ЛИНГВИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ ИНСТРУМЕНТОВ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

LINGUISTIC RESEARCH ON LARGE LANGUAGE MODELS APPLICATIONS

УДК: 81’33 + 004.8 DOI: 10.31249/chel/2025.02.09

Бабина О.И., **Быкова А.С.**

МОДЕЛИРОВАНИЕ ОЦЕНКИ ТЕХНИЧЕСКИХ КОМПЕТЕНЦИЙ С ПРИМЕНЕНИЕМ ОТВЕТОВ АГЕНТА GPT©

Южно-Уральский государственный университет
(*национальный исследовательский университет*),
Россия, *Челябинск*, *babinaoi@susu.ru*, *anastb6@mail.ru*

Аннотация. В статье предложена функциональная схема автоматизации оценки технических компетенций сотрудников IT-компании с применением модели автоматической проверки ответов на вопросы теста технического ассесмента открытого типа. Для получения эталонного ответа на вопрос теста используется генерация с поддержкой найденной релевантной информации на базе большой языковой модели GPT-4o. Далее нейросетевая модель осуществляет оценку компетенций посредством сравнения ответов пользователей с эталонными ответами; все ответы предварительно векторизуются с применением предобученной русскоязычной модели FastText. Разработанная модель сравнения ответов выполнена в форме сиамской нейросети, построенной на базе многослойного перцептрона, которая на выходе имеет нейрон-регрессор, предсказывающий значения в диапазоне от 1 до 10. Экспериментальная часть исследования проводилась на наборе данных, включающем пары вопросов и ответов теста технических компетенций, оцененных экспертами. Результаты показали, что качество модели по метрике среднеквадратичной ошибки (MSE) на тестовой выборке составило 0,036, что свидетельствует о высокой корреляции между оценками модели и экспертными оценками. Разработанная модель и функциональная схема автоматизации техниче-

ского ассесмента могут способствовать оптимизации процессов рекрутинга и мониторинга компетенций сотрудников, а также более глубокому пониманию их реальных знаний и навыков в контексте быстро меняющихся технологий.

Ключевые слова: трансформер; генеративная нейросеть; большие языковые модели; GPT-4o; сиамская нейросеть; многослойный перцептрон; технический ассесмент; генерация с поддержкой релевантной информации.

Получена: 17.07.2024 Принята к печати: 28.12.2024

Babina O.I., Bykova A.S.

Modeling technical assessment using GPT agent's responses©

South Ural State University, Russia, Chelyabinsk,
babinaoi@susu.ru, anastb6@mail.ru

Abstract. The paper proposes a method for automating the technical assessment of employees in an IT company using a model for scoring responses to open-ended technical assessment test questions. The suggested method employs retrieval augmented generation by the large language model GPT-4o to create a reference answer for the test question. Subsequently, a custom neural network model assesses competencies by comparing a user response with the reference answer, both of which have been vectorized using a pre-trained Russian-language model FastText. The developed model is implemented as a Siamese neural network based on a multilayer perceptron, which outputs a regressor predicting values ranging from 1 to 10. The experimental part of the study was conducted on a dataset consisting of pairs of questions and answers from the technical assessment test, scored by experts. The model performance, evaluated using the mean squared error (MSE) on the test split, achieves 0.036, demonstrating a high correlation between the model predictions and expert scores. The developed model and method for automating technical assessment can contribute to optimizing recruitment and monitoring employees' competencies, as well as providing a deeper understanding of their actual knowledge and skills in the context of rapidly changing technologies.

Keywords: transformer; generative neural network; large language models; GPT-4o; Siamese network; multilayer perceptron; technical assessment; retrieval augmented generation.

Received: 17.07.2024 Accepted: 28.12.2024

Введение

В современном мире информационных технологий, где конкуренция в сфере IT-услуг постоянно растет, важно иметь эффективные методы проверки технических компетенций сотрудников.

Традиционные методы оценки знаний могут быть затратными и недостаточно точными. Мы считаем, что применение методов искусственного интеллекта для моделирования проверки компетенций может оптимизировать процесс отбора и оценки готовности персонала выполнять трудовые функции, что, в свою очередь, должно способствовать повышению уровня профессионализма специалистов IT-компаний. Исследование этой проблемы может внести вклад в развитие новых технологий и подходов к оценке профессиональных компетенций, что будет полезно для развития как науки и производства, так и общества в целом.

Теоретические предпосылки

Многие задачи обработки естественного языка сегодня продуктивно решаются с применением нейросетей – см., например, [Advancements in …, 2024; Clark, 2024; Kumar, Singh, 2024; Xu, McAuley, 2023] и др. Эти технологии позволяют эффективно обрабатывать и анализировать большие объемы многоязычных текстовых данных, обеспечивая высокую точность решения поставленной задачи [Łazuka, 2024; Zhu, 2024].

Первые нейронные сети имели архитектуру многослойного перцептрона [Rosenblatt, 1958], который строится из входного слоя, одного или нескольких скрытых полносвязных слоев и выходного слоя, обучающихся с использованием метода обратного распространения ошибки [Rummelhart, 1986]. Архитектура оказывается полезной для задач классификации, регрессии и распознавания образов. С развитием технологий обработки естественного языка появилась новая архитектура нейронных сетей – трансформеры [Attention is all you need, 2017]. Трансформеры используют механизм внимания, который позволяет модели фокусироваться на различных частях входных данных при их обработке, что оказалось особенно эффективно для обработки текстовых последовательностей, при решении задач машинного перевода, суммаризации и генерация текста.

Сегодня трансформеры являются стандартом де-факто при работе с естественным языком, при этом одними из наиболее известных и широко используемых моделей на основе трансформеров являются генеративные модели GPT [Language models …, 2019; Language models …, 2020], разработанные компанией OpenAI. GPT-модели обучаются на больших объемах текстовых

данных и способны генерировать связный и осмысленный текст на основе заданного контекста. Последняя на сегодняшний день вышедшая модель семейства GPT – модель GPT-4o (omni) [GPT-4o System Card, 2024] – обладает улучшенными показателями быстроты и дешевизны обработки данных, обработки языков, отличных от английского, а также дополнена возможностью подавать на вход не только текст, но и изображения[1]. Модели GPT нашли широкое применение в задачах автоматического написания текстов, создания чат-ботов, перевода и многих других.

Однако GPT-модели не лишены недостатков. Во-первых, модели GPT могут генерировать текст, содержащий предвзятость или нежелательный контент, так как они обучаются на больших объемах данных из Интернета, которые могут содержать такие элементы. Во-вторых, модели GPT могут иногда «галлюцинировать», то есть генерировать текст, который кажется осмысленным, но на самом деле не имеет логической связи или содержит фактические ошибки. Подходом для нивелирования недостатков и улучшения качества генерируемого текста является генерация с поддержкой поиска [Retrieval-augmented generation …, 2020], которая сочетает в себе генеративные модели и методы поиска информации по внешним источникам, предписанным базам знаний. Такой подход позволяет модели использовать результаты поиска для улучшения качества генерируемого текста, делая его более точным и информативным, и снижает риск появления предвзятости или нежелательного контента, так как текст генерируется на основе более надежных и проверенных данных.

Для решения задач, опирающихся на выполнение операций сравнения (распознавание лиц, изображений, поиск дубликатов и прочих), используется сиамская архитектура нейросетей [Signature verification …, 1993], которые состоят из двух или более идентичных нейронных сетей, работающих параллельно и имеющих общие веса.

Трансформеры и сиамские нейронные сети обладают уникальными характеристиками и сферами применения, что позволяет решать с их помощью разнообразные задачи. В данном исследовании мы сосредоточились на проблеме моделирования оценки технических компетенций сотрудников IT-компаний и предложили функциональную схему, которая объединяет эти технологические

[1] https://platform.openai.com/docs/models/gpt-4o

решения для достижения цели автоматизации контроля знаний и обеспечивает адекватную оценку уровня компетенций сотрудников, так как строится на оценивании ответов на вопросы теста открытого типа. Таким образом, процесс контроля технических компетенций проводится в условиях естественной коммуникации, что позволяет выявить у тестируемых реальный объем их знаний по актуальному в текущий момент стеку технологий.

Методология исследования

В задачи данной работы входит проведение моделирования системы тестирования для автоматизированной проверки технических компетенций (технического ассесмента) сотрудников IT-компании. Формы проведения технического ассесмента могут быть многообразны – тесты, проектные задания, кодинг-челленджи, интервью и прочие [Dubois, Rothwell, 2010], однако тестирование остается наиболее популярным и простым в исполнении. При моделировании тестирования профессиональных компетенций использование вопросов открытого типа имеет ряд преимуществ по сравнению с тестовыми вопросами с выбором ответа, которые не отражают реальных знаний, а часто лишь проверяют способность логически мыслить, извлекая из ответов подсказки, наличие определенной доли везения (ответ можно угадать), и, кроме того, разработка таких тестов требует затрат, дополнительных усилий. В свою очередь, тесты со свободно-конструируемыми ответами лишены этих недостатков [Карпова, 2010], а значит, могут позволить более достоверно оценить имеющийся набор знаний сотрудника по стеку технологий, при этом форма «вопрос – ответ» является естественной формой коммуникации. Однако автоматизация проверки корректности ответов на вопросы открытого типа требует моделирования понимания высказывания.

Принимая во внимание современный уровень развития технологий обработки естественного языка, имеющиеся предобученные большие языковые модели, мы полагаем естественным решением поставленной задачи моделирование технического ассесмента на основе тестов открытого типа (для повышения достоверности контроля знаний) с применением нейросетевых моделей для автоматизированной проверки корректности ответов.

С учетом изложенного, решение задачи моделирования технического ассесмента осуществлялось в несколько этапов:

1) разработка функциональной схемы проведения автоматизированной оценки компетенций;

2) подготовка набора данных для обучения компонентов схемы оценки технических компетенций;

3) выбор моделей машинного обучения для реализации компонентов функциональной схемы технического ассесмента (в том числе генеративной большой языковой модели для генерации эталонных ответов с поддержкой поиска), применимых для создания программного инструментария;

4) проведение экспериментального обучения моделей на подготовленном наборе данных;

5) анализ результатов моделирования.

Функциональная схема оценки технических компетенций

Мы предлагаем следующий подход для решения задачи автоматизации технического ассесмента. Оценка компетенций должна строиться в режиме диалога. Вопросы для тестирования готовятся заранее и предъявляются сотруднику, который формулирует ответ на естественном языке. Коммуникация ведется на русском языке. Система получает ответ тестируемого, сравнивает полученный ответ с эталонным и ставит оценку на основании степени сходства / различия с эталоном. Эталонный ответ формируется с помощью предобученной большой языковой модели с применением подхода генерации с поддержкой поиска по базе знаний. База знаний включает документы, содержащие информацию, релевантную для ответа на вопрос теста компетенций.

Функциональная схема оценки технических компетенций приведена на рис. 1.

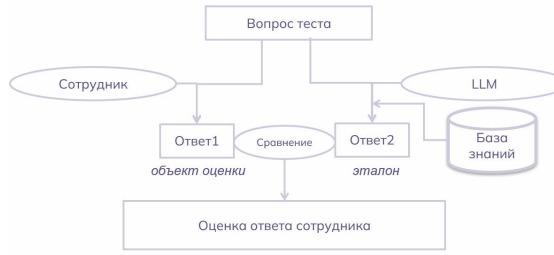

Рис. 1. Функциональная схема оценки технических компетенций

Такая функциональная схема позволяет выполнить реализацию системы технического ассесмента в форме Telegram-бота.

Подготовка набора данных

Предложенная функциональная схема оценки компетенций предопределила подготовку набора данных для обучения различных компонентов реализации функциональной схемы, включающего следующие компоненты: 1) перечень вопросов теста проверки технических компетенций, сопровождающихся текстами материалов, содержащих релевантную теме вопроса информацию, из которых может быть получен правильный ответ на вопрос теста; 2) набор данных, включающий пары «вопрос теста – ответ на вопрос» с числовой оценкой, определяющей степень корректности ответа на поставленный вопрос. В качестве шкалы оценивания была принята шкала с диапазоном оценок от 1 до 10, где 1 – ответ полностью неверный, 10 – ответ совершенно корректный.

Для апробации функциональной схемы были взяты материалы, предоставленные компанией 3divi[1]. Компания является успешным предприятием в сфере разработок в области компьютерного зрения. Для проведения экспериментальной работы компанией были предоставлены вопросы теста технического ассесмента для оценки знаний сотрудников по стеку технологий, включая C++, frontend-разработку, backend-разработку. Для каждого вопроса экспертами компании были определены гиперссылки на доступные в Интернете теоретические материалы по стеку технологий. Каждому вопросу теста были поставлены в соответствие ответы на этот вопрос, в том числе правильный ответ и ответы с различной степенью корректности, составленные сотрудниками компании. Каждому ответу была дана целочисленная оценка экспертом из числа сотрудников компании в диапазоне от 1 (ответ совершенно неверный) до 10 (ответ полностью корректный). Каждый ответ оценивал один эксперт, обладающий опытом разработки на соответствующем стеке технологий. Составленный таким образом первый поднабор данных включал 76 пар вопросов – ответов с экспертными оценками корректности ответа. Количественная характеристика полученных данных приведена в табл. 1.

[1] https://3divi.ru/

Таблица 1

Количественная характеристика исходного поднабора данных

Стек технологий	Количество вопросов по стеку	Количество ответов на каждый вопрос	Всего пар вопросов – ответов
frontend-разработка	11	2	22
C++	11	4	44
backend-разработка	10	1	10
Всего	32	-	76

С целью расширения составленного набора данных был также создан прототип Telegram-бота, реализующий функционал предъявления вопросов технического ассесмента пользователям, без выполнения автоматизированной оценки полученных ответов, но с сохранением полученных ответов пользователей в базу ответов. База вопросов бота была составлена из вопросов первого поднабора данных и расширена за счет вопросов по работе с базами данных и git-репозиторием. Общий объем базы составил 55 вопросов. Логика сценария бота включает: 1) выбор стека компетенций; 2) последовательное предъявление вопросов стека компетенций, предполагающих, что пользователь дает ответы открытого типа; 3) сохранение ответных сообщений пользователей бота в базе ответов. Таким образом, был осуществлен сбор ответов на вопросы теста оценки технических компетенций в процессе опроса сотрудников компании. Составленная база ответов далее передавалась экспертам компании для числовой оценки корректности ответов.

Результатом этапа подготовки данных стал набор данных из двух поднаборов, структура которого включает: 1) вопрос теста компетенций, 2) свободный ответ на вопрос теста, 3) оценку ответа по шкале от 1 до 10. В наборе данных используется 55 различных вопросов по стеку технологий, при этом в общей сложности набор содержит 186 пар вопросов – ответов с оценками.

Имплементация компонентов функциональной схемы

Построенная функциональная схема технического ассесмента предопределяет необходимость выбора технических решений для имплементации каждого из ее компонентов, включая:

1) выбор модели для генерации эталонного ответа;

2) выбор способа векторизации ответа пользователя и эталонного ответа;

3) построение модели, определяющей степень близости векторов и отображающей полученные векторные представления на числовую шкалу, показывающую степень корректности полученного ответа.

Для подготовки эталонного ответа мы выбрали предобученную генеративную языковую модель GPT-4o, версия gpt-4o-2024-05-13, которая в режиме генерации с поддержкой поиска по сформированному индексу формирует ответ на вопрос теста оценки технических компетенций. Для поиска эталонного ответа на вопрос теста технического ассесмента используется документ, полученный по ссылке из набора данных и содержащий информацию, релевантную для ответа на вопрос, таким образом обеспечивая возможность реализации подхода генерации эталонного ответа с поддержкой поиска. Документ подается в компонент для индексирования текста VectorStoreIndex фреймворка LlamaIndex, в результате чего на основе поданной информации формируется индекс. Далее подготовленный промпт вида: «*Напиши на русском языке в нескольких предложениях ответ на вопрос:* “*Q*”», где *Q* – текст вопроса из теста оценки технических компетенций подается на интерфейс взаимодействия с созданным индексом. В ответ на запрос генерируется ответ на основе извлеченной информации с применением модели GPT-4o, для которой установлен параметр температуры равный 0,2. Полученный таким образом ответ рассматривается как эталон правильного ответа на поставленный вопрос *Q*.

Далее текст ответа пользователя и полученный эталон представляются в форме вектора, для чего выбрана предобученная на русских текстах модель FastText [Enriching word …, 2017]. Для векторного представления используется метод get_sentence_vector библиотеки fasttext с подгруженной обученной моделью для русского языка, таким образом, текст ответа сотрудника и текст эта-

лонного ответа модели GPT-4o, сгенерированного по промпту на основе сформированного индекса релевантного теме документа, представляются в форме 300-мерных векторов.

Полученные векторы ответа пользователя и эталонного ответа подаются на вход модели, имеющей структуру сиамской нейросети. Архитектура ветви базовой части модели представляет собой многослойный перцептрон, включающий три полносвязных слоя. В начале сети используется широкий слой с большим количеством нейронов (1024), что позволяет захватывать сложные и высокоуровневые признаки. В середине сети количество нейронов уменьшается (128), сужение способствует выделению более специфических признаков. В конце сети снова увеличивается количество нейронов (1024), что позволяет интегрировать и обобщать извлеченные признаки. Узкий слой активируется функцией ReLU. К широким слоям применяется активационная функция – экспоненциальная линейная единица (ELU), которая задается как:

$$ELU(x) = \begin{cases} x, \text{при} x \geq 0 \\ e^x - 1, \text{при} x < 0 \end{cases}$$

Применение этих функций активации вводит нелинейность в модель и позволяет отслеживать более сложные зависимости. Каждый полносвязный слой сопровождается слоями Dropout для предотвращения переобучения и улучшения обобщающей способности модели.

Векторы признаков, полученные из ветвей базовой сети для обоих входов, вычитаются друг из друга. Полученный вектор разности подается на сжимающий полносвязный слой и далее слой-регрессор.

Архитектура регрессионной модели, выполняющей сравнение двух векторов, представляющих собой выполненные с помощью FastText свертки ответа сотрудника и эталонного ответа, приведена на рис. 2.

Построенная архитектура содержи 637 185 параметров. При обучении модели использовалась функция потерь – среднеквадратичная ошибка. Количество эпох обучения: 500, размер батча: 32.

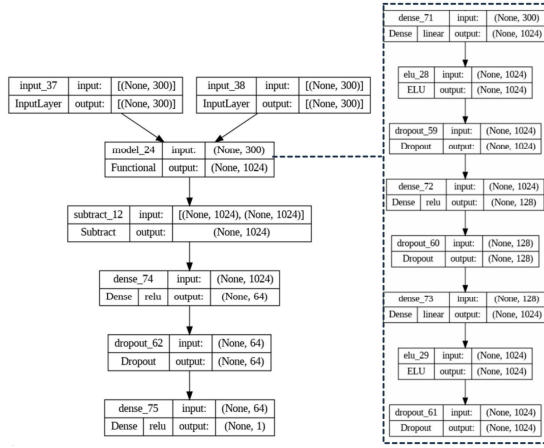

Рис. 2. Архитектура модели оценки технических компетенций

Обучение проводилось на тестовой и валидационной выборках набора данных, выделенных в соотношении 70% и 15% соответственно. Оставшиеся 15% выборки вошли в тестовую часть. Целевая переменная – экспертная оценка корректности ответа – во всех выборках была масштабирована по максимальному значению 10 для ускорения сходимости и повышения стабильности обучения. Таким образом, целевая переменная, подаваемая в модель, принимала значения в диапазоне [0,1; 1].

Результаты

В ходе экспериментальной работы было апробировано несколько вариантов архитектуры и настройки параметров. Описанная в предыдущем пункте архитектура модели показала на валидационной и тестовой выборках наилучший результат. Графики функции потерь обучения (рис. 3) демонстрируют достаточно стабильный процесс обучения модели. Среднеквадратичная ошибка на тестовой выборке составила 0,036, средняя абсолютная ошибка – 0,138. Коэффициент детерминации R2 составил 0,718.

Пример предсказаний модели в сопоставлении с экспертными оценками корректности ответов приведен на рис. 4.

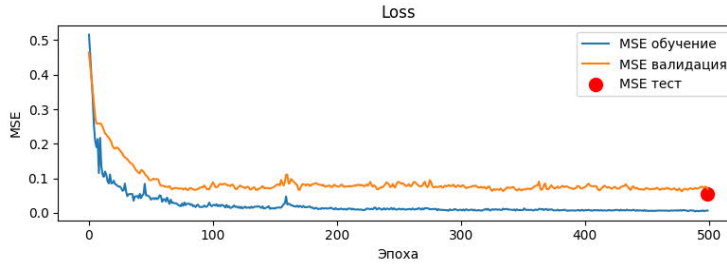

Рис. 3. Изменение функции потерь MSE на обучающей, валидационной и тестовой выборках

que	answer	ethalon	predicted score	actual score
Что такое замыкание (closure)?	Доступ к области памяти, объявленной во внешней функции	Замыкание (closure) — это функция вместе с окружающим её контекстом, который сохраняется и может быть использован при вызове этой функции.	8.637697219848633	8
Что такое move семантика, и где она применяется?	позволяет избавиться от ненужного копирования данных, повысить эффективность использования железа	Move семантика в C++ позволяет передавать ресурсы от одного объекта к другому без копирования, что эффективно используется для оптимизации работы с временными объектами и динамической памятью.	9.973758697509766	10
В чему суть JWT и как происходит проверка что токену можно доверять?	Проверяется подпись и срок жизни токена	Суть JWT заключается в том, что это компактный и самодостаточный токен для передачи информации между сторонами, а проверка доверия к токену происходит через верификацию его подписи с использованием секретного ключа или публичного ключа.	3.811227321624756	2
Какую проблему позволяет решать паттерн saga в микросервисной архитиктуре?	X	Паттерн saga в микросервисной архитектуре позволяет решать проблему управления распределенными транзакциями, обеспечивая согласованность данных и возможность отката операций в случае ошибок.	0.0	1
Какую проблему позволяет решать паттерн saga в микросервисной архитиктуре?	Транзакции	Паттерн saga в микросервисной архитектуре позволяет решать проблему управления распределенными транзакциями, обеспечивая согласованность данных и возможность отката операций в случае ошибок.	9.219715118408203	2

Рис. 4. Фрагмент предсказаний модели на тестовой выборке

В целом модель демонстрирует достаточно высокую корреляцию с оценками ответов, выставленных экспертами, – сравните значения в колонках *predicted score* (оценка модели) и *actual score* (оценка эксперта) на рис. 4. Вместе с тем в отдельных случаях модель показала значительно разнящиеся результаты (см. последний пример на рис. 4). Такое отклонение может быть обусловлено малым объемом и несбалансированностью выборки. Следует отметить, что короткие ответы в целом хуже интерпретируются моделью. Наличие ключевого слова из эталонного ответа, как правило, в значительной степени влияет на его смещение моделью на шкале оценок в сторону правильного ответа. Заметим, что значительные отклонения на нашей выборке всегда демонстрировали ответы, которые моделью оценены лучше, чем экспертом.

Заключение

В статье представлена функциональная схема для автоматизации оценки технических компетенций сотрудников IT-компаний. Разработанная схема тестирования основывается на двух ключевых аспектах: а) использовании предобученной языковой модели для генерации эталонных ответов, полученных по расширенному запросу, включающему информацию, релевантную для ответа, и б) сопоставлении ответов тестируемого с эталоном с помощью сиамской нейросети на базе многослойного перцептрона. Предложенная схема автоматизации технического ассесмента позволяет повысить эффективность процедуры оценки знаний сотрудников. Повышение эффективности связано, во-первых, с тем, что автоматизация оценки ответа в форме текста сокращает время и ресурсы, затрачиваемые на ручную проверку таких ответов. Кроме того, собственно разработка теста с открытыми вопросами менее трудоемка, чем составление вопросов и возможных ответов к ним. Во-вторых, использование открытых вопросов вместо тестов с выбором ответа способствует более глубокой оценке реальных знаний и навыков сотрудников. Внедрение разработанной модели и функциональной схемы тестирования в производственный процесс может способствовать оптимизации процесса отбора персонала и регулярного мониторинга соответствия текущего уровня компетенций сотрудника современному состоянию технологий.

Дальнейшее совершенствование модели оценки компетенций сотрудников может быть проведено в результате апробации возможностей применения ансамблей предобученных языковых моделей в архитектуре нейросети [Ensembling finetuned …, 2024], которая выполняет сопоставление ответов сотрудников с эталонными. Отдельным аспектом для изучения является экспериментальное исследование способов предварительной обработки текстов, которые вполне могут влиять на распознавание семантически значимых компонентов в тексте [Бабина, 2024]. Возможности оптимизации процедур и алгоритмов генерации эталонных ответов с учетом дополнительно найденной релевантной информации требует дальнейшего изучения. Исследование влияния на результаты оценки различных факторов, таких как тип вопроса, лингвистические характеристики релевантной вопросу информации в промпте, эксплицитность / имплицитность представленной в промпте реле-

вантной информации и прочих, также является важным направлением для будущих работ.

Результаты данного исследования могут быть полезны для дальнейшей разработки не только систем контроля знаний в специальной области знаний, но также и обучающих систем, которые смогут помочь новым сотрудникам быстрее адаптироваться к требованиям компании и освоить необходимые технологии.

Список литературы

Бабина О.И., *Зиновьева А.Ю.*, *Неручева Е.Д.* Влияние предварительной обработки набора данных на концептуальную разметку текстовых токенов на основе двунаправленной LSTM // Terra Linguistica. – 2024. – Т. 15, № 3. – С. 109–123. DOI: 10.18721/JHSS.15310

Карпова И.П. Сравнение открытых и выборочных тестов // Открытое образование. – 2010. – № 3. – С. 32–38.

Advancements in natural language processing for text understanding / Basha M.J., Vijayakumar S., Jayashankari J., Alawadi A.H., Durdona P. // Proceedings of the International Conference on Newer Engineering Concepts and Technology (ICONNECT-2023) (Tiruchirappalli, Tamil Nadu, India, 27–28 April 2023). – 2023. – Vol . 399. – Art. no. 04031. DOI: 10.1051/e3sconf/202339904031

Attention is all you need / Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., …, Polosukhin I. // NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, CA, USA, 4–9 Dec 2017). – Red Hook, New York : Curran Associates Inc., 2017. – P. 6000-6010.

Clark A. Enhancing and exploring the use of transformer models in NLP tasks // International Journal of Transcontinental Discoveries. – 2024. – Vol. 11, No. 1. – P. 62–71.

Dubois D.D., *Rothwell W.J*. Competency-Based human resource management: discover a new system for unleashing the productive power of exemplary performers. – London : Nicholas Brealey Publishing, 2010. – 228 p.

Enriching word vectors with subword information / Bojanowski P., Grave E., Joulin A., Mikolov T. // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135–146. – URL: https://aclanthology.org/Q17-1010.pdf

Ensembling finetuned language models for text classification / Arango S.P., Janowski M., Purucker L., Zela A., Hutter F., Grabocka J. // 38th Workshop on Fine-Tuning in Machine Learning (NeurIPS 2024). – 2024. – URL: https://arxiv.org/abs/2410.19889

GPT-4o system card / Hurst A., Lerer A., Goucher A.P., Perelman A., Ramesh A., Clark A., …, Malkov Yu. // arXiv:2410.21276 [cs.CL]. – 2024. – URL: https://arxiv.org/abs/2410.21276

Kumar D., *Singh S.* Advancements in transformer architectures for large language models: from BERT to GPT-3 and beyond // International Research Journal of Modernization in Engineering Technology and Science. – 2024. – Vol. 6, No. 5. – P. 1889–1895. DOI: 10.56726/IRJMETS55985

Language models are few-shot learners / Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal A., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M.,

Wu J., Winter C., Hesse Ch., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner Ch., McCandlish S., Radford A., Sutskever I., Amodei D. // arXiv:2005.14165. – 2020. – URL: https://arxiv.org/abs/2005.14165

Language models are unsupervised multitask learners / Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. // OpenAI Blog. – 2019. – Vol. 1, No. 8. – Art. no. 9. – URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Łazuka M., *Anghel A.*, *Parnell Th.* LLM-pilot: characterize and optimize performance of your LLM inference services // IEEE. – 2024. – URL: https://arxiv.org/abs/2410.02425

Multilingual large language models: a systematic survey / Zhu S., Supryadi Sh., Sun J., Pan L., Cui M., Du J., Jin R., Branco A., Xiong D. // arXiv:2411.11072 [cs.CL]. – 2024. – URL: https://arxiv.org/abs/2411.11072

Retrieval-augmented generation for knowledge-intensive NLP tasks / Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-T., Rocktäschel T., Riedel S., Kiela D. // Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020. – URL: https://discovery.ucl.ac.uk/id/eprint/10100504/1/2005.11401v1.pdf

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // Psychological Review. – 1958. – Vol. 65, No. 6. – P. 386–408.

Signature verification using a "Siamese" time delay neural network / Bromley J., Guyon I., LeCun Y., Säckinger E., Shah R. // Advances in Neural Information Processing Systems. – 1993. – Vol. 6. – P. 737–744.

Xu C., *McAuley J.* A survey on dynamic neural networks for natural language processing // Findings of the Association for Computational Linguistics: EACL 2023. – 2023. – P. 2370–2381. – URL: https://aclanthology.org/2023.findings-eacl.180.pdf

References

Babina, O.I., Zinoveva, A.Yu., Nerucheva, E.D. (2024). Vliyanie predvaritelnoy obrabotki nabora dannykh na kontseptualnuyu razmetku tekstovykh tokenov na osnove dvunapravlennoy LSTM [Dataset preprocessing effects on bi-LSTM-based concept tagging of text tokens]. *Terra Linguistica*, *15*(*3*), 109–123. DOI: 10.18721/JHSS.15310

Karpova, I.P. (2010). Sravnenie otkrytyh i vyborochnyh testov [Comparison of open and multiple choice tests]. *Otkrytoe obrazovanie* [*Open Education*], *3*, 32–38.

Basha, M.J., Vijayakumar, S., Jayashankari, J., Alawadi, A.H., Durdona, P. (2024). Advancements in natural language processing for text understanding. In *Proceedings of the International Conference on Newer Engineering Concepts and Technology* (ICONNECT-2023), 399, 04031. DOI: 10.1051/e3sconf/202339904031

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., …, Polosukhin, I . (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus (eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Red Hook, New York: Curran Associates Inc.

Clark, A. (2024). Enhancing and exploring the use of transformer models in NLP tasks. *International Journal of Transcontinental Discoveries*, *11*(*1*), 62–71.

Dubois, D.D., Rothwell, W.J. (2010). *Competency-based human resource management*: *discover a new system for unleashing the productive power of exemplary performers*. London: Nicholas Brealey Publishing.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from: https://aclanthology.org/Q17-1010.pdf

Arango, S.P., Janowski, M., Purucker, L., Zela, A., Hutter, F., Grabocka, J. (2024). Ensembling finetuned language models for text classification. In *38th Workshop on Fine-Tuning in Machine Learning (NeurIPS 2024)*. Retrieved from: https://arxiv.org/abs/2410.19889

Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark A., …, Malkov, Yu. (2024). GPT-4o System Card. *arXiv preprint*, arXiv:2410.21276 [cs.CL]. Retrieved from: https://arxiv.org/abs/2410.21276

Kumar, D., Singh, S. (2024). Advancements in transformer architectures for large language models: from BERT to GPT-3 and beyond. *International Research Journal of Modernization in Engineering Technology and Science*, *6*(*5*), 1889–1895. DOI: 10.56726/IRJMETS55985

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, A., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, Ch., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, Ch., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint*, arXiv:2005.14165 [cs.CL]. Retrieved from: https://arxiv.org/abs/2005.14165

Radford, A., Wu, J., Child R., Luan D., Amodei D., Sutskever I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(*8*), 9. Retrieved from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Łazuka, M., Anghel, A., Parnell, Th. (2024). LLM-Pilot: characterize and optimize performance of your LLM inference services. *IEEE*. Retrieved from: https://arxiv.org/abs/2410.02425

Zhu, S., Supryadi, Xu, Sh., Sun, H., Pan L., Cui, M., Du, J., Jin, R., Branco, A., Xiong, D. (2024). Multilingual large language models: a systematic survey. *arXiv preprint*, arXiv:2411.11072 [cs.CL]. Retrieved from: https://arxiv.org/abs/2411.11072

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*. Retrieved from: https://discovery.ucl.ac.uk/id/eprint/10100504/1/2005.11401v1.pdf

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(*6*), 386–408.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. *Advances in Neural Information Processing Systems*, *6*, 737–744.

Xu, C., McAuley, J. (2023). A survey on dynamic neural networks for natural language processing. In *Findings of the Association for Computational Linguistics*: *EACL 2023* (pp. 2370–2381). Retrieved from: https://aclanthology.org/2023.findings-eacl.180.pdf

Об авторах

Бабина Ольга Ивановна – кандидат филологических наук, доцент, заведующий кафедрой лингвистики и перевода, Южно-Уральский государственный университет (национальный исследовательский университет), Россия, Челябинск, babinaoi@susu.ru

Быкова Анастасия Сергеевна – магистрант Института лингвистики и международных коммуникаций, Южно-Уральский государственный университет (национальный исследовательский университет), Россия, Челябинск, anastb6@mail.ru

About the authors

Babina Olga Ivanovna – Ph. D. of Philology, Associate Professor, Head of the Department of Linguistics and Translation, South Ural State University, Russia, Chelyabinsk, babinaoi@susu.ru

Bykova Anastassia Sergeevna – Master's Student of the Institute of Linguistics and International Communications, South Ural State University, Russia, Chelyabinsk, anastb6@mail.ru