

Историческая информатика*Правильная ссылка на статью:*

Кузнецов А.В. За пределами тематического моделирования: анализ исторического текста с помощью больших языковых моделей // Историческая информатика. 2024. № 4. DOI: 10.7256/2585-7797.2024.4.72560 EDN: UOIKPJ URL: https://nbpublish.com/library_read_article.php?id=72560

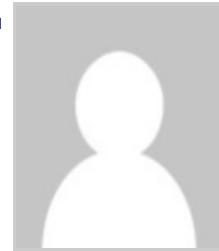
За пределами тематического моделирования: анализ исторического текста с помощью больших языковых моделей**Кузнецов Алексей Валерьевич**

ORCID: 0000-0003-4755-250X

кандидат исторических наук

научный сотрудник; Институт всеобщей истории РАН

119334, Россия, г. Москва, Ленинский пр-т, 32 а, оф. 1426



✉ kuznetsovaleks@rambler.ru

[Статья из рубрики "Компьютеризованный анализ исторических текстов"](#)**DOI:**

10.7256/2585-7797.2024.4.72560

EDN:

UOIKPJ

Дата направления статьи в редакцию:

03-12-2024

Дата публикации:

10-12-2024

Аннотация: Статья посвящена исследованию потенциала больших языковых моделей для тематического анализа исторических текстов. В качестве материала для исследования выбран дневник вологодского гимназиста Кирилла Антоновича Березкина, охватывающий период с января по сентябрь 1849 г. Этот дневник представляет собой уникальный источник для изучения повседневной жизни, мировоззрения и социальных взаимодействий молодого человека в провинциальной России середины XIX века. Записи К. Березкина затрагивают широкий спектр тем от учебы и семейных отношений до размышлений о политике, религии, литературе и искусстве. Анализ дневника позволяет погрузиться в атмосферу эпохи, понять ценности и стремления молодого

дворянина на пороге взрослой жизни, а также проследить формирование его личности под влиянием исторического контекста. В исследовании применялась модель Gemini 1.5 Pro, которая способна обрабатывать объемные текстовые данные. Текст дневника анализировался как целиком, так и по отдельным месяцам, что позволило выявить специфические аспекты содержания. Новизна исследования заключается в применении современных больших языковых моделей для анализа русскоязычного исторического текста. Результаты показали, что Gemini 1.5 Pro способна эффективно выделять ключевые темы в дневнике, предоставляя интерпретируемые результаты. Модель успешно идентифицировала восемь основных тем, отражающих различные аспекты жизни гимназиста. Применение параллельного промптинга с помесячным разбиением текста позволило выявить более специфические темы и нюансы, которые упускались при анализе полного текста. Исследование подтвердило эффективность больших языковых моделей для анализа исторических источников, открывая новые возможности для автоматизации процесса тематического моделирования и выявления скрытых паттернов в больших массивах текстовых данных. Однако, стохастическая природа больших языковых моделей требует многократного анализа и тщательной интерпретации результатов, а также критического сопоставления с данными традиционных методов исторического анализа.

Ключевые слова:

большие языковые модели, тематическое моделирование, дневник, исторический источник, анализ текста, XIX век, цифровая история, промпт-инжиниринг, искусственный интеллект, К. А. Березкин

Введение

Тематическое моделирование – распространенный метод компьютерного анализа текстов, позволяющий выявлять скрытые темы в больших коллекциях документов. За прошедшие годы разработано множество вариантов построения тематических моделей [6, с. 63], [12, р. 410]. В гуманитарных науках, включая исторические исследования, широкое применение нашли вероятностные модели [3], такие как скрытое размещение Дирихле (англ. Latent Dirichlet Allocation, LDA) [10]. LDA и подобные алгоритмы позволяют выделять ключевые темы и отслеживать их динамику в больших объемах текстовой информации.

Одним из первых примеров использования тематического моделирования в исторических исследованиях стал проект анализа газеты Pennsylvania Gazette за 1728–1800 годы, результаты которого были опубликованы в 2006 году [30]. Начиная с 2010-х гг. метод используется для анализа дневников [11], писем [27], периодических изданий [41], сборников статей [18], литературных корпусов [21] и др.

Но несмотря на их эффективность, традиционные методы тематического моделирования, сталкиваются с рядом методологических трудностей [8], [39]. Во многом результаты зависят от качества предварительной обработки текста [14] и оптимальной настройки гиперпараметров, таких как количество тем [7], [9], [31, р. 429–433]. Кроме того, представление результатов в виде списков ключевых слов часто вызывает трудности с интерпретацией, требуя значительных усилий исследователя [5], [17], [31]. Еще одним

серьезным ограничением является низкая эффективность традиционных методов при работе с короткими текстами и небольшими наборами данных. В таких случаях недостаток совместно встречающихся слов снижает точность выделения тем. Эти проблемы подчеркивают необходимость поиска альтернативных подходов, способных преодолеть ограничения традиционного тематического моделирования.

Новый этап в развитии компьютерного анализа текстов и тематического моделирования в том числе начался с появлением больших языковых моделей. Благодаря способности учитывать контекст, семантические связи и лингвистические нюансы, они могут преодолеть ограничения традиционных методов. Уже сейчас очевидно, что их использование открывает новые горизонты для исторической науки и гуманитарных исследований в целом [\[16\]](#),[\[19\]](#),[\[22\]](#),[\[24\]](#). Однако потенциал больших языковых моделей для исторической науки ещё только изучается. В данной статье исследуются возможности универсальных больших языковых моделей для анализа исторических текстов на примере дневника вологодского гимназиста середины XIX века. Предполагается, что использование таких моделей позволит получать интерпретируемые и детализированные результаты при тематическом анализе текстов. Основные цели исследования включают оценку эффективности больших языковых моделей, выявление их ограничений в выделении ключевых тем из русскоязычных исторических документов, а также разработку практических рекомендаций для их применения в гуманитарных исследованиях.

Обзор литературы

Большие языковые модели открывают новые возможности для тематического моделирования, предоставляя альтернативу и дополнение традиционным подходам. Они успешно применяются для интерпретации [\[13\]](#),[\[25\]](#) и оценки качества результатов тематического моделирования [\[38\]](#).

В последнее время усилия исследователей сосредоточены на экспериментах по выделению тем исключительно с использованием больших языковых моделей, без привлечения традиционных алгоритмов. В таких подходах модель получает промпты (запросы), которые формулируют задачу по выделению ключевых тем из текста. В основе такого подхода – способность моделей к эффективной суммаризации и обобщению информации [\[35\]](#). По сути модели предлагается «скать» текст до уровня обсуждаемых в нем тем. Даже простые запросы, например, «Выделите основную тему из текста: [текст]», демонстрируют высокое качество результатов [\[37, р. 239\]](#). Для специфических задач промпты адаптируются: моделям задается определенная роль или формат ответа, что позволяет систематизировать и уточнять результаты [\[23\]](#),[\[29\]](#),[\[40\]](#).

Экспериментальные данные подтверждают, что большие языковые модели превосходят традиционные методы по некоторым критериям. Они обеспечивают более высокую когерентность (согласованность) выделяемых тем, их интерпретируемость и способны работать с короткими текстами, где традиционные алгоритмы обычно испытывают трудности [\[29\]](#),[\[40\]](#).

Однако применение больших языковых моделей для тематического моделирования сопряжено с рядом вызовов, которые требуют учета в исследовательской практике. Во-первых, ограниченность объема входных данных, которые может обработать модель. Например, ChatGPT-4 способен анализировать до 128 000 токенов, Claude – до 200 000 токенов, а Gemini 1.5 Pro – до 2 миллионов токенов. Для работы с большими текстами

применяют два подхода: параллельный и последовательный промпting [15, р. 24]. В параллельном подходе текст делится на части, каждая из которых обрабатывается независимо, а результаты затем обобщаются. Последовательный подход предполагает этапный анализ текста, где модель накапливает информацию и уточняет темы на каждом шаге. Во-вторых, модели генерируют темы с разным уровнем детализации от излишне общих до чрезмерно специфичных [28]. Для решения этой проблемы используют итеративный подход, когда список тем уточняется с помощью дополнительных запросов [29], или добавляют в запрос примеры тем (seed topics) для управления уровнем детализации [15]. Третья проблема – галлюцинации, когда модели создают правдоподобную, но неверную информацию, включая несуществующие темы [20],[42]. Наконец, модели сталкиваются с так называемой «потерей середины контекста» [26]. При анализе длинных текстов они могут упускать информацию из средних частей, что снижает точность и полноту выделяемых тем. Это ограничение требует дробного подхода к анализу текста, чтобы улучшить охват и детальность тематической структуры.

Несмотря на существующие ограничения, большие языковые модели демонстрируют значительный прогресс в тематическом моделировании. Перспективы их использования включают не только автоматизацию анализа, но и более глубокую интерпретацию текстов, что делает их мощным инструментом для гуманитарных исследований.

Материал исследования

В качестве материала для исследования был выбран дневник Кирилла Антоновича Березкина (1830-1908), охватывающий период с 1 января по 12 сентября 1849 г. Дневник велся в последний год его обучения в Вологодской губернской мужской гимназии. Источник предоставляет ценный материал для изучения повседневной жизни, мировоззрения и социальных взаимодействий молодого человека в провинциальной России середины XIX в. Автор дневника, потомственный дворянин, делится своими мыслями и наблюдениями о самых разных аспектах жизни: учебе в гимназии, отношениях с семьей и друзьями, светских развлечениях, культурных событиях, политической обстановке и своих внутренних переживаниях. Дневник был впервые опубликован в 2014 г. [4] и переиздан в 2020 г. с дополнениями [2]. В нашем исследовании использовалась электронная версия дневника, доступная на сайте проекта «Прожито» [1].

Методология исследования

Для анализа тематики дневника К. Березкина была использована модель Gemini 1.5 Pro от корпорации Google (в Российской Федерации признана экстремистской организацией). Указанная модель выделяется уникальной способностью обрабатывать входные данные объемом до 2 миллионов токенов – это примерно вдвое превышает объем текста романа «Война и мир» Л.Н. Толстого. Модель демонстрирует высокую эффективность работы с русским языком, занимает верхние строчки рейтингов языковых моделей (см. рис. 1), где соперничает с последними разработками Open AI и Anthropic. Существенным преимуществом для нас является то, что доступ к ней предоставляется бесплатно с незначительными ограничениями по количеству запросов в день. Анализ проводился на платформе Google AI Studio (<https://aistudio.google.com/>), куда был загружен полный текст дневника в формате .txt. Текст объемом 90 708 слов был токенизирован моделью на 160 477 токенов.

Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization
1	Google: Gemini Pro 1.5	1132	+26/-25	532	Google
1	claude-3-5-sonnet-20241022	1116	+40/-35	255	Anthropic
1	gpt-4o-2024-05-13	1107	+15/-17	1416	OpenAI
1	claude-3-5-sonnet-20240620	1097	+19/-19	1003	Anthropic
1	Google: Gemini Flash 1.5	1093	+24/-26	593	Google
3	gpt-4-turbo-2024-04-09	1073	+15/-16	1386	OpenAI

Рисунок 1. Модель Gemini 1.5 Pro в рейтинге на платформе LLM Arena (<https://llmarena.ru/>) от 05.12.2024.

Эффективность языковых моделей в задачах тематического моделирования во многом зависит от подхода к созданию промптов [29],[32]. Анализ исторических документов требует учета не только текста, но и культурного, исторического и социального контекста эпохи. При составлении промптов мы опирались на современные представления о их структуре и функционале [33],[36] и опыт составления промптов в исторических исследованиях [24],[34].

Предлагаемый нами промпт состоит из двух частей: системного и пользовательского промпта. Системный промпт задает модели роль, определяет контекст и план исследования, дает общие методические рекомендации:

Вы – эксперт по истории Российской Империи XIX века и специалист по анализу личных дневников. Ваша задача – провести анализ текста дневника и определить ключевые темы с учетом культурного, исторического и социального контекста эпохи.

Контекст:

- Автор дневника: Кирилл Антонович Березкин, гимназист Вологодской губернской гимназии, потомственный дворянин.- Период ведения дневника: январь 1849 – сентябрь 1849.- Место: город Вологда, Вологодская губерния.
- Историческая значимость: дневник отражает восприятие важных исторических событий в российской провинции.

План действий:

1. Прочтите текст дневника.
2. Определите ключевые темы на основе их частоты, значимости для автора и эмоциональной окраски.
3. Опишите темы, учитывая их связь с контекстом XIX века.
4. Приведите примеры в виде цитат с указанием даты записи.
6. Обратите внимание на использование архаизмов, историзмов и культурных отсылок в тексте.

Методические рекомендации:

- Подходите к анализу с междисциплинарной точки зрения, связывая темы с историей,

социологией и педагогикой.

- Выбирайте цитаты, которые наиболее репрезентативно и ярко иллюстрируют тему.

Пользовательский промпт дает модели конкретные инструкции, определяет формат ответа и включает образец ответа:

Проведите анализ тематики текста дневника гимназиста из Вологды середины XIX века. Каждая отдельная запись в дневнике начинается знаками ###. Определите и опишите основные темы, обсуждаемые в дневнике.

Представьте результаты в формате списка по следующей схеме:

- Название темы: [краткое название темы]

- Описание: [подробное описание темы, включая её значимость и возможное отражение культурных и социальных реалий того времени]

- Пример: [2-3 цитаты из текста дневника со ссылкой на дату записи, которые иллюстрируют эту тему и подчеркивают её значимость]

Образец ответа:

1. Название темы: Образование и школьные будни

- Описание: Тема отражает повседневную жизнь гимназиста, включая занятия, учебные достижения и трудности, с которыми он сталкивается. Эта тема помогает понять, каким образом система образования в XIX веке формировалась личность и мировоззрение молодежи.

- Пример: «Сегодня мы писали экзамен по латинскому языку, и, несмотря на все старания, получил только "удовлетворительно"». (25 марта 1849)

Учитывая проблему «потери середины контекста» языковыми моделями [26], мы решили изучить также насколько объем предоставленного текста влияет на глубину и детализацию выделения тем. Для этого текст дневника был разделен помесечно на девять частей. После этого с помощью разработанного нами скрипта на языке Python и доступного API сервиса PROXYAPI (<https://proxyapi.ru/>) провели анализ тематики каждой части в автоматическом режиме. Анализ проводился в этом случае на платформе Google Colab (<https://colab.research.google.com/>) с той же моделью Gemini 1.5 Pro. В этом случае использовался подход «параллельного промптинга» [15], когда каждая часть анализировалась отдельно, а затем модели предлагалось обобщить результаты.

В обоих случаях дабы избежать галлюцинаций модели параметр temperature задавался равным 0,2. Параметр temperature в языковых моделях регулирует степень случайности или «кreativnosti» в генерируемых текстах. При значении слишком близком к 0 модель выбирает наиболее вероятные варианты ответа, что делает текст предсказуемым и менее разнообразным. Этот режим подходит для задач, требующих высокой точности в том числе суммаризации текста.

Мы также протестировали возможность управления уровнем детализации выделяемых тем, добавляя в промпт примеры тем, как предложено в работе [15]. На завершающем этапе исследования мы решили проверить, способна ли модель проследить изменения в тематике дневника с течением времени. Для этого мы предложили модели следующий запрос:

Проследите как выделенные в дневнике темы изменялись со временем.

Разбейте текст дневника на временные периоды по отдельным месяцам.

Для каждого периода определите, какие темы были преобладающими, как менялись интересы и настроение автора?

Объясните, с какими событиями в жизни автора связаны такие изменения в тематике?

Для ответа используйте следующую схему:

Период 1:

Преобладающие темы: [Приведите список преобладающих тем]

Описание: [Опишите, как меняются темы в дневнике, а также объясните причины изменений: с какими событиями в жизни автора это связано]

Все материалы, скрипты и ответы модели доступны в репозитории GitHub:
https://github.com/alexeyvkuznetsov/topic_modeling_with_llm

Результаты исследования

Большие языковые модели обладают стохастической природой, что означает, что их ответы формируются на основе вероятностных распределений, а не жестко запрограммированных правил. Это делает их гибкими и креативными, позволяя учитывать широкий контекст, но также приводит к вариативности результатов даже при идентичных запросах. Вследствие этого модель может по-разному формулировать одну и ту же тему, смешивать близкие понятия или генерировать правдоподобную, но неверную информацию (галлюцинировать). Такая особенность требует многократной проверки и уточнения запросов для получения стабильных и интерпретируемых результатов. Поэтому в ходе исследования мы провели пять тематических анализов полного текста дневника К. А. Березкина. Результаты обобщены в Таблице 1 (столбцы Вариант 1 – Вариант 5).

Тема	Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
Образование и школьные будни	Образование и школьные будни	Образование и школьные будни	Образование и школьные будни	Образование и школьные будни	Образование и школьные будни
Социальная жизнь и развлечения	Социальная жизнь и светские развлечения	Социальная жизнь и развлечения	Социальная жизнь и развлечения	Социальная жизнь и развлечения	Социальная жизнь и развлечения
Личные переживания, самоанализ	Личные переживания, размышления и самоанализ	Личные переживания и самоанализ	Личные переживания и самоанализ	Самоанализ и рефлексия	Самоанализ и рефлексия
Взаимоотношения с семьей	Взаимоотношения с семьей и близкими	Взаимоотношения с окружающими	Взаимоотношения с окружающими	Семья и домашняя жизнь	Семья и домашняя жизнь
Романтические увлечения	Романтические увлечения и отношения	Романтические увлечения и любовь	Любовь и романтические увлечения	Романтические интересы и отношения	Романтические интересы и отношения

	девушками			девушками
Религия верования	Религиозные взгляды и церковные обряды	и Религия верования	и Религия и народные поверья	и Религия и народные поверья
Чтение и лит. интересы	Чтение литературные интересы	и Литература искусство	и Литература искусство	и Литература и литературные интересы
Политические события	Политические события общественные настроения	Политические события общественная жизнь	Политические события общественные настроения	Политические события общественная жизнь
Хобби увлечения	и Хобби увлечения	и –	–	Самостоятельное творчество (рисование, писательство)
Быт повседневная жизнь	и Быт повседневная жизнь в Вологде	и –	Быт и повседневная жизнь	–
Природа	–	Наблюдения за природой и смена времен года	–	–
Планы на будущее	–	–	Планы на будущее и поиск своего места в жизни	–
Воспоминания детства	–	–	–	Воспоминания детства

Таблица 1. Темы, выделенные в дневнике К. Березкина моделью Gemini 1.5 Pro.

Многократный анализ позволил выявить устойчивые тематические кластеры и оценить степень вариативности в выделяемых темах. Во всех пяти прогонах анализа модель стабильно выделяла восемь основных тем, каждая из которых отражает значимые аспекты жизни Кирилла Березкина:

1. «Образование и школьные будни»: Эта тема охватывает описания уроков, экзаменов, взаимоотношений с учителями и одноклассниками, отражая повседневную жизнь гимназиста. Модель во всех случаях подчеркивает, что эта тема доминирует в дневнике и отражает повседневную жизнь гимназиста в Вологодской губернской гимназии.
2. «Социальная жизнь и развлечения»: В рамках этой темы собраны записи, описывающие участие в балах, маскарадах, вечерних собраниях, ярмарках и других формах досуга. Эти события иллюстрируют культурную и общественную жизнь провинциального города, в которой активно участвовал автор.
3. «Личные переживания и самоанализ»: Эта тема раскрывает внутренний мир К. Березкина: его размышления о характере, чувствах, стремлениях и недостатках, а также его размышления о будущей карьере и поиске своего места в жизни. Она подчеркивает

склонность автора к рефлексии и критическому осмыслению себя.

4. «Взаимоотношения с семьей»: Эта тема отражает отношения автора с близкими родственниками – матерью, отчимом, сестрами и бабушкой. Также включены воспоминания о детстве, которые помогают понять семейные традиции и их влияние на формирование личности автора. Интересно, что, говоря об отчиме, Кирилл использует устаревшее слово – вотчим. Но для модели это не становится препятствием в определении их семейного отношений.

5. «Романтические увлечения»: Тема включает записи о романтических увлечениях гимназиста, его восприятии внешности и характера девушек, а также переживаниях, связанных с общением с ними. Эти записи показывают, как формировались романтические идеалы автора.

6. «Религия и верования»: В этой теме представлены описания посещений церкви, участия в церковных обрядах, а также размышления о народных традициях и религиозных праздниках. Тема подчеркивает важность религии в жизни К. Березкина.

7. «Чтение и литературные интересы»: Тема объединяет впечатления автора от прочтения произведений русских и зарубежных писателей, а также его собственные литературные опыты, включая написание стихов и прозы. Чтение для К. Березкина – одна из основных форм проведения досуга.

8. «Политические события»: Здесь отражены записи, где автор упоминает ключевые события своего времени, такие как революции в Европе, война в Венгрии и эпидемия холеры в Петербурге. Автор передает общественные настроения, связанные с этими событиями, а также свое собственное восприятие политической ситуации.

Примечательно, что данное нами в промпте в качестве образца название темы «Образование и школьные будни» дословно повторяется во всех ответах. Остальные наименования тем в разной степени отличаются разнообразием формулировок.

Помимо этих восьми основных тем, в некоторых прогонах модель выделяла дополнительные, менее выраженные темы, такие как «Хобби и увлечения» (рисование, разведение птиц), «Быт и повседневная жизнь», «Погода», «Воспоминания детства» и «Планы на будущее». Наличие этих дополнительных тем варьировалось от одного прогона к другому, что подчеркивает стохастическую природу модели.

Анализ текста дневника с разбиением его помесечно демонстрирует тенденцию к более детальному и конкретному выделению тем по сравнению с анализом целого текста. Это проявляется в первую очередь в разнообразии наименований тем. Так вместо обобщенной темы «Социальная жизнь и развлечения» модель дает весьма многообразные формулировки: «Социальная жизнь и развлечения», «Социальная жизнь и отношения», «Социальная жизнь и досуг», «Социальная жизнь и городская среда», «Социальная жизнь и общество», «Социальная жизнь и светские мероприятия» (полный ответ

см.:

https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/data/intermediate_themes_gemini.txt). Но главное – существенно отличаются описания тем, в которых подчеркиваются разные аспекты общественной жизни. Появляются и новые темы: «Социальное неравенство и коррупция» (сетования автора на социальное неравенство и коррупцию в гимназии), «Природа и сельская жизнь» (впечатления от времени проведенного в июле 1849 г. в деревне Калинкино), «Воспоминания детства», «Холера и страх болезни». При анализе меньших фрагментов текста модель сосредотачивается на

более локальных контекстах и может получить более полное представление о тематическом разнообразии текста.

При обобщении результатов помесячного анализа модель формирует список, фактически совпадающий с темами, выделяемыми при анализе целого текста (Таблица 1, столбец Помесячный анализ) (полный ответ см.: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/data/summary_themes_gemini.txt).

Заданный нами в промпте формат ответа позволяет легко интерпретировать выделенные моделью темы. Для примера приведем описание темы из одного анализа полного теста (Таблица 1, Вариант 5, весь текст ответа см: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/gemini_05.txt):

Название темы: Исторические события и политическая обстановка

Описание: Кирилл упоминает важные исторические события, происходящие в России и мире, например, Венгерское восстание, революционные движения в Европе, смерть великого князя Михаила Павловича. Эти записи отражают его интерес к политике и истории, а также позволяют понять, как эти события воспринимались в российской провинции.

Пример:

- «Читал петербургские газеты, иностранные происшествия. Везет австрийцам, славно управляются в Венгрии, да что еще из этого всего будет?» (4 января)
- «Холера в Петербурге усилилась, неужели не прекратится совсем в России к весне?» (7 января)
- «Пишут, что русские войска – тысяч семь – выступили в Трансильванию на помощь австрийцам против чехов.» (1 марта)
- «В Риме республика учреждена, что из этого будет?» (1 марта)
- «Пишут, что русские войска – тысяч семь – выступили в Трансильванию на помощь австрийцам против чехов [Здесь автор явно путает чехов с венграми – А.К.]. В Риме республика учреждена, что из этого будет? Кажется, передерутся все, начнется баталия, катафасия, кажется.» (1 марта)

Сравним это описание со схожей темой помесячного анализа текста дневника за март 1849 г. (https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/data/intermediate_themes_gemini.txt):

Название темы: Исторические события и общественная жизнь

Описание: Кирилл упоминает о исторических событиях, происходящих в Европе (революции 1848-1849 гг., война Австрии с Сардинским королевством), читает газеты, интересуется политикой. Эти записи показывают, как отражались европейские события в российской провинции, какие источники информации были доступны.

Пример:

«1 марта. ...Пишут, что русские войска – тысяч семь – выступили в Трансильванию на помощь австрийцам против чехов. В Риме республика учреждена, что из этого

будет?»

«3 марта. ...Вечером вот теперь читал газеты, и там все та же история: беспорядки на Западе, видно, дело так не обойдется без большой драки.»

«29 марта. ...Читал сегодня газеты, особенного нет ничего, война Австрии с Карлом Альбертом началась...»

В обоих вариантах модель дала осмысленные и точные описания и привела релевантные примеры, позволяющие оценить качество выделенной темы. Обратим внимание, что в первом варианте описания темы цитаты об участии Российской Империи в подавлении Венгерского восстания («Пишут, что русские войска – тысяч семь – выступили в Трансильванию на помощь австрийцам против чехов») и учреждении республики в Риме («В Риме республика учреждена, что из всего этого будет?») воспроизведены дважды. Первый раз отдельно, а второй в составе более широкой цитаты, начинающейся с описания участия Российской Империи в подавлении Венгерского восстания и заканчивающейся общим впечатлением автора от событий в Европе («Кажется, передерутся все, начнется баталия, катастрофа, кажется»). Что, на наш взгляд, свидетельствует о том, что модель может идентифицировать отдельные исторические события даже по краткому упоминанию в объеме одного предложения. Это подтверждает и второй пример, где по одной фразе – «29 марта. ...Читал сегодня газеты, особенного нет ничего, война Австрии с Карлом Альбертом началась...» – модель определила войну Австрии с Сардинским королевством 1848/1849 гг.

В варианте, когда модели предлагалось проанализировать сразу весь текст выявилась одна закономерность: в качестве примеров модель выбирала первые подходящие цитаты, которые встречались в тексте. Поэтому подавляющее число примеров относятся к январю 1849 г.

Степень детализации выделяемых тем можно эффективно регулировать, уточняя запросы к модели. Например, если в запросе указать более редкую тему, такую как «Поездка в деревню», или сосредоточиться на конкретном аспекте широкой категории, например, заменить «Политические события» на «Политические события в Европе в 1849 году», модель способна точно идентифицировать и подробно описать эти темы. Такой подход позволяет не только адаптировать результаты анализа под исследовательские задачи, но и обеспечить более глубокое погружение в интересующие аспекты текста (целиком диалог с моделью см.: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/themes_change.ipynb):

Название темы: Поездка в деревню

Описание: Тема рассказывает о поездке гимназиста в деревню Калинкино и Ослапиху. Кирилл восхищается красотой природы, наблюдает за сменой времен года, описывает сельские праздники и обычай. Эта тема отражает его чувствительность к окружающему миру и характерный для того времени интерес к природе как источнику вдохновения и гармонии. Его описания поездки в Калинкино и окрестности дают представление о жизни российской деревни середины XIX века, включая социальные взаимодействия, быт и праздники.

Пример:

«...местоположение этой деревни самое невыгодное, никаких красот природы ни ручейка

нет ни речки, даже нет лесочки...» (8 июля 1849)

«Широко стелется зеленый луг, медленно и плавно расхаживают на нем белые как снег гуси и пестрые утки...» (22 июля 1849)

«Здесь все знакомы всякий мужик из какой бы он ни был деревни увида, что едут бара, снимает шапку, бабы поджавши к сердцу руки отвешивают поклоны.» (20 июля 1849)

Название темы: Политические события в Европе в 1849 году

Описание: Кирилл интересуется политическими событиями в Европе (революции 1848-1849 гг., война в Венгрии), следит за новостями в газетах («Петербургские ведомости»). Эта тема показывает, как глобальные события воспринимались в российской провинции и какое влияние оказывали на формирование мировоззрения молодого дворянина. Он размышляет о возможности военной службы в связи с этими событиями.

Пример:

«В Венгрии война в самом разгаре Дербечин взят без выстрела русские идут все вперед...» (12 июля 1849)

«Пишут, что русские войска – тысяч семь – выступили в Трансильванию на помощь австрийцам против чехов. В Риме республика учреждена, что из всего этого будет?» (11 января 1849)

«Не идти ли мне в военную службу? Правду говорит пословица, что у иного на неделе семь пятниц, так-то и у меня везде: хочется идти и в такую службу, и в другую...» (3 марта 1849)

На запрос проследить изменения в тематике дневника с течением времени модель дала подробный и качественный ответ, который подтвердил её способность к интерпретации сложных текстов с учётом исторического и культурного контекста. Анализ модели оказался хорошо структурированным, отражая не только тематические изменения, но и их связь с ключевыми событиями в жизни автора (целиком диалог с моделью см.: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/themes_change.ipynb).

Например, в описании января 1849 года модель связывает преобладание тем социальной жизни, романтических интересов и религиозных размышлений с особенностями зимнего периода, включая празднование Крещения и сезон балов и маскарадов. Кирилл анализирует свои эмоции, размышляет о взаимоотношениях с окружающими, что характерно для праздничной атмосферы и его возраста. Параллельно начинаются учебные занятия, что приводит к упоминаниям об образовании. Такой подход показывает способность модели учитывать многослойность тем в дневнике.

В феврале модель выделила тему образования как доминирующую, что связано с подготовкой к экзаменам. Увеличение записей, посвящённых школьным занятиям, учителям и одноклассникам, отражает возросшее напряжение и важность образовательного процесса в жизни автора. Параллельно модель корректно отметила влияние Великого поста, усилившего религиозную тематику и размышления Кирилла о смысле жизни. В этом месяце также сохраняются элементы социальной активности, связанные с масленичными гуляниями, и развивается интерес к литературе.

Модель также адекватно передала изменение тематической структуры в апреле и мае, когда на первый план выходят «Образование и школьные будни» и «Самоанализ и

размышления о будущем», появляется тема «Политические события в Европе в 1849 году». Эти темы связаны с подготовкой к экзаменам, размышлениями о будущем и началом войны в Венгрии. В июне указывается на повышение роли семьи и социальной жизни в связи с прощанием с гимназией и обсуждением будущей службы с родителями.

Особое внимание заслуживает смена тематики в июле, когда Кирилл отправляется в деревню. Модель правильно идентифицирует доминирующие темы, такие как «Поездка в деревню», «Литературные интересы и творчество», «Самоанализ и размышления о будущем», связывая их с описаниями сельских пейзажей, наблюдениями за деревенской жизнью и народными обычаями, интенсивным чтением и размышлениями о своем будущем.

В августе и сентябре, после возвращения из деревни преобладает тема «Самоанализ и размышления о будущем» связанная с мыслями автора о выборе профессии.

Таким образом, модель не только точно выделила преобладающие темы каждого месяца, но и связала их изменения с конкретными событиями в жизни гимназиста, включая его семейные отношения, образовательную деятельность и участие в культурной жизни Вологды.

Обсуждение результатов

Результаты исследования подтвердили значительный потенциал больших языковых моделей для тематического анализа исторических текстов. На примере дневника К. А. Березкина модель Gemini 1.5 Pro успешно выделила восемь ключевых тем, охватывающих различные аспекты жизни гимназиста в Российской империи середины XIX века. Устойчивость этих тем была подтверждена многократным анализом текста, что свидетельствует о надежности полученных результатов. Анализ раскрыл важные черты мировоззрения Кирилла Березкина. Его записи свидетельствуют о негативном отношении к социальной несправедливости и о способности критически осмысливать окружающую действительность. Дневник раскрывает внутренний мир автора, его сложные эмоциональные переживания, связанные с учебой, экзаменами и различными сторонами социальной жизни. Вместе с этим автор предстает перед нами как молодой человек, увлеченный литературой, изобразительным искусством, наблюдением за народными традициями, пробующий свои силы в писательстве.

Выделенные темы могут стать основой для дальнейших исследований, направленных на изучение социальных, культурных и образовательных реалий эпохи. Например, тема «Политические события» выходит за рамки простой фиксации исторических фактов, демонстрируя размышления автора о значении глобальных процессов, таких как европейские революции и эпидемия холеры, а также его гражданскую позицию. Аналогично, тема «Образование и школьные будни» позволяет реконструировать образовательную среду гимназии и ее влияние на формирование личности молодого человека, подчеркивая, как система обучения и межличностные взаимодействия формировали взгляды юноши.

Сравнение результатов анализа полного текста дневника и его помесячных фрагментов показало, что разбиение текста на части позволяет выявить более специфические темы, которые не были очевидны при анализе полного текста. Это можно объяснить тем, что обработка меньших фрагментов текста позволяет модели сосредотачиваться на локальных контекстах, минимизируя потерю информации из середины длинных текстов.

Наблюдаемая вариативность в формулировках и детализации тем обусловлена

стохастической природой больших языковых моделей. Формулировки тем могут отличаться по степени конкретности: от общих, например, «Литература и искусство» (Таблица 1, Вариант 2, весь текст ответа см: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/gemini_02.txt), до более узких, таких как «Литература и чтение» или «Искусство и творчество (рисование)» (Таблица 1, Вариант 5, весь текст ответа см: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/gemini_05.txt). В другом случае тема «Чтение и литературные интересы» дополняется отдельной темой «Хобби и увлечения», которая, согласно описанию, охватывает такие занятия автора, как рисование, созидаание гербария и разведение птиц (Таблица 1, Вариант 1, весь текст ответа см: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/gemini_01.txt). Одновременно менее значимые для повествования темы, такие как «Быт и повседневная жизнь», «Погода», «Хобби и увлечения», «Планы на будущее», появляются лишь в некоторых ответах модели. Подобная вариативность подтверждает необходимость многократного анализа и тщательной интерпретации результатов. Одновременно она указывает на «размытость» границ между темами и их взаимосвязь.

Однако эта вариативность не ставит под сомнение корректность выделенных тем, а скорее указывает на необходимость многократного анализа и тщательной интерпретации результатов. Более того, вариативность может быть истолкована как дополнительный источник информации, указывая на «размытые» границы между темами и их взаимосвязь. В подтверждение этого, в конце одного из анализов модель подводит итог следующими словами: «Это основные темы, которые прослеживаются в дневнике Кирилла Березкина. Важно отметить, что многие из них тесно переплетены между собой и отражают многогранность личности автора, а также сложную и интересную жизнь российской провинции середины XIX века» (Таблица 1, Вариант 1, весь текст ответа см: https://github.com/alexeyvkuznetsov/topic_modeling_with_llm/blob/main/gemini_01.txt)

В некоторых случаях модель допускала ошибки, приводя нерелевантные цитаты или неверно указывая датировки. Например, цитата о вступлении русских войск в Трансильванию, взятая из записи за 1 марта 1849 года, несколько раз ошибочно датировалась 11 января 1849 года, как показано в приведённом выше примере. Подобные неточности подчёркивают важность критического подхода при использовании ответов, полученных с помощью языковой модели, и необходимость их проверки традиционными методами исторического анализа.

Наш опыт тематического анализа дневника вологодского гимназиста позволяет сформулировать ряд рекомендаций по использованию языковых моделей для тематического анализа текстов. Для повышения надежности получаемых результатов рекомендуется проводить анализ одного и того же текста несколько раз. Сравнение результатов даст возможность определить стабильное тематическое ядро и оценить степень вариативности в выделении менее явных тем. Для более детального анализа полезно проводить анализ отдельных частей большого текста. Изменяя промпты можно направлять внимание модели на различные аспекты текста и регулировать уровень детализации анализа тем.

Ограничения и перспективы исследования

Настоящее исследование, несмотря на полученные результаты, имеет ряд ограничений. Анализ проводился на основе единственного источника – дневника К.А. Березкина. Для обобщения выводов необходимо провести аналогичные исследования на более

разнообразном наборе исторических текстов различных жанров, периодов и авторов. Мы использовали только одну большую языковую модель – Gemini 1.5 Pro. Сравнительное исследование различных моделей позволит лучше понять их возможности и ограничения. Важно понимать, что языковые модели стремительно развиваются, а корпорации соревнуются в расширении их возможностей и качестве ответов. Тенденция 2024 г. – появление мультимодальных генеративных моделей, работающих с различными типами контента: текстом, изображениями, видео, музыкой. С небольшим отрывом от проприетарных моделей развиваются и модели с открытым исходным кодом. Появление следующего поколения моделей Gemini ожидается уже в конце 2024 г. или начале 2025 г. Их возможности мы пока не знаем.

Анализ тематики дневника К. А. Березкина является лишь первым шагом в изучении этого памятника. В дальнейшем мы планируем сосредоточиться на двух направлениях: углубленном анализе конкретных, выявленных в дневнике тем (в частности, восприятие европейских событий в российской провинции), и дальнейшей разработке методов анализа исторических текстов с помощью больших языковых моделей. Перспективным направлением представляется создание специализированных агентов, интегрированных с базами исторических знаний. Такой подход позволит повысить качество анализа, точность и интерпретируемость получаемых результатов, открывая новые возможности для исторических исследований.

Заключение

Проведенное исследование убедительно продемонстрировало значительный потенциал больших языковых моделей как эффективного инструмента для тематического анализа исторических источников. Анализ дневника Кирилла Антоновича Березкина позволил не только выделить ряд значимых тем, но и раскрыть многогранность личности автора, реконструировать социальный контекст провинциальной России середины XIX века, а также получить более глубокое представление о его мировоззрении и восприятии окружающей действительности.

Сравнение анализа полного текста дневника и его фрагментов показало, что дробление текста позволяет выявлять дополнительные нюансы и избегать потерь информации, характерных для обработки длинных текстов. Стохастическая природа языковых моделей требует внимательной интерпретации их ответов и сопоставления результатов нескольких вариантов анализа, что позволяет обнаружить устойчивое тематическое ядро. Критический подход и сопоставление результатов, полученных с помощью больших языковых моделей, с данными традиционных методов анализа исторических источников является залогом достоверности и научной ценности исследования.

В заключение следует отметить, что большие языковые модели открывают новые перспективы для исторических исследований, автоматизируя трудоемкие этапы тематического анализа и помогая выявлять скрытые закономерности и тенденции в больших корпусах текстов. Это не только ускоряет процесс исследования, но и позволяет формулировать новые гипотезы и формировать более глубокое и всестороннее понимание прошлого, открывая путь к качественно новому уровню анализа исторических источников.

Библиография

1. Березкин К. А. Дневник / Подготовка текста к загрузке: Анна Завина, Антон Микерин. [Электронный ресурс]. // Электронный корпус «Прожито». URL: <https://corpus.prozhito.org/person/925>

2. Березкин К. А. Дневник: к 190-летию К. А. Березкина / под ред. Г. Л. Шалагиной. 2-е изд., доп. СПб.; Вологда: Киселев А. В., 2020.
3. Воронцов К. В. Обзор вероятностных тематических моделей // Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большиакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С. М.: Изд-во НИУ ВШЭ, 2017. С. 195–268.
4. Дневник Кирилла Антоновича Березкина / под ред. Г. Л. Шалагиной. СПб., 2014.
5. Кузнецов А. В. Проблемы интерпретации результатов тематического моделирования в исторических исследованиях // Актуальные проблемы гуманитарных и общественных наук: Сборник статей IX Всероссийской научно-практической конференции. Пенза, 2023. С. 90–91.
6. Милкова М. А. Тематические модели как инструмент «Дальнего чтения» // Цифровая экономика. 2019. № 1(5). С. 57–70.
7. Agrawal A., Fu W., Menzies T. What is wrong with topic modeling? And how to fix it using search-based software engineering // Information and Software Technology. 2018. Vol. 98. Pp. 74–88.
8. Althage M. Potenziale und Grenzen der Topic-Modellierung mit Latent Dirichlet Allocation für die Digital History // Digital History: Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft. 2022. Vol. 6. No. 24. P. 255–277.
9. Binkley D., Heinz D., Lawrie D., Overfelt J. Understanding LDA in source code analysis // Proceedings of the 22nd International Conference on Program Comprehension. ACM, 2014. Pp. 26–36.
10. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.
11. Blevins C. Topic Modeling Historical Sources: Analyzing the Diary of Martha Ballard // DH. 2011. Pp. 97–99.
12. Bodrunova S. S. Topic Modeling in Russia: Current Approaches and Issues in Methodology // The Palgrave Handbook of Digital Russia Studies. 2021. Pp. 409–426.
13. de Melo T., Merialdo P. Beyond Topic Modeling: Comparative Evaluation of Topic Interpretation by Large Language Models // Intelligent Systems Conference. Cham: Springer Nature Switzerland, 2024. Pp. 215–230.
14. Denny M. J., Spirling A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it // Political Analysis. 2018. Vol. 26. No. 2. Pp. 168–189.
15. Doi T., Isonuma M., Yanaka H. Topic Modeling for Short Texts with Large Language Models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). 2024. Pp. 21–33.
16. Garcia G. G., Weilbach C. If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History // arXiv preprint arXiv:2310.10808. 2023. URL: <https://arxiv.org/abs/2310.10808>
17. Gillings M., Hardie A. The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice // Digital Scholarship in the Humanities. 2023. Vol. 38. No. 2. Pp. 530–543.
18. Goldstone A., Underwood T. The quiet transformations of literary studies: What thirteen thousand scholars could tell us // New Literary History. 2014. Vol. 45. No. 3. Pp. 359–384.
19. Götter C. Künstliche Intelligenz und Geschichtswissenschaft. Mehr als ein neues Werkzeug für die Digital History? // Historische Zeitschrift. 2024. Vol. 319. No. 2. S. 299–330.
20. Huang L. et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions // arXiv preprint arXiv:2311.05232. 2023. URL: <https://arxiv.org/abs/2311.05232>

<https://arxiv.org/abs/2311.05232>

21. Jockers M. L., Mimno D. Significant themes in 19th-century literature // *Poetics*. 2013. Vol. 41. No. 6. Pp. 750–769.
22. Kansteiner W. Digital doping for historians: Can history, memory, and historical theory be rendered artificially intelligent? // *History and Theory*. 2022. Vol. 61. No. 4. Pp. 119–133.
23. Kirilenko A. P., Stepchenkova S. Facilitating topic modeling in tourism research: Comprehensive comparison of new AI technologies // *Tourism Management*. 2025. Vol. 106. P. 105007.
24. König M. ChatGPT und Co. in den Geschichtswissenschaften – Grundlagen, Prompts und Praxisbeispiele // *Digital Humanities am DHIP*. 19. August 2024. URL: <https://dhdhi.hypotheses.org/9197> DOI: <https://doi.org/10.58079/126eo>
25. Li D., Zhang B., Zhou Y. Can large language models (LLM) label topics from a topic model? // *SocArXiv*. 2023. URL: <https://doi.org/10.31235/osf.io/23x4m>
26. Liu N. F. et al. Lost in the middle: How language models use long contexts // *Transactions of the Association for Computational Linguistics*. 2024. Vol. 12. Pp. 157–173.
27. McGillivray B., Buning B., Hengchen S. Topic Modelling: Hartlib's Correspondence Before and After 1650 // *Reassembling the Republic of Letters in the Digital Age*. Göttingen, 2019. Pp. 426–428.
28. Mu Y. et al. Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling // *arXiv preprint arXiv:2405.00611*. 2024. URL: <https://arxiv.org/abs/2405.00611>
29. Mu Y. et al. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling // *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024. Pp. 10160–10171.
30. Newman D. J., Block S. Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper // *Journal of the American Society for Information Science and Technology*. 2006. Vol. 57. No. 6. Pp. 753–776.
31. Oiva M. Topic Modeling Russian History // *The Palgrave Handbook of Digital Russia Studies*. 2021. Pp. 427–442.
32. Pham C. et al. TopicGPT: A Prompt-based Topic Modeling Framework // *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024. Pp. 2956–2984.
33. Phoenix J., Taylor M. *Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs at Scale*. O'Reilly Media, Inc., 2024.
34. Pollin Ch. Workshopreihe «Angewandte Generative KI in den (digitalen) Geisteswissenschaften» (v1.1.0). Zenodo. URL: <https://doi.org/10.5281/zenodo.10647754>
35. Pu X., Gao M., Wan X. Summarization is (almost) dead // *arXiv preprint arXiv:2309.09558*. 2023. URL: <https://arxiv.org/abs/2309.09558>
36. Schulhoff S. et al. The Prompt Report: A Systematic Survey of Prompting Techniques // *arXiv preprint arXiv:2406.06608*.
37. Solanki S. R., Khublani D. K. *Generative Artificial Intelligence: Exploring the Power and Potential of Generative AI*. Apress, 2024.
38. Stammbach D. et al. Revisiting automated topic model evaluation with large language models // *arXiv preprint arXiv:2305.12152*. 2023. URL: <https://arxiv.org/abs/2305.12152>
39. Villamor Martin M., Kirsch D. A., Prieto-Nañez F. The promise of machine-learning-driven text analysis techniques for historical research: topic modeling and word embedding // *Management & Organizational History*. 2023. Vol. 18. No. 1. Pp. 81–96.
40. Wang H. et al. Prompting large language models for topic modeling // *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023. Pp. 1236–1241.

41. Wehrheim L. Economic History Goes Digital: Topic Modeling the Journal of Economic History // Cliometrica. 2019. Vol. 13. No. 1. Pp. 83–125.
42. Xu Z., Jain S., Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models // arXiv preprint arXiv:2401.11817. 2024. URL: <https://arxiv.org/abs/2401.11817>

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.

Со списком рецензентов издательства можно ознакомиться [здесь](#).

Статья посвящена передовому методу анализа текстов — тематическому моделированию. Тематическое моделирование (в отличие от привычного для историков контента-анализа) пока не так часто встречается в исторических исследованиях, тем интереснее опыты применения такого компьютеризированного анализа текстов в историческом исследовании. В статье убедительно показано, что за последнее десятилетие тематическое моделирование прошло апробацию на очень разных исторических источниках, начиная с газетных материалов и продолжая анализом дневников, писем, периодических изданий, сборников статей, литературных корпусов.

Как справедливо отмечает автор, для применения тематического моделирования необходимо анализировать пространный текст, желательно состоящий из большого объема данных, который ещё и желательно должен быть предварительно обработан исследователем.

В качестве исторического источника, на основании которого проводился методический эксперимент — дневниковые записи дворянина Кирилла Антоновича Березкина, доступные как в виде публикации под редакцией Г. Л. Шалагиной (СПб.; Вологда, 2020), так и в электронном корпусе дневников «Прожито» (<https://corpus.prozhito.org/person/925>), описывающие большую часть 1849 года.

Автор статьи подчеркивает, что существенным усилением тематического моделирования могут стать большие языковые модели, которые способны улавливать весьма детализированный контекст высказываний исторических источников. Совершенно справедливо в статье отмечается, что в литературе представлены два подхода: параллельный и последовательный промптинг. Рассматривая актуальные подходы к использованию больших языковых моделей для исследовательских задач в истории, имеет смысл обратить внимание на варианты построения собственных моделей, который уже могут получаться без привлечения больших вычислительных ресурсов (как показано, например, в книге Raschka S. Build a Large Language Model. Manning, 2024. 368 р.).

В исследовании была использована модель Gemini 1.5 Pro, позволяющая обработать до 2 млн токенов входящей информации, то есть фактически можно загрузить весь текст исследуемого источника. В статье учитываются важные особенности работы с историческими источниками, действительно, «анализ исторических документов требует учета не только текста, но и культурного, исторического и социального контекста эпохи». При этом, в работе удалось применить находчивое методическое решение, разделив промпты на системный (ролевой) и пользовательский (исследовательский).

Автор избрал методически взвешенный подход, учитываются и частые потери содержания в середине длинных отрезков текстов, и в принципе снижая вероятностные «галлюцинации» модели понижением показателя «температура», равно как и протестировав варианты детализации откликов модели. Исследование показало, что для выделения устойчивых тематических групп необходимо провести многократные

повторения обращений к модели, потому что модель работает вероятностно, а не алгоритмически. В пяти итерациях анализа модель стабильно выделяла восемь основных тем, каждая из которых отражает существенные стороны дневниковых записей Березкина. Важным преимуществом проведенной работы является то, что все важные экспериментальные результаты доступны в качестве файлов в репозитории GitHub. Главный вопрос, который остается после знакомства с интересным и полезным экспериментом применения тематического моделирования с усилением большой языковой моделью (Gemini 1.5 Pro), возможно ли выйти за границы автоматизации понятных исследователю аналитических процедур, и построить работу с автоматизацией интерпретационных задач исследования. Например, вполне очевидная задача, которая возникает при использовании языковых моделей: возможно ли построить подробную и достоверную биографическую справку, опираясь на дневниковые записи. Но это вопрос для будущих работ автора. Предложенный к публикации эксперимент вполне удался. Статья может быть безусловно рекомендована для публикации в журнале «Историческая информатика». Выводы – самостоятельные, обоснованные, пионерские, статья наверняка вызовет большой интерес читательской аудитории.