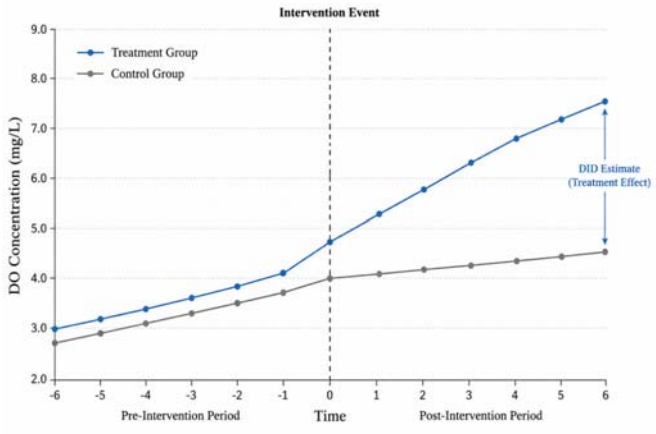


Fig. 3. Dynamic changes in dissolved oxygen concentration before and after environmental events and DID effect analysis diagram

In summary, this IoT monitoring system provides a high-timeliness and high-precision data foundation for the refined management of watershed water environment. Its analysis results can directly serve the formulation of ecological protection policies and the scientific assessment of soil and water environment.

References

1. Assessment and analysis of surface water environmental quality in counties within national key ecological functional zones / D. Zhai, X. Gao, H. Liu [et al.] // Environmental Monitoring in China. – 2022. – N 38 (4). – P. 56–64.
2. Assessing the effect of the joint governance of transboundary pollution on water quality: Evidence from China // Frontiers in Environmental Science. – 2022.
3. Dissolved oxygen level measurement in water using IoT and machine learning // Journal of Theoretical and Applied Information Technology. – 2025. – N 103 (4).

ПРОЕКТ УЧЕБНОГО ЦЕНТРА ПО ТЕХНОЛОГИЯМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА С ИСПОЛЬЗОВАНИЕМ УДАЛЕННЫХ СЕРВЕРОВ С ТЕХНОЛОГИЕЙ ЛОКАЛЬНОГО ЗАПУСКА AI

А. В. Федорович, А. Э. Шлыков

Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь

Научные руководители: Ю. В. Крышнев, А. Е. Запольский

Рассмотрена концепция и архитектура учебного центра по технологиям искусственного интеллекта, объединяющего удаленные GPU-серверы и технологии локального развертывания AI-моделей. Предложена гибридная инфраструктурная модель, обеспечивающая практическое обучение работе с большими языковыми моделями, генеративным AI и системами машинного обучения при полном контроле над данными.

Ключевые слова: искусственный интеллект, большие языковые модели, локальный инференс, GPU-инфраструктура, MLOps, гибридная архитектура, учебный центр, open-source модели.

Стремительное развитие технологий искусственного интеллекта (ИИ) и, в частности, больших языковых моделей (LLM), формирует устойчивый спрос на специалистов,

способных не только использовать готовые облачные сервисы, но и развертывать, настраивать и оптимизировать AI-системы самостоятельно [1]. Традиционные образовательные программы, ориентированные на теоретическую подготовку, не обеспечивают формирования практических компетенций работы с реальной вычислительной инфраструктурой.

Существующие подходы к обучению AI-технологиям имеют ряд ограничений: зависимость от проприетарных облачных API, отсутствие контроля над конфиденциальностью данных, высокая стоимость токенов при масштабном использовании и невозможность работы с моделями в условиях ограниченного сетевого доступа [2]. В связи с этим актуальной является задача создания образовательной среды, в которой обучающиеся получают опыт полного цикла работы с AI-моделями – от загрузки и квантования до развертывания в производственной инфраструктуре.

Целью данной работы является представление концепции и технической архитектуры учебного центра, реализующего гибридную модель обучения: локальные рабочие станции для начального знакомства с инструментами и удаленные GPU-серверы для ресурсоемких задач обучения и инференса.

Предлагаемая концепция основана на принципе «локальный запуск как базовый навык». Обучающиеся осваивают полный цикл работы с AI-моделями: выбор модели, квантование, развертывание, интеграция через API, тонкую настройку и эксплуатацию в нагруженной среде с применением методов эффективной адаптации [3]. Целевая аудитория центра включает разработчиков программного обеспечения, специалистов по данным, ML-инженеров, IT-менеджеров, студентов технических специальностей и предпринимателей, внедряющих AI в бизнес-процессы. Концептуальная схема учебного центра представлена на рис. 1.

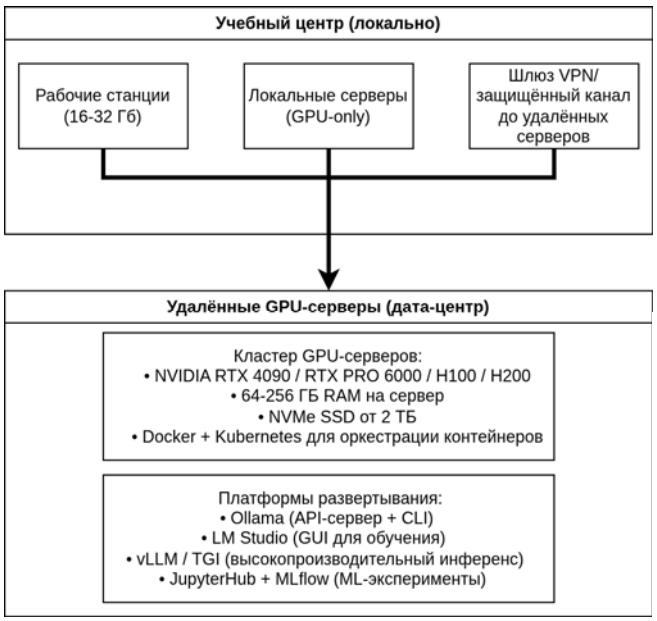


Рис. 1. Концептуальная схема гибридной архитектуры учебного центра

Ключевыми стратегическими целями проекта являются: формирование компетенций в области локального развертывания и тонкой настройки AI-моделей; обеспечение практического опыта работы с вычислительной инфраструктурой промыш-

ленного уровня; создание экосистемы для разработки AI-решений с соблюдением требований информационной безопасности; подготовка кадров без зависимости от зарубежных облачных провайдеров.

Технологический стек центра построен преимущественно на open-source решениях, что обеспечивает воспроизводимость, прозрачность и отсутствие лицензионных ограничений [5]. Основные компоненты локального запуска AI представлены в таблице.

Компоненты локального запуска AI и их назначение

Компонент	Назначение	Использование в обучении
Ollama [6]	CLI-инструмент для запуска LLM	Базовое знакомство с моделями
LM Studio	GUI-платформа	Обучение, тестирование моделей
llama.cpp	Низкоуровневый инференс	Оптимизация, квантование
vLLM [4]	Высокопроизводительный инференс	Нагруженное обслуживание
Docker [7]	Изолированные контейнерные среды	ML-Ops, воспроизводимость
OpenWebUI	Веб-интерфейс к локальным моделям	Создание чат-ботов

Взаимодействие между компонентами инфраструктуры реализуется по многоуровневой схеме (рис. 2).

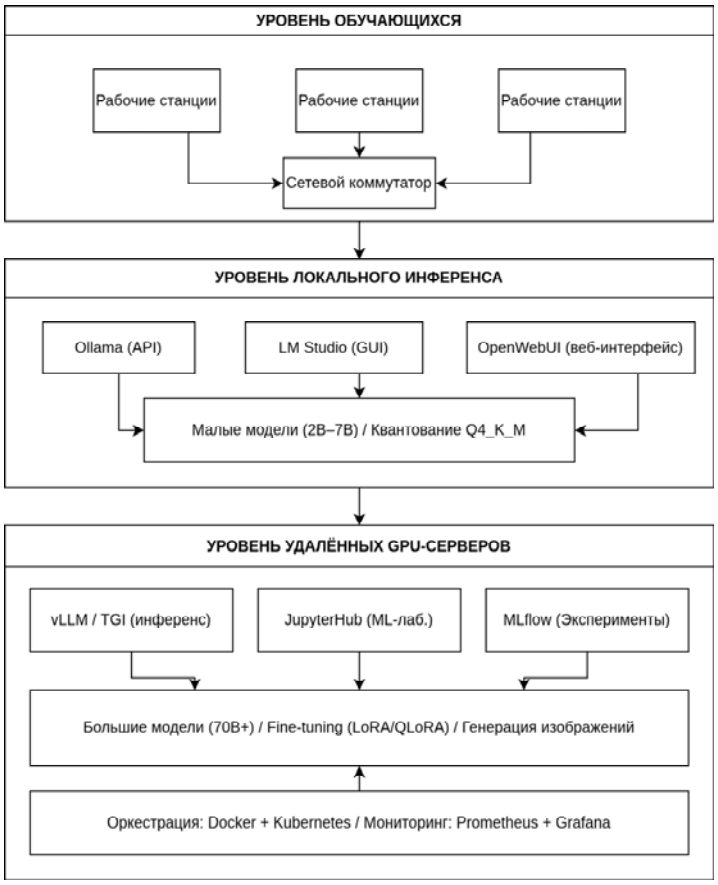


Рис. 2. Многоуровневая схема взаимодействия компонентов инфраструктуры

Обучающиеся подключаются к локальной сети центра через рабочие станции. Локальные серверы обеспечивают базовый инференс малых моделей (2–7 млрд параметров) для начального знакомства. Удаленные GPU-сер-веры предоставляют вычислительные мощности для тонкой настройки моделей, работы с большими моделями (70+ млрд параметров), параллельного инференса и генерации изображений [4].

Программный стек центра включает: операционные системы на базе ядра Linux; платформы развертывания моделей (Ollama, vLLM) [6, 4]; ML-фреймворки (PyTorch, Hugging Face Transformers) [5]; инфраструктурные инструменты (Docker, Kubernetes) [7]; системы мониторинга (Prometheus, Grafana).

Литература

1. Vaswani, A. Attention Is All You Need / A. Vaswani // Advances in Neural Information Processing Systems. – 2017. – URL: https://arxiv.org/abs/1706.03762 (date of access: 29.07.2026).
2. Touvron, H. Llama: Open and Efficient Foundation Language Models / H. Touvron. – 2023. – URL: https://arxiv.org/abs/2302.13971 (date of access: 29.07.2026).
3. Hu, E. J. LoRA: Low-Rank Adaptation of Large Language Models / E. J. Hu. – 2021. – URL: https://arxiv.org/abs/2106.09685 (date of access: 29.07.2026).
4. Kwon, W. Efficient Memory Management for Large Language Model Serving with PagedAttention / W. Kwon // Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP). – 2023. – URL: https://docs.vllm.ai (date of access: 29.07.2026).
5. Hugging Face – The AI community building the future. – URL: https://huggingface.co (date of access: 29.07.2026).
6. Ollama – Run large language models locally. – URL: https://ollama.com (date of access: 29.07.2026).
7. Docker – Accelerated Container Application Development. – URL: https://www.docker.com (date of access: 29.07.2026).

АВТОМАТИЗИРОВАННАЯ СИСТЕМА РЕГИСТРАЦИИ СВАРОЧНЫХ ПРОЦЕССОВ

М. А. Проскуряков

Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь

Научный руководитель Ю. В. Крышнев

Рассмотрена разработанная автоматизированная система регистрации сварочных процессов.

Ключевые слова: дуговая сварка, регистратор, сварочные процессы, датчик температуры, контроль, технология Zigbee, база данных, программное обеспечение, сервер, речевой информатор.

В настоящее время производятся различные системы контроля режимов сварочных работ, при этом некоторые сварочные аппараты уже оснащены отдельными элементами такого контроля. Однако они не предполагают ведения базы данных, реализации удаленного контроля за процессом сварки и выполнения углубленного многофакторного анализа.

Разрабатываемая автоматизированная система регистрации сварочных процессов предназначена для контроля за соблюдением технологии при проведении сварочных работ, документального подтверждения качества каждого сварного стыка трубопровода, создания отчетов о работе электросварщика и сварочного оборудования, контроля за расходом материалов и ресурсов.