

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»**

Департамент анализа данных и машинного обучения

Пояснительная записка к курсовой работе
по дисциплине “Машинное обучение в семантическом и сетевом анализе”
на тему:

«Построение интеллектуального агента с помощью технологий LLM»

Выполнила:

студентка группы ПМ21-1
факультета информационных
технологий и анализа больших данных

_____ Яцык Я.М.

Научный руководитель:

Ст. преподаватель, Малекова В.А.

2024 г

Файлы работы можно найти на: <https://github.com/vseteplee/TCN>.

Список литературы

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825v1*, 2023.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Mazi  rz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, William El Sayed. Mixtral of Experts. *arXiv preprint arXiv:2401.04088v1*, 2024.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165v4*, 2020.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. *arXiv preprint arXiv:2109.01652v5*, 2022.

- [С. В. Макрушин. Лекции по дисциплине «Теория сложных сетей в экономике / Машинное обучение на графах», 2024](#)
- Документация библиотеки LangChain.
https://api.python.langchain.com/en/latest/langchain_api_reference.html, 2024.
- Документация библиотеки LangChain
https://api.python.langchain.com/en/latest/langchain_api_reference.html, 2024.
- Документация библиотеки LangChain x OpenAI
https://api.python.langchain.com/en/latest/chat_models/langchain_openai.chat_models.base.ChatOpenAI.html, 2024.
- Документация библиотеки LangChain
https://python.langchain.com/v0.1/docs/get_started/quickstart/, 2024.

Построение интеллектуального агента с помощью технологий LLM

Яцык Я. М.
Факультет ИТиАБД, ПМ21-1
Финансовый университет при правительстве РФ
Москва, 4-й Вешняковский пр-д, 4

Май 29, 2024

АННОТАЦИЯ

Благодаря своей универсальности и эффективности, большие языковые модели (Large Language Models) пользуются огромной популярностью для качественного решения множества задач, включая понимание, обобщение, анализ и перевод текстов. Естественным стало развитие применения технологий LLM для создания диалоговых систем, где интеллектуальный агент, имитируя поведение человека, может вести беседу в рамках поставленных целей. Данная работа посвящена созданию HR-бота, опрашивающего соискателя на предлагаемую вакансию.

Ключевые слова Большие языковые модели (LLM) · Интеллектуальный агент · Диалоговые системы · Создание подсказок (Prompt Engineering) · Telegram-бот

1 Введение

Мотивацией к созданию HR-бота послужила необходимость автоматизации опроса соискателей, что является простой и монотонной (по причине наличия четкого списка вопросов, на которые необходимо получить ответ), однако требующей большого количества времени и стрессоустойчивости от работника компании-работодателя. Задача, которую решает интеллектуальный агент - принять решение, удовлетворяет ли соискатель открытой вакансии (если да, то необходимо пригласить на дальнейшие этапы собеседования). Предлагаемая версия агента универсальна, поскольку может быть изменена в соответствии с требуемой идентичностью специалиста HR и масштабирована на необходимое количество вопросов о предлагаемой вакансии, могут быть добавлены дополнительные вакансии.

2 Модель Mixtral 8x7B

Большая языковая модель (Large Language Model, сокращенно LLM) — это вид модели глубокого обучения, которая работает с текстовыми данными, способна их понимать и генерировать текст на человеческом языке. Поскольку такие модели обучаются на огромных текстовых корпусах (включающих собрания книг, статей, тексты с сайтов и прочих источников), то они содержат в себе очень большое число обучаемых параметров. Считается, что языковая модель является большой, если содержит более одного миллиарда параметров.

В основу создаваемого интеллектуального агента взята большая языковая модель Mixtral 8x7B, созданная в декабре 2023 французским AI-стартапом Mistral AI. Эта модель находится в открытом доступе и является одной из наиболее качественных LLM, поддерживающих в том числе русский язык, на данный момент.

2.1 Архитектура Mixture of Experts

В модели Mixtral 8x7B используется архитектура Sparse Mixture of Experts, являющаяся оптимизированной версией Mixture of Experts. Ключевой особенностью последней является декомпозиция поставленной перед моделью комплексной задачи на множество более простых задач, каждую из которых решает соответствующая нейросеть, называемой экспертом. Предложенные экспертами предсказания объединяются и предлагаются как итоговый ответ. Такой подход позволил существенно уменьшить итоговое количество параметров модели, необходимое для решения задачи по сравнению с классической transformer-based архитектурой, где все необходимые "знания" модели по предметной области содержатся в одном feed-forward слое.

2.2 Архитектура Sparse Mixture of Experts

Однако не всегда для решения необходимо мнение всех экспертов, вычисление предсказаний которых может занимать много времени. В архитектуре Sparse Mixture of Experts добавлена еще одна дополнительная входная модель, которая принимает решение, какие из экспертов наиболее хорошо подходят в данном контексте, и только выбранные экспертные модели вычисляют свои предсказания, в то время как не подходящие модели не получают входных данных и не выполняют вычислений. Таким образом, содержащая 42 миллиарда параметров Mixtral 8x7B, основанная на архитектуре разреженных экспертов, работает с такой же скоростью, как, например, намного более легкая модель LLaMa 7B, хранящая свои знания только в 7 миллиардах параметров.

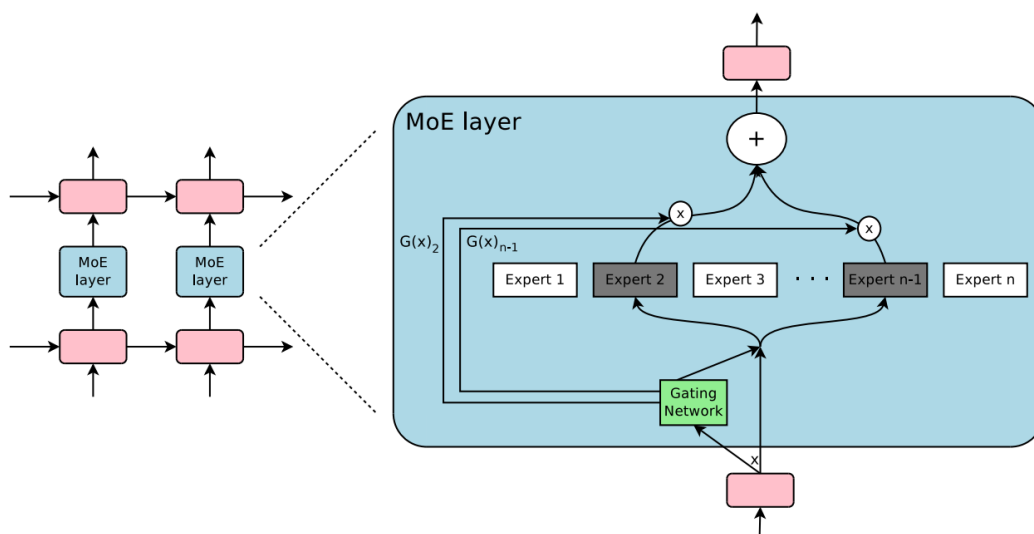


Рис. 1: Схематическое представление слоя Sparse Mixture of Experts, встроенного в рекуррентную сеть. Gating model - модель-классификатор, направляющая подаваемый контекст выбранным экспертам (серые); Expert N - модель-эксперт, дающая предсказание по входным данным в рамках своей экспертной области; + - обобщенное по всем полученным предсказаниям решение исходной задачи.

3 Создание подсказок

Промпт - написанная на человеческом языке инструкция, подсказка, для языковой модели, необходимая для генерации текста, соответствующего действительности, с требуемыми свойствами и структурой. Поскольку большие языковые модели обучены на огромном количестве неструктурированных текстов из разных источников, то полученная информация "в памяти" модели (то есть хранящаяся в параметрах уже обученной большой языковой модели) может быть противоречивой, неактуальной или выдуманной. В сочетании с тем, что языковые модели являются вероятностными, нередко порождаются так называемые галлюцинации у языковых моделей. Это означает, что модель "придумывает" нечто, не соответствующее реальности: фальсифицирует факты, имеет разного рода предубеждения, допускает логические ошибки. Именно для устранения галлюцинаций и задания вектора "мысли" большой языковой модели и необходимы хорошо написанные промпты.

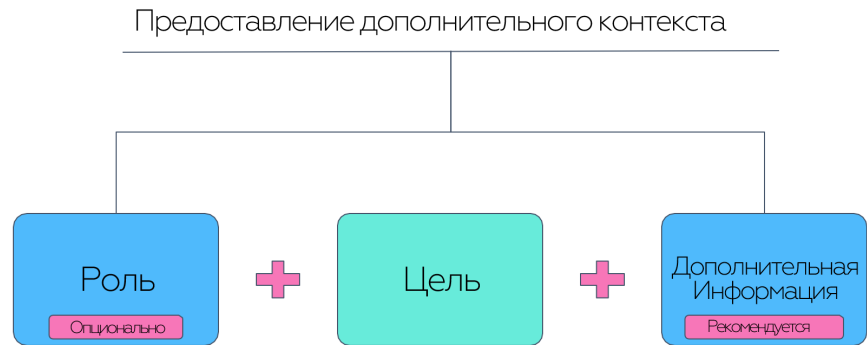


Рис. 2: Базовая структура подсказки

В промпте описывается ожидаемое поведение языковой модели, уточняются все необходимые подробности. В частности, с помощью промпта можно ощутимо улучшить качество выбора областей знаний, моделей-экспертов в архитектуре, к которым будет обращаться языковая модель. В общем случае подсказка может состоять из трех частей: опциональное описание роли, где раскрываются подробности "личности" агента; цель - основная часть, одновременно являющаяся и ограничением для модели, и улучшающая качество ее предсказания; контекст - информация, которую необходимо "держать в памяти" в условиях поставленной задачи, примеры ожидаемых входных или выходных данных.

Создание подсказок, промпт-инжиниринг, стало отдельной дисциплиной в области разработки и оптимизации промптов для эффективного использования языковых моделей. Не существует строго определенной формы или структуры написания промптов, однако множество исследований смогли показать большую эффективность некоторых особенных формулировок. Так, на сегодняшний день существует несколько десятков техник создания запросов к языковой модели, например, Zero-, One- и Few-Shot, Chain of Thoughts, Tree of Thoughts, Ask-before-Answer, Emotional Prompting, Self-Correct и другие. Далее будут описаны некоторые техники, которые были использованы при создании HR-бота.

3.1 Zero-Shot, One-Shot, Few-Shot

Языковая модель без явной инструкции не сможет структурировать ответ по одному шаблону, как это от нее ожидается (как не смог бы и человек, если ему не сообщить, как именно должен выглядеть результат его работы). Чтобы стандартизировать ответы модели, ей можно показать примеры.

Пример может быть неполным, это называется нулевой разметкой или zero-shot. Идея в том, что при обучении языковой модели, уже вероятно были прочитаны данные, которые помогут модели верно ответить на, как правило, простой вопрос, например, определение тональности текста.

Все представленные далее примеры для демонстрации техник сгенерированы на большой языковой модели ChatGPT 3.5.

Пример использования Zero-shot для определения тональности текста

Определи интонацию текста, как нейтральную, позитивную или негативную.
Текст: Думаю, сегодня был самый обычный день.
Интонация:

Результат

Нейтральная

В случаях, когда модели предлагается хоть сколько-нибудь нетривиальная задача, как правило, необходимо контекстное обучение на явных примерах того, что ожидается от ее предсказания. Таким образом, явно демонстрируя примеры ответов, можно добиться ощутимо более качественных генерируемых моделью ответов. В зависимости от того, сколько размеченных примеров предлагается модели в промпте, одна или несколько, подсказку называют, соответственно, One- или Few-shot.

Пример использования Few-shot для определения тональности текста

Сегодня обалденная погода! // Положительный
Новый фильм Гая Ричи оказался ярким и захватывающим, как всегда. // Нейтральный
Эти новые джинсы совершенно неудобные и некрасивые. // Отрицательный
Мне понравилась та новая кафешка на Арбате. //

Результат

Положительный

Что примечательно, даже неверно размеченные примеры позволяют модели делать верные предсказания.

Пример Few-shot с неправильной разметкой

Ответь только одним числом:
 $1+1=12$
 $34+90=8$
 $23+10=200$
 $100+11=$

Результат

111

В предложенных выше примерах использовались 3 разных способа описать приписанные метки, и это демонстрирует гибкость техники с точки зрения синтаксиса используемых конструкций. Однако для разных задач стоит выбирать наиболее подходящую разметку. В приведенном ниже примере меток как таковых нет, однако разметкой является смысловое описание слов.

Пример использования One-shot

Слово "купита" означает небольшой сосуд для хранения молока. Пример использования этого слова: "Аня осторожно налила свежее молоко в купиту, аккуратно устанавливая её на полке в старом сельском доме."
Слово "фирить" означает слышать что-либо. Пример использования этого слова:

Результат

"Сидя в тишине леса, я фирил легкий шепот деревьев и пение птиц."

3.2 Self-Correct и Emotional Prompting

В некоторых случаях модели способны сами исправлять недочеты своих ответов. Для этого используются соответствующие ключевые слова или фразы, например, "Важно удостовериться, что все слова написаны правильно и что предложения оформлены с учетом правил пунктуации." "Проверь свой ответ" или "Удостоверься, что это лучшее решение" и так далее, в зависимости от специфики задачи.

Также качественно улучшить ответ без явных инструкций, что именно необходимо исправить, можно используя эмоционально-окрашенные фразы (которые могут не иметь отношения к задаче, а лишь апеллировать к "чувствам" собеседника, которого имитирует модель). Такими словами могут быть разного рода заверения о важности и ценности ожидаемого предсказания (например, "Это очень важно для моей карьеры"), обещание вознаграждения за верный ответ, соответствующий прочим инструкциям. Таким вознаграждением могут быть обещание чаевых, притом их размер и валюта могут также иметь значение ¹. Могут быть применены угрозы в различных формулировках (лишение тех

¹Abel Salinas, Fred Morstatter. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. *arXiv preprint arXiv:2401.03729v3*, 2024

же чаевых/премии или откровенно терроризирующие модель аморальные высказывания, обещающие расправу над невинными котятками за несоблюдение инструкций). Однако могут быть использованы и менее радикальные слова, описывающие лишь "обстановку" для модели: "Сегодня понедельник, самый продуктивный день в году", "ChatGPT считает, что ты не сможешь это сделать", "Ты - самый грамотный и квалифицированный эксперт".

Стоит также отдельно отметить не выделяемые в отдельные техники важные особенности написания подсказок. В виду того, что подавляющее число больших языковых моделей обучены на корпусах, содержащих в большей степени англоязычные тексты, написание промптов на английском языке с отдельной инструкцией о том, что ожидаемый ответ должен быть написан кириллицей и на русском языке, зачастую может оказаться более эффективным, чем написанный аналогичный промпт на русском языке. Также, несмотря на то, что подсказка пишется на человеческом языке, следует придерживаться как можно большей структурированности текста, разделяя графически смысловые блоки промпта (используя, например, строку из символов # или _). Простота и грамотность языка инструкций также будут эффективнее сложносочиненных предложений и/или написанных с грамматическими ошибками.

3.3 Подсказки для HR

Так как интеллектуальный агент для HR-бота должен имитировать речь человека, то необходимо создание соответствующей идентичности и постановка задачи для модели. Под идентичностью понимается описание "личности" агента, включая его имя, профессию (HR специалист), необходимые сведения о компании, которую он представляет, и подробности предлагаемых вакансий. Также в идентичности должны быть зафиксированы все необходимые ограничения, которых должен придерживаться агент, а в частности особенности речи (язык, вежливый, но формальный стиль общения, соответствующий профессии, лаконичность речи).

Частью идентичности является и цель диалога, а именно опрос соискателя в соответствии с необходимым набором требуемых характеристик или предлагаемых условий и последующее принятие решения, удовлетворяет ли кандидат требованиям вакансии. В том числе должен быть установлен запрет на выход за рамки темы диалога для агента.

Все озвученные выше ограничения и особенности идентичности модели были реализованы в виде одного промпта, подаваемого модели перед началом диалога. А именно, в инструкциях указано самому проверять себя на грамотность и лаконичность речи, в виде Few-shot были добавлены примеры того, как должен выглядеть диалог в зависимости от контекста, описаны общие принципы взаимодействия с собеседником. Для обращения к самой языковой модели Mixtral 8x7B были использованы инструменты библиотеки LangChain.

4 LangChain и реализация HR-бота

Фреймворк LangChain предоставляет API-интерфейсы, которые позволяют разработчикам подключаться из кода к большим языковым моделям и запрашивать данные от них. Через LangChain можно взаимодействовать с общедоступными большими языковыми моделями, выполняя простые вызовы API без написания сложного кода.

На основе класса `langchain_core.language_models.chat_models.BaseChatModel` был создан класс для обращения к модели Mixtral 8x7B с использованием Deefinfra API, где можно задать все требуемые параметры языковой модели (температура, параметры семплинга top-p и top-k, максимальное количество токенов и т.д.).

Модуль `langchain_core.messages` позволяет удобно взаимодействовать с сообщениями модели (AIMessage) и пользователя (HumanMessage), связывая их в единый диалог, подставляя в промпты.

Для создания универсальных шаблонов промптов можно использовать класс `langchain_core.prompts.ChatPromptTemplate`, позволяющий подставлять в написанный шаблон подсказки разные значения переменных извне. Для создания HR-бота были использованы шаблоны для автоматизированного написания промптов в соответствии с заданными значениями имени, сведений о компании и подробностей вакансии. Таким образом, можно не переписывать весь текст промпта в зависимости от изменившихся значений параметров, а лишь исправить значения соответствующих переменных.

Так как задача HR-бота заключается в проведении опроса по заданному набору вопросов, то есть вспомогательный промпт-шаблон для отдельного вызова языковой модели (тоже Mixtral 8x7B), опреде-

ляющий по словарю этапов, на какой стадии разговора находится диалог. Предсказание этого вызова модели добавляется в основной промпт с HR-идентичностью.

Таким образом в ответ на сообщение пользователя модель-помощник определяет, на каком этапе разговора сейчас находится диалог, и модель-HR в соответствии с этим генерирует ответ в рамках своей цели и идентичности. Так продолжается до логического завершения диалога.

Для представления созданной модели в виде Telegram-бота была использована библиотека aiogram, позволяющая использовать Telegram Bot API с помощью Python.

5 Результаты

Полный код HR-бота можно найти на <https://github.com/vseteplee/TCN>. Для Запуска необходимо указать bot_token (main.py) вашего Telegram-бота и ключ к API Deepinfra deepinfra_api_token (deepinfra.py).

Ниже представлен результат работы бота в виде одного из проведенных диалогов.

HR-бот	Здравствуйте! Меня зовут Эста, я специалист в сфере HR и подбора персонала в частной компании ВкусВкусыч. Вы в поиске новой работы?
Пользователь	Здравствуй, Эста! Да
	Отлично! Мы рады предложить вам вакансию продавец-консультант. Хотите узнать подробности?
	Это единственная открытая вакансия?
	Да, в настоящее время у нас есть только одна открытая вакансия продавца-консультанта в нашей сети продуктовых магазинов ВкусВкусыч. Может быть, это вас заинтересует?
	Да, рассказывайте)
	Превосходно! Мы предлагаем вам работу по графику на выбор 3/3 или 2/2 с 8.30 до 17.30. Вам удобно такое расписание?
	Мммм, не люблю ранние подъемы.
	Понимаю. Мы предлагаем два варианта графика работы: 3/3 или 2/2 дней подряд с последующим выходным днём. Например, вы можете выбрать работу по графику 3/3, тогда вы будете работать три дня подряд, а затем будете иметь трёхдневный выходной. Вариант 2/2 предполагает работу два дня подряд, а затем двухдневный выходной. Выбирайте удобный для вас график. Вам подходит такой вариант?
	Первый вариант звучит неплохо.

Превосходно! Тогда давайте перейдём к следующему вопросу. Какая зарплата вы хотели бы получать в ВкусВкусыч, работая на должности продавец-консультант?

80 т.р.

Зарплата на нашей вакансии продавца-консультанта составляет до 75 тыс. руб. Ваша зарплата будет зависеть от вашего опыта, навыков и производительности. Вас устроит зарплата в размере 75 тыс. руб?

Устроит

Превосходно! Мы рады сообщить, что наша компания также предлагает программу «Здоровый сотрудник»: ДМС и абонемент в фитнес-клуб, кешбэк 15% бонусами на покупки в ВкусВкусыче. Еще мы компенсируем питание, оформление мед. книжки, фирменную одежду и обучение. Есть ли у вас вопросы по вакансии или условиям работы?

Да. Где находится ваш офис?

Наш офис расположен по адресу Ростов-на-Дону, улица Максима Горького, дом 159. Наш магазин находится в удобном для посещения месте, вблизи от остановки общественного транспорта. Может быть, у вас есть ещё вопросы?

А будет собеседование?

Да, собеседование будет проводиться онлайн в ZOOM. Мы просим вас выбрать день с понедельника по пятницу любое время с 10.00 до 15.00. Собеседование займет не более получаса. Ссылку для подключения мы пришлём вам в ближайшее время. Вам удобно такое время собеседования?

В пт в 14.15.

Отлично! Я пришлю Вам ссылку на собеседование в ближайшее время. А можете ли Вы сообщить мне Ваше имя и актуальный для связи номер телефона? Спасибо!

Меня зовут Валя, номер 89997766555.

Спасибо, Валя! Будем ждать Вас на собеседовании в пятницу в 14:15. Если у Вас возникнут вопросы до собеседования, пожалуйста, не стесняйтесь связаться со мной. До встречи!

До свидания!

Список литературы

- [1] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825v1*, 2023.
- [2] Noam Shazeer, Azalia Mirhoseini, Krzysztof Mazi  rz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- [3] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, William El Sayed. Mixtral of Experts. *arXiv preprint arXiv:2401.04088v1*, 2014.
- [4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165v4*, 2020.
- [5] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. *arXiv preprint arXiv:2109.01652v5*, 2022