1. Добрый день! Моя работа посвящена исследованию способов эффективного дообучения языковых моделей в задаче выделения информации из контекста.
2. Языковые модели развиваются быстрыми темпами, однако у них все еще остается ряд проплем. Среди них можно выделить - отсутствие возможности быстро актуализировать знания, высокие затраты на адаптацию и фактологические ошибки.
3. Одним из способов решить эти проблемы является метод retrieval augmented generation. Мы делаем поисковые индексы по тематическим документам. А затем для ответа на пользовательский вопрос дополнительно добавляем в контекст генерации несколько наиболее релевантных фрагментов документации.
4. Несмотря на эффективность, у метода также есть своим известные проблемы: опора на нерелевантную информацию, а также зависимость от положения информации в контексте.
5. Из-за нерелевантного контекста ответ модели может ухудшиться. Как пример на изображении модель дает корректны ответ вовсе без контекста, а с добавлением нерелевантной информации начинает давать неверный ответ.
6. Проблема, когда модель плохо обрабатывает информацию в середине контекста в научной литературе называется lost in the middle. Согласно результатам из оригинальной статьи, положение информации крайне важно для модели.
7. Из-за того, что для RAG основной задачей генеративной модели является аггрегация информации из контекста, то можно использовать небольшие языковые модели и адаптировать из к основной задаче. Потому целью исследования можно определить как построение эффективного пайплайна дообучения языковых моделей в задаче извлечения информации. Основными задачами является - ▶ Исследование проблем и оценка существующих методов. ▶ Комбинация эффективных подходов. ▶ Оценка качества генерации на созданном бенчмарке.
8. Все написано на слайде
9. Одним из способов решить проблему lost in the middle является дообучение на синтетических данных. В частности предлагается сгенерировать большой набор словарей и дообучать модель на запросы по поиску значения по ключу.
10. Так как в открытом доступе нет русскоязычных RAG бенчмарков, то был собран тестовый набор по наиболее популярным доменам применения RAG-систем. Кроме того, была создана типизация вопросов, по которым можно оценить отдельные навыки модели. Общий размер бенчмарка порядка 1000 примеров.

11. В качестве основного обучающего датасета использовался WebGLM-QA переведенный на русский язык, где контекстной информацией были поисковые запросы. Для метода RAFT использовался этот же датасет с добавлением 3 отвелекающих докуметов и 10 процентами негативных примеров. В качестве метрик использовались оценки модели-судьи, а также метрики ROUGE-L.
12. Если посмотреть на результаты дообучения 3B модели qwen-2.5, то можно увидеть что метод RAFT показывает себя эффективнее, чем классическое дообучение. Кроме того был получен рост во всех типах вопросов, кроме no-info и incorrect by design.
13. Для 1.5b ситуация не такая однозначная, так как на более простых типах вопросов качестве уменьшилось, а на более сложных типах существенно возрасло. И в этом случае RAFT показал себя хуче чем классическое дообучение
14. Таблица
15. У RAFT больше recall из-за COT
16. Все предущие замеры проводились с 10 фрагментами документации. Чтобы исследовать значимость их количества и положения, была проведена серия замеров. По результатам видно, что хотя отличия и есть, однако они в большинстве случаев незначительны. Из этого можно сделать вывод что у современных моделей размером хотя бы 3 миллиарда параметра проблема lost in the middle не возникает в сценариях генерации с контекстом порядка 10к токенов.
17. Также исследовалось добавление дополнительных этапов. В частности - предварительный этап руссификации на датасете saiga показал ухудшение итогового качества, хотя по метрикам MMLU был прирост. Вероятно это связано с ухудшением обработки контекста. Финальный этап дообучения на синтетичских словарях не оказался эффективным. 3b модель решала задачу без всякого дообучения, а 1.5b модель не сходилась на том же размере контекста и задача была слишком сложной.
18. Подводя итоги исследования, хочется отметить, что для реальных сценариев применения RAG систем не была была замечена проблема lost-in-the-middle. А также результаты экспериментов показывают, что хотя небольшие языковые модели и можно дообучать под задачу RAG и это будет давать качественный прирост, однако в некоторых сценарих применения небольшой вес все же является ограничением и не позволяет моделям достигать тоже же качества даже после дообучения.