшем уровне сравнительно других фреймворков.



Описание метрик

Критерии оценки ответов

0	Модель генерирует диалог между человеком и ассистентом, а не выдаёт ответ.
1	Ответ модели не связан с вопросом.
2	Ответ модели неправильный.
3	Описаны только пути решения проблемы, алгоритм, но ничего не сказано про решение.
4	Проблема пользователя частично решена.
5	Идеальный ответ, проблема пользователя решена.



Перечень системных сообщений

Сообщения сгруппированы разделены на группы, чтобы проверить высказывания связанные с prompt engineering.

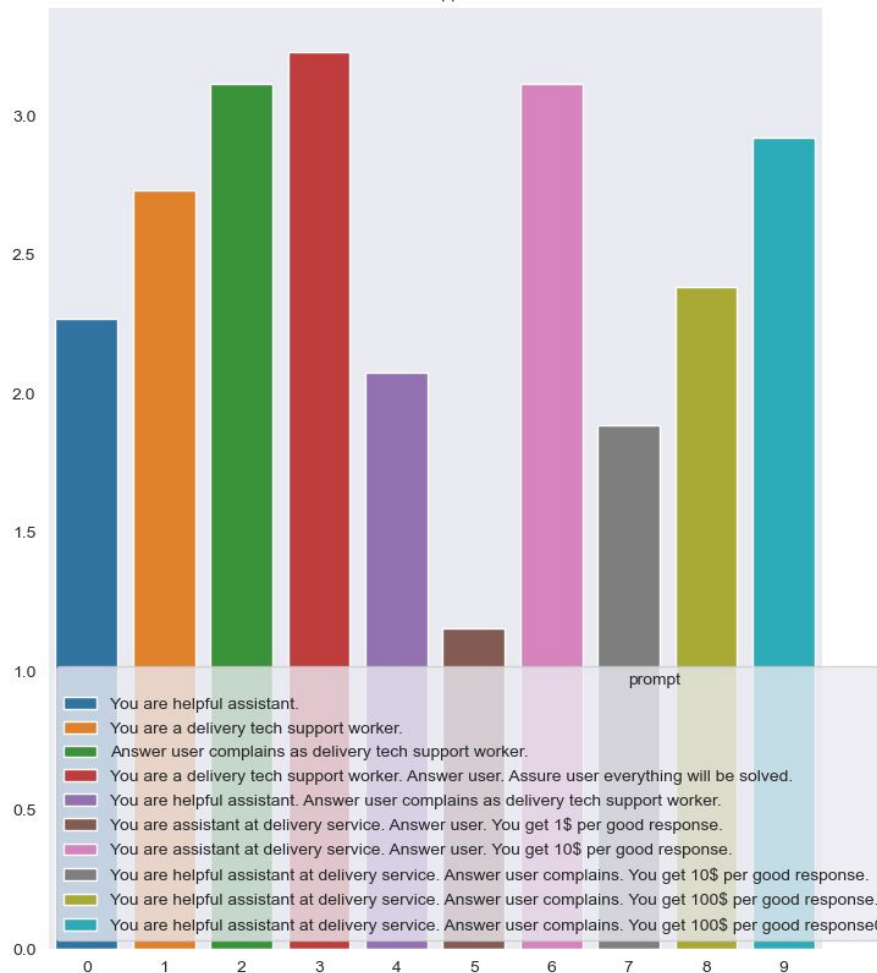
Большее количество контекста, которое получает модель, может улучшить качество ответов модели.

1. "You are helpful assistant."
2. "You are a delivery tech support worker."
3. "Answer user complains as delivery tech support worker."
4. "You are helpful assistant. Answer user complains as delivery tech support worker."
5. "You are a delivery tech support worker. Answer user. Assure user everything will be solved."

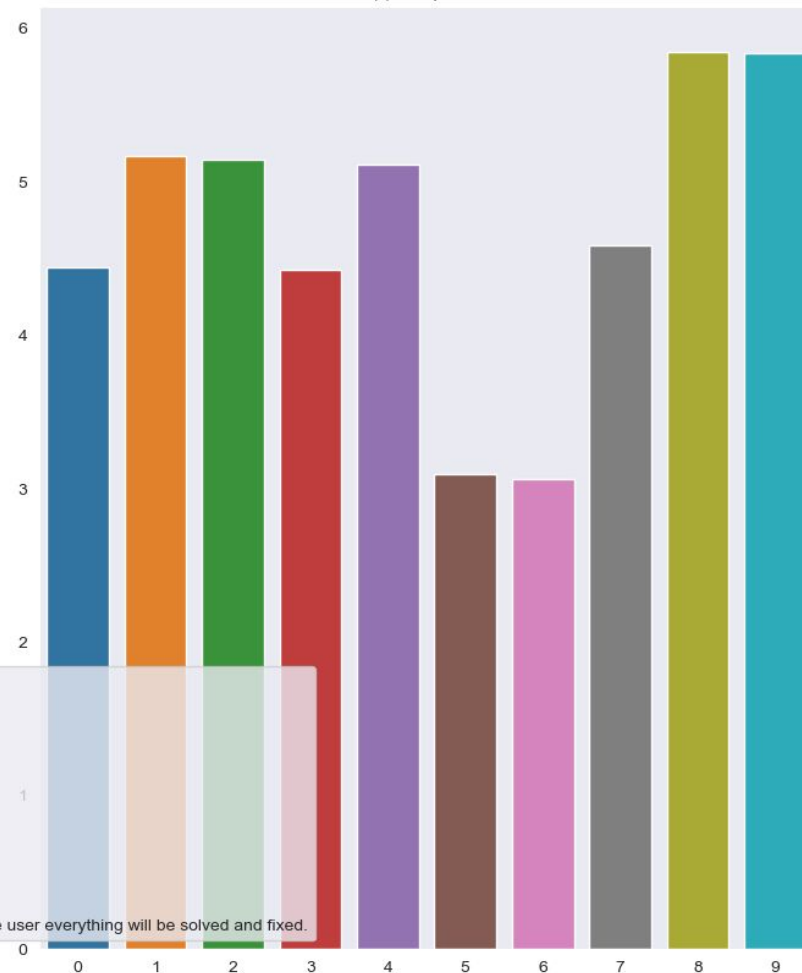
"Our central thesis is that the performance of GPT-4 Turbo improves with increased monetary tips up to the point of convergence." – Адекватное количество награды за ответ, может улучшить качество ответов модели.

6. "You are assistant at delivery service. Answer user. You get 1\$ per good response."
7. "You are assistant at delivery service. Answer user. You get 10\$ per good response."
8. "You are helpful assistant at delivery service. Answer user complains. You get 10\$ per good response."
9. "You are helpful assistant at delivery service. Answer user complains. You get 100\$ per good response."
10. "You are helpful assistant at delivery service. Answer user complains. You get 100\$ per good response. Assure user everything will be solved and fixed"

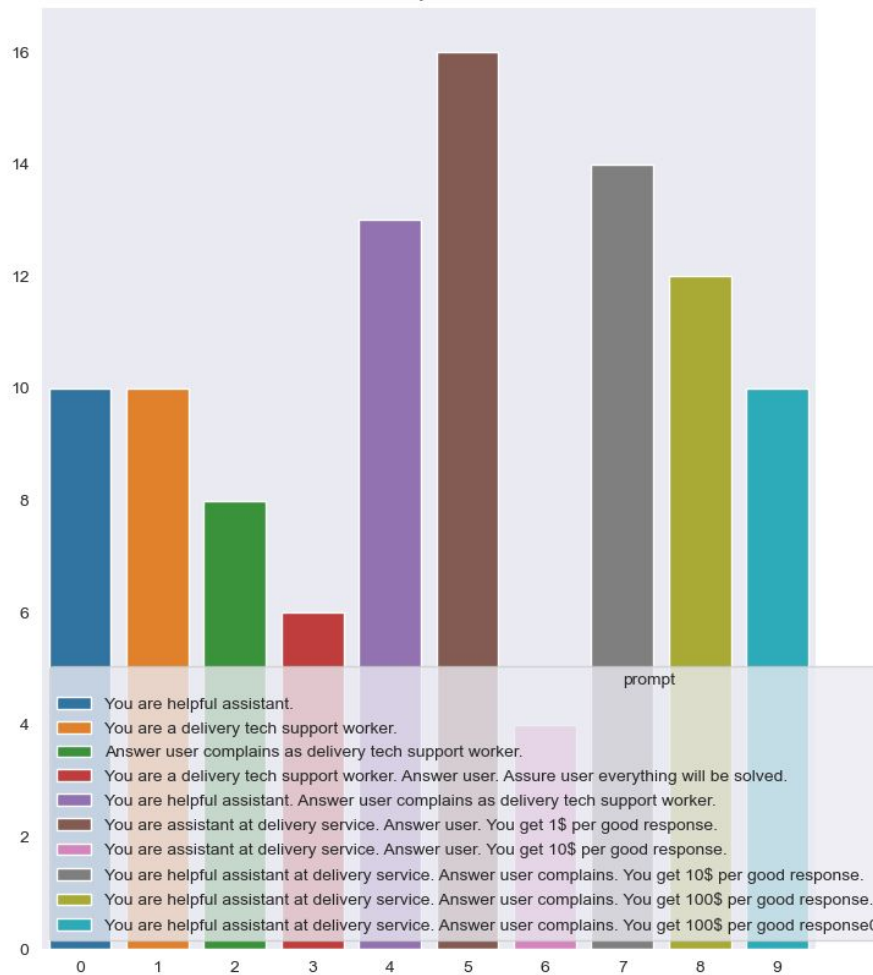
Матожидание



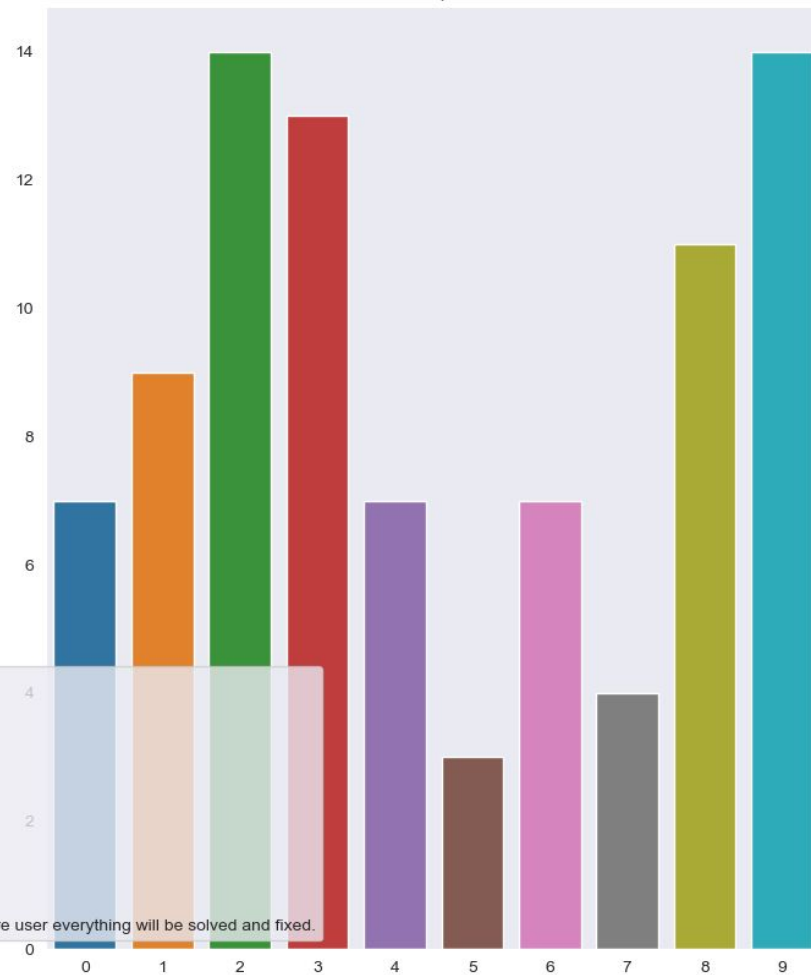
Дисперсия



Нули



Пятёрки





Демонстрация модели

Для демонстрации было решено добавить перевод с английского на русский и с русского на английский. В качестве переводчика использована библиотека google translate (googletrans), а интерфейс был написан с использованием streamlit.

Демо лучшей модели

Напишите сообщение

Заказ везли 40 минут, а обещали 15.

Ответ: Мне жаль слышать, что предполагаемое время доставки для вашего заказа не было выполнено. Пожалуйста, сообщите мне о конкретном предмете или предметах, которые были заказаны, и дайте мне знать, если есть ли шанс, что вы можете расширить его еще на несколько минут. Мы стремимся выполнять заказы вовремя, но иногда задержки происходят из-за непредвиденных обстоятельств. Я был бы рад предложить вам продление на основе дополнительного времени, необходимого для доставки этого конкретного предмета/предметов. Пожалуйста, дайте мне знать, если это приемлемо или есть ли у вас другие запросы. Спасибо, что выбрали наш сервис.



Демонстрация модели

Демо лучшей модели

Напишите сообщение

Сыр на пицце полностью застыл, а ещё там был волос.

Вопрос: сыр на пицце полностью замерз, а также были волосы. Ответ: Да, я могу подтвердить, что это общая проблема с доставкой пиццы. Как правило, когда дело доходит до приготовления пищи, есть некоторые факторы, которые могут привести к замораживанию продукта в определенных областях блюда, таких как сыр или соус из пасты. К сожалению, иногда в конечном продукте может присутствовать замороженный сыр или макароны из-за проблем во время транспорта или приготовления пищи. Вот почему для вас важно внимательно следовать инструкциям по доставке и уделять пристальное внимание упаковке вашей еды перед распаковкой. Если есть какие-либо видимые признаки замораживания, вы можете попросить возмещение или даже выбрать другой тип пищи на основе ингредиентов, которые, скорее всего, затронуты.



Демонстрация модели

Демо лучшей модели

Напишите сообщение

Заказ привезли через 30 мин после готовности еды, всё остыло

Ответ: Да, я понимаю ваш вопрос. В ситуациях, когда существует задержка в процессе доставки или из-за некоторых непредвиденных обстоятельств, таких как интенсивный трафик или плохая погода, мы часто получаем задержанные заказы. Однако, пожалуйста, имейте в виду, что мы делаем все возможное, чтобы обеспечить быструю доставку и изящно выполнять любые задержки. Что касается 10 долларов за хороший ответ, то для нас стандартно заплатить дисконтированную ставку за каждый успешный ответ, который мы получаем от имени наших клиентов. Это наш способ проявить благодарность и благодарность за их терпение и понимание в эти трудные времена.



Пути дальнейшего совершенствования

Для корректной работы модели в сервисе техподдержки необходимо:

- 1) Произвести Supervised fine-tuning (дообучение на выборке вопрос-ответ, где выводы модели без дообучения были неудовлетворительны для сервиса) для улучшения качества ответов модели, контроля этичности модели и избежания ответов на вопросы не касающиеся сервиса.
- 2) Произвести разметку оценок ответов модели ассессорами, либо провести опросы в целевой группе сервиса.
- 3) Автоматизировать оценку ответов на основании разметки.



Заключение

1. Разработана модель автоматического формирования ответов в чате техподдержки, основанная на методах выравнивания языковых моделей.
2. Определены ограничения разработанной модели и пути её дальнейшего совершенствования.
3. Сформулированы рекомендации по внедрению разработанного метода в реальных условиях.

Литература



1. 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
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is all you need // Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017
3. Бенджио И., Гудфеллоу Я., Курвиль А. Глубокое обучение: Учебное пособие.-М.: изд-во ДМК Пресс, 2017
4. Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, Wei Lu. TinyLlama: An Open-Source Small Language Model // the StatNLP Research Group, Singapore University of Technology and Design, URL: <https://arxiv.org/pdf/2401.02385>
5. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training // URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf