МЕЖДУНАРОДНЫЙ НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ (МНИИПУ)
Международная научно-исследовательская лаборатория стратегии управления социально-технологических систем

КУРСОВАЯ РАБОТА

на тему:

«Искусственный интеллект в палеографии: обзор методов и доверенная фиксация результатов расшифровки на платформе Academic Gate»

Автор:
Усманов Руслан Гаязович,
научный сотрудник МНИИПУ,
автор проекта «Academic Gate»

Научный руководитель:
Кибальников Сергей Владимирович,
д-р техн. наук, профессор,
академик РАЕН

Оценка: _____
Подпись: _____

Москва 2026

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....3
ГЛАВА 1. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ПАЛЕОГРАФИИ: ЗАДАЧИ, МЕТОДЫ И СОВРЕМЕННОЕ СОСТОЯНИЕ.....5
1.1. Палеография и задачи, которые решает ИИ.....5
1.2. Методы.....5
1.3. Зарубежные эталоны.....6
1.4. Российские проекты.....7
1.5. Ограничения существующих систем и выводы по главе 1.....9
ГЛАВА 2. ДОВЕРЕННАЯ ФИКСАЦИЯ И СРАВНЕНИЕ РАСШИФРОВОК НА ПЛАТФОРМЕ ACADEMIC GATE.....11
2.1. Постановка задачи.....11
2.2. Архитектура платформы Academic Gate.....11
2.3. Метод сравнения текстов.....13
2.4. Апробация на паре «исходный текст — переработанная монография» 14
2.5. Апробация на модельном палеографическом примере.....15
2.6. Протокол «эксперт + ИИ» для расшифровки рукописи.....16
2.7. Выводы по главе 2.....17
ЗАКЛЮЧЕНИЕ.....18
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....20
ПРИЛОЖЕНИЕ А. СОСТАВ ЗАКЛЮЧЕНИЯ AG-CMP.....22
ПРИЛОЖЕНИЕ Б. ПРОГРАММА КОНТРОЛЬНОГО РАСЧЁТА ПОКАЗАТЕЛЕЙ.....23
ПРИЛОЖЕНИЕ В. ПЕРЕЧЕНЬ РАБОТ ПО ПРОЕКТУ ACADEMIC GATE. 24

ВВЕДЕНИЕ

Актуальность темы. Российские архивы, библиотеки и музеи хранят миллионы страниц рукописей — от берестяных грамот и богослужебных книг до делопроизводства XVIII века и метрических книг. Их прочтение требует редкой квалификации палеографа, и ручная расшифровка не успевает за темпами оцифровки. Методы искусственного интеллекта (ИИ) — распознавание рукописного текста, датировка и атрибуция по почерку, восстановление утраченных фрагментов — за последние пять лет перешли из лабораторий в практику: зарубежные системы Ithaca и Enoch опубликованы в ведущих научных журналах, в России запущены проекты МИФИ и Института русского языка РАН, Новгородского университета, Сбера и Яндекса.

Вместе с тем массовое применение ИИ порождает новую проблему — проблему доверия. Модель может ошибиться, а разные исследователи и разные модели дают разные чтения одного текста. Возникают вопросы: кто и когда первым получил расшифровку, чем машинное чтение отличается от экспертного, как количественно сравнить две версии и как сделать результат сравнения проверяемым. Обзор литературы показывает, что существующие системы сосредоточены на качестве распознавания и почти не решают задачу фиксации и проверки результатов.

Цель работы — проанализировать современное состояние применения ИИ в палеографии и разработать методику доверенной фиксации и количественного сравнения расшифровок рукописей на основе платформы Academic Gate и блокчейна Erachain.

Задачи:

– описать задачи палеографии, которые решаются методами ИИ, и типовой конвейер обработки рукописи;

– проанализировать зарубежные эталонные системы и российские проекты, сопоставив их по материалу, задачам и результатам;

– выявить нерешённые проблемы, общие для существующих систем;

– описать архитектуру платформы Academic Gate и её инструменты, применимые к палеографии;

– формализовать метод сравнения базового и преобразованного текста и его показатели;

– провести апробацию метода на реальной паре текстов и на модельном палеографическом примере;

– предложить протокол работы «эксперт + ИИ» для расшифровки рукописи.

Объект исследования — методы и системы искусственного интеллекта для обработки исторических рукописей.

Предмет исследования — способы проверки, количественного сравнения и доверенной фиксации результатов машинной и экспертной расшифровки.

Методы исследования: аналитический обзор научных публикаций и открытых источников, сравнительный анализ, методы обработки текстов (шинглирование, коэффициент Жаккара, расстояние Левенштейна), вычислительный эксперимент.

Практическая значимость. Метод сравнения текстов реализован как модуль платформы Academic Gate, выдаёт номерные заключения с отпечатком SHA-256 и может закрепляться в блокчейне Erachain. Предложенный протокол позволяет архивам и исследовательским группам фиксировать приоритет расшифровки и прозрачно сопоставлять машинные и экспертные чтения.

Структура работы. Курсовая работа состоит из введения, двух глав, заключения, списка литературы и трёх приложений. В первой главе рассмотрены задачи и методы ИИ в палеографии и выполнен обзор зарубежных и российских систем. Во второй главе описаны платформа Academic Gate, метод сравнения текстов и результаты его апробации.

ГЛАВА 1. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ПАЛЕОГРАФИИ: ЗАДАЧИ, МЕТОДЫ И СОВРЕМЕННОЕ СОСТОЯНИЕ

1.1. Палеография и задачи, которые решает ИИ

Палеография изучает историю письма и устанавливает по внешним признакам текста, когда, где и кем он был написан. Традиционно эту работу выполняет эксперт, сопоставляя начертания букв, сокращения, линовку и материал. Искусственный интеллект не заменяет эксперта, а ускоряет и масштабирует четыре его задачи:

1. Распознавание — перевод рукописного или печатного изображения в машиночитаемый текст (HTR, OCR).

2. Датировка — определение времени создания документа по почерку и языку.

3. Атрибуция — определение места создания, школы или конкретного писца.

4. Восстановление — заполнение утраченных фрагментов текста.

Для исторических рукописей все четыре задачи осложнены малым объёмом размеченных данных, вариативностью начертаний и плохой сохранностью носителей. Поэтому успешные системы, как правило, специализированы под одну письменную традицию.

1.2. Методы

Современный конвейер обработки рукописи состоит из нескольких этапов.

1. Предобработка изображения — выравнивание, удаление шума, бинаризация.

2. Сегментация — выделение на странице областей текста, строк и отдельных знаков.

3. Распознавание — свёрточные и рекуррентные нейросети, а в последние годы трансформеры переводят строку изображения в

последовательность символов. Для рукописей используется термин HTR (handwritten text recognition) в отличие от OCR для печатного текста.

4. Классификация и атрибуция — модели обучаются отличать почерки, школы и периоды по признакам начертания.

5. Языковое моделирование — модели текста восстанавливают пропуски и исправляют ошибки распознавания с учётом контекста.

Качество распознавания принято оценивать долей ошибок на уровне знаков (CER, character error rate) и слов (WER, word error rate). Обе метрики основаны на расстоянии Левенштейна — минимальном числе вставок, удалений и замен, переводящих одну строку в другую [16].

Ключевое ограничение методов — потребность в размеченных выборках. Для каждой письменной традиции, особенно для кириллических рукописей с титлами, лигатурами и вязью, их приходится создавать заново.

1.3. Зарубежные эталоны

Ithaca (DeepMind, Оксфордский университет и др.) решает три задачи для древнегреческих надписей: восстановление утраченного текста, определение места и даты. По данным авторов, модель самостоятельно восстанавливает текст с точностью 62 %, определяет регион в 71 % случаев и датирует надпись в пределах 30 лет; совместная работа историка с моделью повышает точность восстановления с 25 до 72 % [1]. Это главный довод в пользу схемы «эксперт + ИИ», а не замены эксперта.

Enoch (Гронингенский университет и др.) датирует рукописи Мёртвого моря, сочетая радиоуглеродный анализ с анализом почерка нейросетью. Средняя ошибка датировки составила около 28–31 года, а ряд рукописей оказался старше традиционных палеографических оценок [2]. Пример показывает, что ИИ-датировка становится сильнее, когда опирается на независимые естественнонаучные данные.

KuroNet распознаёт кудзусидзи — скорописный японский текст домодерной эпохи — методами глубокого обучения [3]. Проект важен как

пример работы с письменностью, в которой начертание знаков сильно варьирует, — задача, близкая к древнерусскому письму.

Transkribus — платформа распознавания рукописного текста для архивов, где исследователь сам обучает модель на своих документах. Авторы показывают, что HTR стал практически применим для массовой обработки архивов, но упирается в финансирование и ресурсы [4]. Для кириллических рукописей Transkribus остаётся ориентиром, однако требует собственных обучающих выборок.

1.4. Российские проекты

«Рукописное наследие Древней Руси» (НИЯУ МИФИ и Институт русского языка им. В. В. Виноградова РАН). Проект создаёт поисковую систему по оцифрованным рукописям из российских библиотек и музеев; руководитель — доцент кафедры кибернетики МИФИ Д. Демидов. На май 2024 г. база включала 245 миней объёмом около 500 страниц каждая. Обработка включает сегментацию страницы на элементы и знаки и классификацию с обучением нейросети. Разработчики учитывают, что одна буква встречается в написании до 150 вариантов, а старый алфавит насчитывал 46 букв против 33 современных; пользователь ищет по древнерусскому слову или фразе, а ошибки распознавания проверяют эксперты [5].

Новгородский государственный университет. Магистрант Политехнического института НовГУ И. Филиппов разработал программу, которая распознаёт и толкует берестяные грамоты за несколько секунд. Система обучена на пяти наиболее объёмных уже расшифрованных грамотах; по словам разработчиков, точность после доработок достигла 98 %, а обучающая выборка пополняется новыми расшифровками пользователей; в дальнейшем систему планируют применить к рукописям [6].

Сбер и Российское историческое общество — «Цифровой Пётр». Пилотный проект по расшифровке почерка Петра I на документах РГАДА;

алгоритм обучали на опубликованных «Письмах и бумагах» императора. Из 143,8 тыс. документов петровского наследия к марту 2023 г. отсканировано и передано около 2,5 тыс. (около 1,5 %); проект рассматривается как образец для работы ИИ с другими архивными источниками [7].

Яндекс — «Поиск по архивам». Сервис запущен в январе 2023 г. и расшифровывает документы XVIII — начала XX в. в дореформенной орфографии: метрические книги, исповедные ведомости, ревизские сказки. Модель обучена на сотнях тысяч рукописных строк и десятках миллионов сгенерированных примеров; при запуске было доступно более 2,5 млн страниц с построчной расшифровкой, первым фондом стал Главархив Москвы [8]. К 2026 г. сервис содержит более 23 млн листов архивных документов XVIII — начала XX в. [9].

Таким образом, российские проекты различаются по объекту: богослужебные книги (МИФИ — ИРЯ РАН), берестяные грамоты (НовГУ), делопроизводство XVIII в. (Сбер — РИО) и массовые архивные записи для генеалогии (Яндекс). Сводное сравнение приведено в таблице 1.

Таблица 1 — Сравнение систем ИИ для обработки рукописей

Система	**Материал**	**Главная задача**	**Ключевой результат**	**Источник**
Ithaca	Древнегреческие надписи	Восстановление, место, дата	62 % восстановления; 71 % по региону; дата ±30 лет	[1]
Enoch	Рукописи Мёртвого моря	Датировка	Ошибка около 28–31 года	[2]
KuroNet	Японская скоропись кудзусидзи	Распознавание	Глубокое обучение для вариативной письменности	[3]
Transkribus	Архивные рукописи	Распознавание (HTR)	Массовая обработка архивов	[4]
МИФИ — ИРЯ РАН	Древнерусские минеи	Распознавание и поиск	245 миней, до 150 вариантов буквы	[5]
НовГУ	Берестяные грамоты	Распознавание и толкование	98 % по данным разработчиков	[6]

Сбер — РИО	Документы Петра I	Распознавание почерка	Пилот, ~2,5 тыс. документов	[7]
Яндекс	Метрические книги, ревизские сказки	Распознавание и поиск	Более 23 млн листов (2026)	[8; 9]

1.5. Ограничения существующих систем и выводы по главе 1

Общее ограничение всех систем — зависимость от размеченных данных. Зарубежные модели обучены на конкретных традициях письма и не переносятся на кириллические рукописи без собственных выборок. Для древнерусского письма это особенно остро из-за вариативности начертаний (до 150 вариантов одной буквы [5]), сокращений и надстрочных знаков.

Второе ограничение — проверяемость. Лучшие результаты получены там, где модель работает вместе с экспертом (Ithaca [1]) или опирается на независимые данные (радиоуглеродный анализ в Enoch [2]). Цифры точности российских проектов в основном приводятся по словам разработчиков в СМИ и требуют подтверждения в рецензируемых публикациях.

Третье ограничение — отсутствие механизма фиксации результатов. Ни одна из рассмотренных систем не отвечает на вопросы, кто и когда первым получил расшифровку, как менялось прочтение и чем машинное чтение отличается от экспертного в измеримых показателях.

Выводы по главе 1:

1. Мировой уровень задают системы, решающие сразу несколько задач — восстановление, датировку и атрибуцию.

2. Российские проекты пока сосредоточены на распознавании и поиске; задачи датировки и атрибуции писцов для древнерусских рукописей остаются открытыми.

3. Перспективное направление — схема «эксперт + ИИ» с прозрачной проверкой результатов и открытыми обучающими выборками. Для её реализации необходим слой доверия: фиксация авторства и даты каждого результата и количественное сравнение версий. Этому посвящена глава 2.

ГЛАВА 2. ДОВЕРЕННАЯ ФИКСАЦИЯ И СРАВНЕНИЕ РАСШИФРОВОК НА ПЛАТФОРМЕ ACADEMIC GATE

2.1. Постановка задачи

Из обзора в главе 1 следуют четыре требования к инструменту, который дополнит модели распознавания:

1. Приоритет: результат (расшифровка, комментарий, заключение) должен получать неизменяемую отметку автора и времени.

2. История: каждая правка чтения сохраняется как новая версия, предыдущие версии не теряются.

3. Измеримость: различие двух версий текста выражается проверяемыми числовыми показателями, а не только словесной оценкой.

4. Подчинённость ИИ фактам: если заключение пишет языковая модель, она не должна противоречить вычисленным показателям; итоговую оценку подписывает эксперт.

Эти требования реализуются на платформе Academic Gate (academicgate.su), разработанной в Международной научно-исследовательской лаборатории стратегии управления социально-технологических систем МНИИПУ [10].

2.2. Архитектура платформы Academic Gate

Платформа интегрирована через API с блокчейном Erachain и работает в сайдчейне лаборатории IP-lab [11]. Применительно к палеографии она даёт четыре инструмента.

Фиксация приоритета расшифровки. Отпечаток SHA-256 файла с расшифровкой или комментарием эксперта записывается в блокчейн. Запись неизменяема и привязана к удостоверенной личности автора, поэтому подтверждает, кто и когда получил результат. Принцип восходит к доверенной отметке времени С. Хабера и У. Сторнетты [12] и к блокчейну Биткоина [13]. На платформе этим способом уже закреплены книги автора, а также настоящая курсовая работа.

Версионирование чтений. Каждая правка расшифровки сохраняется как новая версия с отпечатком, поэтому видно, как менялось прочтение и кто вносил изменения.

Проверка по схеме «эксперт + ИИ». Модуль сравнения текстов сначала вычисляет проверяемые показатели, затем языковая модель GigaChat готовит заключение, получая эти показатели как установленные факты. Заключение сохраняется под номером с отпечатком SHA-256 и может быть закреплено в Erachain; итоговую оценку подписывает эксперт. Метод подробно описан в параграфе 2.3.

Распознавание речи и текста. Сервис расшифровки на основе GigaChat переводит аудиозаписи в текст (на платформе выполнено 57 расшифровок); эту же связку можно применить к надиктованным экспертом чтениям рукописей.

Помимо инструментов, непосредственно применимых к палеографии, в 2026 году на платформе реализован ряд модулей, образующих общую инфраструктуру доверенной научной работы: СКВ-карточка и «семидольник» проекта с фиксацией разделов в Erachain (запись № 3360914-1), патентный ассистент, ИИ-помощник, защищённые файлы с шифрованием ключами Erachain, визитки исследователей с книгами и услугами, сертификаты авторства, модули «Срез К1–К4» и «Датасет 3 × 3» для сбора и проверки эмпирических данных, ИИ-боты в Telegram и MAX, новостной канал с фиксацией публикаций в блокчейне. Полный перечень проделанной работы с указанием назначения каждого компонента и его значения для палеографии приведён в приложении В.

Таким образом, платформа не заменяет модели распознавания из параграфов 1.3–1.4, а дополняет их слоем доверия: каждый результат ИИ и эксперта получает автора, дату и неизменяемый след.

2.3. Метод сравнения текстов

Модуль сравнения сопоставляет базовый текст и его преобразованную версию — например, два чтения одной рукописи, машинную и экспертную расшифровку или две редакции издания. Обработка проходит шесть шагов.

1. Извлечение текста. Из PDF текст извлекается утилитой pdftotext, из EPUB — разбором HTML-страниц книги. Если слов меньше 150, документ считается сканом без текстового слоя и сравнение не выполняется.

2. Нормализация. Текст приводится к нижнему регистру и разбивается на слова кириллицы и латиницы; список литературы в конце документа исключается, чтобы ссылки не искажали показатели.

3. Поиск дословных совпадений. Каждый текст разбивается на шинглы — последовательности из шести подряд идущих слов [14]. Совпадение шингла означает дословно перенесённую фразу.

4. Структура. Заголовками считаются строки вида «Глава N», «Введение», «Заключение» и нумерованные подразделы «1.1.»; повторы из оглавления отбрасываются.

5. Терминология. Слова длиной от семи букв сравниваются по основе из первых шести букв, поэтому «биткоины» и «биткоином» считаются одним термином; месяцы, служебные слова и имя автора исключаются.

6. Заключение. По показателям строятся автоматические выводы по фиксированным правилам, затем языковая модель GigaChat пишет заключение и получает эти выводы как факты, которым нельзя противоречить.

Основные показатели вычисляются так. Пусть Sb и Sn — множества шинглов базового и преобразованного текста, Vb и Vn — их словари, а Ln — список всех шинглов преобразованного текста. Тогда

$$K = (\text{число шинглов из Ln, входящих в Sb}) / |Ln| \times 100\ \%, \quad (1)$$

$$R = |Sb \cap Sn| / |Sb| \times 100\ \%, \quad (2)$$

$$J = |Vb \cap Vn| / |Vb \cup Vn| \times 100\ \%, \quad (3)$$

где K — доля преобразованного текста, совпадающая с базовым дословно; R — доля базового текста, перенесённая в новый; J — общность словаря (коэффициент Жаккара [15]). Доля нового или переписанного текста равна 100 % − K.

Автоматические выводы формируются по правилам: при K < 5 % текст признаётся написанным заново, при 5 % ≤ K < 30 % — переписанным в основной части, при K ≥ 30 % — требующим переработки; J < 30 % трактуется как признак глубокой переработки, а не низкой оригинальности. Если преобразованный текст помечен как черновик, сравнение объёма не оценивается. Эти правила понадобились на практике: без них модель ошибочно описывала низкую общность словаря как заимствование и делала выводы о «сокращении объёма» черновика.

Заключение вместе с показателями сериализуется в JSON, получает номер вида AG-CMP-NNNN и отпечаток SHA-256; отпечаток можно закрепить в Erachain, а итоговую оценку подписывает эксперт. Состав полей заключения приведён в приложении А.

2.4. Апробация на паре «исходный текст — переработанная монография»

Первое заключение платформы, № AG-CMP-0001, сравнивает исходную книгу — подборку материалов «Биткоин: краткое руководство» (2018) [17] — и черновик переработанной монографии того же автора, подготовленной в МНИИПУ. Результаты приведены в таблице 2.

Таблица 2 — Заключение № AG-CMP-0001

Показатель	**Базовый текст**	**Преобразованный текст**
Объём, слов	11 695	1 740 (черновик)
Дословные совпадения K	—	0 %
Общность словаря J	—	9,3 %
Ссылок на источники	0	17
Разделов и подразделов	0	17

Для проверки воспроизводимости показатели были независимо пересчитаны отдельной программой (приложение Б) по тем же файлам.

Получено: объём 12 034 и 1 901 слово, K = 0,0 %, R = 0,0 %, J = 10,0 %. Расхождения в объёме и J (около 3 % и 0,7 п. п.) объясняются различиями токенизации — в контрольной программе цифры считаются словами, а «ё» приводится к «е»; на главный показатель K они не влияют. Вывод заключения — текст полностью переписан, с новой структурой и научным аппаратом — подтверждён независимым расчётом.

2.5. Апробация на модельном палеографическом примере

Чтобы оценить применимость метода к палеографии, был поставлен модельный эксперимент. В качестве экспертного чтения взят начальный фрагмент «Повести временных лет» по Лаврентьевскому списку [18]: «Се повѣсти времяньных лѣт, откуду есть пошла руская земля, кто въ Киевѣ нача первѣе княжити, и откуду руская земля стала есть». В качестве машинного чтения смоделирована расшифровка с типичными для HTR ошибками: замена «ѣ» на «е», потеря «ъ», искажение отдельных букв («временьных», «кнажити»). Пример модельный: он иллюстрирует поведение показателей и не является результатом работы конкретной системы распознавания. Результаты приведены в таблице 3.

Таблица 3 — Показатели сравнения машинного и экспертного чтения (модельный пример)

Показатель	**Значение**	**Интерпретация**
CER, ошибки на уровне знаков	5,6 % (7 правок на 126 знаков)	Высокое сходство
WER, ошибки на уровне слов	33,3 % (7 из 21 слова)	Каждое третье слово отличается
K при шинглах из 6 слов	12,5 %	Ложный вывод «текст переписан»
K при шинглах из 3 слов	42,1 %	Ближе к реальному сходству
J, общность словаря	41,7 %	Занижена из-за орфографии
CER после нормализации графики	1,6 %	Остаются только содержательные ошибки

Эксперимент выявил важное методическое ограничение. Показатели, рассчитанные на уровне слов и шестисловных шинглов, разработаны для обнаружения заимствований в современных текстах. В палеографии почти

одинаковые чтения различаются отдельными буквами, и одна ошибка распознавания «ломает» сразу шесть шинглов: при CER 5,6 % показатель K составил лишь 12,5 %, что по правилам из параграфа 2.3 привело бы к ложному выводу о «переписанном» тексте.

Отсюда следуют три доработки метода для палеографических задач:

1. Нормализация графики перед сравнением: приведение «ѣ» к «е», «і» к «и», снятие конечного «ъ», раскрытие титл по словарю сокращений. После такой нормализации CER снизился с 5,6 до 1,6 %, и остались только содержательные расхождения.

2. Основной показатель — CER на уровне знаков, дополнительный — WER; шинглы использовать с длиной 2–3 слова.

3. Пороговые правила автоматических выводов задавать отдельно для режима «два чтения одной рукописи» — например, по CER: менее 2 % — чтения совпадают, 2–10 % — расхождения в отдельных знаках, свыше 10 % — существенно разные чтения. Пороги требуют калибровки на размеченной выборке.

2.6. Протокол «эксперт + ИИ» для расшифровки рукописи

На основе параграфов 2.2–2.5 предлагается следующий протокол работы с рукописью на платформе:

1. Загрузка изображения страницы; его отпечаток SHA-256 фиксируется в Erachain как исходный объект.

2. Машинное распознавание внешней моделью (Transkribus, модели проектов из параграфа 1.4 или собственная); результат сохраняется как версия 1 с указанием модели.

3. Экспертное чтение — правка текста или диктовка с последующей расшифровкой речи; результат сохраняется как версия 2 с указанием эксперта.

4. Сравнение версий модулем сравнения в палеографическом режиме (нормализация графики, CER, WER, короткие шинглы); список расхождений по знакам.

5. Заключение языковой модели, подчинённое вычисленным показателям, и подпись эксперта.

6. Фиксация заключения AG-CMP-NNNN и итоговой версии в Erachain.

7. Пополнение обучающей выборки: пары «изображение — экспертное чтение» с зафиксированным авторством передаются в открытый датасет.

Протокол отвечает на все четыре требования из параграфа 2.1 и одновременно решает проблему, названную в главе 1 ключевой, — нехватку размеченных данных: каждая проверенная экспертом расшифровка становится атрибутированным обучающим примером.

2.7. Выводы по главе 2

1. Платформа Academic Gate дополняет модели распознавания слоем доверия: фиксацией автора и времени результата в блокчейне Erachain, версионированием чтений и проверяемыми заключениями.

2. Метод сравнения текстов формализован в виде трёх показателей (K, R, J) и правил автоматических выводов, которым подчинено заключение языковой модели.

3. Апробация на реальной паре текстов (заключение № AG-CMP-0001) подтверждена независимым пересчётом: K = 0 %, текст полностью переписан.

4. Модельный палеографический эксперимент показал, что для сравнения чтений рукописей метод необходимо дополнить нормализацией графики и показателем CER; без этого шестисловные шинглы дают ложный вывод о существенном расхождении.

5. Предложен протокол «эксперт + ИИ» из семи шагов, превращающий каждую проверенную расшифровку в атрибутированный обучающий пример.

ЗАКЛЮЧЕНИЕ

В курсовой работе рассмотрено применение искусственного интеллекта в палеографии и разработана методика доверенной фиксации и количественного сравнения расшифровок рукописей.

В первой главе описаны четыре задачи палеографии, которые решаются методами ИИ, — распознавание, датировка, атрибуция и восстановление, — и типовой конвейер обработки рукописи. Обзор зарубежных систем (Ithaca, Enoch, KuroNet, Transkribus) и российских проектов (МИФИ — ИРЯ РАН, НовГУ, Сбер — РИО, Яндекс) показал, что мировой уровень задают системы, решающие несколько задач одновременно, а российские проекты пока сосредоточены на распознавании и поиске. Выявлены три общих ограничения: зависимость от размеченных данных, недостаточная проверяемость заявленных результатов и отсутствие механизма фиксации приоритета и истории чтений.

Во второй главе описана платформа Academic Gate, разработанная в МНИИПУ и интегрированная с блокчейном Erachain, и формализован метод сравнения текстов с показателями K, R и J. Апробация на паре «исходный текст — переработанная монография» (заключение № AG-CMP-0001) подтверждена независимым пересчётом. Модельный палеографический эксперимент выявил ограничение метода — чувствительность шестисловных шинглов к единичным ошибкам распознавания — и позволил сформулировать доработки: нормализацию графики, использование CER и отдельные пороги для режима сравнения чтений. Предложен протокол «эксперт + ИИ» из семи шагов.

Таким образом, цель работы достигнута, поставленные задачи решены. Научная новизна состоит в соединении методов сравнения текстов с блокчейн-фиксацией результатов применительно к палеографии и в обосновании палеографического режима сравнения на основе вычислительного эксперимента.

Направления дальнейшей работы: реализация палеографического режима в модуле сравнения; калибровка пороговых правил на размеченной выборке древнерусских рукописей; апробация протокола на реальных документах совместно с архивом или библиотекой; подтверждение заявленных точностей российских систем по рецензируемым публикациям.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Assael Y., Sommerschield T., Shillingford B., Bordbar M., Pavlopoulos J., Chatzipanagiotou M., Androutsopoulos I., Prag J., de Freitas N. Restoring and attributing ancient texts using deep neural networks // Nature. 2022. Vol. 603. P. 280–283. DOI: 10.1038/s41586-022-04448-z.

2. Popović M., Dhali M. A., Schomaker L., van der Plicht J., Rasmussen K. L., La Nasa J., Degano I., Colombini M. P., Tigchelaar E. Dating ancient manuscripts using radiocarbon and AI-based writing style analysis // PLOS ONE. 2025. Vol. 20, No. 6. e0323185. DOI: 10.1371/journal.pone.0323185.

3. Clanuwat T., Lamb A., Kitamoto A. KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning // 2019 International Conference on Document Analysis and Recognition (ICDAR). 2019. P. 607–614. DOI: 10.1109/ICDAR.2019.00103.

4. Muehlberger G., Seaward L., Terras M. et al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study // Journal of Documentation. 2019. Vol. 75, No. 5. P. 954–976. DOI: 10.1108/JD-07-2018-0114.

5. Учёные создали поисковую систему по старославянским рукописям на основе ИИ // Хабр. 29.05.2024. URL: https://habr.com/ru/news/818059/ (дата обращения: 01.11.2026).

6. В Великом Новгороде ИИ учится читать древнерусские тексты // Рамблер. 29.10.2025. URL: https://weekend.rambler.ru/read/55541753-v-velikom-novgorode-ii-uchitsya-chitat-drevnerusskie-teksty/ (дата обращения: 01.11.2026).

7. Будет ли продолжен проект по расшифровке сложного почерка Петра Первого // Российская газета. 22.03.2023. URL: https://rg.ru/2023/03/22/budet-li-prodolzhen-proekt-po-rasshifrovke-slozhnogo-pocherka-petra-pervogo.html (дата обращения: 01.11.2026).

8. «Яндекс» научил нейросети расшифровывать архивные записи со сложной орфографией // Лайфхакер. 25.01.2023. URL: https://lifehacker.ru/poisk-po-arxivam-yandeks/ (дата обращения: 01.11.2026).

9. О Поиске по архивам [Электронный ресурс] // Яндекс. URL: https://yandex.ru/archive/help (дата обращения: 01.11.2026).

10. Academic Gate — единый научный шлюз [Электронный ресурс]. МНИИПУ, 2026. URL: https://academicgate.su (дата обращения: 01.11.2026).

11. Erachain: документация [Электронный ресурс]. URL: https://gitlab.com/erachain (дата обращения: 01.11.2026).

12. Haber S., Stornetta W. S. How to time-stamp a digital document // Journal of Cryptology. 1991. Vol. 3, No. 2. P. 99–111.

13. Nakamoto S. Bitcoin: A Peer-to-Peer Electronic Cash System. 2008. URL: https://bitcoin.org/bitcoin.pdf (дата обращения: 01.11.2026).

14. Broder A. Z. On the resemblance and containment of documents // Proceedings of Compression and Complexity of Sequences 1997. IEEE, 1997. P. 21–29.

15. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et du Jura // Bulletin de la Société vaudoise des sciences naturelles. 1901. Vol. 37. P. 547–579.

16. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. 1965. Т. 163, № 4. С. 845–848.

17. Биткоин: краткое руководство. Подборка лучшего материала. Cryptomir.online, 2018. 65 с.

18. Полное собрание русских летописей. Т. 1. Лаврентьевская летопись. 2-е изд. Л.: Изд-во АН СССР, 1926–1928.

ПРИЛОЖЕНИЕ А. СОСТАВ ЗАКЛЮЧЕНИЯ AG-CMP

Таблица А.1 — Поля номерного заключения модуля сравнения

Поле	**Содержание**
Номер	AG-CMP-NNNN, присваивается последовательно
Дата и автор запроса	Учётная запись исследователя на платформе
Исходные документы	Названия и отпечатки SHA-256 базового и преобразованного текстов
Объём	Число слов в каждом тексте; признак черновика
Показатели	K, R, J (формулы 1–3)
Структура	Список разделов и подразделов каждого текста
Терминология	Появившиеся и ушедшие термины
Автоматические выводы	Результат применения пороговых правил
Заключение	Текст языковой модели, подчинённый выводам
Отпечаток	SHA-256 сериализованного JSON заключения
Фиксация	Номер записи в блокчейне Erachain (при закреплении)
Подпись эксперта	Итоговая оценка и подпись ответственного эксперта

ПРИЛОЖЕНИЕ Б. ПРОГРАММА КОНТРОЛЬНОГО РАСЧЁТА ПОКАЗАТЕЛЕЙ

Ниже приведена программа на языке Python, использованная для независимого пересчёта показателей в параграфах 2.4 и 2.5.

```
import re

def words(t):
    t = t.lower().replace('ё', 'е')
    return re.findall(r'[а-яa-z0-9ѣъіѳv]+', t)

def strip_refs(t):
    # список литературы в конце документа не сравнивается
    i = max(t.rfind('Список литературы'), t.rfind('СПИСОК ЛИТЕРАТУРЫ'))
    return t[:i] if i > len(t) * 0.5 else t

def shingles(w, n=6):
    return [' '.join(w[i:i + n]) for i in range(len(w) - n + 1)]

def metrics(base, new, n=6):
    wb, wn = words(strip_refs(base)), words(strip_refs(new))
    Sb = set(shingles(wb, n)); Ln = shingles(wn, n); Sn = set(Ln)
    K = sum(s in Sb for s in Ln) / max(1, len(Ln)) * 100
    R = len(Sb & Sn) / max(1, len(Sb)) * 100
    J = len(set(wb) & set(wn)) / len(set(wb) | set(wn)) * 100
    return K, R, J

def levenshtein(a, b):
    prev = list(range(len(b) + 1))
    for i, ca in enumerate(a, 1):
        cur = [i]
        for j, cb in enumerate(b, 1):
            cur.append(min(prev[j] + 1, cur[j - 1] + 1,
                           prev[j - 1] + (ca != cb)))
        prev = cur
    return prev[-1]

def cer(machine, expert):
    return levenshtein(machine, expert) / len(expert) * 100
```

ПРИЛОЖЕНИЕ В. ПЕРЕЧЕНЬ РАБОТ ПО ПРОЕКТУ ACADEMIC GATE

Платформа Academic Gate (academicgate.su) разработана автором в 2026 году в Международной научно-исследовательской лаборатории стратегии управления социально-технологических систем МНИИПУ под научным руководством С. В. Кибальникова. Полный цикл семидольника проекта принят наставником — проф. С. В. Кибальниковым — и закреплён в блокчейне Erachain: 7 записей фиксации с отпечатками SHA-256 и электронными подписями автора и наставника. В таблице В.1 перечислены реализованные компоненты.

Таблица В.1 — Реализованные компоненты платформы Academic Gate

Компонент	**Назначение**	**Значение для палеографии**
Фиксация в Erachain	Отпечаток SHA-256 и подпись автора записываются в блокчейн (сайдчейн IP-lab); пример — запись № 3360914-1	Приоритет расшифровки и комментария эксперта
СКВ-карточка и «семидольник»	Описание проекта по 7 разделам жизненного цикла с фиксацией каждого раздела	Структурирование исследовательского проекта по рукописи
Модуль сравнения текстов	Показатели K, R, J, номерные заключения AG-CMP с отпечатком; первое — № AG-CMP-0001	Количественное сравнение чтений (глава 2)
Сервис транскрибации	Перевод аудиозаписей в текст на основе GigaChat; выполнено 57 расшифровок	Расшифровка надиктованных экспертом чтений
ИИ-помощник и патентный ассистент	Ответы на вопросы, подготовка материалов патентной заявки	Подготовка описаний и заявок по результатам
Защищённые файлы	Шифрование файлов ключами Erachain, передача адресату	Работа с неопубликованными материалами архивов
Визитки исследователей	Профиль, услуги, книги (PDF/EPUB до 50 МБ) с фиксацией	Публикация расшифровок и изданий с подтверждённым авторством
Сертификаты авторства	Документ с номером, отпечатком и ссылкой на запись в блокчейне	Подтверждение вклада эксперта
Модуль «Срез К1–К4»	Анонимные срезы по кодам участников, группы А/Б,	Сбор экспертных оценок без персональных данных

	выгрузка CSV, отпечаток данных в Erachain	
Модуль «Датасет 3 × 3»	Расчёт $P = \Sigma YZ/\Sigma Z^2$, $R = 1/P$, R^2, проверка прогноза, фиксация датасета	Пример воспроизводимого эмпирического датасета
Команды и кабинет	Совместная работа групп, кабинет с вкладками, раздел «Исследования», мобильная версия (PWA)	Работа распределённой группы палеографов
ИИ-боты и каналы	Боты в Telegram и MAX отвечают о платформе; канал t.me/academicgate и канал в MAX с фиксацией новостей в блокчейне	Информирование о результатах
Регистрация и пилоты	Пакет для регистрации программы для ЭВМ в Роспатенте; пакет пилотов для лабораторий, пилот с МФТИ	Правовая охрана и внедрение