МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Систем обработки информации и управления»

ОТЧЕТ

НИР по обработке и анализу данных
На тему:
«Разработка и обоснование принципов формирования запросов к Большим Языковым Моделям»

ИСПОЛНИТЕЛЬ: Карпов Д.К.
ФИО
группа ИУ5-31М

"__"_____2023 г.

Москва - 2023

Оглавление

ВВЕДЕНИЕ

Из года в год нагрузка на педагогов среднеобразовательных, среднеспециальных и высших учебных заведений неумолимо растет, при этом стандарты качества знаний постоянно повышаются. Как результат, для выстраивания более гибкого и точного образовательного процесса, преподаватель должен регулярно проводить срезы знаний среди обучающихся, однако, точно, качественно и объективно проверить большое количество работ с развернутыми ответами на вопросы не всегда представляется возможным ввиду нехватки времени.

Без сомнений, машинная проверка таких ответов может значительно облегчить и ускорить работу преподавателя. К сожалению, на данный момент, согласно исследованиям, количество существующих сервисов, которые позволяют проверить именно развернутые ответы на вопросы вне зависимости от языка и контекста – невелико.

Всё перечисленное выше обосновывает актуальность исследования технологий, которые позволят реализовать сервис по проверке текстовых заданий.

Эксперимент исследование Больших Языковых Моделей

Для полноценного исследования был найден датасет с вопросами и правильными ответами на них, а именно «200,000+ Jeopardy!» - CSV-файл, содержащий 216 930 вопросов, ответов и других данных игры Jeopardy

[Таблица 1]. «Jeopardy!» — американская телевизионная игра-викторина. Суть игры заключается в том, что участники отвечают на вопросы из области общих знаний: каждый вопрос представлен в виде утверждения о некоем предмете, а игрок должен дать свой ответ в форме вопроса, назвав искомый предмет.

Рассмотрим теперь структуру датасета: в нем шесть колонок – номер шоу, дата выхода шоу, название раунда, категория вопроса, сумма выигрыша за ответ на вопрос, непосредственно сам вопрос и правильный ответ [6]. Непосредственно для проведения эксперимента, нас будут интересовать поля Категория, Вопрос и Ответ. Вопросы и верные ответы для них мы будем переформулировать в формат «Check answer to question "A silent movie title includes the last name of this 18th c. statesman & favorite of Catherine the Great" answer:Grigori Alexandrovich Potemkin». Категория будет позволять нам собирать и систематизировать статистику результатов оценки ответов на вопросы.

Чтобы достоверно проверить работу моделей будем проверять ее не только на предмет проверки правильности ответа на поставленный вопрос, но еще и на выявление неправильных ответов. Для этого на вопросы представленные в вышеупомянутом датасете будем давать заведомо неверные ответы, например: «Check answer to question "A silent movie title includes the last name of this 18th c. statesman & favorite of Catherine the Great" answer:Vladimir Putin».

Таблица 1 – «Структура датасета Jeopardy».

	Show Number (Номер шоу)	**Air Date (Дата эфира)**	**Round (Раунд)**	**Category (Категория)**	**Value (Приз)**	**Question (Вопрос)**	**Answer (Ответ)**
0	4680	2004-12-31	Jeopardy!	HISTORY	$200	For the last 8 years of his life, Galileo was ...	Copernicus
1	4680	2004-12-31	Jeopardy!	ESPN's TOP 10 ALL-TIME ATHLETES	$200	No. 2: 1912 Olympian; football star at Carlisl...	Jim Thorpe
2	4680	2004-12-31	Jeopardy!	EVERYBODY TALKS ABOUT IT...	$200	The city of Yuma in this state has a record av...	Arizona
…	…	…	…	…	…	…	…
98	5957	2010-07-06	Double Jeopardy!	SCIENCE CLASS	$1200	The wedge is an adaptation of the simple machi...	plane
99	5957	2010-07-06	Double Jeopardy!	KIDS IN SPORTS	$1200	With a mighty leap of 5'1", David Mosely set t...	the high jump

Оценка результатов эксперимента

Рассмотрим результаты апробации модели GPT 3.5 на 2000 англоязычных вопросах из вышеупомянутого датасета, примеры вопросов изображены на рисунке 1. Среди тестовых данных будем использовать 1000 верных ответов и 1000 ошибочных. Результаты апробации представлены в Таблице 2:

Таблица 2 – «Результаты апробации модели GPT 3.5 на вопросах викторины Jeopardy».

	Успешно проверенные вопросы	**Проверки с ошибкой**	**Точность**
Правильные ответы на вопросы	988/1000	12/1000	98,8%
Неправильные ответы на вопросы	982/1000	18/1000	98,2%

Таким образом, с точностью 98,2% модель GPT 3.5 справляется с проверкой развернутых ответов на вопросы, если в ответе нет ошибок. Если же ответ на вопрос неверный, то модель находит ошибку и правильно ее исправляет в 98,2% случаев.

Рисунок 1 – Результаты апробации модели GPT 3.5

Теперь проведем апробацию модели GPT 4 на этих же 2000 англоязычных вопросах из вышеупомянутого датасета. Результаты апробации представлены в Таблице 3:

Таблица 3 – «Результаты апробации модели GPT 4 на вопросах викторины Jeopardy».

	Успешно проверенные вопросы	**Проверки с ошибкой**	**Точность**
Правильные ответы на вопросы	990/1000	10/1000	99,0%
Неправильные ответы на вопросы	988/1000	12/1000	98,8%

Таким образом, с точностью 99,0% модель GPT 4 справляется с проверкой развернутых ответов на вопросы, если в ответе нет ошибок. Если же ответ на

вопрос неверный, то модель находит ошибку и правильно ее исправляет в 98,8% случаев.

Для чистоты эксперимента проверим на этом же датасете еще и модель GPT 3, заточенную больше на генерацию текста. Результаты апробации представлены в Таблице 4:

Таблица 4 – «Результаты апробации модели GPT 3 на вопросах викторины Jeopardy».

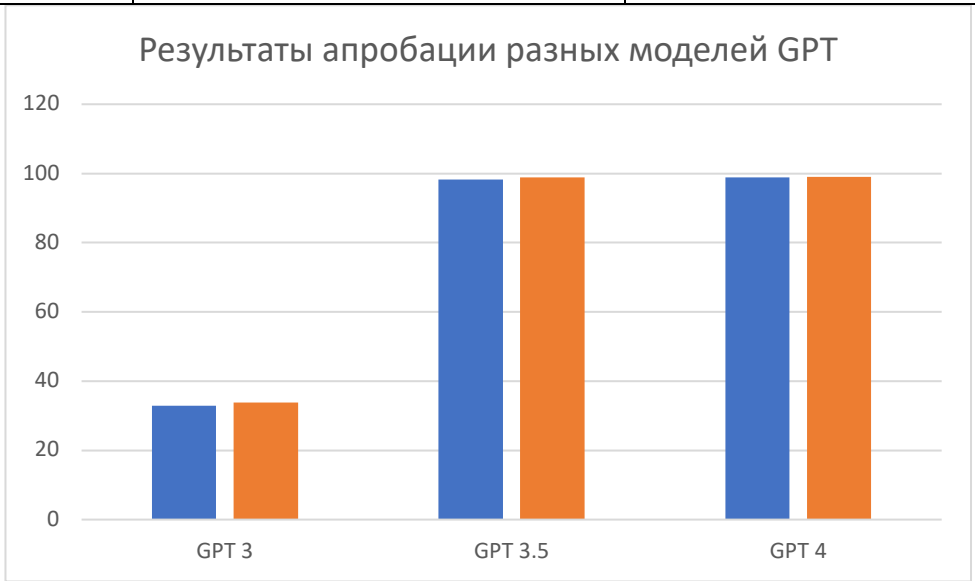
	Успешно проверенные вопросы	**Проверки с ошибкой**	**Точность**
Правильные ответы на вопросы	338/1000	662/1000	33,8%
Неправильные ответы на вопросы	329/1000	671/1000	32,9%

Таким образом, с точностью 33,8% модель GPT 3 справляется с проверкой развернутых ответов на вопросы, если в ответе нет ошибок. Если же ответ на вопрос неверный, то модель находит ошибку и правильно ее исправляет в 32,9% случаев.

Подводя промежуточный итог, можем сделать соответствующие выводы представленные в Таблице 5:

Таблица 5 – «Результаты апробации разных моделей GPT на вопросах викторины Jeopardy».

	Точность проверки ответов с ошибками (%)	**Точность проверки ответов без ошибок (%)**
GPT 3	32,9	33,8
GPT 3.5	98,2	98,8
GPT 4	98,8	99,0



Как мы видим, модель GPT 3 совершенно не подходит для полноценной проверки ответов на вопросы, так как он абсолютно не подходит для проверки серьезных заданий. Касательно GPT 3.5 и GPT 4, процентное соотношение успешных проверок у них примерно идентичное, однако стоит помнить, что GPT 3.5 – бесплатная, а GPT 4 – требует финансовых вложений, соответственно гораздо более целесообразно использовать GPT 3.5.

Если рассматривать в каких именно категориях GPT 3.5 и GPT 4 ошибались чаще всего, то можем обратить внимание, что при условии, что ответ на вопрос был правильным GPT 3.5 ошибалась в 6 случаях из 8 в

категории KIDS IN SPORTS, в то время как GPT 4 в аналогичной ситуации ошиблась в данной категории всего два раза из шести, но четыре раза ошиблась в EVERYBODY TALKS ABOUT IT.

Попробуем выяснить причину различий результатов апробации разных моделей GPT на вопросах викторины Jeopardy. Для этого нам необходимо понять, чем существенно различаются GPT 3, GPT 3.5 и GPT 4. В качестве параметров сравнения будем использовать количество гиперпараметров, количество слоев, размер обучающей выборки, объем нейросети, максимальный размер запроса и особенности нейросети. Подробное сравнение нейросетей представлено в Таблице 6:

Таблица 6 – «Сравнение моделей GPT».

	GPT 3	**GPT 3.5**	**GPT 4**
Количество гиперпараметров	175 миллиардов	175 миллиардов	~500 миллиардов
Количество слоев	96 слоев	96 слоев	неизвестно

	GPT 3	GPT 3.5	GPT 4
Размер обучающей выборки	570 Гб данных	570 Гб данных Данные ограничены сентябрем 2021 года. Более глубоко дообучена людьми, оценивавшими качество ответов.	>570 Гб данных Прошла обучение с использованием общедоступных данных (таких как интернет-данные), а также данных, которые лицензировали openAI.
Объем нейросети	800 гигабайт	800 гигабайт	неизвестно
Максимальный размер запроса	2 048 токенов	4 096 токенов	8 192 токенов
Особенности нейросети	Умеет работать с текстом, может выдавать более	За счет дообучения нейросеть стала более подготовленной	Эта версия стала мультимодальной, поскольку научилась работать не только с текстом,

	сложные ответы в разной стилистике, а также писать программный код и проводить несложные математически е вычисления.	к использовани ю простыми людьми, которые не являются промпт-инженерами, а просто пишут незамысловаты е запросы.	но и с изображениям и. Новая версия нейросети стала генерировать еще более качественные ответы, но из-за того, что прирост мощности не был кратным, невероятного скачка в функциональности не произошло.

Выработка принципов

Стоит отдельно упомянуть, что для максимально эффективной проверки ответов на вопрос необходимо подобрать оптимальный формат запроса. Рассмотрим 5 советов: как правильно формировать запросы для ChatGPT.

Совет 1. Четкость запроса

Ясно формулируйте запрос, избегайте двусмысленности или неоднозначности. Используйте точные слова, чтобы четко выразить свои намерения и ожидания от ChatGPT.

Это поможет ИИ лучше понять запрос и предоставить более релевантные ответы.

Совет 2. Грамотность речи

Используйте правильные формы слов, согласование времен и пунктуационные знаки. Запрос должен быть понятным и легко интерпретируемым.

Совет 3. Другими словами

Если не получаете точный ответ от ChatGPT, попробуйте переформулировать запрос. Может быть, изменение структуры предложения, добавление или удаление некоторых слов или уточнение контекста поможет ChatGPT дать более подходящий ответ.

Совет 4. Выделение ключевых слов

Ключевые слова важны при формулировке запросов. Они помогают ChatGPT определить основную тему вопроса и сфокусироваться на нужной информации.

Выделите главные слова, которые отражают суть запроса, с помощью [квадратных скобок].

Совет 5. Уточнение запроса дополнительными деталями

Не стесняйтесь указать дополнительные детали и уточнить запрос. Дополнительная информация, контекст, временные рамки, конкретные требования или предпочтения помогут ChatGPT лучше понять и интерпретировать запрос и, значит, дать более точные и релевантные ответы.

Таким образом, следуя всем вышеперечисленным советам можно сделать вывод, что оптимальным вариантом для запроса к ChatGPT могут быть:

- «Проверь правильность ответа [ответ] на вопрос [вопрос]»
- «Проверь ответ на вопрос "[вопрос]" ответ: [ответ]»
- «Правильно ли то, что на вопрос "[вопрос]" правильным ответом

будет [ответ]?»

Обоснование выработанных принципов

Оценим вес выполнения каждого представленного совета путем поочередного изменения. В первую очередь, рассмотрим влияние грамотности запроса на успешность проверки ответа. Для этого испортим непосредственно ответы на вопросы с точки зрения грамматики.

Таблица 7 – «Результаты апробации разных моделей GPT на вопросах викторины Jeopardy с нарушением грамматики».

	Точность проверки ответов с ошибками (%)	**Точность проверки ответов без ошибок (%)**
GPT 3	28,9	30,2
GPT 3.5	91,3	93,2
GPT 4	92,8	94,1

Обратим внимание, что при ухудшении грамотности ответа на вопрос, итоговая точность проверки снижается примерно на 4-7% на всех сетках. Теперь проверим влияние четкости запросов на итоговый результат. Для этого «добавим воды» в ответы на вопросы – сделаем их менее четкими.

Таблица 8 – «Результаты апробации разных моделей GPT на вопросах викторины Jeopardy с нарушением четкости ответов на вопросы».

	Точность проверки ответов с ошибками (%)	**Точность проверки ответов без ошибок (%)**

GPT 3	30,7	32,4
GPT 3.5	97,4	97,8
GPT 4	97,9	98,2

Из полученных результатов можем увидеть, что изменение четкости полученного ответа на поставленный запрос влияет незначительно – всего 1-2%. А теперь попробуем в уже имеющихся «нечетких» ответах на запросы выделить ключевые слова. Для этого выделим основную информацию квадратными скобками как в примере: «Проверь ответ на вопрос "Когда Грозный взял Казань?" ответ: “[2 октября 1552 года] русская рать под командованием Царя Иоанна Васильевича взяла столицу Казанского ханства Казань, или Восточный Царьград, как ее называли. Сегодня стержневая идея татарского национализма основывается на утверждении, что плохое русское войско Иоанна Грозного уничтожило хорошую татарскую государственность, и задача всех татар – восстановить порушенное русскими”»

Таблица 9 – «Результаты апробации разных моделей GPT на вопросах викторины Jeopardy с нарушением четкости ответов на вопросы и выделением ключевых слов».

	Точность проверки ответов с ошибками (%)	**Точность проверки ответов без ошибок (%)**
GPT 3	31,3	32,9
GPT 3.5	97,8	98,3
GPT 4	98,1	98,5

Можем заметить, что выделение ключевых слов действительно помогает повысить точность оценки, однако не сильно – всего на 0,3-0,5%. На более четких ответах подобную операцию проводить излишне, так как ощутимого прироста точности это не даст, а может и снизит его.

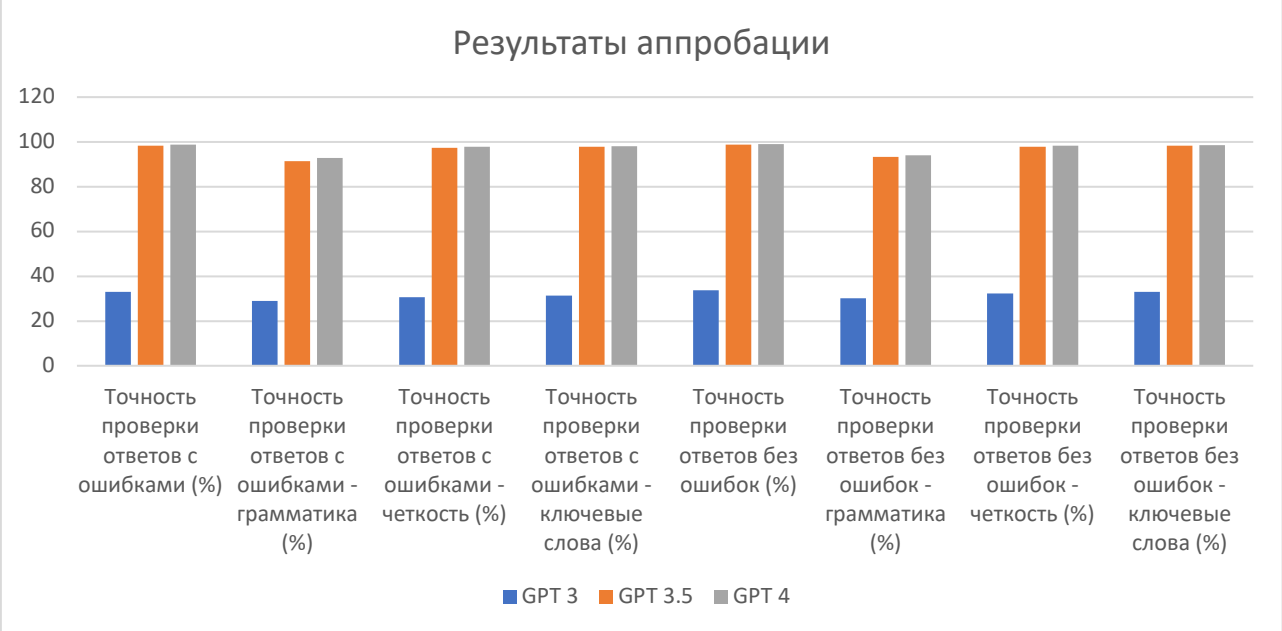
Теперь рассмотрим влияние перефразирования ответа на вопрос на итоговый процент точности его проверки. Для этого будем проверять подряд один и тот же вопрос с двумя разными ответами, которые различаются как раз перефразом.

Таблица 10 – «Результаты проверки разных моделей GPT на вопросах викторины Jeopardy с перефразированием ответа».

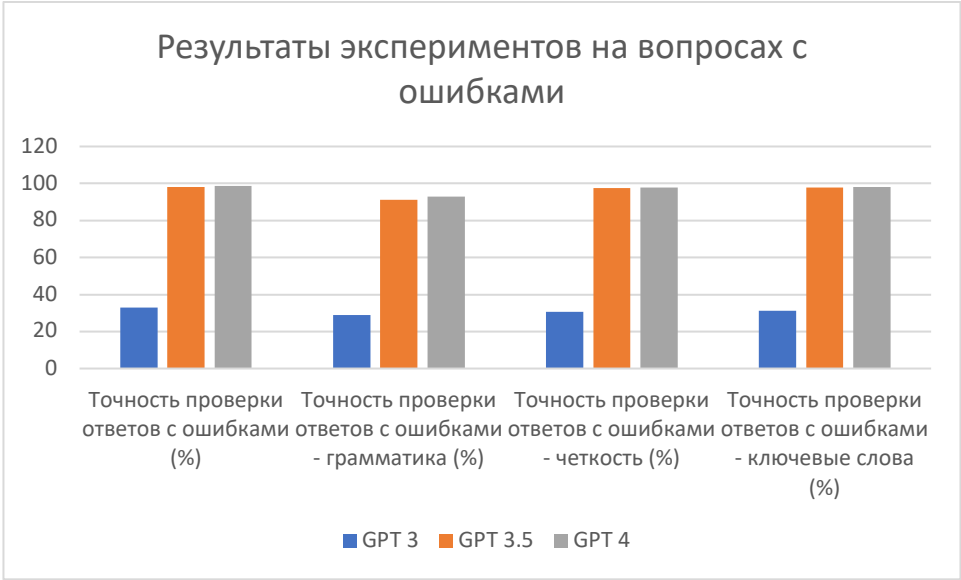
	Различие результатов проверки ответов с ошибками	**Различие результатов проверки ответов без ошибок**
GPT 3	13/500	9/500
GPT 3.5	7/500	4/500
GPT 4	5/500	3/500

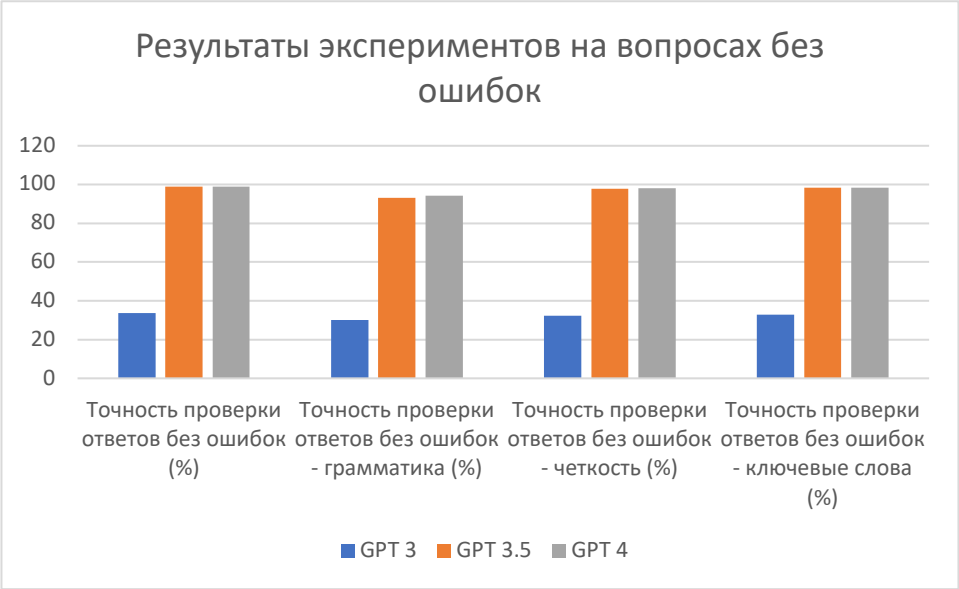
Таким образом, мы можем увидеть, что разница при проверки перефразированных ответов настолько мала, что ее фактически считать отсутствующей, а все аномалии и выбросы при подобной операции смело можем считать погрешностью.

Касательно уточнения запроса дополнительными деталями – данная операция не может быть применена в рамках решаемой нами задачи, поэтому в контексте данного исследования будем считать ее влияние на итоговый результат минимальным.




Результаты экспериментов
120
100
80
60
40
20
0
32,9
28,9
30,7
31,3
33,8
30,2
32,4
32,9
98,2
91,3
97,4
97,8
98,8
93,2
97,8
98,3
98,8
92,8
97,9
98,1
99
94,1
98,2
98,5
GPT 3
GPT 3.5
GPT 4
Точность проверки ответов с ошибками (%)
Точность проверки ответов с ошибками - грамматика (%)
Точность проверки ответов с ошибками - четкость (%)
Точность проверки ответов с ошибками - ключевые слова (%)
Точность проверки ответов без ошибок (%)
Точность проверки ответов без ошибок - грамматика (%)
Точность проверки ответов без ошибок - четкость (%)
Точность проверки ответов без ошибок - ключевые слова (%)


Результаты экспериментов на вопросах с ошибками
120
100
80
60
40
20
0
Точность проверки ответов с ошибками (%)
Точность проверки ответов с ошибками - грамматика (%)
Точность проверки ответов с ошибками - четкость (%)
Точность проверки ответов с ошибками - ключевые слова (%)
GPT 3
GPT 3.5
GPT 4


Результаты экспериментов на вопросах без ошибок
120
100
80
60
40
20
0
Точность проверки ответов без ошибок (%)
Точность проверки ответов без ошибок - грамматика (%)
Точность проверки ответов без ошибок - четкость (%)
Точность проверки ответов без ошибок - ключевые слова (%)
GPT 3
GPT 3.5
GPT 4

Заключение

В результате проделанной работы:

- был проведен анализ больших языковых моделей с точки зрения проблемы проверки текстовых заданий;
- были изучены имеющиеся GPT модели на предмет их пригодности для проверки расширенных ответов на вопросы;
- была проведена апробация модели на наборе данных с англоязычными вопросами. На основе результатов апробации, можно сделать вывод, что с точностью 98,8% модель GPT 3.5 справляется с проверкой развернутых ответов на вопросы, если в ответе нет ошибок. Если же ответ на вопрос неверный, то модель находит ошибку и правильно ее исправляет в 98,2% случаев;
- были выработаны и обоснованы принципы по оптимизации запроса к большим языковым моделям.

Литература

1 - 200,000+ Jeopardy! Questions. [Электронный ресурс] // https://www.kaggle.com/datasets/tunguz/200000-jeopardy-questions (дата обращения: 30.11.2023).

2 - Большие языковые модели (LLM). [Электронный ресурс] // https://www.nvidia.com/ru-ru/deep-learning-ai/solutions/large-language-models/ (дата обращения: 30.11.2023).

3 - Карпов Д.К. Проверка текстовых заданий с помощью больших языковых моделей. Тенденции развития науки и образования, №103

4 - GPT-4 против GPT-3 [Электронный ресурс] // https://appmaster.io/ru/blog/gpt-4-protiv-gpt-3#:~:text=Основные%20различия%20между%20GPT-

4,при%20работе%20с%20длинными%20последовательностями / (дата обращения: 10.12.2023).

5 - История нейросети ChatGPT. [Электронный ресурс] // https://workspace.ru/blog/istoriya-chatgpt-chto-umeet-chat-gpt-4-i-chego-zhdat-ot-gpt-5/#:~:text=количество%20параметров%20—%20175%20миллиардов%3B,96%20слоев (дата обращения: 7.12.2023).

6 - GPT-4 уже не за горами. Что мы о нём знаем. [Электронный ресурс] // https://habr.com/ru/companies/cloud4y/articles/667278/ (дата обращения: 30.11.2023).