УДК 004.8:681.5

***Ли Цзымин***1,

аспирант

***Ковальчук Елена Александровна***2,

специалист 1 категории

***Мишра Аншуман***3,

аспирант

***Потехин Вячеслав Витальевич***4,

доцент, **канд. техн. наук.**

**ИНТЕРАКТИВНЫЙ МЕТОД ДИАГНОСТИКИ ПРОМЫШЛЕННЫХ НЕИСПРАВНОСТЕЙ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ**

1,3,4 Россия, г. Санкт-Петербург, ФГАОУ ВО «Санкт-Петербургский политехнический университет Петра Великого», 1 li66.ts@edu.spbstu.ru,

3 mishra.a@edu.spbstu.ru, 4 slava.potekhin@spbstu.ru

2 Россия, г. Санкт-Петербург, ООО «СТЦ», ekovalchuk@stc-spb.ru

***Аннотация.*** В ответ на сложные задачи диагностики неисправностей в эпоху Индустрии 4.0 в данной статье предлагается интеллектуальная интерактивная система на основе крупных языковых моделей или Large Language Model (LLM). Система использует механизм RAG (Retrieval-Augmented Generation) для интеграции знаний из промышленной сферы и получения текстовой информации из различных источников, включая базы данных и системы распознавания изображений. Это позволяет точно диагностировать неисправности. Благодаря предоставлению структурированных диагностических результатов данное решение призвано значительно повысить эффективность управления оборудованием и возможности прогнозирования технического обслуживания. В настоящее время успешно создана архитектура LLM на основе Google gemini-1.5-flash и ведется оптимизация системы RAG.

***Ключевые слова:*** искусственный интеллект; большие языковые модели; диагностика неисправностей; предиктивное обслуживание; RAG-системы; интерактивные методы.

UDC 004.8:681.5

***Ziming Li***1,

Phd Student;

***Elena A. Kovalchuk*** 2

1st Category Specialist;

***Anshuman Mishra***3,

PhD student;

***Vyacheslav V. Potekhin***4***,***

Associate Professor, PhD of Technical Sciences

**INTERACTIVE METHOD FOR DIAGNOSING INDUSTRIAL FAULTS BASED ON LARGE LANGUAGE MODELS**

1,3,4 Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia, 1 li66.ts@edu.spbstu.ru, 3 mishra.a@edu.spbstu.ru, 4 slava.potekhin@spbstu.ru

2 "STC" LLC, St. Petersburg, Russia, ekovalchuk@stc-spb.ru

***Abstract.*** In response to the complex challenges of fault diagnosis in the Industry 4.0 era, this article proposes an intelligent interactive system based on large language models (LLMs). The system uses the RAG (Retrieval-Augmented Generation) mechanism to integrate industrial knowledge bases and obtain textual information from various sources, such as databases and image recognition systems, for accurate fault diagnosis. By providing structured diagnostic results, this solution is designed to significantly improve equipment management efficiency and predictive maintenance capabilities. Currently, an LLM architecture based on Google gemini-1.5-flash has been successfully created, and the RAG system is being optimized.

***Keywords:*** artificial intelligence; large language models; fault diagnosis; predictive maintenance; RAG systems; interactive methods.

В эпоху быстрого развития Индустрии 4.0 предприятия сталкиваются с беспрецедентными возможностями и вызовами. Из-за своего непосредственного влияния на стабильность производственных процессов, проблемы диагностики и обслуживания, являются одними из наиболее приоритетных для предприятий любого масштаба. Ввиду стремительного развития технологий, традиционные методы диагностики и обслуживания зачастую становятся недостаточными для удовлетворения потребностей современного производства. [1]

Традиционные методы диагностики и обслуживания, промышленного оборудования обычно базируются на заранее заданных расписаниях технического обслуживания и регулярных проверках. Эти методы часто оказываются неэффективными в предсказании неожиданных сбоев или выявлении начальных стадий деградации. Также, эти методы не учитывают изменения условий эксплуатаций. [2]

Ранее существующие методы диагностики промышленных неисправностей играли важную роль. [3], но в условиях взрывного роста объемов разнородных данных (таких как данные датчиков, записи о техническом обслуживании, журналы операций и т. д.) и потребности в более интеллектуальной и активной диагностике их традиционный режим работы постепенно показывает свои ограничения [4].

Некоторые методы прогнозирования неисправностей с использованием ИИ уже существуют [5, 6], но они в значительной степени опираются на данные, собранные датчиками, такими как температура и давление, и являются более сложными для пользователей в практическом применении. Руководство по устранению неисправностей, предоставленное производителем, носит слишком общий характер и не позволяет провести детальный анализ с использованием конкретных данных. LLM может эффективно снизить порог входа для пользователей, обрабатывая текстовую информацию на естественном языке, а также изображения, базы данных и другие данные.

Рис. 1. Рабочий процесс системы

Основная цель этого проекта – разработка интеллектуальной системы, способной точно диагностировать и прогнозировать неисправности промышленного оборудования. Эта система построена на основе больших языковых моделей (LLM), которые служат центральным интеллектом для интерпретации сложных описаний неисправностей оборудования на естественном языке.

Важным элементом, способствующим эффективности системы, является ее специализированная база промышленных знаний, заключенная в настраиваемый набор данных industrial\_faults.csv. Данная база хранит структурированную информацию о различных промышленных неисправностях и способах их диагностирования, структура представляется в виде:

1. Типа оборудования: конкретное промышленное оборудование.
2. Симптомов неисправности: наблюдаемые признаки проблем.
3. Возможных причинах: распространенные основные причины неисправностей.
4. Этапах диагностики: рекомендуемые процедуры для выявления основных причин.
5. Решения: практические меры по ремонту или техническому обслуживанию.

Чтобы значительно повысить точность диагностики и обеспечить ее актуальность, мы внедряем структуру Retrieval-Augmented Generation (RAG). Этот передовой подход позволяет LLM динамически извлекать и синтезировать информацию из нашей базы знаний industrial\_faults.csv. Предоставляя LLM контекстуально релевантные данные, RAG гарантирует, что диагнозы не только генерируются интеллектуально, но и прочно основаны на устоявшихся промышленных практиках и исторических данных о неисправностях.

图形用户界面, 文本, 应用程序

描述已自动生成

Рис. 2. Веб-страница для ввода и вывода данных

Была успешно создана базовая архитектура на основе LLM и заполнена специализированная база знаний о промышленных неисправностях. Первоначальные испытания подтверждают способность системы предоставлять релевантные и последовательные диагностические рекомендации. Также была разработана веб-страница, которая облегчает ввод информации и отображение результатов в процессе тестирования, как показано на рисунке 2. В дальнейшем ее можно будет модифицировать в соответствии с требованиями API для подключения к другим программным системам.

В ходе тестирования была создана база знаний (cases.csv – удобный для редактирования онлайн-документ) на основе раздела по диагностике неисправностей в руководстве по CLICK ПЛК [7], которая строго следует шестиколоночной структуре: «Название устройства», «Тип неисправности», «Симптомы неисправности», «Возможные причины», «Этапы диагностики» и «Решение», в соответствии с информацией из руководства в базу знаний добавлены 20 строк с информацией о неисправностях, как показано на рисунке 3.

图片包含 日历

描述已自动生成

Рис. 3. База знаний (выдержка)

Затем было проведено тестирование функциональности и эффективности интерактивной диагностической системы на основе модели LLM и архитектуры RAG через веб-интерфейс. Promt, который вводится в LLM, показан на рисунке 4. В ходе тестирования была проведена непосредственная проверка базы знаний. Для этого были выбраны типичные случаи неисправностей, такие как неработающий индикатор питания (PWR) CLICK ПЛК и работающий индикатор выходного модуля без выхода. Соответствующие описания симптомов из руководства были точно введены в систему.文本

描述已自动生成

Рис. 4. Promt на ввод

В рамках исследования было проведено нечеткое тестирование, включающее описание неисправностей с использованием естественного языка (более разговорного, выполненное носителем русского языка). Это позволило имитировать ввод данных пользователем в реальных условиях эксплуатации. Результаты тестирования зафиксированы в журнале сервера, пример которого приведен в таблице 1.

Таблица 1

**Пример проверки валидности базы знаний**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Тестовый объект | Информация в базе знаний | Фаза тестирования | Введите информацию о неисправности | Обзор выходных данных | Результат точный/неточный |
| ПЛК CLICK | Индикатор PWR (питание) не горит | Прямое тестирование | Индикатор ERR (ошибка) МИГАЕТ | Соответствовать базе знаний | точный |
| Нечеткое соответствие | Лампочка ошибки светится красным без остановки | Светодиод ошибки мигает; лампочки ошибки периодично загорается красным и потухает; лампочка ошибки моргает красным цветом | точный |

Результат вывода показан на рисунке 5. В ходе тестирования были последовательно проверены все 20 строк данных в базе знаний. Точность прямого тестирования составила 100%, а точность нечёткого соответствия — 85%. Результаты показали, что информация, содержащаяся в диагностическом отчёте, сгенерированном LLM, в значительной степени совпадает с данными, представленными в базе знаний. Это подтверждает способность системы RAG точно извлекать и использовать информацию из файла cases.csv. Это свидетельствует о том, что LLM не выдвигает независимые гипотезы, а эффективно применяет имеющиеся структурированные знания в качестве фундамента для анализа и диагностики.

Результаты тестирования подтвердили эффективность предложенной интерактивной диагностической системы на основе LLM и RAG при использовании структурированной внешней базы знаний (БЗ). Экспериментально доказан потенциал архитектуры RAG для повышения точности LLM в промышленных приложениях и ключевая роль полноты БЗ в производительности системы. При диагностике неисправностей, отсутствующих в БЗ (например, «аномальное срабатывание всех выходных реле без программной команды»), LLM, опираясь на общие инженерные знания, предлагает лишь общие гипотезы (помехи, аппаратные сбои), уступая в специфике диагностике на основе БЗ. Предварительное тестирование загрузки файлов (например, фото индикаторов ПЛК) и влияния системных промптов показало, что, хотя LLM пытается учесть файловую информацию, глубокая интеграция мультимодальной информации требует дальнейшей оптимизации.

图形用户界面, 文本, 应用程序

描述已自动生成

Рис. 5. Пример результата диагностики

Будущая работа будет направлена на оптимизацию стратегии поиска RAG, расширение охвата БЗ и углубление интеграции мультимодальной информации.

Исследование было профинансировано Китайским советом по стипендиям (CSC) (202409010292).

**Список литературы**

1. Бондаренко М.Э. Анализ экспериментальных исследований активной комбинированной опоры ротора / М.Э. Бондаренко, Р.Н. Поляков, М.А. Токмакова, А.Д. Серебренников // Фундаментальные и прикладные проблемы техники и технологии - Орёл: ОГУ имени И.С. Тургенева – №1(357) – 2023. С.133 – 140.
2. Родичев А.Ю. Применение методов машинного обучения в диагностике промышленного оборудования / А.Ю. Родичев, И.В. Родичева, К.В. Васильев // XI всероссийская научно-практическая конференция имени академика А.Г. Шипунова – 2024 – сб.ст. – Орел: ОГУ имени И.С. Тургенева - 2024. С.10 -14.
3. Albeshri M.A., Khemakhem M.A., Eassa, F.E. Multimodal Large Language Model-Based Fault Detection and Diagnosis in Context of Industry 4.0 // Electronics. 2024. Vol. 13, no. 24, 4912. DOI: 10.3390/electronics13244912.
4. Yang H., Siew M., Joe-Wong C. An LLM-Based Digital Twin for Optimizing Human-in-the-Loop Systems // 2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys). Hong Kong, Hong Kong, 2024. С. 26-31. DOI: 10.1109/FMSys62467.2024.00009.
5. Потехин В.В. Анализ методов диагностирования и прогнозирования технического состояния газотурбинной установки / В. В. Потехин, Я. В. Лудищев, С. В. Болотов // Вопросы системного технологического перехода : Сборник научных трудов по материалам конференции «Технологическая перспектива: новые рынки и точки экономического роста» 2021-2022. В 2-х томах, Санкт-Петербург, 10–11 ноября 2022 года. – Санкт-Петербург: Центр научно-информационных технологий "Астерион", 2023. – С. 100–107. – EDN RTVYJI.
6. Потехин В.В. Использование методов машинного обучения для предиктивного анализа при эксплуатации производственных систем / В. В. Потехин, Д. Д. Лядский, С. В. Болотов // Вопросы системного технологического перехода : Сборник научных трудов по материалам конференции «Технологическая перспектива: новые рынки и точки экономического роста» 2021-2022. В 2-х томах, Санкт-Петербург, 10–11 ноября 2022 года. – Санкт-Петербург: Центр научно-информационных технологий "Астерион", 2023. – С. 90–100. – EDN YZSUMN.
7. CLICK PLC User Manual [Электронный ресурс] / AutomationDirect. – URL: https://cdn.automationdirect.com/static/manuals/c0userm/c0userm.html (дата обращения: 04.06.2025).