УДК 004.82

НОРМАЛИЗАЦИЯ ТЕКСТА, РАСПОЗНАННОГО ПРИ ПОМОЩИ ТЕХНОЛОГИИ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ, С ИСПОЛЬЗОВАНИЕМ ЛЕГКОВЕСНЫХ LLM

В. К. Вершинин[1] [0009-0001-9425-0881], И. В. Ходненко[2] [0009-0003-7787-7126], С. В. Иванов[3] [0000-0002-1128-2942]

[1-3]*Университет ИТМО, г. Санкт-Петербург, Россия*

[1]vershinin@itmo.ru, [2]svivanov@itmo.ru,[3]Ivan.Khodnenko@itmo.ru

Аннотация

Несмотря на значительный прогресс, технологии оптического распознавания символов (OCR) для исторических газет по-прежнему допускают 5–10% ошибок на уровне символов. В работе представлена полностью автоматизированная система нормализации пост-OCR, объединяющая легкие языковые модели (LLM) объемом 7–8 млрд параметров, обученные по инструкциям и квантизованные до 4 бит (INT4), с небольшим набором регулярных выражений. На наборе данных BLN600 (600 страниц британских газет XIX в.) лучшая модель YandexGPT-5-Instruct Q4 снижает Character Error Rate (CER) с 8.4% до 4.0% (–52.5%) и Word Error Rate (WER) с 20.2% до 6.5% (–67.8%), повышая при этом семантическое сходство до 0.962. Система работает на потребительском оборудовании (RTX-4060 Ti, 8 ГБ VRAM) со скоростью около 35 секунд на страницу и не требует дополнительного обучения или параллельных данных. Полученные результаты показывают, что компактные INT4-LLM являются практичной альтернативой крупным моделям для постобработки OCR исторических документов.

Ключевые слова*: оптическое распознавание символов, пост-OCR-коррекция, исторические газеты, большие языковые модели, квантизация, INT4, конвейер нормализации, ошибка на уровне символов, семантическое сходство, регулярные выражения, YandexGPT-5, легкие модели, обработка естественного языка, цифровые гуманитарные науки, оцифровка документов.*

ВВЕДЕНИЕ

Масштабная оцифровка архивных коллекций опирается на системы оптического распознавания символов (OCR) [1], качество работы которых напрямую определяет полнотекстовый поиск, аналитические исследования и извлечение структурированных данных. Для исторических печатных источников уровень ошибок по-прежнему остается значительным: согласно современным исследованиям, средний показатель Character Error Rate (CER) находится в диапазоне 5–10% [2, 3], что существенно снижает качество поиска и автоматического извлечения фактов.

Метрики и целевые уровни

В работе использованы следующие три стандартные метрики.

Character Error Rate (CER): вычисляется по формуле

$$\mathrm{CER} = \frac{S+I+D}{N},$$

где *S* — количество замен, *I* — вставок, *D* — удалений, *N* — общее количество символов в эталонном тексте.

Word Error Rate (WER) определяется аналогично, но на уровне слов:

$$\mathrm{WER} = \frac{Sw+Iw+Dw}{Nw}.$$

Semantic Similarity (SS) — косинусное сходство между векторными представлениями MiniLM для эталонного текста и гипотезы.

Практические ориентиры из литературы показывают, что для крупномасштабной оцифровки «хорошее» качество OCR соответствует точности 98–99% (т. е. CER ≈ 1–2%), «среднее» — 90–98% (CER ≈ 2–10%), «плохое» — менее 90% (CER > 10%) [4]. Для задач извлечения информации и поиска наблюдается заметное ухудшение результатов при уменьшении точности ниже 70–80% [5]. На основании этого, а также эмпирических наблюдений, согласно которым при CER ≈ 2–3% обычно фиксируется WER ≈ 8–12% [6], в настоящей работе приняты следующие пороговые значения: CER < 0.05, WER < 0.10 и SS > 0.90 – как консервативные критерии «пригодности к использованию».

Исторические газеты XIX в. представляют особую трудность: сложная многоколоночная верстка, смешение шрифтов и износ бумаги приводят к ошибкам сегментации и распознавания, повышая исходные уровни CER и WER по сравнению с современными публикациями. Крупные проекты по оцифровке газет (например, Europeana Newspapers) прямо указывают эти факторы как основные причины снижения качества OCR для исторических периодических изданий, подчеркивая влияние сложной верстки и низкого качества оригиналов [7–10]. Это также отражено в наших экспериментах: для корпуса BLN600 исходные значения составляют CER = 0.084 и WER = 0.202 (см. табл. 1, строка Baseline OCR), которые уменьшаются после нормализации (см. разд. 4).

Определение

Нормализация текста – это процесс преобразования текста в стандартизованную каноническую форму. Она включает исправление орфографических ошибок, раскрытие аббревиатур, устранение сокращений, нормализацию пунктуации, регистра и других языковых вариаций с целью обеспечения согласованного и однородного представления текстовых данных [11].

Последние исследования показывают, что посткоррекция с использованием больших языковых моделей (LLM) позволяет снизить CER ниже 5% и существенно повысить пригодность текста для последующих задач [2, 12, 13]. Однако большинство предложенных решений опирается на модели объемом 13–70 млрд параметров, требующие серверов с объемом видеопамяти ≥ 40 ГБ и/или трудоемкой донастройки на параллельных корпусах [14, 15]. Эти требования делают технологию малодоступной для региональных и институциональных архивов.

Постобучающая квантизация весов (INT8/INT4; GPTQ, AWQ, NF4 и др.) резко снижает потребление памяти и часто сохраняет качество, близкое к полновесным моделям, для задач с коротким и средним контекстом [16]. Однако при сверхдлинных входных данных (> 64 К токенов) агрессивная 4-битная квантизация может заметно ухудшить качество, тогда как 8-битная остается практически без потерь [17]. В нашей задаче (OCR-фрагменты на уровне страниц) длины контекста малы, что позволяет использовать компактные 4-битные модели класса 7–8 В на потребительских видеокартах (≤ 8 ГБ VRAM).

В настоящем исследовании рассмотрено, насколько далеко можно продвинуть нормализацию пост-OCR текстов исторических газет в условиях таких ограничений по ресурсам. Реализован минимальный конвейер: OCR → предобработка → нормализация LLM → постобработка → текст и проведена его оценка на открытом корпусе BLN600 [18]. Показано, что достижение целевых порогов качества возможно при 4-битном выводе на GPU с 8 ГБ видеопамяти, что делает данный подход практически применимым для широкого круга архивных учреждений.

1. ОБЗОР СВЯЗАННЫХ РАБОТ

Vision-LLM, использующийся как прямой OCR, демонстрирует возможности GPT-4V для распознавания текстовых изображений [12]. На рукописном наборе данных IAM они сообщили о значениях CER равных 3.32% и 13.75% на уровне страницы и строки соответственно, в то время как лучшая специализированная CTC-модель достигала 2.89% и 6.52% [12, табл. 4].

На многоязычном наборе данных уличных знаков MLT19 качество резко падает для нелатинских алфавитов: F1-score снижается с 82 (EN) до 1–11 (AR, KO, [12, табл. 2].

Seq2seq-коррекция

LSTM seq2seq-модель применяется после базового OCR и снизили CER с 7–9 % до 4–5 % на корпусе ECCO-TCP [14, разд. 4.2]. Метод требует крупного параллельного корпуса пар «сырой OCR / эталон», что затруднительно для маломасштабных коллекций [15].

Инструкционное дообучение трансформеров

Показано, что Llama-2-7B, дообученная в режиме следования инструкциям, снижает CER с 20% до 9% для газет XIX века (для BART – 15%; [2, табл. 1]), что подчеркивает преимущество моделей, обученных по инструкциям, над ранними seq2seq-трансформерами [2].

Синтетический шум и адаптация при инференсе (TTA)

Добавление марковского шума с последующим дообучением снижает CER с 5–7% до 2–3% [3, разд. 4.3]. Подход SCN-TTA дает сопоставимые 2–3%, начиная с 9% [13, табл. 6].

Многовидовое объединение

Комбинирование нескольких версий OCR одного и того же документа снижает WER с 8–10% до 6–7% [19, табл. 5], что дополняет подходы seq2seq для языков с ограниченной аннотацией [15].

Выводы

Существуют три эффективных направления:

- использование крупных или дообученных LLM;
- генерация синтетического шума и адаптация при инференсе (TTA);
- многовидовое объединение.

Все они демонстрируют CER < 5 %, но требуют либо очень больших моделей (> 13 B) и GPU > 40 ГБ, либо крупных корпусов для дообучения. Наш подход следует идее (i), но использует квантизованные LLM 7–8 B без дополнительного обучения, помещающиеся в 8 ГБ VRAM, устраняя тем самым главный инфраструктурный барьер.

2. МЕТОД

Мощные языковые модели с размером 13–70 млдр параметров успешно уменьшают ошибки OCR, однако требуют ≥ 40 ГБ видеопамяти или трудоемкого дообучения на параллельных корпусах, что недопустимо для региональных архивов с ограниченными ресурсами и небольшими наборами данных.

В настоящей работе показано, что компактные 7–8 B модели, квантизованные до 4 бит и работающие «из коробки», по точности (CER/WER) не уступают крупным моделям, при этом потребляют в несколько раз меньше вычислительных ресурсов.

Как показано на рис. 1, входной поток состоит из сканов газет (в форматах JPEG/PNG/PDF, с разрешением 150–300 dpi). Если исходное OCR-распознавание

не предоставлено, применяется Tesseract 5.3 с языковыми пакетами eng/ru и параметрами oem 3 – psm 4.

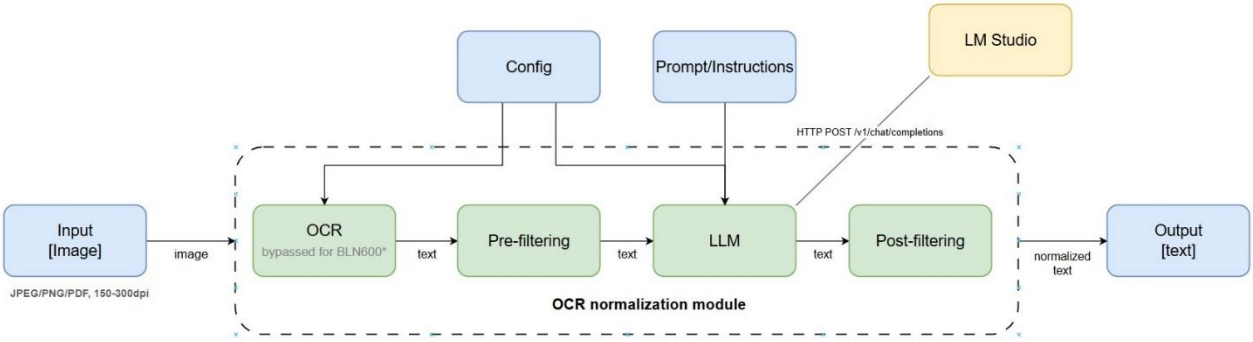


Рис. 1. Схема конвейера нормализации от OCR к LLM-модулю.

Предварительная фильтрация

Наивный слой регулярных выражений (RegEx) удаляет шум, с которым языковая модель (LLM) справляется слабо (см. Листинг 1).

Листинг 1. Правила RegEx, используемые при предварительной и последующей фильтрации

```
def prefilter(text: str) -> str:
    text = text.lower()
    text = re.sub(r"[^\\w\\s]", " ", text)
    text = re.sub(r"\\s+", " ", text)
    return text.strip()

def postfilter(text: str) -> str:
    text = re.sub(r"[()\\[\\]']", "", text)
    text = re.sub(r"\\s+", " ", text)
    return text.strip()
```

Как видно из Листинга 1, набор регулярных выражений, используемых для пред- и постфильтрации, намеренно минимален. Он служит только для удаления шумовых символов, с которыми языковые модели работают плохо, оставляя основную часть нормализации самой модели. Такой подход позволяет избежать переобучения на конкретных артефактах OCR и гарантирует, что улучшения, приведенные в табл. 1, обусловлены именно этапом LLM-нормализации, а не ручной предобработкой.

Нормализация с помощью LLM

Предобработанный текст передается в LM Studio (REST-интерфейс, запрос POST /v1/chat/completions). По умолчанию используется модель YandexGPT-5-Lite-8B-Instruct Q4_K_M (4,9 ГБ); альтернативно — Mistral-7B-Instruct Q4_K_M (4,4 ГБ). Параметры инференса задаются в YAML-конфигурации: temperature = 0.2, top_p = 0.7, top_k=50, max_tokens=4096. Используется промпт под названием «correction» (см. Листинг 2), который явно запрещает галлюцинации и добавление нового содержания.

Листинг 2. Шаблон промпта, используемого для корректировки текста

```
You are an expert text corrector, specialized in fixing OCR error.
1) Fix spelling/grammar.
2) Keep original wording if correct.
3) Do NOT add new sentences.
4) Remove or repair only garbled words.
Text to correct:
"{input_text}"
```

Постфильтрация

Заключительный проход регулярных выражений удаляет оставшиеся лишние символы и двойные пробелы, формируя итоговый текстовый файл .txt.

3. ИНСТРУМЕНТЫ И СРЕДА ВЫПОЛНЕНИЯ

Программное обеспечение: Python 3.12, transformers 4.51.3, sentence-transformers 4.1.0, LM Studio 0.3.17 (build 10).

Аппаратное обеспечение: RTX 4060 Ti (8 ГБ VRAM), Ryzen 5 5600G, 36 ГБ оперативной памяти, CUDA 11.8.

Подсчет семантического сходства выполняется с использованием модели all-MiniLM-L6-v2, при этом вычисляется косинусное сходство после L2-нормализации.

Как показано на рис. 2, YAML-конфигурация описывает используемый набор данных, выбранную языковую модель (LLM) и параметры сэмплирования. Скрипт run_inference.py последовательно обрабатывает пары «сырой OCR / эталонный

текст», сохраняя тексты после каждого этапа (журнал конвейера). Скрипт evaluate_metrics.py вычисляет значения WER, CER и SS для финального результата, а evaluate_aggregates.py усредняет метрики по всему корпусу.

Эксперименты проводились исключительно на корпусе BLN600 (600 страниц британских газет XIX в.), включающем pdf-изображения, результаты «сырого» OCR и эталонный (размеченный) текст. Промежуточные метрики после OCR и обработки LLM используются для анализа вклада каждого этапа.

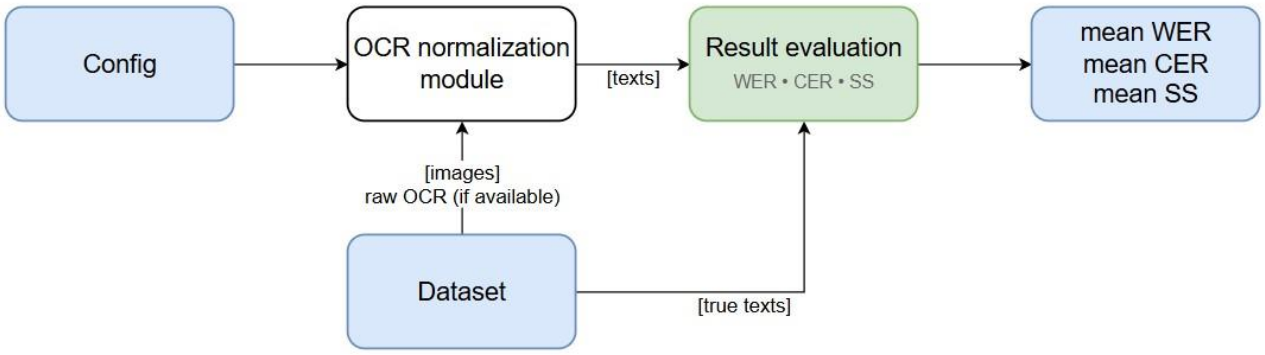


Рис. 2. Схема процесса оценки на наборе данных BLN600

Пример конфигурации

Как показано в Листинге 3, YAML-файл конфигурации определяет набор данных, выбранную языковую модель и параметры сэмплирования.

Листинг 3. YAML-конфигурация для YandexGPT-5 на корпусе BLN600

```
dataset:
  name: BLN600
  path: TMP/bln600
llm:
  model_name: yandexgpt-5-lite-8b-instruct
  host: http://localhost:1234/v1/chat/completions
  temperature: 0.2
  top_p: 0.7
  top_k: 50
  max_tokens: 4096
experiment:
  output_path: results/bln600_yandex.yaml
```

Унифицированный формат результатов

Скрипты для оценки ожидают наличие ключей:

«ground_truth, ocr_text, corrected_text».

Таким образом, предложенный конвейер может выполняться на графических процессорах с объемом памяти ≤ 8 ГБ, не требует дополнительных данных для дообучения и легко переиспользуется благодаря YAML-конфигурациям и API-сервису LM Studio.

4. РЕЗУЛЬТАТЫ

В табл. 1 представлены метрики моделей без дообучения (результаты, полученные нашим подходом).

Табл. 1. Качество на корпусе BLN600 после нормализации (GPU 8 ГБ)

Модель	**CER ↓**	**WER ↓**	**SS ↑**
Базовый OCR	0.0840	0.2020	0.8455
Mistral-7B-Q4	0.0921	0.1240	0.9315
YandexGPT-5-Q4	**0.0399**	**0.0650**	**0.9616**
Llama-2-7B-Q4	0.1490	0.1650	0.9279
Llama-2-13B-Q4	0.4205	0.4060	0.8400

В табл. 2 представлены метрики моделей с дообучением, взятые из источников.

Табл. 2. Данные, приведенные в литературе

Модель (источник)	**CER ↓**	**WER ↓**	**SS ↑**
Llama-2-13B-Q8*	0.038	n/a	n/a
Llama-2-7B-Q8*	0.048	n/a	n/a
Llama-3-8B-F16**	0.080	0.190	n/a

* Модели дообучены на исторических данных OCR [2].

** Данные из [3].

Основные выводы

Модель YandexGPT-5-Q4 снижает Character Error Rate (CER) на 52% и Word Error Rate (WER) на 68% по сравнению с исходным OCR, при этом помещаясь в 4.9 ГБ VRAM.

Модель Mistral-7B-Q4 демонстрирует умеренные улучшения при том же объеме памяти.

Обе модели обеспечивают семантическое сходство выше 0.93, что превышает установленный порог пригодности к использованию (0.90).

Динамика ошибок

На рис. 3–5 приведены значения метрик до и после нормализации на корпусе BLN600 для моделей, используемых в нашем пайплайне, а также для моделей из литературы.

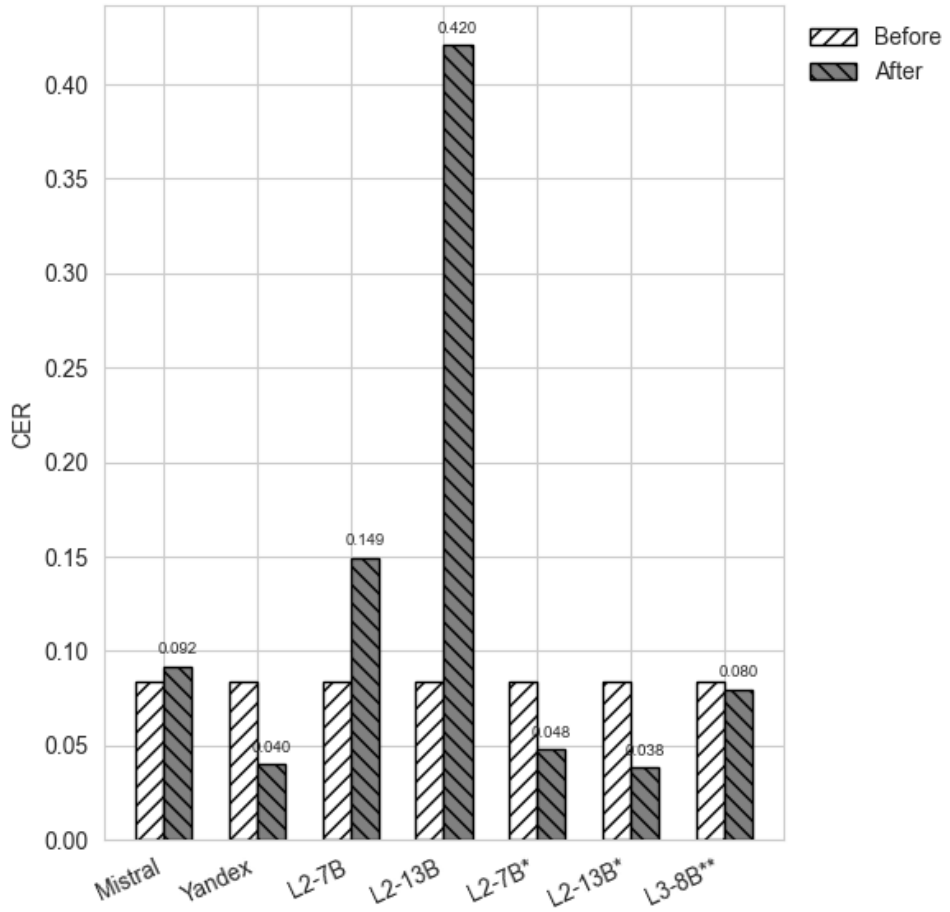


Рис. 3. Ошибка на уровне символов (CER) до и после нормализации на корпусе BLN600.

Нестержневые (без звездочек) столбцы — наши запуски zero-shot; звездочками отмечены результаты, приведенные в литературе. Столбец «До» (CER = 0.084) общий для всех моделей.

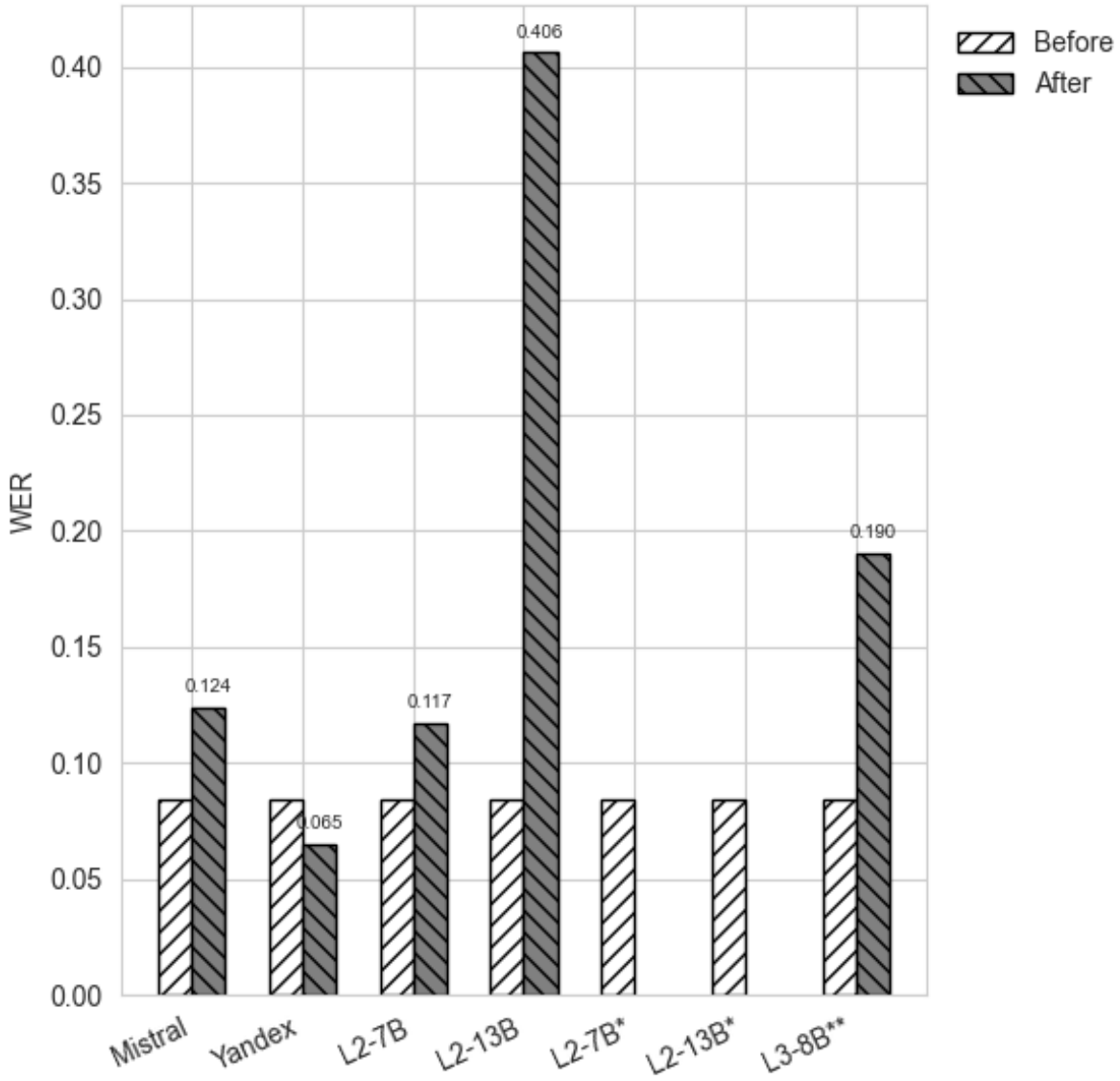


Рис. 4. Ошибка на уровне слов (WER) до и после нормализации на BLN600. Нестержневые столбцы — наши результаты без дообучения; звездочками отмечены литературные базовые модели.

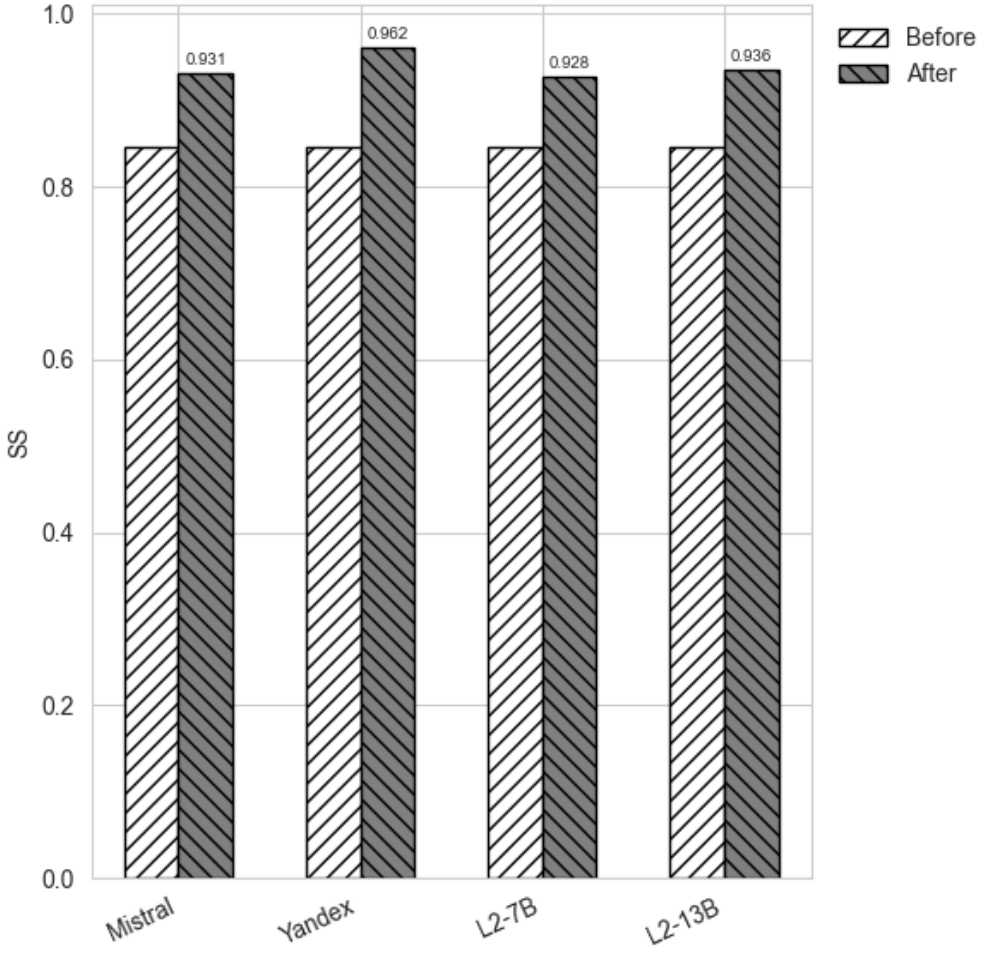


Рис. 5. Семантическое сходство (MiniLM) до и после нормализации для наших zero-shot запусков. Дообученные модели из литературы не предоставляют значения SS и поэтому не показаны.

Время выполнения

В табл. 3 представлено время инференса моделей на указанной видеокарте.

Табл. 3. Задержка и использование памяти на RTX 4060 Ti (8 ГБ)

Модель	VRAM (ГБ)	Время (с/страница)
Mistral-7B-Q4	4.1	32.7
YandexGPT-5-Q4	**4.6**	**35.3**
Llama-2-7B-Q4	3.6	38.3
Llama-2-13B-Q4	6.9	112.5

5. ОБСУЖДЕНИЕ

Важно отметить, что предложенный нами метод не оценивался на таких бенчмарках, как IAM или MLT19, которые ориентированы на OCR-задачи построчного или пословного уровня, а также на распознавание уличных знаков. Эти наборы данных включают очень короткие входные контексты – часто ограниченные одним словом или одной строкой.

В подобных условиях наш подход, основанный на использовании LLM для нормализации фрагментов на уровне целой страницы, вероятно, показал бы худшие результаты, поскольку успешная коррекция в минимальном контексте обычно требует специализированных словарей, лексиконов или дополнительных внешних источников, обеспечивающих недостающий контекст.

В противоположность этому наш конвейер изначально спроектирован для OCR-фрагментов на уровне страницы, где доступен более широкий текстовый контекст. Это позволяет LLM использовать соседние токены, чтобы корректно интерпретировать и исправлять ошибки OCR.

Когда контекстное окно сужается — до длины строки или слова, языковой модели не хватает информации для надежной коррекции, если она опирается только на внутренние вероятностные представления. В таких условиях модели обычно выигрывают от использования доменных словарей, ограниченного декодирования или правил постобработки.

Таким образом, эффективность предложенного нами конвейера нормализации обусловлена наличием достаточного контекстного окружения и не может напрямую распространяться на задачи с крайне коротким входным текстом.

Для полноты и воспроизводимости исследования мы включили наборы IAM и MLT19 в сопровождающий GitHub-репозиторий настоящей работы. Эти датасеты не входят в основную часть оценки, однако предоставлены как вспомогательные ресурсы — для будущих сравнительных исследований и для исследователей, заинтересованных в адаптации нашего подхода к OCR-задачам с коротким контекстом.

Таким образом, эффективность предложенного конвейера нормализации напрямую связана с доступностью достаточного контекста для анализа, и его применение может быть ограничено задачами, где длина входных данных крайне

мала. В дальнейшем работа будет направлена на проверку обобщающей способности подхода при сокращенном контексте и в многоязычных корпусах.

ЗАКЛЮЧЕНИЕ

Представлен zero-shot конвейер с 4-битной квантизацией, способный устранять ошибки OCR при использовании потребительского оборудования.

На бенчмарке BLN600 предложенный конвейер снижает показатели ошибок CER с 8.4% до 4.0% и WER с 20.2% до 6.5% при использовании модели YandexGPT-5-Instruct-Q4 (8B), а также CER до 9.2% и WER до 12.4% при модели Mistral-7B-Instruct-Q4.

Обе модели помещаются в видеопамять объемом ≤ 5 ГБ и обрабатывают одну страницу газеты примерно за 35 секунд на RTX 4060 Ti, демонстрируя результаты, сопоставимые или превосходящие значительно более крупные дообученные модели — без необходимости дополнительного обучения или использования параллельных данных.

Укажем возможные направления дальнейшей работы:

- расширение на многоязычные данные (кириллица + латиница);
- применение легковесных архитектур типа Mixture-of-Experts и chain-of-thought prompting для сложных макетов страниц;
- создание открытого российского бенчмарка с административными формами для стимулирования открытых исследований.

В целом компактные 4-битные LLM представляют собой практичную и экономичную альтернативу крупным моделям для постобработки OCR, открывая возможности масштабного цифрового восстановления исторических и специализированных архивов. Исходный код и скрипты для оценки результатов доступны в открытом репозитории: https://github.com/Kerysfel/OCRNorm

СПИСОК ЛИТЕРАТУРЫ

1. *Memon J., Sami M., Khan R.A.* Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR) // IEEE Access. 2020. Vol. 8. P. 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

2. *Thomas A., Gaizauskas R., Lu H.* Leveraging LLMs for Post-OCR Correction of Historical Newspapers // Proceedings of the 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). 2024. P. 116–121. https://doi.org/10.18653/v1/2024.lt4hala-1.6

3. *Bourne J.* Scrambled text: training Language Models to correct OCR errors using synthetic data // arXiv preprint. 2024. arXiv:2409.19735. https://doi.org/10.48550/arXiv.2409.19735

4. *Holley R.* How Good Can It Get? Analysing and Improving OCR Accuracy in Large-Scale Historic Newspaper Digitisation Programs // D-Lib Magazine. 2009. Vol. 15, No. 3/4. https://doi.org/10.1045/march2009-holley

5. *van Strien D., Beelen K., Coll Ardanuy M., Hosseini K., McGillivray B., Tolfo G.S.* Assessing the Impact of OCR Quality on Downstream NLP Tasks // Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). 2020. P. 484–496. https://doi.org/10.5220/0009169004840496

6. *Drobac S., Friberg Heppin K., Wirén M., Lindén K.* Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods // International Journal on Document Analysis and Recognition (IJDAR). 2020. Vol. 23. P. 279–295. https://doi.org/10.1007/s10032-020-00359-9

7. *Neudecker C., Antonacopoulos A.* Making Europe's Historical Newspapers Searchable // DAS 2016 Workshop / Europeana Newspapers. 2016.(Workshop paper). URL: https://www.primaresearch.org/www/files/das2016/Europeana%20Newspapers.pdf.

8. *Boillet M., Kermorvant C., Paquet T.* Robust text line detection in historical documents: learning and evaluation methods // International Journal on Document Analysis and Recognition (IJDAR). 2022. Vol. 25. P. 95–114. https://doi.org/10.1007/s10032-022-00395-7

9. *Ermakova L., Tolfo G.S., Hosseini K.* On the Impact of OCR Quality on Named Entity Extraction from Historical Newspapers // DH Benelux 2021 (Extended abstracts). 2021. URL: https://dhbenelux.org/wp-content/uploads/booklet2021.pdf#page=66

10. *Kettunen K.* Optical Character Recognition Quality Affects Perceived Usefulness and Trust // arXiv preprint. 2022. arXiv:2209.08222.

11. *Sreelekha S., Sumam A.R., Nair R.R.* Systematic Review on Text Normalization Techniques and Its Approach to Non-Standard Words // Preprint. 2023 (ResearchGate). URL: https://www.researchgate.net/publication/373877004.

12. *Shi Y., Peng D., Liao W., Lin Z., Chen X., Liu C., Zhang Y., Jin L.* Exploring OCR Capabilities of GPT-4V(ision): A Quantitative and In-Depth Evaluation // arXiv preprint. 2023. arXiv:2310.16809.

13. *Guan S., Xu C., Lin M., Greene D.* Effective Synthetic Data and Test-Time Adaptation for OCR Correction // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). 2024. P. 15412–15425 (ACL Anthology).

14. *Kanerva J., Ledins C., Käpyaho S., Ginter F.* OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches // Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025). 2025. Tallinn, Estonia (ACL Anthology).

15. *Rijhwani S., Anastasopoulos A., Neubig G.* OCR Post-Correction for Endangered Language Texts // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 5931–5942.
https://doi.org/10.18653/v1/2020.emnlp-main.478

16. *Jin R., Du J., Huang W., Liu W., Luan J., Wang B., Xiong D.* A Comprehensive Evaluation of Quantization Strategies for Large Language Models // Findings of ACL 2024 (also arXiv preprint). 2024. arXiv:2402.16775.
https://doi.org/10.48550/arXiv.2402.16775

17. *Mekala A., Atmakuru A., Song Y., Karpinska M., Iyyer M.* Does Quantization Affect Models' Performance on Long-Context Tasks? // arXiv preprint. 2025. arXiv:2505.20276. https://doi.org/10.48550/arXiv.2505.20276.

18. *Booth C.W., Thomas A., Gaizauskas R.* BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth-Century Newspaper Texts // Proceedings of LREC-COLING 2024. 2024. P. 2440–2446.
https://doi.org/10.15131/shef.data.25439023.

19. *Gupta H., Del Corro L., Broscheit S., Hoffart J., Brenner E.* Unsupervised Multi-View Post-OCR Error Correction with Language Models // Proceedings of the

2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. P. 8647–8652. https://doi.org/10.18653/v1/2021.emnlp-main.680

NORMALIZATION OF TEXT RECOGNIZED BY OPTICAL CHARACTER RECOGNITION USING LIGHTWEIGHT LLMs

V. K. Vershinin[1] [0009-0001-9425-0881], I. V. Khodnenko[2] [0009-0003-7787-7126], S. V. Ivanov[3] [0000-0002-1128-2942]

[1–3]*ITMO University, Saint-Petersburg, Russia*

[1]vershinin@itmo.ru, [2]Ivan.Khodnenko@itmo.ru, [3]svivanov@itmo.ru

Abstract

Despite recent progress, Optical Character Recognition (OCR) on historical newspapers still leaves 5–10% character errors. We present a fully automated post-OCR normalization pipeline that combines lightweight 7–8B instruction-tuned LLMs quantized to 4-bit (INT4) with a small set of regex rules. On the BLN600 benchmark (600 pages of 19th-century British newspapers), our best model YandexGPT-5-Instruct Q4 reduces Character Error Rate (CER) from 8.4% to 4.0% (–52.5%) and Word Error Rate (WER) from 20.2% to 6.5% (–67.8%), while raising semantic similarity to 0.962. The system runs on consumer hardware (RTX-4060 Ti, 8 GB VRAM) at about 35 seconds per page and requires no fine-tuning or parallel training data. These results indicate that compact INT4 LLMs are a practical alternative to large checkpoints for post-OCR cleanup of historical documents.

Keywords*: optical character recognition, post-OCR correction, historical newspapers, large language models, quantization, INT4, normalization pipeline, character error rate, semantic similarity, regex rules, YandexGPT-5, lightweight models, natural language processing, digital humanities, document digitization.*

REFERENCES

1. *Memon J., Sami M., Khan R.A.* Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR) // IEEE Access. 2020. Vol. 8. P. 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

2. *Thomas A., Gaizauskas R., Lu H.* Leveraging LLMs for Post-OCR Correction of Historical Newspapers // Proceedings of the 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). 2024. P. 116–121. https://doi.org/10.18653/v1/2024.lt4hala-1.6

3. *Bourne J.* Scrambled text: training Language Models to correct OCR errors using synthetic data // arXiv preprint. 2024. arXiv:2409.19735. https://doi.org/10.48550/arXiv.2409.19735

4. *Holley R.* How Good Can It Get? Analysing and Improving OCR Accuracy in Large-Scale Historic Newspaper Digitisation Programs // D-Lib Magazine. 2009. Vol. 15, No. 3/4. https://doi.org/10.1045/march2009-holley

5. *van Strien D., Beelen K., Coll Ardanuy M., Hosseini K., McGillivray B., Tolfo G.S.* Assessing the Impact of OCR Quality on Downstream NLP Tasks // Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). 2020. P. 484–496. https://doi.org/10.5220/0009169004840496

6. *Drobac S., Friberg Heppin K., Wirén M., Lindén K.* Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods // International Journal on Document Analysis and Recognition (IJDAR). 2020. Vol. 23. P. 279–295. https://doi.org/10.1007/s10032-020-00359-9

7. *Neudecker C., Antonacopoulos A.* Making Europe's Historical Newspapers Searchable // DAS 2016 Workshop / Europeana Newspapers. 2016.(Workshop paper). URL: https://www.primaresearch.org/www/files/das2016/Europeana%20Newspapers.pdf.

8. *Boillet M., Kermorvant C., Paquet T.* Robust text line detection in historical documents: learning and evaluation methods // International Journal on Document Analysis and Recognition (IJDAR). 2022. Vol. 25. P. 95–114. https://doi.org/10.1007/s10032-022-00395-7

9. *Ermakova L., Tolfo G.S., Hosseini K.* On the Impact of OCR Quality on Named Entity Extraction from Historical Newspapers // DH Benelux 2021 (Extended abstracts). 2021.
URL: https://dhbenelux.org/wp-content/uploads/booklet2021.pdf#page=66

10. *Kettunen K.* Optical Character Recognition Quality Affects Perceived Usefulness and Trust // arXiv preprint. 2022. arXiv:2209.08222.

11. *Sreelekha S., Sumam A.R., Nair R.R.* Systematic Review on Text Normalization Techniques and Its Approach to Non-Standard Words // Preprint. 2023 (ResearchGate). URL: https://www.researchgate.net/publication/373877004.

12. *Shi Y., Peng D., Liao W., Lin Z., Chen X., Liu C., Zhang Y., Jin L.* Exploring OCR Capabilities of GPT-4V(ision): A Quantitative and In-Depth Evaluation // arXiv preprint. 2023. arXiv:2310.16809.

13. *Guan S., Xu C., Lin M., Greene D.* Effective Synthetic Data and Test-Time Adaptation for OCR Correction // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). 2024. P. 15412–15425 (ACL Anthology).

14. *Kanerva J., Ledins C., Käpyaho S., Ginter F.* OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches // Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025). 2025. Tallinn, Estonia (ACL Anthology).

15. *Rijhwani S., Anastasopoulos A., Neubig G.* OCR Post-Correction for Endangered Language Texts // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 5931–5942.
https://doi.org/10.18653/v1/2020.emnlp-main.478

16. *Jin R., Du J., Huang W., Liu W., Luan J., Wang B., Xiong D.* A Comprehensive Evaluation of Quantization Strategies for Large Language Models // Findings of ACL 2024 (also arXiv preprint). 2024. arXiv:2402.16775.
https://doi.org/10.48550/arXiv.2402.16775

17. *Mekala A., Atmakuru A., Song Y., Karpinska M., Iyyer M.* Does Quantization Affect Models' Performance on Long-Context Tasks? // arXiv preprint. 2025. arXiv:2505.20276. https://doi.org/10.48550/arXiv.2505.20276.

18. *Booth C.W., Thomas A., Gaizauskas R.* BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth-Century Newspaper Texts // Proceedings of LREC-COLING 2024. 2024. P. 2440–2446. https://doi.org/10.15131/shef.data.25439023.

19. *Gupta H., Del Corro L., Broscheit S., Hoffart J., Brenner E.* Unsupervised Multi-View Post-OCR Error Correction with Language Models // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. P. 8647–8652. https://doi.org/10.18653/v1/2021.emnlp-main.680

СВЕДЕНИЯ ОБ АВТОРАХ

ВЕРШИНИН Владислав Константинович – ассистент факультета технологий искусственного интеллекта, инженер лаборатории систем поддержки принятия решений, Университет ИТМО. Окончил образовательную программу «Big Data and Machine Learning» Университета ИТМО.

Область научных интересов: RAG-системы, автоматизация цифровой обработки изображений, машинное обучение, обработка естественного языка.

Vladislav Konstantinovich VERSHININ – Teaching Assistant at the Faculty of Artificial Intelligence Technologies and Engineer at the Laboratory of Decision Support Systems, ITMO University. Graduate of the “Big Data and Machine Learning” program at ITMO University.

Research interests: Retrieval-Augmented Generation (RAG) systems, automated digital image processing, machine learning, natural language processing.

email: vershinin@itmo.ru

ORCID: 0009-0001-9425-0881

ИВАНОВ Сергей Владимирович – доцент факультета технологий искусственного интеллекта, Университет ИТМО; руководитель образовательной программы «Инженерия искусственного интеллекта». Кандидат технических наук (2008).

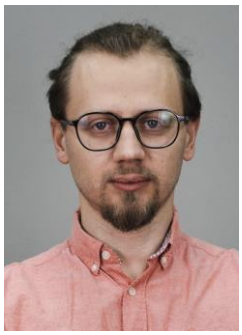
Область научных интересов: машинное обучение, искусственный интеллект, математическое моделирование, оптимизация.

Sergey Vladimirovich IVANOV – Associate Professor at the Faculty of Artificial Intelligence Technologies, ITMO University; Head of the “Artificial Intelligence Engineering” educational program. PhD in Engineering Sciences (2008).

Research interests: machine learning, artificial intelligence, mathematical modeling, optimization.

email: svivanov@itmo.ru

ORCID: 0000-0002-1128-2942



ХОДНЕНКО Иван Владимирович — старший научный сотрудник и старший преподаватель, Национальный центр когнитивных разработок, Университет ИТМО. Окончил Волгоградский государственный технический университет (бакалавр) и Университет ИТМО (магистратура, аспирантура). Кандидат технических наук (2022).

Область научных интересов: машинное обучение, компьютерное зрение, распознавание документов.

Ivan Vladimirovich KHODNENKO – Senior Researcher and Senior Lecturer at the National Center for Cognitive Research, ITMO University. Holds a Bachelor’s degree from Volgograd State Technical University and completed Master’s and PhD studies at ITMO University. PhD in Technical Sciences (2022).

Research interests: machine learning, computer vision, document recognition.

email: Ivan.Khodnenko@itmo.ru

ORCID: 0009-0003-7787-7126

Материал поступил в редакцию 10 октября 2025 года