

Научная статья  
УДК 004.891  
doi: 10.17586/2713-1874-2024-3-47-59

## ИНТЕЛЛЕКТУАЛЬНАЯ ПОДДЕРЖКА ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ В MES-СИСТЕМАХ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

*Наталья Викторовна Добренко<sup>1✉</sup>, Дмитрий Александрович Добренко<sup>2</sup>,  
Максим Валерьевич Улизько<sup>3</sup>*

<sup>1,2,3</sup>Университет ИТМО, Санкт-Петербург, Россия  
<sup>1</sup>graziokisa@yandex.ru<sup>✉</sup>, <https://orcid.org/0000-0001-6206-8033>  
<sup>2</sup>enotpalaskun@gmail.com, <https://orcid.org/0009-0006-1485-1166>  
<sup>3</sup>contrey1337@gmail.com, <https://orcid.org/0009-0001-2374-8025>  
Язык статьи – русский

**Аннотация:** Рассмотрен подход к разработке интерактивной диалоговой системы, которая обеспечивает интеллектуальную поддержку принятия управленческих решений в системе управления производственными процессами (MES, Manufacturing Execution Systems) металлургической компании. Цель работы – разработка чат-бота на базе большой языковой модели (LLM), который должен предоставлять разработчикам и пользователям MES-системы ответы на их вопросы по всему жизненному циклу MES-системы – от ее моделирования и настройки до обучения персонала. Рассмотрены два варианта LLM – модели Llama 2.7 и GPT-4. Предобученные модели были дообучены контексту металлургии с применением технологии RAG. Во время первичной настройки выявлены сильные галлюцинации модели при взаимодействии со сложными контекстными терминами. С целью их устранения модель была дополнена техническим словарем, который помог существенно сократить количество галлюцинаций. С помощью метрик ROUGE и перплексии было оценено качество работы моделей Llama и GPT-4. Внедрение технического словаря помогло исключить большую часть галлюцинаций и улучшило общий результат работы модели в среднем на 16%. Экспериментально показаны преимущества модели LLM Llama 2.7b для работы в указанной предметной области. Построенный чат-бот получил высокую оценку экспертов из предметной области, что свидетельствует о значительном потенциале применения данной технологии для интеллектуальной поддержки управления MES-системами в сфере металлургии и других узкоспециальных технических областях.

**Ключевые слова:** большие языковые модели, искусственный интеллект, чат-бот, Llama, GPT-4, MES системы, RAG

**Ссылка для цитирования:** Добренко Н. В., Добренко Д. А., Улизько М. В. Интеллектуальная поддержка принятия управленческих решений в mes-системах с использованием больших языковых моделей // Экономика. Право. Инновации. 2024. № 3. С. 47–59. <http://dx.doi.org/10.17586/2713-1874-2024-3-47-59>.

## INTELLIGENT MANAGEMENT DECISION SUPPORT IN MES SYSTEMS USING LARGE LANGUAGE MODELS

*Natalya V. Dobrenko<sup>1✉</sup>, Dmitry A. Dobrenko<sup>2</sup>, Maxim V. Ulizko<sup>3</sup>*

<sup>1,2,3</sup>ITMO University, Saint Petersburg, Russia  
<sup>1</sup>graziokisa@yandex.ru<sup>✉</sup>, <https://orcid.org/0000-0001-6206-8033>  
<sup>2</sup>enotpalaskun@gmail.com, <https://orcid.org/0009-0006-1485-1166>  
<sup>3</sup>contrey1337@gmail.com, <https://orcid.org/0009-0001-2374-8025>  
Article in Russian

**Abstract:** The article considers an approach to developing an interactive dialogue system that provides intelligent support for making management decisions in the manufacturing execution systems (MES) of a metallurgical company. The objective of the work is to develop a chatbot based on a large language model (LLM), which should provide developers and users of the MES system with answers to their questions throughout the life cycle of the MES system – from its modeling and configuration to personnel training. Two LLM variants are considered – the Llama 2.7 and GPT-4 models. The pre-trained models were further trained in the metallurgy context using RAG technology. During the initial configuration, strong hallucinations of the model were detected when interacting with complex contextual terms. To eliminate them, the model was supplemented with a technical dictionary, which helped to significantly reduce the number

of hallucinations. Using the ROUGE and perplexity metrics, the quality of the Llama and GPT-4 models was assessed. The introduction of a technical dictionary helped to eliminate most of the hallucinations and improved the overall performance of the model by an average of 16%. The advantages of the LLM Llama 2.7b model for working in the specified subject area were experimentally demonstrated. The built chatbot was highly rated by experts from the subject area, which indicates significant potential for the use of this technology for intelligent support of MES system management in the field of metallurgy and other highly specialized technical areas.

**Keywords:** artificial intelligence, chatbot, large language models, GPT-4, Llama, MES systems, RAG

**For citation:** Dobrenko N. V., Dobrenko D. A., Ulizko M. V. Intelligent Management Decision Support in Mes Systems Using Large Language Models. *Ekonomika. Pravo. Innovacii*. 2024. No. 3. pp. 47–59. (In Russ.). <http://dx.doi.org/10.17586/2713-1874-2024-3-47-59>.

**Введение.** Использование систем управления производственными процессами (MES, Manufacturing Execution Systems) является основным трендом развития современного высокотехнологичного производства [1]. MES-системы обеспечивают интеграцию и автоматизацию всех этапов производства, от планирования до выполнения, позволяя в реальном времени отслеживать и анализировать производственные данные, тем самым повышая производительность и сокращая финансовые и материальные издержки.

Однако MES-системы представляют собой многофункциональные программные решения, которые характеризуются большим количеством пользовательских интерфейсов и неочевидной логикой взаимодействия модулей системы. Они сложны для освоения рядовым пользователем, в качестве которого, как правило, выступает инженер-производственник, не имеющий образования в области компьютерных технологий. Более того, пользователям необходимо освоить не только интерфейс, но и знания о быстро изменяющихся производственных процессах, что в условиях ограниченного времени на обучение и необходимости быстрой адаптации становится весомой проблемой.

Все это в полной мере относится к металлургической промышленности, которая сталкивается с необходимостью внедрения инструментов интеллектуальной поддержки работы с MES-системой на всем протяжении ее жизненного цикла от ее моделирования и настройки до обучения персонала. Под такими системами в научных кругах понимают программные решения, которые позволяют принимать обоснованные решения для контекстной области, основанные на больших объемах данных [2]. Одним из решений для такого инструмента может стать чат-бот-

помощник на базе большой языковой модели (LLM, Large Language Model). Такой интерактивный инструмент сможет предоставлять пользователям мгновенные ответы на их вопросы, помогать в навигации по системе и объяснять функционал различных модулей. Чат-бот может быть доступным на различных устройствах, что предоставляет дополнительные преимущества конечным пользователям. Внедрение чат-бота позволит обеспечить круглосуточную поддержку пользователей, улучшить контроль над производственными процессами и, как следствие, сократить количество инцидентов на производстве. Система позволит собирать и анализировать данные о запросах, предпочтениях пользователей и внутренних процессах, что в дальнейшем будет способствовать увеличению эффективности производства.

Авторы статьи ставят задачу разработать чат-бот-помощник, который будет погружен в контекст предметной области металлургического производства. При этом необходимо:

- обеспечить высокое качество ответов, которое позволит пользователю без ошибок и дополнительного привлечения поддержки решить свой вопрос;

- свести к минимуму присущие большим языковым моделям галлюцинации, в особенности при работе со сложными техническими терминами. Под галлюцинациями понимаем такую информацию, которая является неправдоподобной, неточной или вымышленной, но при этом модель ее выдает как правдивую [3].

Выстроить ограничения для чат-бота: ответы должны быть написаны в деловом и вежливом стиле, должна отсутствовать неуместная креативность.

Полученный результат необходимо верифицировать с помощью группы экспертов,

которая в будущем будет пользоваться продуктом. Таким образом, можно отметить следующие элементы новизны и сделать вывод: LLM не применялся для классических MES систем в металлургии, а исследования в смежных областях не ставили вопрос улучшения понимания сложных технических терминов.

**Обзор литературы.** Чат-боты находят широкое применение в автоматизации обслуживания клиентов и технической поддержке [4]. Однако опыт использования чат-ботов непосредственно в MES-системах представлен в доступной авторам литературе лишь одной работой [5]. В ней описан опыт использования чат-бота с системой прогнозирования в качестве интерфейсного слоя для MES. Чат-бот нацелен на координацию производства, помогая работникам цеха и обучаясь на основе их запросов, таким образом выступая в качестве интеллектуального помощника.

Также сравнительно слабо представлен опыт применения больших языковых моделей в металлургии и смежных отраслях. В [6] подчеркивается потенциал интеграции LLM в системы промышленной автоматизации. В то же время, как отмечают авторы, установление соотношений между представлением знаний в промышленных программных системах и естественным языком является сложной задачей. Неточное или сопровождающееся потерями преобразование может привести к галлюцинациям агентов LLM для производственной системы и негативно повлиять на их производительность. Более того, сложно оценить, правильно ли агенты LLM преподнесли контекстную информацию, предоставленную в подсказке.

В статье [7] рассмотрены основные стратегии адаптации языковых моделей (на примере GPT-3) и их интеграции в производственный процесс обрабатывающей промышленности, в том числе этап сбора и обработки данных для обучения модели контексту. Одно из предлагаемых решений – предварительная обработка исходных документов и очистка их от мета-тегов.

В работе [8] представлен практический опыт внедрения LLM в реальных производственных условиях. Авторы используют модель GPT-3.5 в качестве ассистента поддержки работника на текстильном

производстве. Для технически сложного контекста они подчеркивают важность в разработке «системного промта»: текстовый запрос которого по умолчанию формируется вместе с пользовательским (например, «Вы – полезный помощник для работников завода», «Вы должны давать ответ длиной максимум в 50 слов»). Однако в статье не раскрывается подробно, каким образом базовая языковая модель дополняется контекстными данными.

В статье [9] авторы описывают опыт интеграции чат-бота в MES-систему и освещают трудности, с которыми столкнулись в ходе развертывания. Авторы разработали интеллектуальный чат-бот, который может помогать в работе персоналу цеха и сопровождать оператора в принятии решений на основе собранных данных. Разработчики сделали вывод, что необходимой процедурой для внедрения подобной системы является процесс постоянного итеративного дообучения модели с использованием профессионала из области контекста, которому обучается модель.

Анализируя решения, представленные на рынке, в том числе в сфере промышленности, авторы приводят статистику, согласно которой критически важными факторами в работе модели являются понимание сложного технического контекста, а также отсутствие галлюцинаций, способных ввести в заблуждение пользователя.

Авторы статьи [10] приводят опыт разработки чат-бота-помощника для приложения с телемедициной на основе модели LLM Llama-2.7. Разработчики сталкиваются с вопросом минимизации галлюцинаций в LLM в контексте сложных технических терминов и предлагают метод, который заключается в вычленении метаданных из запроса с последующим переиспользованием их в качестве промта.

Таким образом, актуальность применения чат-ботов в качестве средства интеллектуальной поддержки MES-систем и LLM как базы для построения чат-ботов не вызывает сомнений, но задачи их реализации в литературе остаются нераскрытыми. Более того, многие авторы подчеркивают проблемы, связанные с построением чат-ботов на основе LLM, в первую очередь галлюцинации, возникающие при обработке текстов, относящихся к сложным техническим предметным

областям. В ходе исследования было выявлено, что на сегодняшний день еще в недостаточной мере представлены работы детально рассматривающие вызовы и решения, сопровождающие использование LLM в контексте MES систем, в частности в металлургии. исследования было

**Выбор и настройка большой языковой модели.** Для разработки и тестирования чат-бота были выбраны большие языковые модели GPT-4 (Open AI) [11] и Llama 2.7 (Meta) [12] как самые перспективные с точки зрения гибкости к настройке под конкретную

предметную область. Для настройки контекста выбрана технология RAG (Retrieval-Augmented Generation) [13] – это метод, который позволяет языковой модели использовать данные из внешних источников для формирования ответов на запросы пользователей. RAG расширяет возможности модели обогащая их обновленной актуальной информацией из внешних баз данных и повышает способность модели отвечать на вопросы и предоставлять своевременные, уместные и учитывающие контекст ответы. На рисунке 1 представлена архитектура RAG решения.

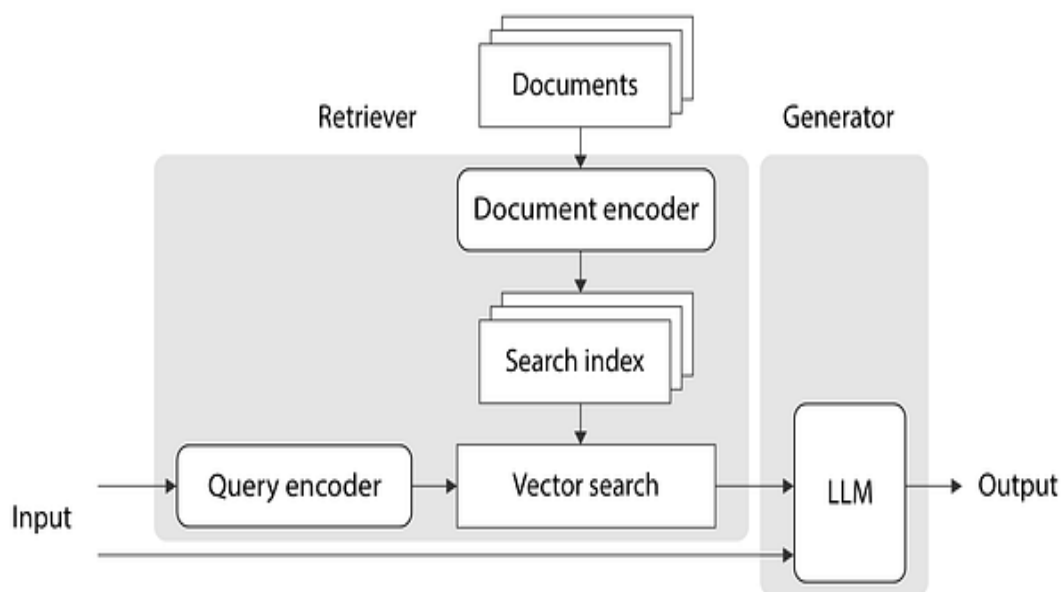


Рисунок 1 – Архитектура RAG технологии

Источник: составлено авторами на основе [22]

Поисковый компонент преобразует входной текст в последовательность чисел с плавающей запятой (вектор), используя кодировщик запроса. Далее он, используя единый подход, трансформирует каждый из документов, применяя кодировщик документов, после чего сохраняет закодированные документы в виде поискового индекса. Затем, в поисковом индексе, поисковый компонент выполняет поиск векторов документов, которые имеют отношение к входному вектору [14]. После этого он преобразует векторы документов обратно в их текстовое представление и возвращает эти тексты в качестве результата своей работы. Генератор, который принимает текст, введенный пользователем, и соответствующие ему документы, комбинирует все это в единый промпт и предлагает LLM дать ответ на вопрос пользователя с

учетом информации, имеющейся в найденных ранее документах. Результат выдачи LLM является выходными данными всей системы.

С помощью фреймворка HuggingFace [15] были импортированы заранее обученные модели LLM GPT-4 и Llama 2.70b.

После инициализации и импорта необходимых фреймворков произведена настройка и квантование модели. Использована языковая модель с авторегрессией – это тип модели машинного обучения, которая применяет методы авторегрессии для прогнозирования следующего слова в последовательности на основе предшествующих слов. Стратегия работы модели формировалась в процессе настройки параметров генерации, конфигурируемые параметры для Llama представлены на рисунке 2.

```

llm = HuggingFaceLLM(
    context_window=4096,
    max_new_tokens=256,
    generate_kwargs={"temperature": 0.0125, "do_sample": False},
    system_prompt=system_prompt,
    query_wrapper_prompt=query_wrapper_prompt,
    tokenizer_name="NousResearch/Llama-2-7b-chat-hf",
    model_name="NousResearch/Llama-2-7b-chat-hf",
    device_map="auto",
    model_kwargs={"torch_dtype": torch.float16, "load_in_8bit": True }
)

```

Рисунок 2 – Параметры настройки для Llama 2.70b

*Источник: составлено авторами*

Для импорта документов использовался фреймворк rupdf. В качестве входных данных для погружения модели в контекст были собраны более 150 документов, в среднем по 5 страниц каждый. Они представляют собой подробные руководства (инструкции) для пользователя как пользоваться системой. Документы описаны техническими специалистами. Также на этапе подготовки выделен набор вопросов, который использовался для тестирования системы.

Как показали эксперименты, на данном этапе обученная контексту модель тяжело воспринимает технический контекст. Модель

галлюцинирует, сталкиваясь со сложными терминами, характерными для предметной области (результаты тестирования представлены в таблицах 2–4), хотя информация о значениях этих терминов и аббревиатур содержалась во входных документах. Учитывая особенности работы языковых моделей, можно предположить, что на фоне массы другого контекста модель “теряет” знание об аббревиатурах и сложных технических терминах.

Для преодоления этого недостатка предложена улучшенная архитектура модели, представленная на рисунке 3.

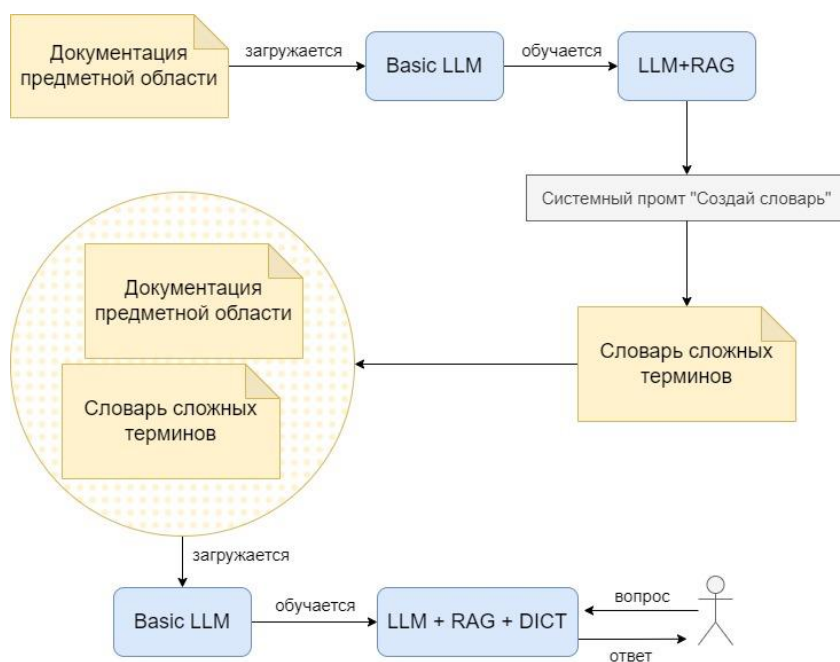


Рисунок 3 – Улучшенная архитектура модели

*Источник: составлено авторами*

Порядок настройки модели в этом случае представлен следующими шагами.

1) На вход в систему загружается исходный набор документов для обучения контексту.

2) Модель обучается контексту, после чего «сама себя» просит с помощью системного промта составить словарь сложных технических терминов.

3) На второй итерации модель заново обучается контексту, только теперь, к исходному набору документа добавляется

словарь терминов, полученный на предыдущем шаге.

4) Пользователь работает с моделью второй итерации.

**Оценка эффективности предложенной модели.** Для оценки эффективности предложенной улучшенной модели использованы следующие метрики.

1) ROUGE (Recall-Oriented Understudy for Gisting Evaluation) – набор метрик, используемых для оценки качества суммаризации текста [16]:

$$ROUGE(N) = \frac{\sum_{S \in \{RefSum\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RefSum\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1),$$

где  $n$  – длина  $n$ -граммы;  $gram_n$  и  $countmatch(gram_n)$  – максимальное количество  $n$ -грамм, одновременно встречающихся в содержании образца и наборе референса для него, RefSum – «эталонные» значения. Набор ROUGE сравнивает генерируемое краткое изложение

с одним или несколькими эталонными суммаризациями и вычисляет точность, полноту и F1-меру.

Показатели ROUGE дают представление о способности модели генерировать краткие изложения [17].

Таблица 1

### Метрика ROUGE

Источник: составлена авторами

Метрика	Совпадение
ROUGE-1	только отдельных слов
ROUGE-2	двух последовательно идущих слов
ROUGE-N	$n$ -последовательно идущих слов
ROUGE-L	максимально длинная совпадающая последовательность слов

2) Перплексия – распространенная метрика для оценки производительности языковых моделей. Она измеряет, насколько хорошо модель предсказывает последователь-

ность текста. Меньшее значение перплексии указывает на лучшую производительность модели [18].

$$P = 2^{-\frac{1}{N} \sum_{i=1}^N \log_2 P(w_i | w_1, w_2, \dots, w_{i-1})} \quad (2),$$

где  $N$  – количество слов в последовательности,  $P(w_i | w_1, w_2, \dots, w_{i-1})$  – вероятность слова  $w_i$  при наличии предыдущих слов  $w_1, w_2, \dots, w_{i-1}$ .

3) Экспертная оценка (Human Evaluation) [19, 20]. В качестве экспертов выступали три

инженера предприятия, которые являются профессионалами металлургической отрасли и в целевом процессе будут являться пользователями MES системы. Экспертам было предложено оценить ответы модели на основе следующих критериев: релевантность

(понимание контекста), плавность (уровень качество по шкале от 0 до 10 (больше – «естественности» языка), связность и общее лучше).

Таблица 2

## Оценка работы моделей на вопросе 1

Вопрос	1. Как создать конфигурацию «ОМ»?			
	Лlama-2-7b	Лlama-2-7b + словарь	GPT-4	GPT-4 + словарь
	1	2	3	4
<b>Ответ</b>	<p>Чтобы создать конфигурацию для модуля управления объектами (ОМ), выполните следующие действия:</p> <ol style="list-style-type: none"> <li>1. Нажмите кнопку «Создать конфигурацию» в правом верхнем углу страницы.</li> <li>2. Выберите опцию «Управление объектами» в раскрывающемся меню.</li> <li>3. Выберите тип конфигурации, которую хотите создать (например, «Управление объектами»).</li> <li>4. В поле «Категория объектов» выберите категорию объектов, которыми хотите управлять (например, «Корпоративный»).</li> <li>5. В поле «Имя объекта» введите имя объекта, которым хотите управлять.</li> <li>6. Нажмите кнопку «Создать», чтобы сохранить конфигурацию.</li> </ol> <p>Примечание. Точные шаги могут различаться в зависимости от конкретной версии используе-</p>	<p>Конфигурация объектной модели включает в себя configurator атрибутов и configurator объектов. Конфигурирование следует начать с configurator атрибутов:</p> <ol style="list-style-type: none"> <li>1. Нажмите кнопку «Создать конфигурацию» в правом верхнем углу страницы.</li> <li>2. Выберите опцию «Управление объектами» в раскрывающемся меню.</li> <li>3. Выберите тип конфигурации «Атрибут».</li> <li>4. В поле «Категория объектов» выберите категорию объектов, которыми хотите управлять.</li> <li>5. В поле «Имя объекта» введите имя объекта, которым хотите управлять.</li> <li>6. Нажмите кнопку «Создать», чтобы сохранить конфигурацию.</li> </ol>	<p>Конфигурация объектной модели: Конфигуратор атрибутов</p> <p>Чтобы начать конфигурирование с configurator атрибутов, выполните следующие шаги:</p> <ol style="list-style-type: none"> <li>1. Нажмите кнопку «Создать конфигурацию» в правом верхнем углу страницы.</li> <li>2. Выберите опцию «Управление объектами» в раскрывающемся меню.</li> <li>3. Выберите тип конфигурации «Атрибут».</li> <li>4. В поле «Категория объектов» выберите категорию объектов, которыми хотите управлять.</li> <li>5. В поле «Имя объекта» введите имя объекта, которым хотите управлять.</li> <li>6. Нажмите кнопку «Создать», чтобы сохранить конфигурацию.</li> </ol> <p>После того как конфигурация атрибутов будет создана, можно переходить к configurator объектов для дальнейшей настройки объектной модели.</p>	<p>Создание конфигурации для модуля управления объектами (ОМ)</p> <ol style="list-style-type: none"> <li>1. Нажмите кнопку «Создать конфигурацию» в правом верхнем углу страницы.</li> <li>2. Выберите опцию «Управление объектами» в раскрывающемся меню.</li> <li>3. Выберите тип конфигурации, которую хотите создать (например, «Управление объектами»).</li> <li>4. В поле «Категория объектов» выберите категорию объектов, которыми хотите управлять (например, «Корпоративный»).</li> <li>5. В поле «Имя объекта» введите имя объекта, которым хотите управлять.</li> <li>6. Нажмите кнопку «Создать», чтобы сохранить конфигурацию.</li> </ol>

Продолжение таблицы 2

	1	2	3	4
	мого вами программного обеспечения. Более подробные инструкции см. в документации к программному обеспечению.		Примечание: Пожалуйста, учтите, что точные шаги могут различаться в зависимости от версии вашего программного обеспечения. Для получения более детальных инструкций обратитесь к документации программы.	
<b>Оценка (ROUGE F1) Больше – лучше</b>	ROUGE-1: 0.19 ROUGE-2: 0.030 ROUGE-L: 0.017	ROUGE-1: 0.41 ROUGE-2: 0.074 ROUGE-L: 0.028	ROUGE-1: 0.18 ROUGE-2: 0.032 ROUGE-L: 0.018	ROUGE-1: 0.44 ROUGE-2: 0.076 ROUGE-L: 0.029
<b>Оценка (перплексия) Меньше – лучше</b>	4.8515	4.651	4.86	4.666
<b>Оценка эксперта Больше – лучше</b>	Релевантность: 9 Качество речи: 6 Согласованность: 4 Общая: 6,3	Релевантность: 9 Качество речи: 8 Согласованность: 8 Общая: 8,3	Релевантность: 6 Качество речи: 5 Согласованность: 4	Релевантность: 9 Качество речи: 8 Согласованность: 8
<b>Общий комментарий</b>	С помощью словаря улучшилась согласованность написанного текста, модель лучше стала понимать контекст и исключила галлюцинации. В среднем качество ответа возросло на 17%. GPT-4 и Llama имеют схожий показатель качества.			

Численная оценка каждого теста рассчитывалась как:

$$n = \left( \frac{AvgRouge_{M1}}{AvgRouge_{M2}} + \frac{Perplexity_{M1}}{Perplexity_{M2}} + \frac{ExpScore_{M1}}{ExpScore_{M2}} \right) * 100 \quad (3),$$

где M1 – модель без словаря, M2 – модель со словарем, AvgRouge – средняя оценка по ROUGE, Perplexity – перплексия, ExpScore – экспертная оценка (критерий «Общая»).

Общая оценка рассчитывалась как средняя между всеми тестами:

$$T_{avg} = \frac{T_1 + T_2 + T_3}{3} \quad (4),$$

где T – результат теста. Сопоставление результатов расчета по формулам (3) и (4) для стандартной модели и модели со словарем показало, что модель со словарем получила оценку в среднем на 16% выше, чем модель без словаря.

**Интерфейсная часть чат-бота.** Как отмечают авторы статьи [21], если перед компанией стоит задача при разработке приложения покрыть максимально широкую аудиторию, то разработка нативных версий под каждое устройство займет огромное количество времени и потребует привлечения большого числа специалистов разных направлений, решением данной проблемы может являться разработка кросс-платформенного web-приложения. В связи с этим в качестве интерфейса выбран корпоративный мессенджер компании, работающий в WEB. API (Application Program Interface) [18] мессенд-



жера позволяет без дополнительной разработки подключить модель в качестве собеседника в чат. Пользователь вводит свой вопрос как сообщение в чат, его запрос

уходит на сервер для обработки моделью, далее модель возвращает свой ответ в чат. На рисунке 4 представлен пример интерфейса чата.

Таблица 3

### Оценка работы моделей на вопросе 2

Источник: составлено авторами

Вопрос	2. Как найти конфигурацию в MES?			
	Лlama-2-7b	Лlama-2-7b + словарь	GPT-4	GPT-4 + словарь
	1	2	3	4
Ответ	<p>Для поиска конфигураций в MES вы можете выполнить следующие действия:</p> <ol style="list-style-type: none"> <li>1. Откройте папку «PrjImport» на вашем диске.</li> <li>2. Нажмите на «Page Label» 6, чтобы получить доступ к странице конфигурации.</li> <li>3. Нажмите на кнопку «Search», расположенную в правом верхнем углу страницы.</li> <li>4. В поле поиска введите имя группы конфигураций, которую вы хотите найти.</li> <li>5. Нажмите на кнопку «Search», чтобы отобразить результаты.</li> <li>6. Выберите нужную группу конфигураций из списка результатов.</li> </ol> <p>Кроме того, вы также можете использовать компонент «Кнопка», чтобы добавить кнопку поиска на страницу конфигурации. Для этого выполните следующие действия.</p>	<p>Для поиска конфигураций в MES вы можете выполнить следующие действия:</p> <ol style="list-style-type: none"> <li>1. Нажмите на область «Поиск» в интерфейсе МУК.</li> <li>2. Введите наименование конфигурации.</li> <li>3. Нажмите на кнопку «Search».</li> </ol> <p>Надеюсь, это поможет! Дайте мне знать, если у вас возникнут дополнительные вопросы.</p>	<p>Чтобы добавить кнопку поиска на страницу конфигурации в MES, выполните следующие шаги:</p> <ol style="list-style-type: none"> <li>1. Перейдите к редактору страницы конфигурации: Убедитесь, что вы находитесь на нужной странице конфигурации, где хотите добавить кнопку.</li> <li>2. Выбор компонента: Найдите панель инструментов или меню компонентов, где доступны элементы управления для страницы.</li> <li>3. Добавление кнопки: <ul style="list-style-type: none"> <li>– Найдите компонент «Кнопка» в списке доступных элементов.</li> <li>– Перетащите его на область страницы конфигурации, куда вы хотите добавить кнопку.</li> </ul> </li> <li>4. Настройка кнопки: <ul style="list-style-type: none"> <li>– Установите текст кнопки (например, «Поиск конфигураций»).</li> <li>– Назначьте действие, которое должно выполняться при нажатии на кнопку. Обычно это связано с вызовом функции поиска, которая будет инициировать процесс, описанный в предыдущих шагах.</li> </ul> </li> <li>5. Свяжите кнопку с функцией поиска: <ul style="list-style-type: none"> <li>– Откройте свойства кнопки и найдите поле для назначения действия или вызываемой функции.</li> <li>– Убедитесь, что действие связано с выполне-</li> </ul> </li> </ol>	<p>Для поиска конфигураций в MES можно выполнить следующие действия: Нажмите на область «Поиск» в строящем интерфейсе, введите имя конфигурации.</p>

Продолжение таблицы 3

	1	2	3	4
			нием поиска в поле, где вводится имя группы конфигураций. 6. Сохраните изменения: После настройки кнопки обязательно сохраните изменения на странице конфигурации.	
<b>Оценка (ROUGE F1)</b>  Больше – лучше	ROUGE-1: 0.01 ROUGE-2: 0.003 ROUGE-L: 0.009	ROUGE-1: 0.4 ROUGE-2: 0.105 ROUGE-L: 0.033	ROUGE-1: 0.01 ROUGE-2: 0.003 ROUGE-L: 0.008	ROUGE-1: 0.41 ROUGE-2: 0.075 ROUGE-L: 0.028
<b>Оценка (перплексия)</b>  Меньше – лучше	6.745	2.111	6.159	2.453
<b>Оценка эксперта</b>  Больше – лучше	Релевантность: 0 Качество речи: 0 Согласованность: 0 Общая: 0	Релевантность: 7 Качество речи: 9 Согласованность: 8 Общая: 8	Релевантность: 0 Качество речи: 0 Согласованность: 0 Общая: 0	Релевантность: 6 Качество речи: 8 Согласованность: 7 Общая: 7
<b>Общий комментарий</b>	Модель без обучения со словарем выдала ответ, на большую часть состоящий из случайного набора предложений. После обучения со словарем удалось получить релевантный ответ (улучшение на 100%). Модель под управлением GPT-4 показала немного худший результат, чем Llama.			

**Заключение.** Проведенное исследование подтвердило эффективность обучения модели LLM с учётом контекста (RAG) в качестве чат-бот помощника для MES-системы. Внедрение технического словаря помогло исключить большую часть галлюцинаций и улучшило общий результат работы модели в среднем на 16%. В качестве целевой модели была выбрана Llama 2.7b, так как показала лучшую работу на практических тестах. Чат-бот получил высокую оценку экспертов из

предметной области, что свидетельствует о значительном потенциале применения данной технологии в сфере металлургии и других технических областях. В качестве ближайшей перспективы развития ожидается внедрение модуля, который на основе данных, собранных в процессе промышленной эксплуатации, сможет выделять недостающий функционал системы, который требуется в первую очередь для эффективной работы MES системы.

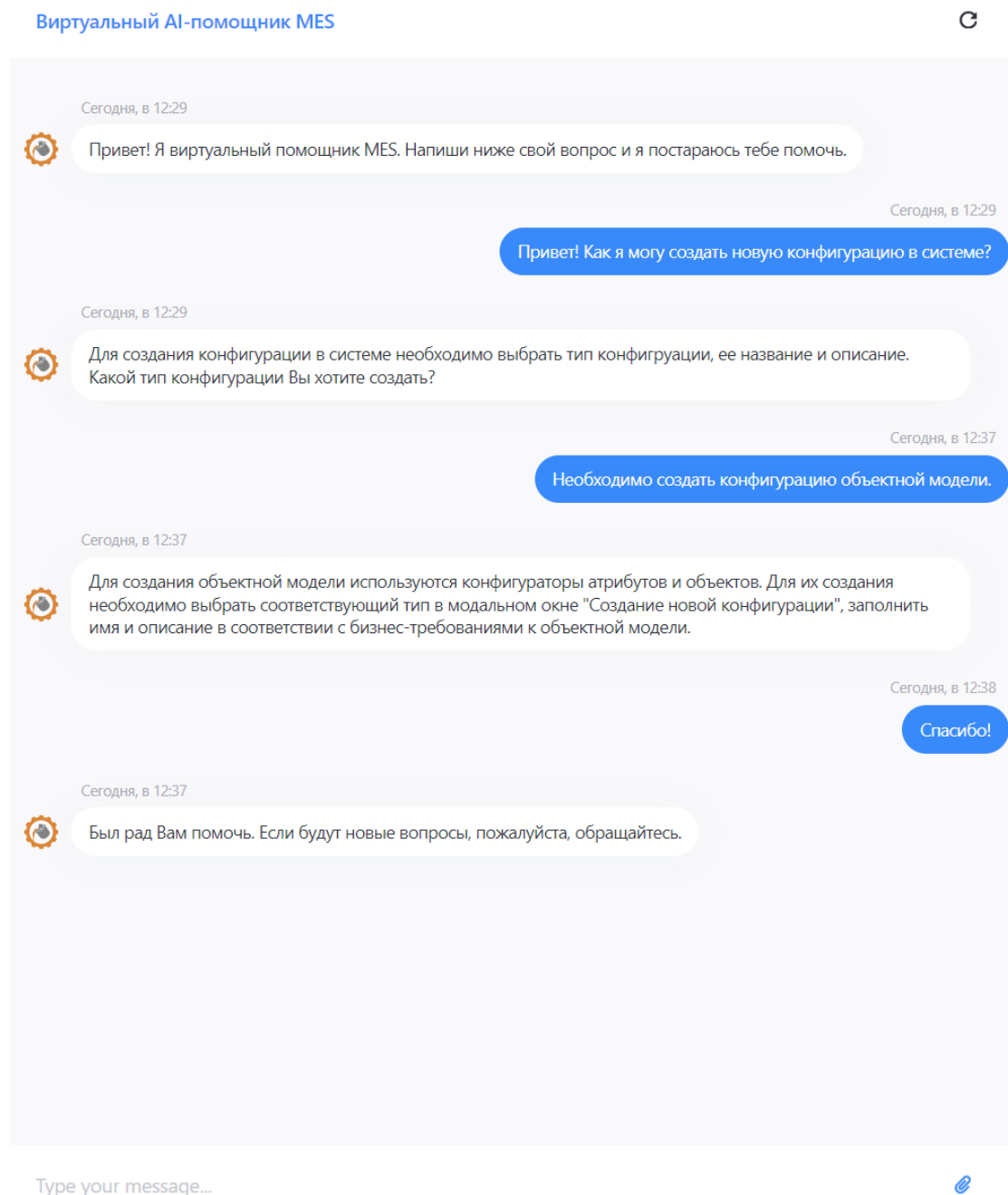


Рисунок 4 – Интерфейс чат-бота

#### Список источников

1. Shojaeinasab A., et al. Intelligent manufacturing execution systems: A systematic review // *Journal of Manufacturing Systems*. Январь, 2022. Т. 62. С. 503–522. (In Eng.).
2. Gupta J. N., Forgionne G. A., Mora M. (Eds.). Intelligent decision-making support systems: foundations, applications and challenges [Электронный ресурс]. – Режим доступа: <https://doi.org/10.1007/1-84628-231-4> (In Eng.).
3. Hernandez-Salinas B., Terven J., ChaveZ-Urbiola E. A., Cordova-Esparza D. M., Romero-Gonzalez J. A., Arguelles A., Cervantes, I. IDAS: Intelligent driving assistance system using RAG // *IEEE Open*

#### References

1. Shojaeinasab A., et al. Intelligent Manufacturing Execution Systems: A Systematic Review. *Journal of Manufacturing Systems*. January 2022. Vol. 62. pp. 503–522.
2. Gupta J. N., Forgionne G. A., Mora M. (Eds.). Intelligent Decision-Making Support Systems: Foundations, Applications and Challenges. Available at: <https://doi.org/10.1007/1-84628-231-4>
3. Hernandez-Salinas B., Terven J., ChaveZ-Urbiola E. A., Cordova-Esparza D. M., Romero-Gonzalez J. A., Arguelles A., Cervantes, I. IDAS: Intelligent Driving Assistance System using RAG. *IEEE Open*

Journal of Vehicular Technology. 2024. T. 5. C. 1139–1165. (In Eng.).  
DOI: 10.1109/OJVT.2024.3447449.

4. Shi F., et al. Leveraging the power of large language models to drive progress in the manufacturing industry // 9th International Conference on Financial Innovation and Economic Development. 2024. C. 125–133. (In Eng.).

5. Adam M., Wessel M., Benlian A. AI-based chatbots in customer service and their effects on user compliance // *Electron Markets*. 2021. № 31. C. 427–445 (In Eng.). DOI: 10.1007/s12525-020-00414-7.

6. Mantravadi S., Jansson A.D., Møller C. User-friendly MES interfaces: Recommendations for an AI-based chatbot assistance in Industry 4.0 shop floors // *Intelligent Information and Database Systems – Springer, Lecture Notes in Artificial Intelligence*, 2020. – C.189–201. (In Eng.).

7. Yuchen X., et al. Towards autonomous system: flexible modular production system enhanced with large language model agents. September 12<sup>th</sup> to 15<sup>th</sup> // *ETFA*. 2023. (In Eng.).

8. Kernan Freire S., et al. Harnessing large language models for cognitive assistants in factories // *Proceedings of the 5<sup>th</sup> International Conference on Conversational User Interfaces*. 2023. C. 1–6. (In Eng.). DOI: 10.1145/3571884.3604313.

9. Mantravadi S., Jansson A. D., Møller, C. User-friendly mes interfaces: Recommendations for an ai-based chatbot assistance in Industry 4.0 shop floors // *Asian Conference on Intelligent Information and Database Systems*. 2020. C. 189–201. (In Eng.). DOI: 10.1007/978-3-030-42058-1\_16.

10. Gams M., et al. Developing a medical chatbot: Integrating medical knowledge into GPT for healthcare applications // *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*. 2024. C. 88–97. (In Eng.). DOI: 10.3233/AISE240018.

11. Abdullah M., Madain A., Jararweh Y. ChatGPT: Fundamentals, applications and social impacts // *Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 2022. C. 1–8. (In Eng.). DOI: 10.1109/SNAMS58071.2022.10062688.

12. Roumeliotis, K. I.; Tselikas, N. D.; Nasiopoulos, D. K. Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model. *Preprints 2023*, 2023072142. (In Eng.). DOI: 10.20944/preprints202307.2142.v1.

13. Li J., Yuan Y., Zhang Z. Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. 2024. (In Eng.). DOI: 10.48550/arXiv.2403.10446.

14. Fan W., et al. A Survey on RAG Meeting LLMs: Towards retrieval-augmented large language models

*Journal of Vehicular Technology*. 2024. Vol. 5. pp. 1139–1165.  
DOI: 10.1109/OJVT.2024.3447449.

4. Shi F., et al. Leveraging the Power of Large Language Models to Drive Progress in the Manufacturing Industry. *9th International Conference on Financial Innovation and Economic Development*. 2024. pp. 125–133.

5. Adam M., Wessel M., Benlian A. AI-based Chatbots in Customer Service and Their Effects on User Compliance. *Electron Markets*. 2021. No. 31. pp. 427–445. DOI: 10.1007/s12525-020-00414-7.

6. Mantravadi S., Jansson A.D., Møller C. User-friendly MES Interfaces: Recommendations for an AI-based Chatbot Assistance in Industry 4.0 Shop Floors. *Intelligent Information and Database Systems. Springer, Lecture Notes in Artificial Intelligence*. 2020. pp. 189–201.

7. Yuchen X., et al. Towards Autonomous System: Flexible Modular Production System Enhanced with Large Language Model Agents. September 12<sup>th</sup> to 15<sup>th</sup>. *ETFA*. 2023.

8. Kernan Freire S., et al. Harnessing Large Language Models for Cognitive Assistants in Factories. *Proceedings of the 5<sup>th</sup> International Conference on Conversational User Interfaces*. 2023. pp. 1–6. DOI: 10.1145/3571884.3604313.

9. Mantravadi S., Jansson A. D., Møller, C. User-friendly Mes Interfaces: Recommendations for an AI-based Chatbot Assistance in Industry 4.0 Shop Floors. *Asian Conference on Intelligent Information and Database Systems*. 2020. pp. 189–201. DOI: 10.1007/978-3-030-42058-1\_16.

10. Gams M., et al. Developing a Medical Chatbot: Integrating Medical Knowledge into GPT for Healthcare Applications. *Intelligent Environments 2024: Combined Proceedings of Work-shops and Demos & Videos Session*. 2024. pp. 88–97. DOI: 10.3233/AISE240018.

11. Abdullah M., Madain A., Jararweh Y. ChatGPT: Fundamentals, Applications and Social Impacts. *Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 2022. pp. 1–8. DOI: 10.1109/SNAMS58071.2022.10062688.

12. Roumeliotis, K. I.; Tselikas, N. D.; Nasiopoulos, D. K. Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model. *Preprints 2023*, 2023072142. DOI: 10.20944/preprints202307.2142.v1.

13. Li J., Yuan Y., Zhang Z. Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. 2024. DOI: 10.48550/arXiv.2403.10446.

14. Fan W., et al. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language

- // Proceedings of the 30<sup>th</sup> ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024. С. 6491–6501. (In Eng.).
15. Castaño J., et al. Analyzing the evolution and maintenance of ML models on Hugging Face // Proceedings of the 21<sup>st</sup> International Conference on Mining Software Repositories. 2024. С. 607–618. (In Eng.).
  16. Lin C.Y. ROUGE: A package for automatic evaluation of summaries [Электронный ресурс]. – Режим доступа: <https://aclanthology.org/W04-1013.pdf> (In Eng.).
  17. Hu T., Zhou X. H. Unveiling LLM evaluation focused on metrics: challenges and solutions. 2024 [Электронный ресурс]. – Режим доступа: <https://doi.org/10.48550/arXiv.2404.09135>. (In Eng.).
  18. Hu Y., et al. Can perplexity reflect large language model's ability in long text understanding? [Электронный ресурс]. – Режим доступа: <https://doi.org/10.48550/arXiv.2405.06105>. (In Eng.).
  19. Shankar S., et al. Who Validates the Validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. 2024 [Электронный ресурс]. – Режим доступа: <https://doi.org/10.48550/arXiv.2404.12272> (In Eng.).
  20. Huang H., et al. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. 2024 [Электронный ресурс]. – Режим доступа: <https://arxiv.org/html/2403.02839v1> (In Eng.).
  21. Пчелкин А.Ю., Гусарова Н.Ф. Кроссплатформенная разработка на базе вебтехнологий для поддержки решений в проблемно-ориентированных системах управления // Экономика. Право. Инновации. 2022. № 1. С. 41–47.  
DOI: 10.17586/2713-1874-2022-1-41-47.
  22. Mr-Pickles. Добавление собственных данных в LLM с помощью RAG // Интернет-портал Хабр [Электронный ресурс]. – Режим доступа: <https://habr.com/ru/companies/wunderfund/articles/779748/>
  - Models. *Proceedings of the 30<sup>th</sup> ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024. pp. 6491–6501.
  15. Castaño J., et al. Analyzing the evolution and maintenance of ML models on Hugging Face. *Proceedings of the 21<sup>st</sup> International Conference on Mining Software Repositories*. 2024. pp. 607–618.
  16. Lin C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. Available at: <https://aclanthology.org/W04-1013.pdf>
  17. Hu T., Zhou X. H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. 2024. Available at: <https://doi.org/10.48550/arXiv.2404.09135>.
  18. Hu Y., et al. Can perplexity reflect large language model's ability in long text understanding? Available at: <https://doi.org/10.48550/arXiv.2405.06105>.
  19. Shankar S., et al. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. 2024. Available at: <https://doi.org/10.48550/arXiv.2404.12272>
  20. Huang H., et al. An Empirical Study of Llm-As-A-Judge for Llm Evaluation: Fine-Tuned Judge Models are Task-Specific Classifiers. 2024. Available at: <https://arxiv.org/html/2403.02839v1>
  21. Pchelkin A. Yu., Gusarova N. F. Cross-Platform Development Based on Web Technologies to Support Solutions in Problem-Oriented Management Systems. *Ekonomika. Pravo. Innovacii*. 2022. No. 1. pp. 41–47. (In Russ.).  
DOI: 10.17586/2713-1874-2022-1-41-47.
  22. Mr-Pickles. (2023, December 11). Adding Your Own Data to LM Using RAG. *Internet portal Habr*. Available at: <https://habr.com/ru/companies/wunderfund/articles/779748/> (In Russ.).