**DOI: 10.19181/inter.2026.18.2.8**
**EDN: XBLIML**

# Экспертная валидация тематического моделирования (BERTopic) и последующее применение LLM к выделенным темам в качественном исследовании в рамках тематического анализа

**Ссылка для цитирования:**
*Червоткина П. И.* Экспертная валидация тематического моделирования (BERTopic) и последующее применение LLM к выделенным темам в качественном исследовании в рамках тематического анализа // Интеракция. Интервью. Интерпретация. 2026. Т. 18. № 2. С. 91–125. https://doi.org/10.19181/inter.2026.18.2.8 EDN: XBLIML

**For citation:**
Chervotkina P. I. (2026) Expert Validation of Thematic Modeling (BERTopic) and Subsequent Application of LLM to Selected Topics in Qualitative Research Within the Framework of Thematic Analysis. *Interaction. Interview. Interpretation*. Vol. 18. No. 2. P. 91–125. https://doi.org/10.19181/inter.2026.18.2.8

**Червоткина Полина Игоревна**
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
E-mail: bcur1119@gmail.com

*Статья посвящена применению методов машинного обучения и генеративных языковых моделей в анализе качественных интервью. Цель работы — разработка и валидация смешанной методологии, сочетающей тематическое моделирование (BERTopic) с контролем эксперта. Эмпирическую базу составили 20 транскриптов полуформализованных интервью с членами семей о практиках заботы о здоровье. На первом этапе с помощью BERTopic, UMAP и HDBSCAN выделены кластеры реплик на основе биграмм и триграмм, а затем — с добавлением униграмм. Далее проведен подбор гиперпараметров. Второй этап включал валидацию кластеров экспертом и построение управляемой (guided) модели с добавлением тем от эксперта (seed-темы). Третий этап — визуализация связей тем через графы совместной встречаемости терминов. На четвертом этапе выполнены обобщение и группировка тем на основе результатов предыдущих этапов с использованием DeepSeek,*

*полученные данные сравнивались с работой кодировщика по темам модели BERT и транскриптам интервью.*

*Установлено, что BERTopic с биграммами и триграммами демонстрирует более высокое разнообразие тем (0,96) и разделимость (силуэтный коэффициент 0,69) по сравнению с моделью, включающей и униграммы (разнообразие тем 0,84, силуэтный коэффициент 0,56). Управляемые модели позволили снизить долю шумовых реплик, выявить темы, не обнаруженные при изначальном запуске модели (например, ошибки устройств) и повысить метрики разделимости тем (силуэтный коэффициент 0,72 для модели с биграммами и триграммами, 0,73 для модели с униграммами). LLM демонстрирует ограниченную способность самостоятельно формулировать темы, не заданные в промпте, однако ее работа умеренно согласуется с кодировкой исследователя. В работе предлагаются методы снижения галлюцинаций и повышения качества тематического обобщения со стороны LLM.*

**Ключевые слова:** качественный анализ интервью; тематическое моделирование; BERTopic; большие языковые модели; тематический анализ; машинное обучение

## Введение

Использование готовых языковых моделей при анализе качественных интервью сегодня активно осмысляется исследователями. На мой взгляд, применение ИИ в анализе качественных данных можно разделить на 3 способа: использование больших языковых моделей (LLM) через диалоговый интерфейс или API, применение специализированных методов машинного обучения для тематического моделирования, классификации или кластеризации и, наконец, использование специальных приложений и программ с готовыми алгоритмами обработки. Все перечисленные способы имеют свои ограничения и достоинства, однако в рамках данной работы рассматривается именно использование специализированных методов машинного обучения и предлагается способ их валидации с помощью эксперта.

Исследования по сравнению качества кодирования языковыми моделями и человеком-исследователем показывают, что ИИ демонстрирует достойный уровень работы. При анализе интервью в сфере здравоохранения сравнивалась согласованность кодировок GPT и исследователя, коэффициент Каппа Коэна составил 0.4, что соответствует умеренной согласованности экспертов [Li, Fernandez, Schwartz et al., 2024], при этом ИИ отлично выделил основные категории, но менее распространенные коды, в отличие от исследователя, он не выявил. При использовании больших языковых моделей общего доступа также встает вопрос о конфиденциальности данных: загруженные транскрипты интервью могут использоваться нейросетью дальше. В некоторых исследованиях используются модели, которые работают независимо от основной, не сохраняют введенные данные и не обучаются на их основе [Li, Fernandez,

Schwartz et al., 2024]. Использование API-обращений в облачной среде программирования (Google Colab) или локально на своем компьютере частично может решить этот вопрос, хотя политика обращения с данными пользователя может меняться, что оставляет риск того, что данные могут перейти третьим лицам. Также, чтобы избежать раскрытия данных, генеративные модели можно использовать на заключительных этапах: назвать уже сформулированные темы или коды более емко или их обобщить [Gamieldien, Case, Katz et al., 2023]. Это имеет смысл с той точки зрения, что генеративные модели (GPT (OpenAI), DeepSeek и т. д.) являются трансформерами-декодерами и способны эффективно генерировать текст самостоятельно, а не просто выбирать данные из текста. К основным преимуществам работы с большими языковыми моделями относится скорость обработки, что позволяет исследователю получить предварительные результаты анализа и углубляться в материал уже исходя из этих данных [Hitch, 2024] или получить новые коды, которые не были замечены человеком.

Для улучшения результатов работы с языковыми моделями важно понимать специфику их работы, разделять задачи кодирования и обобщения уже полученных кодов. Некоторые авторы [Anakok, Katz, Chew et al., 2025] предлагают последовательное применение различных генеративных моделей под каждый этап тематического моделирования и дают конкретные советы по созданию промптов. Также минимальная настройка моделей, например, установление температуры, может помочь повысить воспроизводимость результатов или, наоборот, креативность [De Paoli, 2024; Anakok, Katz, Chew et al., 2025].

Важно отметить, что подход к использованию ИИ в анализе качественных данных не подчиняется единому шаблону, и исследователи экспериментируют с воспроизведением разных методов кодирования тематического анализа [De Paoli S., 2024; Anakok, Katz, Chew et al., 2025], обоснованной теории [Nelson, 2020], контент-анализа [Renz, Carrington, Badger, 2018]. В рамках обоснованной теории в работе с ИИ сохранялась индуктивная логика: математически выделялись паттерны, далее они интерпретировались исследователем, и снова с помощью компьютеризированных алгоритмов проверялась их распространенность [Nelson, 2020], при этом автор статьи сама адаптировала математические методы под свои задачи и корректировала их. Несмотря на то, что применение машинного обучения с написанием кода может быть довольно сложным для исследователя, именно этот способ позволяет комбинировать различные методы и при этом контролировать процесс обработки, меняя параметры. В отличие от этого большие языковые модели остаются «черным ящиком» и способны галлюцинировать. Воспроизводимость результатов здесь также затруднена, это возможно лишь при фиксации параметров, когда большая языковая модель разворачивается через API.

Другие авторы [Renz, Carrington, Badger, 2018] с помощью программы Linguistic Inquiry and Word Count (LIWC) выделили характеристики того, как медсестры описывают свой опыт использования электронной записи, расширив изначальную интерпретацию. В этом случае стратегию можно назвать дедуктивной, так как анализ проводился с выявлением уже изначально

заданных характеристик в программе. Применение готовых программ для компьютерного анализа текста удобно с точки зрения готового интерфейса, но ограничивает исследователя в методах. Кроме того, поскольку сфера анализа естественного языка развивается быстро, программное обеспечение может устаревать и не обеспечивать такую же высокую скорость, как большие языковые модели и методы машинного обучения.

Некоторые исследователи стремятся анализировать ситуацию контекста интервью и фокусируются на самом взаимодействии между интервьюером и информантом. Они предлагают анализировать интервью не как набор изолированных высказываний, а как диалогическую практику, в которой значимы не только ответы, но и реплики интервьюера, их последовательность и взаимное влияние [Quillivic, Payet, 2024]. Для этого используется комбинация мультиязычных эмбеддингов предложений (Sentence-BERT) для измерения лингвистического выравнивания между репликами, а также классификация диалогических актов и тем с помощью обучения на малом числе примеров. Это позволяет выявлять, как характеристики участников (например, пол) или психологические особенности информантов (например, симптомы ПТСР) влияют на структуру беседы и стиль вопросов интервьюера, что дополняет содержательный анализ данных.

В рамках нашего исследования ставится задача проверки эффективности методов машинного обучения для кодирования данных в рамках тематического анализа, при этом компьютерные методы проверяются, корректируются исследователем. Результаты работы моделей машинного обучения дополнительно пересматриваются LLM и исследователем для корректировки тем, при этом исследователь может использовать полные транскрипты, а не только отдельные реплики. Это позволяет оценить, насколько результаты машинного обучения полезны для последующего кодирования как генеративной моделью, так и кодировщиком-человеком. Использованные алгоритмы выложены в открытый доступ на платформе GitHub[1]. Эффективность методов машинного обучения подразумевает выделение интерпретируемых тем с высокими метриками разделимости (силуэтный коэффициент > 0,6), а также удобство для исследователя (сокращение времени на первичное ознакомление с материалом) и для LLM (отсутствие галлюцинаций) при использовании этих результатов для обобщения тем.

## Методология

Ниже представлено подробное описание этапов применения методов.

**Этап 1.** Выделение кластеров с помощью BERTopic и подбор гиперпараметров. Все реплики информантов лемматизируются с сохранением частицы «не». Для каждой реплики вычисляются векторные представления (BERT), после

[1] Pachpol8. Healthcare-Topics-Bertopic. *GitHub*. URL: https://github.com/Pachpol8/healthcare-topics-bertopic (дата обращения: 10.03.2026).

чего выполняется снижение размерности (UMAP) и кластеризация (HDBSCAN). Ключевые слова тем извлекаются на основе TF-IDF через CountVectorizer с диапазоном от биграмм до триграмм. Затем модель запускается еще раз, но вместе с биграммами и триграммами извлекаются еще и униграммы. С помощью подбора параметров выбирается лучшая модель с точки зрения как метрик, так и интерпретируемости тем на основе ключевых словосочетаний для обоих запусков. Результат — набор тем с описанием через ключевые n-граммы (словосочетаний) и матрица принадлежности реплик. Этот этап включает в себя и этап выделения кодов и тем [Braun, Clarke, 2006], так как n-граммы фиксируют смысловые единицы (о чем говорят информанты), а темы (кластеры) отражают более крупные смысловые фрагменты. Формирование тем исключительно индуктивное.

**Этап 2.** Валидация кластеров и формирование seed-тем. Исследователь анализирует темы, полученные на этапе 1: изучает ключевые n-граммы и примеры реплик, выявляет значимые для исследования словосочетания, которые определяют конкретный код. Выявляются неохваченные моделью коды (словосочетания), которые важны для анализа. Формируется список из дополнительных кодов — набор семантически связанных слов и словосочетаний (униграммы, биграммы, триграммы), которые модель будет искать в данных. В этот список добавляются словосочетания, как выделенные экспертом самостоятельно, так и те, которые были получены моделью, но которые, по мнению эксперта, необходимы для анализа. Далее проводится подбор наилучших гиперпараметров и для модели с униграммами, и для модели только с биграммами и триграммами. В результате получаются две финальные тематические структуры: в первой теме они описываются отдельными словами либо фразами; во второй — только словосочетаниями. Проводится сравнение метрик с базовой моделью: оценивается улучшение силуэта, когерентности и интерпретируемости для каждого варианта. Обе модели сохраняются для последующего анализа. Этот этап служит аналогом этапа проверки тем [Braun, Clarke, 2006], дополнение тем идет не только индуктивно, но и дедуктивно.

**Этап 3.** Визуализация связей тем и ключевых слов. Строятся два типа графов: граф связей тем через общие ключевые словосочетания (узлы — номера тем, ребра — ключевые словосочетания, встречающиеся в темах) и граф совместной встречаемости терминов в репликах (узлы — ключевые униграммы по модели 2, ребра — совместная встречаемость в одной реплике, вес — количество таких реплик).

**Этап 4.** Создание тем второго порядка. Исследователь на основе списка тем, ключевых словосочетаний, примеров реплик, транскриптов и графов объединяет близкие темы в категории более высокого порядка. Параллельно тот же набор тем, примеров реплик, ключевых словосочетаний передается в DeepSeek с промптом на группировку. Сравниваются обе классификации; расхождения анализируются как дополнительный источник гипотез или повод для выявления ограничений применения машинного обучения. Этот этап служит продолжением этапа проверки тем [Braun, Clarke, 2006], когда темы укрупняются и объединяются, указываются аналитические связи между темами.

## Эмпирическая база исследования

Статья опирается на материалы, собранные автором, включающие 20 транскриптов полуформализованных интервью продолжительностью от 30 до 50 минут с членами семей на тему практик заботы о здоровье в семьях. Период проведения интервью с 7 февраля по 8 марта 2026 года. Возраст информантов — от 20 до 84 лет. Интервью проводилось с каждым членом семьи по отдельности, в отсутствие других членов семьи, чтобы исключить их влияние на ответы респондента и получить не только индивидуальное представление о практиках заботы о здоровье, но и целостное понимание этих практик в масштабе всей семьи. Выбор данного корпуса обусловлен его доступностью для исследователя (он собран самостоятельно), а также тем, что тема заботы о здоровье может порождать схожие тематические слова, однако высказывания при этом будут различаться по семантике, что обеспечивает эффективную проверку метода. Семейная забота о здоровье выбрана как область с высокой долей сложных кодов, связанных с переговорами, конфликтами между различными членами семьи, которые могут выражаться косвенно. Это создает вызов и для алгоритмов машинного обучения, и для LLM и позволяет оценить границы их применимости в качественном анализе. В итоговый корпус вошли также пять интервью с единственными представителями семей. С точки зрения исследователя, их позиция представляет самостоятельный интерес и описывает индивидуальное восприятие заботы о здоровье в семье.

Отбор информантов шел методом доступных случаев в первую очередь из окружения исследователя. Часть информантов была отобрана по методу снежного кома. Дополнительно для поиска использовалась рассылка по Телеграм-каналам, связанным со здоровьем, поддержанием здорового образа жизни. Критериями для возможности включения в исследование были:

- совместное проживание на одной территории не менее одного года (допустимы небольшие перерывы до 1–2 месяцев). Исключение — недавно рожденные дети;
- информанты не должны были иметь профессионального медицинского образования, однако могли проходить дополнительные курсы и иметь знакомых с профессиональным образованием.

При построении выборки реализовывался отбор по гетерогенности семей в рамках доступного случая: например, и однополые, и разнополые, включающие несколько поколений или только одно. Так как целью исследования был охват разных ситуаций практик заботы о здоровье и их влияния на семейную рутину, то более широкая выборка позволяла охватить максимальное число возможных случаев.

### *Реализация этапов*

*Этап 1. Выделение кластеров с помощью BERTopic и подбор гиперпараметров*

На первом этапе код загружает транскрипты интервью, извлекает только реплики информанта (игнорируя вопросы интервьюера) и лемматизирует

их с сохранением отрицаний (частица «не» не удаляется). Затем осуществляется переход от отдельных слов к биграммам и триграммам (например, «не заниматься», «высокое давление»): это сделано для того, чтобы уменьшить пересечение тем, поскольку, например, частотные глаголы в виде униграмм слишком похожи во многих ответах и размывают границы между кластерами. Кроме того, из анализа исключаются слишком короткие реплики (менее 12 символов), которые не несут содержательной нагрузки. Далее с помощью Sentence-BERT каждая реплика преобразуется в смысловой вектор (числовой отпечаток фразы), размерность которого снижается алгоритмом UMAP, после чего HDBSCAN группирует эти векторы в кластеры (темы), автоматически отбрасывая шумовые высказывания. Наконец, для каждого кластера методом TF-IDF вычисляются наиболее характерные n-граммы, которые и представляют содержательную суть темы. Такой подход позволяет увидеть не просто набор слов, а устойчивые смысловые паттерны в речи человека, включая отрицания и фразовые обороты. Лемматизация осуществлялась с помощью библиотеки pymorphy3. Стоп-слова, то есть высокочастотные слова, которые несут мало смысловой нагрузки, включали стандартный набор NLTK и расширенный пользовательский список.

BERTopic — готовая библиотека, которая реализует алгоритм тематического моделирования на основе эмбеддингов. В основе ее лежит двунаправленная трансформерная архитектура BERT (Bidirectional Encoder Representations from Transformers): она преобразует каждое слово или предложение в плотный вектор (эмбеддинг), который учитывает не только само слово, но и все окружение — слова слева и справа, порядок, зависимость смысла от контекста [Grootendorst, 2022]. Этот метод доказывает свою эффективность на различной длине анализируемых данных: и на коротких корпусах с комментариями [Egger, Yu, 2022], и на текстах интервью [Ionescu, Han, Suasnabar et al., 2026]. В данной работе используется модель для создания эмбеддингов paraphrase-multilingual-MiniLM–L12-v2[2], которая указана по умолчанию. Алгоритм UMAP используется для того, чтобы алгоритмы кластеризации могли работать с данными эффективно. Он снижает размерность многомерных данных (в данном случае векторные представления текстов) и сжимает их в пятимерное пространство, стараясь сохранить структуру: близкие точки остаются близкими, а далекие — далекими [McInnes, Healy, Melville et al., 2018]. HDBSCAN — это алгоритм кластеризации, который ищет плотные области в данных и автоматически определяет количество кластеров [Campello, Moulavi, Sander et al., 2013]. Он может помечать выбросы (точки, которые не относятся ни к одному кластеру) и не требует заранее задавать число групп.

Использование биграмм и триграмм вместо униграмм для первой модели обосновано тем, что устойчивые словосочетания передают более специфичный и контекстно-зависимый смысл, чем отдельные слова, это также позволяет уловить отрицания. В отличие от униграмм, где частотные глаголы или

[2] *Hugging Face*. URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (дата обращения: 05.04.2026).

существительные нередко повторяются в разных темах и создают ложные семантические пересечения, биграммы захватывают уникальные паттерны, характерные для конкретной темы, что уменьшает взаимное «зашумление» и позволяет модели четче разделять кластеры. Кроме того, в разговорной речи о здоровье многие значимые понятия выражены именно фразами («мерить давление», «болеть голова»), потеря которых при переходе к униграммам привела бы к размыванию тем и снижению интерпретируемости результатов. Добавление униграмм в модель было сделано с целью обнаружения связности тем и поиска уникальных, единичных слов в темах.

В процессе итеративной настройки модели BERTopic перебирались ключевые гиперпараметры: размер локальной окрестности UMAP (n_neighbors = 15–21), минимальный размер кластера HDBSCAN (min_cluster_size = 10–15), минимальная частота n-граммы (min_df = 2–3) и размер словаря (max_features = 2000–4000).

Параметр UMAP n_neighbors определяет размер локальной окрестности при построении низкоразмерного представления. Меньшие значения усиливают локальную структуру, что приводит к более компактным кластерам и повышению силуэта, но может увеличить количество мелких тем. Большие значения делают проекцию более глобальной, сглаживая шум, но связаны с риском потерь специфичных микросемантических различий. n_components (фиксированно 5) — размерность целевого пространства.

Меньший min_cluster_size дает больше мелких тем, однако их интерпретируемость может быть осложнена, больший — укрупняет кластеры. Параметр min_df определяет минимальное число документов, в которых должна встретиться n-грамма, чтобы попасть в словарь. Увеличение min_df отсекает редкие и шумовые биграммы, снижая разнообразие тем, но может повысить их разделимость.

В свою очередь, max_features ограничивает размер словаря наиболее частотными n-граммами. Слишком низкий порог делает темы общими, снижая разнообразие; слишком высокий включает редкий шум, что может снизить разделимость кластеров.

Оценка качества велась по силуэтному коэффициенту (разделимость тем — topic diversity), то есть разнообразию тем (уникальности ключевых слов) и доле шума. Под долей шума понимается процент реплик, которые не попали ни в одну тему и отметились выбросами. Разнообразие тем рассчитывается по 10 наиболее популярным словосочетаниям в каждой теме.

Гиперпараметр max_features никак не влиял на метрики кластеризации и был выставлен на значении 2000. В модель было загружено 956 реплик уже с учетом фильтрации. Метрики модели 1 BERTopic с биграммами и триграммами отражены в таблице П1 в приложении. Результаты всех запусков показывают высокое разнообразие тем (выше 0,9) и силуэт выше 0,5, что указывает на хорошую разделимость кластеров [Rousseeuw, 1987]. Для уточнения интерпретации и валидации полученных тем модель была повторно запущена с расширенным диапазоном n-грамм: были включены униграммы, биграммы и триграммы. Это позволило сравнить результаты моделей,

валидизировать и уточнить темы уже на дальнейших этапах. Лучшие метрики модели с униграммами отражены в таблице П2 в приложении, был выбран вариант 1 в пользу большего разнообразия тем.

В таблице П3 в приложении указаны примеры ключевых словосочетаний для модели 1 (только с биграммами и триграммами) с указанием в скобках реплик, которые относятся к этой теме. Выбор сделан в пользу модели 3, так как разнообразие тем является самым высоким. Большее разнообразие позволяет рассмотреть больше различных словосочетаний и отобрать наилучшие для этапа тематического моделирования.

*Этап 2. Валидация кластеров и формирование seed-тем*

Управляемое тематическое моделирование — метод, позволяющий пользователю задать набор ключевых слов (сидов) для каждой темы, которую он хочет получить. Модель использует эти слова, чтобы «подтолкнуть» процесс кластеризации: она сравнивает эмбеддинги документов с эмбеддингами сидов и назначает документы в соответствующие темы, а также искусственно повышает значимость (IDF) сидов, чтобы они с большей вероятностью попадали в итоговое представление темы[3]. На практике если исходные темы не существуют или могут быть разделены на более мелкие темы, то они не будут моделироваться. Также необходимо точное или максимально точное совпадение слов, чтобы тема была действительно создана. Для того чтобы создать этот корпус, я обратилась к таким стратегиям:

1) перечитывание вопросов гайда для уточнения тех тем, которые не были охвачены алгоритмом;
2) чтение реплик, которые были выделены на первом этапе моделирования, для выявления важных с точки зрения исследователя слов и словосочетаний;
3) примерное соотнесение текущих тем с укрупненными темами.

Было установлено, что часть выделенных моделью тем имеет сильные пересечения и даже на основе примеров из текста сложно понять, к какой категории их можно отнести. Также были добавлены биграммы из прошлого моделирования, которые показались самыми интерпретируемыми и теоретически насыщенными. Наиболее сложным для выделения являлись категории описания эмоционального восприятия данных и конфликтов, разногласий с другими членами семьи. Зачастую биграммы и триграммы упоминались лишь один раз во всех интервью, что означает, что они не будут учитываться при формировании темы.

Лучшие значения метрик у управляемой модели 1 (только биграммы и триграммы) были достигнуты при трех наборах параметров, указанных в таблице П4 в приложении. Лучшей моделью была выбрана модель с параметрами 2. Данный выбор обусловлен тем, что другие наборы параметров приводили к появлению тем со смешанным содержанием, либо к чрезмерному уровню шума.

[3] Guided Topic Modeling. *BERTopic Documentation*. URL: https://maartengr.github.io/BERTopic/getting_started/guided/guided.html (дата обращения: 16.04.2026).

Для управляемой модели с униграммами лучшей стала модель с параметрами, указанными в таблице П5 в приложении.

Итоговое распределение тем с репликами представлено ниже, в таблице 1. Названия тем созданы исследователем на основе ключевых n-грамм и примеров реплик.

*Таблица 1*

**Итоговое распределение тем по модели 1 Guided BERTopic и 2 Guided BERTopic**

| № | Словосочетания модели 1 (биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 1) | Словосочетания модели 2 (униграммы, биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 2) |
|---|---|---|---|---|
| 1<br>1 | Не пользоваться, отслеживать шаг, будний день, интересно посмотреть, количество шаг, сидеть дом, не помнить (94) | Шаги и часы | Врач, сходить, анализ, доверять, рекомендация, беспокоить, доктор, друг, проблема (102) | Доверие к врачам |
| 22 | Заниматься спорт, спорт заниматься, ходить тренировка, не заниматься, профессиональный спорт, ходить зал, не спорт, ходить спортзал (81) | Спорт / физическая активность | Часы, сон, день, спать, телефон, час, пульс, просыпаться (94) | Сон, пульс, шаги |
| 33 | Не доверять, сходить врач, доверять врач, врач большой, врач анализ, врач сходить, хороший врач, рекомендация врач (67) | Доверие к врачам | Спорт, заниматься, ходить, заниматься спорт, упражнение, фитнес, спина, тренировка, зал (92) | Спорт / физическая активность |
| 44 | Проблема решать, поэтому следить, начало изучать, личный инициатива, принимать участие, проходить курс, ходить плавать, табличка ее, жена соответственно, зал ходить (59) | Фиксация и обсуждение данных | Давление, тонометр, пульс, лекарство, мерить, мерить давление, таблетка, померить, высокий (63) | Давление |
| 55 | Кандидат наука, врач доверять, вместе смотреть, муж работать, не обсуждать, смотреть врач, не обращать внимание (57) | Совместный выбор врача | Вес, килограмм, взвешиваться, набрать, потерять, весы, месяц, допустить (46) | Контроль веса |

*Продолжение табл. 1*

| № | Словосочетания модели 1 (биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 1) | Словосочетания модели 2 (униграммы, биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 2) |
|---|---|---|---|---|
| 66 | Не обращать внимание, не никто, не никто не, не бывало, не обращать, обращать внимание (51) | Отсутствие практик или их смена | Талия, взвешивать, следить, курс, думать, голова, пользоваться, шутить, пробовать, разумный (44) | Получение информации о здоровье |
| 77 | Мерить давление, вести дневник, голова болеть, давление тонометр, повышенный давление, давление высокий, давление врач, период нужно (49) | Давление | Особо, не никто не, не никто, не бывало, жрать, против (43) | Отсутствие практик или их смена |
| 88 | Взвешиваться месяц, пользоваться приложение, набрать вес, не набирать, ради интерес, не меняться (48) | Контроль веса | Интернет, блогер, научный, статья, почитать, попадаться, телевизор, доверять (40) | Источники информации о здоровье |
| 99 | Здоровый питание, белка жир, жир углевод, здоровый питание не, белка жир углевод, правильно питаться, не считать, сидеть дом (45) | Здоровое питание | Еда, питание, питаться, продукт, стараться, калория, готовить, весы, жирный (40) | Здоровое питание |
| 110 | Мама бабушка, образ жизнь, член семья, не принимать, папа мама, течение день, шаг сделать, допустить мама, далекий родственник, вместе жить (29) | Контроль родственников | Мама, родственник, бабушка, родитель, папа, семья, ребенок, старший, каждый, мама бабушка (39) | Забота о других |
| 111 | Не доверять, не особо доверять, особо доверять, обратить внимание, принцип тема, здравый смысл, не смотреть, информация читать (28) | Источники информации о здоровье | Чувствовать, ощущение, анализ, тело, самочувствие, информация, очередь, показатель, первый (39) | Ощущения и анализы |
| 112 | Каждый месяц, не переживать, отслеживать давление, не отслеживать, сдавать кровь, следить здоровье, анализ кровь, посещать врач (25) | Женское здоровье / беременность | Таблетка, пить, врач, здоровье, мама, следить, нужно, контролировать, заболеть (34) | Обсуждение здоровья в семье и контроль родственников |

*Продолжение табл. 1*

| № | Словосочетания модели 1 (биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 1) | Словосочетания модели 2 (униграммы, биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 2) |
|---|---|---|---|---|
| 113 | Здоровый питание, следить здоровье, член семья, режим день, физический активность, не болеть, здоровье член, нужно заниматься спорт, нужно заниматься (22) | Здоровая семья (понятие) | Муж, врач, мнение, жена, заставлять, декрет (32) | Семейные разногласия |
| 114 | Тема здоровье, зрение гормон, здоровье таки, здоровье достаточно, здоровье здоровье, туда ходить, сдавать анализ кровь, спортзал ходить, состояние здоровье (21) | Цели практик и анализы | Шаг, количество шаг, количество, врать, приложение, километр, прогулка, гулять, посмотреть (22) | Ошибки устройств |
| 115 | Скинуть килограмм, не испытывать, тысяча шаг, первый эмоция, чувство вина, видеть плюс, думать нужно, естественно первый, начинать чувствовать беспокоить, начинать чувствовать (19) | Эмоции и переговоры | Беременность, отслеживать, гинеколог, сдавать, чекап, месяц, врач, сходить, анализ (17) | Беременность / женское здоровье |
| 116 | Хороший чувствовать, полезный здоровье, результат анализ, показатель физический, первый эмоция, показатель физический тело, отслеживать самочувствие (18) | Ощущения vs данные | Здоровье, жизнь, давно, минздрав, рекомендация, рекомендация минздрав, стол, делаться, использовать, подойти (17) | Официальные рекомендации |
| 117 | Пить таблетка, не контролировать, желудочный кишечный, давление периодически, не слушать, подростковый возраст, не сходить (17) | Забота о других | Семья, здоровый, здоровый семья, питание, здоровье, привычка, следить, активность, следить здоровье, здоровый питание (16) | Здоровая семья (понятие) |
| 118 | Не анализ, хороший наоборот, анализ хороший, заставлять вести, доверять большой, не информативный, не вести (16) | Отношение к анализам | Полгода, месяц, год, назад, посмотреть, пара месяц, откладывать месяц, полгода точно, месяц год назад (11) | Частота визитов ко врачу и других практик |

*Окончание табл. 1*

| № | Словосочетания модели 1 (биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 1) | Словосочетания модели 2 (униграммы, биграммы и триграммы), в скобках указано число реплик | Названия темы (модель 2) |
|---|---|---|---|---|
| 119 | Высокий образование, работа учеба, год учиться, жить муж, не сидеть, дом заниматься, заниматься дом, выходить гулять, третий курс (15) | Образование и распорядок дня | Скинуть, радовать, сожаление, килограмм, скинуть килограмм, сверху, прислушиваться, надеяться, показывать, поясница целое (11) | Эмоции и переговоры |
| 220 | Определенный количество, тысяча шаг, количество шаг, касаемо ребенок, ребенок маленький, тысяча стараться, давление высокий, дать приложение, не мерить, врач показывать (13) | Ребенок и шаги | Образование, год, жить, высокий, учиться, высокий образование, муж, магистратура, год высокий образование, жить муж (11) | Образование и статус |
| 221 | Месяц год назад, откладывать месяц, пара месяц, захотеться посмотреть, месяц год, полгода точно, стараться полгода, год назад год, назад год (11) | Частота визитов к врачу и других практик | Самостоятельно, сначала, предложить, начало, адаптироваться, проблема решать, начало изучать, принимать участие, друг (11) | Старт практик заботы о здоровье |

По многим темам (спорт, давление, контроль веса, здоровое питание, забота о родственниках, женское здоровье, понятие здоровой семьи, отношение к анализам) модели сходятся, что подтверждает устойчивое наличие таких тем в текстах. При этом модель 1 лучше выделила темы с эмоциями, чувствами по отношению к мониторингу, а модель 2 с униграммами выделила отдельную уникальную тему с ошибками устройств. Можно сказать, что модели Guided показывают более отличительные и содержательные темы, чем базовая модель при небольшом снижении метрики разнообразия тем.

*Этап 3. Визуализация связей тем и ключевых слов*

Граф совместной встречаемости по модели 1 только из биграмм и триграмм, где узлы представляют словосочетания, не имел ни одного ребра, что означает, что ни одна пара словосочетаний из выбранного набора не встречается вместе ни в одной реплике информанта. Это означает, что внутри каждого отдельного ответа информанта используются непересекающиеся наборы фраз из модели 1. В таком случае более информативен может быть граф связей тем

через общие ключевые слова. В этом графе узлы — это сами темы (кластеры реплик). Ребро появляется, если одно и то же ключевое словосочетание (биграмма или триграмма) входит в топ-15 ключевых словосочетаний для двух разных тем. Такие словосочетания часто встречаются именно в этой теме, но редко встречаются в других. Наиболее связанной является тема 12 («Женское здоровье / беременность») и 13 («Здоровая семья»). Такой граф показывает связи практик, на которые исследователь может в дальнейшем обратить внимание при анализе: женские практики заботы о здоровье, в том числе в период беременности и в период восстановления после нее, связаны с регулярным взвешиванием каждый месяц (тема 8, «контроль веса»), сдачей анализов (тема 12, «анализ кровь, сдавать кровь»). Слова, указанные на ребрах между узлами, показывают некоторую схожесть стратегий поведения в отношении здоровья — «не обсуждать», «не обращать внимание», не «обращать» — между категориями 5 («Совместный выбор врача») и 6 («Отсутствие практик или их смена»).

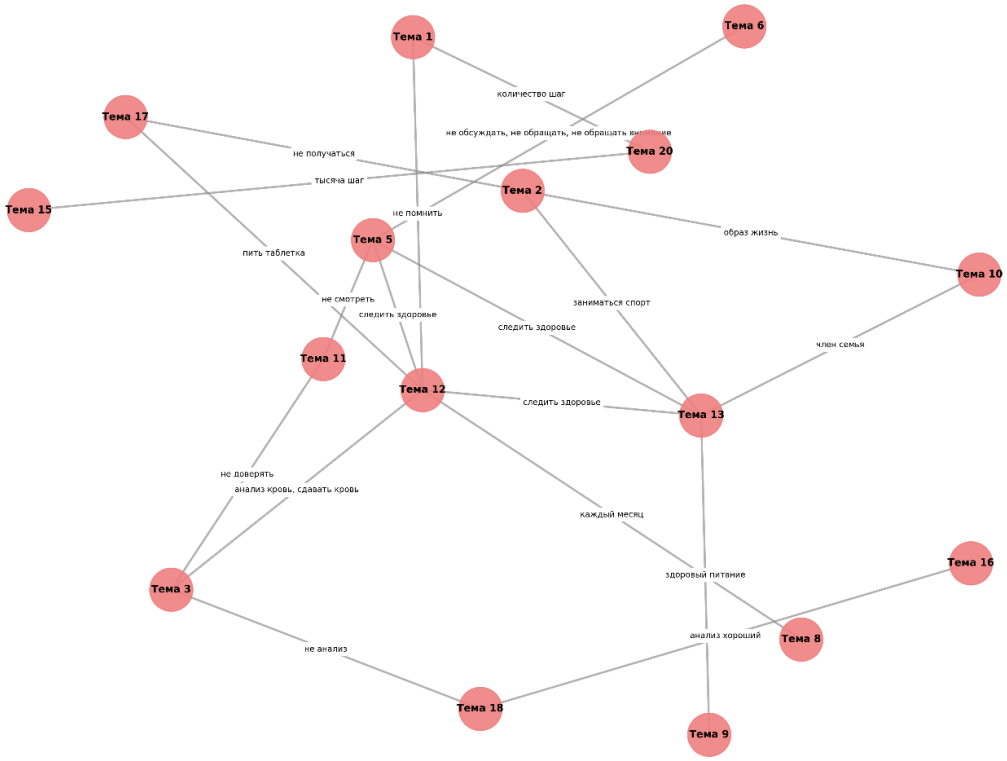



**Рисунок 1.** Граф совместной встречаемости (для модели 2 Guided BERTopic).

Для модели 2 с униграммами был составлен граф совместной встречаемости униграмм в репликах информантов. Этот граф показывает отдельные слова, которые встречаются в документах коллекции. В этом графе для всех документов, не являющихся шумом, из каждой темы извлекаются

топ-15 наиболее частотных униграмм. Каждая такая униграмма становится узлом, которому присваиваются тема и частота. Затем для каждой реплики проверяется, какие из этих узлов в них встречаются; если в одной реплике присутствуют две или более таких фраз, между соответствующими узлами добавляется ребро, вес которого равен числу реплик, где эта пара встретилась совместно. После этого были оставлены только 3% самых сильных связей, так как иначе граф было сложно интерпретировать из-за большого числа связей. Узлы раскрашиваются по темам, а их размер отражает частоту фразы в своей теме. Толщина ребра отражает силу связи, а более крупный размер узла показывает ключевые униграммы. Такой граф позволяет посмотреть связи как внутри темы, так и между темами через отдельные слова. Например, относительно сильной связью обладает группа из трех узлов — «доверять», «интернет», «врач», чуть менее сильной является цепочка «пульс» — «часы» — «сон» — «отслеживать» — «обсуждать». В центре графа сосредоточено большое число связей, однако их сложно интерпретировать. Таким образом, оба графа дополняют друг друга и позволяют понять, какие темы близки друг с другом на уровне семантики, а также какая внутренняя структура слов находится внутри темы.

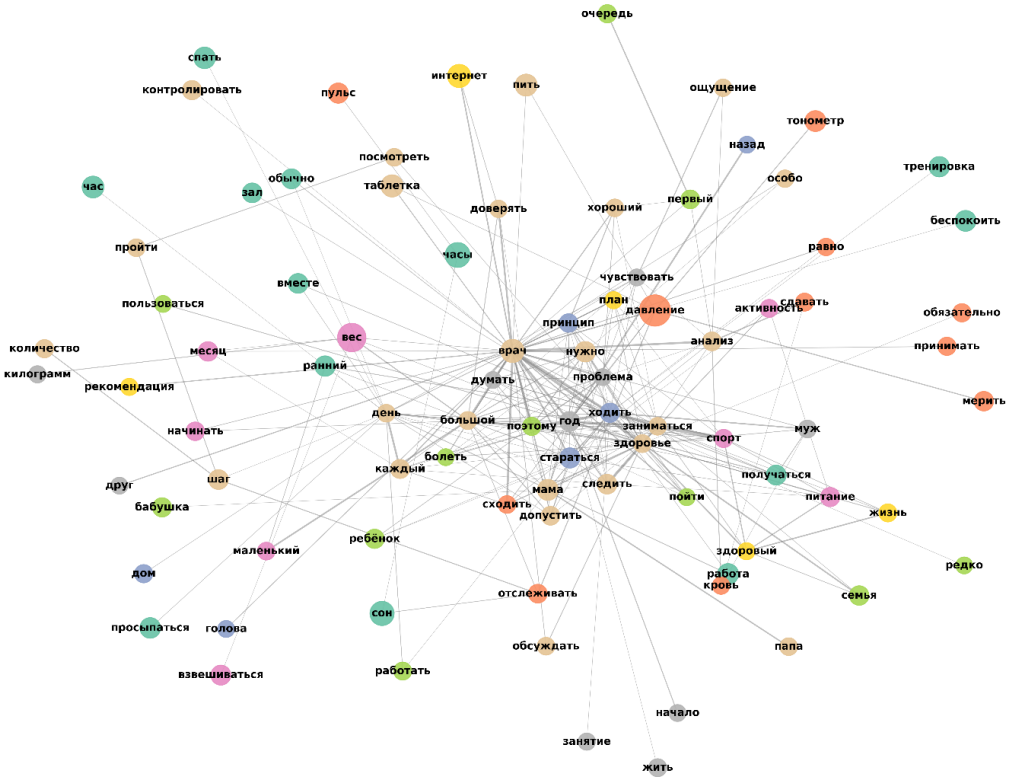



**Рисунок 2.** Граф связей тем через общие ключевые слова (для модели 2 Guided BERTopic)

*Этап 4. Создание тем второго порядка*

Далее необходимо перейти к проверке и обобщению созданных тем. Так как предметом изучения были практики заботы о здоровье, при уточнении тем необходимо было описать как условия их возникновения, так и характер, способ реализации этих практик в семейном контексте и групповую семейную динамику, формирующуюся в связи с этими практиками.

Файлы с примерами реплик информантов по темам были предварительно анонимизированы: из текстов удалены имена участников, названия конкретных медицинских учреждений. Первоначально планировалось развернуть модель локально на собственном компьютере, чтобы обеспечить максимальную конфиденциальность и снять ограничения на объем обрабатываемых данных. Однако из-за недостаточной вычислительной мощности ПК (ограниченный объем оперативной памяти) и жестких ограничений по максимальной длине контекста из отправленных документов (количеству токенов, которое модель может обработать за один раз) локальный запуск оказался невозможным, поэтому применялась нейросеть DeepSeek через диалоговый интерфейс[4]. В чат загружались анонимизированные файлы с описаниями тем и примерами реплик обеих моделей, а также гайд интервью. С помощью специально составленного промпта LLM ставилась задача оценить предварительные темы из двух моделей, объединить, разделить, переименовать их, сгруппировать итоговые темы в тематические блоки, создать комплексное описание данных, соотнести с исследовательскими задачами. Полный промпт указан в приложении. Сравнение тем, выделенных моделью и кодировщиком, представлено в таблице 2. Столбцы «Номера исходных тем (модель 1)» и «Номера исходных тем (модель 2)» заполнены LLM. При оценке совпадений подтем учитывались как описания, предложенные LLM, так и цитируемые ею реплики. Под полным совпадением подразумеваются те темы, где совпадают отобранные цитаты и описание темы. Частичное совпадение фиксировалось в тех случаях, когда описания темы у LLM и у исследователя не совпадали в полной мере. Низкий уровень совпадения свидетельствовал либо об ошибке LLM, либо о том, что модель не смогла в полной мере передать основной смысл, объединяющий сразу несколько тем, выделенных исследователем.

Исследовательский анализ включал просмотр примеров реплик для каждой выделенной темы, обращение к транскриптам интервью и выделение более обобщенных или содержательных тем. Исследователь первоначально опирался на примеры реплик из тем и ключевые словосочетания, и если этой информации было недостаточно, то он обращался к полным транскриптам.

Использование предварительных тем, полученных с помощью Guided BERTopic, ускорило работу исследователя по проверке, особенно при выделении подтем, связанных с форматами взаимодействия с устройствами мониторинга, с пониманием здоровой семьи и с конфликтными ситуациями. Реплики по этим тематикам оказались в более однородных кластерах, что позволило выделить все подтемы практически без дополнительных усилий. Кроме того,

[4] *DeepSeek*. URL: https://www.deepseek.com/ (дата обращения: 16.04.2026).

*Таблица 2*

**Темы и подтемы, выделенные языковой моделью и кодировщиком на основе тем, полученных от моделей 1 и 2 BERTopic**

| Номер | Главная тема | Подтема LLM | Соответствующие подтемы исследователя | Номера исходных тем (модель 1 с биграммами и триграммами) | Номера исходных тем (модель 2 с униграммами, биграммами и триграммами) | Совпадения |
|---|---|---|---|---|---|---|
| 1 | **Цифровой и нецифровой мониторинг** | Функционально-контролирующий тип мониторинга | «Устройство как сопровождение при хроническом заболевании», «устройство как сопровождение в активности или достижении цели» | 2 (частично), 4, 5, 19, 6, 9, 1 | 2, 4, 5, 9, 14 | Частичное |
| 2 | | Исследовательско-любопытствующий тип мониторинга | «Устройство-компаньон» | | | Полное |
| 3 | | Тревожно-избегающий тип мониторинга | «Устройство как зеркало себя» | | | Частичное |
| 4 | | Социально-демонстративный тип мониторинга | «Устройство как зеркало себя» | | | Частичное |
| 5 | | Отказ от мониторинга | «Эмоции от трекеров и данных» | | | Низкое |

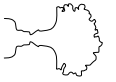
*Продолжение табл. 2*

| Номер | Главная тема | Подтема LLM | Соответствующие подтемы исследователя | Номера исходных тем (модель 1 с биграммами и триграммами) | Номера исходных тем (модель 2 с униграммами, биграммами и триграммами) | Совпадения |
|---|---|---|---|---|---|---|
| 6 | **Практики питания и физической активности** | Контроль питания (калории, диета) | «Здоровое питание», «подсчет калорий» | 9, 3 | 9, 3 | Полное |
| 7 | | Спорт, растяжка, прогулки, бассейн | «Спорт», «прогулки», «массаж» | | | Частичное |
| 8 | **Семейная координация практик заботы о здоровье** | Единый центр (матриархальный/патриархальный) | «Индивидуальные практики», «ответственный за здоровье» | 12, 13, 21, 13, 10 | 12, 13, 21, 13, 10 | Полное |
| 9 | | Распределенная координация (кооперативная) | «Индивидуальная практика с обсуждением» | | | Частичное |
| 10 | | Активное включение (один инициирует, другой выполняет) | «Коллективные практики» | | | Частичное |
| 11 | | Конфликтная координация | «Разногласия», «стратегии навязывания позиции» | | | Частичное |
| 12 | | Пожилая опека (взрослые дети контролируют пожилых родителей) | «Коллективные практики» | | | Полное |

*Продолжение табл. 2*

| Номер | Главная тема | Подтема LLM | Соответствующие подтемы исследователя | Номера исходных тем (модель 1 с биграммами и триграммами) | Номера исходных тем (модель 2 с униграммами, биграммами и триграммами) | Совпадения |
|---|---|---|---|---|---|---|
| 13 | **Условия возникновения практик** | Социальное давление / пример окружения | «Влияние окружения», «влияние СМИ и интернета» | 4, 1, 2, 5, 6, 9, 2, 8, 3 | 4, 14, 2, 5, 9, 2, 8, 3, 15, 3 | Полное |
| 14 | | Предписание врача / медицинские показания | «Институциональное давление» | | | Полное |
| 15 | | Жизненные переходы (беременность, роды, переезд, выход на пенсию) | «Смена статуса семьи» | | | Полное |
| 16 | | Острый эпизод болезни | «Личный опыт болезни», «болезнь близкого» | | | Частичное |
| 17 | | Хроническое заболевание | «Личный опыт болезни», «болезнь близкого» | | | Частичное |
| 18 | | Цифровая доступность и любопытство | «Интерес к технологиям и инновациям» | | | Частичное |
| 19 | | Укрепление внутрисемейных связей (совместные активности) | «Эмоциональные эффекты на коллективном уровне» | | | Полное |
| 20 | | Рост самостоятельности и ответственности (особенно у детей) | Нет | | | Нет |





*Продолжение табл. 2*

| Номер | Главная тема | Подтема LLM | Соответствующие подтемы исследователя | Номера исходных тем (модель 1 с биграммами и триграммами) | Номера исходных тем (модель 2 с униграммами, биграммами и триграммами) | Совпадения |
|---|---|---|---|---|---|---|
| 21 | | Появление новых конфликтов / напряжение | «Разногласия», «стратегии навязывания позиции» | | | Частичное |
| 22 | | Снижение тревожности (через контроль) или ее усиление | «Эмоциональные эффекты на индивидуальном уровне» | | | Частичное |
| 23 | | Изменение семейной рутины (встраивание практик) | «Закрепление практик», «нахождение компромиссной практики» | | | Частичное |
| 24 | **Представления о здоровье и нормативные модели** | Представления о здоровой семье | «Умеренность», «постоянство», «забота о близких», «отслеживание самочувствия», «наличие ресурсов» (понимание здоровой семьи через данные подтемы). «Состояние здесь и сейчас», «долгосрочная инвестиция» (подтемы понимания здоровья) | 1, 13, 17, 15 | 1, 13, 17, 15 | Низкое |

*Окончание табл. 2*

| Номер | Главная тема | Подтема LLM | Соответствующие подтемы исследователя | Номера исходных тем (модель 1 с биграммами и триграммами) | Номера исходных тем (модель 2 с униграммами, биграммами и триграммами) | Совпадения |
|---|---|---|---|---|---|---|
| 25 | | Оценка самочувствия: ощущения vs анализы | «Опора на ощущения» | | | Полное |
| 26 | | Недоверие к рекомендациям врача и самостоятельная коррекция | «Доверие ко врачу», «доверие к информации» | | | Низкое |
| 27 | | Выбор врача и критерии доверия | «Доверие ко врачу», «доверие к информации» | | | Частичное |

наличие конкретных цитат для каждого кластера сделало обращение к полным транскриптам незатруднительным. Анализ начинался с практик, затем исследователь перешел к понятию здоровой семьи и далее — к темам статуса членов семьи и распорядка дня. Обращение к транскрипту требовалось лишь в тех случаях, когда реплика была очень короткой или когда необходимо было рассмотреть процесс переговоров и взаимодействия между разными членами одной семьи.

Исследователем было выявлено три способа реализации практик заботы о здоровье в семье: когда член семьи использует индивидуальные практики и не делится результатами с другими; когда существует коллективная практика; и когда данные и результаты индивидуальной практики обсуждаются с другими членами семьи. При этом расширение практики на других членов семьи может проявляться как относительно субъекта (того, кто поддерживает здоровье), так и относительно объекта (ради блага кого эта практика реализуется). Такое разделение позволяет зафиксировать активных и пассивных участников, а также тех, кто получает преимущество от расширения практик. Аспект семейных ролей, на котором LLM построила свою классификацию практик, в исследовательских подтемах выделен в самостоятельную подтему «Ответственный за здоровье». Также LLM не выделила поиск компромисса в практиках как отдельную подтему, однако, с точки зрения исследователя, именно поиск компромисса является основой для закрепления практики на уровне всей семьи.

Условия появления практик заботы о здоровье совпадают у LLM и исследователя, однако исследователь подчеркивает, что наиболее распространенными условиями возникновения практик заботы о здоровье являются именно внутренние условия, например, набор веса из-за беременности, возникновение проблемы со здоровьем в связи с возрастом, болезнь близкого, LLM не делала разделения на внутренние и внешние условия.

Исследователь выделил способы взаимодействия с устройствами мониторинга на основе целей использования, ситуации использования и эмоций от использования. Полученная типология оказалась более детальной, чем у LLM: в одном случае использование устройства при управлении хроническим заболеванием не вызывало эмоций, но при этом было постоянным; в другом — устройство использовалось как сопровождение в активности или для достижения цели, использование носило периодический характер и могло вызывать положительные эмоции. У LLM оба этих типа были объединены в одну подтему «Функционально-контролирующий тип мониторинга».

Также LLM выделила подтему понимания здоровой семьи, однако, в отличие от исследователя, не смогла выделить в ней отдельные паттерны восприятия. Исследователь обнаружил, что разные семьи придерживаются чаще одного или двух конкретных взглядов на здоровую семью. Женщины склонны рассматривать здоровье как долгосрочную инвестицию, требующую постоянных вложений сегодня ради качества жизни в будущем, мужчины чаще воспринимают здоровье как текущий ресурс, ценность которого актуализируется преимущественно в ситуации «здесь и сейчас». Тема 20 не была выделена

кодировщиком, но может быть добавлена и расширена при сравнении поведения родителей и детей.

Для обнаружения галлюцинаций исследователь проверял каждую приведенную LLM цитату на соответствие исходным транскриптам и оценивал релевантность цитаты теме. Некоторые подтемы LLM были некорректны, так как LLM опиралась на отдельную цитату и начинала «додумывать» смысл, который вкладывал информант. Такими темами оказались подтема 13 (LLM интерпретировала упоминание блогера как ключевой фактор перехода к здоровому образу жизни, сам информант лишь сообщил о симпатии к этому блогеру, не связывая с ним начала своих практик), подтема 18 (цитата информанта была придумана), подтема 26 (LLM использовала цитату об экспертах и врачах из интернета, тогда как речь шла о врачах учреждений; при недоверии к врачам информанты не корректируют назначения сами, а обращаются к нескольким специалистам).

В случае галлюцинации модели или ее неверной интерпретации темы предлагается применять такие методы:

– попросить модель выдать примеры цитат из разных интервью, которые относятся к этой теме с указанием номера транскрипта. Далее эти цитаты исследователь может проверить самостоятельно;
– в самом промпте стоит указать, чтобы модель опиралась только на приложенные материалы, и честно признавалась в том, что ей не удалось выделить какие-то коды;
– сделать дополнительную генерацию с этим же промптом. Это позволит проверить совпадение тем, а также выбрать наилучшие формулировки тем или цитат;
– перейти в режим беседы с LLM, что позволит уточнить и скорректировать существующие темы. Итеративная работа с промптами позволяет также избежать генерации кодов на единичных цитатах [Salazar et al., 2025]. Последующие промпты можно задавать LLM с позиции критика, который должен проверить результаты предыдущей генерации как работу профессионального исследователя.

Полное совпадение подтем, выделенных LLM и исследователем, зафиксировано в 8 из 27 случаев, частичное — в 15, низкое — в 3, и одна подтема не была обозначена исследователем. Эти результаты демонстрируют хорошее пересечение, если учесть, что исследователь также использовал машинные темы в качестве отправной точки, но мог дополнять анализ с использованием транскриптов. Передача в LLM уже сгруппированных реплик по тематическим кластерам позволяет модели учитывать весь массив данных, обходя технические ограничения на объем загружаемых файлов.

LLM оказалась ограничена в формулировании тем, связанных с переговорами и причинами разногласий между членами семьи относительно понимания ценности практик заботы о здоровье. Потенциальным способом улучшения качества ее работы видится переход в роль собеседника-критика или рецензента: в диалоге с исследователем модель последовательно уточняет темы, находит недостатки в собственных предыдущих ответах и предлагает

альтернативные интерпретации. Дообучение в данном случае затруднительно, поскольку требует больших объемов размеченных данных, специфичных для каждого конкретного исследования.

## Заключение

Сочетание методов машинного обучения (BERTopic, UMAP, HDBSCAN) с контролем эксперта обеспечивает более качественное тематическое моделирование по сравнению с использованием только алгоритмов, даже при оптимизации гиперпараметров.

Шумовая тема, которую алгоритм HDBSCAN отмечал меткой –1, не попала в визуализацию и не была включена LLM в анализ. В основном туда попали длинные реплики, в которых поднимается сразу несколько тем или тема «скрыта» для модели из-за нетипичных словосочетаний:

*«Для меня в первую очередь это способность концентрироваться. То есть если я, например, понимаю, что мне сложно сконцентрироваться, чтобы что-то вот, ну, на работе думать, то я понимаю, что что-то идет не так, и нужно выспаться получше, поесть получше, или наоборот, там, сходить развеяться, то есть как-то я загналась. В первую очередь это вот эта вот рабочая работоспособность. Да, на втором месте, наверное, это вот вес, потому что, ну, когда там я вижу плюс какой-то на весах, я понимаю, что это отек, отек откуда-то взялся, значит, надо что-то вот тоже менять»* (ж., 31, Москва, замужем, живет с мужем);

*«Мне нравится технология. Я вообще ими мечтал бы обвешаться, этим всем, только чтобы оно не натирало, не мешало как-то и не отвлекало в целом. Нравятся всякие лампочки, бимпочки, то, что это все бегает. Но не могу сказать, что это информативно. То есть от них не устаешь, а просто за какой-то бессмысленностью через какое-то время этот энтузиазм теряется»* (м., 30, Санкт-Петербург, женат, живет с женой, дочерью).

Это подчеркивает важность предварительной обработки транскриптов, например, дополнительного разделения реплик на отдельные смысловые части.

Кроме того, многие из объемных тем содержали реплики, не соответствующие основной тематике, даже несмотря на высокие значения метрик. Роль исследователя здесь заключается в том, чтобы использовать полученные темы лишь как примерный ориентир, сохраняя за собой право формировать новые категории — как путем объединения нескольких тем, так и на основе отдельных показательных реплик. Применение модели BERT, UMAP и HDBSCAN с извлечением ключевых слов и словосочетаний позволяет

оценить распространенность темы и выборочно ознакомиться с наиболее характерными для нее репликами. Использование предварительных выделенных тем ускоряет процесс пересмотра тем исследователем, так как позволяет просматривать цитаты сразу по конкретной тематике. Применение модели BERT видится эффективным на большом корпусе текстов, а дальнейшее выборочное уточнение и раскрытие тем исследователем производится уже на основе выборки реплик.

Использование результатов модели BERT для LLM позволяет генеративной модели использовать весь объем материала и достичь достаточно хорошего уровня кодирования, но не решает вопрос галлюцинаций и ошибок. Для улучшения работы LLM предлагается итеративный промптинг: назначение LLM роли критика после роли исследователя. Заданный в промпте примерный перечень тем направлял LLM, что могло ограничить появление неожиданных категорий.

## Литература / References


Anakok I., Katz A., Chew K.J., Matusovich H. (2025) Leveraging Generative Text Models and Natural Language Processing to Perform Traditional Thematic Data Analysis. *International Journal of Qualitative Methods*. Vol. 24. P. 1–13. DOI: https://doi.org/10.1177/16094069251338898

Braun V., Clarke V. (2016) Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*. Vol. 3. No. 2. P. 77–101. DOI: https://doi.org/10.1191/1478088706qp063oa

Campello R.J.G.B., Moulavi D., Sander J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates. *Lecture Notes in Computer Science*. Vol. 7819. P. 160–172. DOI: https://doi.org/10.1007/978-3-642-37456-2_14

De Paoli S. (2024) Performing an Inductive Thematic Analysis of Semi–Structured Interviews with a Large Language Model: An Exploration and Provocation on the Limits of the Approach. *Social Science Computer Review*. Vol. 42. No. 4. P. 997–1019. DOI: https://doi.org/10.1177/08944393231220483

Egger R., Yu J.A (2022) Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*. Vol. 7. P. 1–16. DOI: https://doi.org/10.3389/fsoc.2022.886498

Gamieldien Y., Case J.M., Katz A. (2023) Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding. *SSRN*. P. 1–54. DOI: https://doi.org/10.2139/ssrn.4487768

Grootendorst M. (2022) BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv preprint*. P. 1–10.

Hitch D. (2024) Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qualitative Health Research*. Vol. 34. No. 7. P. 595–606. DOI: https://doi.org/10.1177/10497323231217392

Ionescu T.C., Han L., Suasnabar J.H., Stiggelbout A., Verberne S. (2026) Analyzing Cancer Patients' Experiences with Embedding-Based Topic Modeling and LLMs. *arXiv preprint*. P. 1–26.

Li K.D., Fernandez A.M., Schwartz R., Rios N., Carlisle M.N., Amend G.M., Patel H.V., Breyer B.N. (2024) Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study. *Journal of Medical Internet Research*. Vol. 26. P. 1–13. DOI: https://doi.org/10.2196/56500

McInnes L., Healy J., Melville J. (2018) UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. Vol. 3. No. 29. P. 861. DOI: https://doi.org/10.21105/joss.00861

Nelson L.K. (2020) Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*. Vol. 49. No. 1. P. 3–42. DOI: https://doi.org/10.1177/0049124117729703

Quillivic R., Payet C. (2024) *Semi–Structured Interview Analysis: A French NLP Approach for Social Sciences*. Brussels: JADT.

Renz S. M., Carrington J. M., Badger T. A. (2018) Two Strategies for Qualitative Content Analysis: An Intramethod Approach to Triangulation. *Qualitative Health Research*. Vol. 28. No. 5. P. 824–831. DOI: https://doi.org/10.1177/1049732317753586

Rousseeuw P. J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*. Vol. 20. P. 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7

Salazar M., Chaw M., Hellier Y., Hsia S., Gruenberg K. (2025) Comparison of Qualitative Analyses Conducted by Artificial Intelligence Versus Traditional Methods. *American Journal of Pharmaceutical Education*. P. 1–5. DOI: https://doi.org/10.1016/j.ajpe.2025.101882

**Сведения об авторе:**

**Червоткина Полина Игоревна —** студентка бакалавриата, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. **E-mail:** bcur1119@gmail.com. **РИНЦ Author ID:** 1340883; **ORCID ID:** 0009-0006-3586-7087.

**Статья поступила в редакцию:** 17.04.2026
**Принята к публикации:** 30.04.2026

## Expert Validation of Thematic Modeling (BERTopic) and Subsequent Application of LLM to Selected Topics in Qualitative Research within the Framework of Thematic Analysis

**DOI: 10.19181/inter.2026.18.2.8**

***Polina I. Chervotkina*** *Saint Petersburg State University, St. Petersburg, Russia*
*E-mail: bcur1119@gmail.com*

*This article explores the application of machine learning and generative language models to the analysis of qualitative interviews. The aim of the study is to develop and validate a mixed methodology combining topic modeling (BERTopic) with expert supervision. The empirical base consisted of 20 transcripts of semi-structured interviews with family members about health care practices. In the first stage, BERTopic, UMAP, and HDBSCAN were used to identify clusters of replicas based on bigrams and trigrams, and then with the addition of unigrams. Hyperparameter selection was then performed. The second stage involved expert validation of the clusters and the construction of a guided model with the addition of expert-generated themes (seed themes). The third stage involved visualizing the relationships between themes through term co-occurrence graphs. In the fourth stage, the results of the previous stages were summarized and grouped using DeepSeek. The results were compared with the encoder’s performance on the BERT model topics and interview transcripts.*

*It was found that BERTopic with bigrams and trigrams demonstrates higher topic diversity (0.96) and separability (silhouette coefficient of 0.69) compared to the model including unigrams (topic diversity of 0.84, silhouette coefficient of 0.56). Guided models made it possible to reduce the proportion of noisy utterances, identify topics not detected during the initial model run (e.g., device errors), and improve topic separability metrics (silhouette coefficient of 0.72 for the model with bigrams and trigrams, 0.73 for the model with unigrams). LLM demonstrates limited ability to independently formulate topics not specified in the prompt, but its performance is moderate with researcher coding. The paper proposes methods for reducing hallucinations and improving the quality of thematic generalization in LLM.*

**Keywords:** qualitative interview analysis; topic modeling; BERTopic; large language models; thematic analysis; machine learning

**Author bio:**

**Polina I. Chervotkina —** Bachelor's Student, Saint Petersburg State University, St. Petersburg, Russia. **E-mail:** bcur1119@gmail.com. **RSCI Author ID:** 1340883; **ORCID ID:** 0009-0006-3586-7087.

**Received:** 17.04.2026
**Accepted:** 30.04.2026

## ПРИЛОЖЕНИЕ

*Таблица П1*

**Метрики модели 1 BERTopic (биграммы и триграммы)**

| | min_ cluster_ size | n_ neighbors | min_df | Число тем | Шум,% | Силуэт | Разнообразие тем |
|---|---|---|---|---|---|---|---|
| 11 | 12 | 18 | 2 | 22 | 23,7 | 0,6398 | 0,9320 |
| 22 | 15 | 21 | 2 | 14 | 25,3 | 0,6901 | 0,9512 |
| 33 | 12 | 19 | 2 | 21 | 18,2 | 0,5690 | 0,9592 |

*Таблица П2*

**Метрики модели 2 BERTopic (униграммы, биграммы и триграммы)**

| | min_ cluster_ size | n_ neighbors | min_df | Число тем | Шум,% | Силуэт | Разнообразие тем |
|---|---|---|---|---|---|---|---|
| 11 | 12 | 15 | 2 | 25 | 25,5 | 0,5627 | 0,8354 |
| 22 | 11 | 20 | 2 | 24 | 18,8 | 0,6304 | 0,8319 |

*Таблица П3*

**Примеры тем и ключевых слов модели 1 BERTopic (биграммы и триграммы), в скобках указано число реплик, относящихся к этой теме**

| | Параметр 1 | Параметр 2 | Параметр 3 |
|---|---|---|---|
| 1 | Количество шаг, часть здоровье, не бывало, не обсуждать, не следить, не обсуждать не (156) | Сходить врач, посещать врач, рекомендация врач, анализ сдавать, хороший врач, не врач (85) | Сходить врач, посещать врач, анализ сдавать, хороший врач, рекомендация врач, не врач, не болеть (84) |
| 2 | Заниматься спорт, спорт заниматься, ходить спортзал, профессиональный спорт, ходить зал, не заниматься, не спорт (74) | Начало изучать, личный инициатива, думать далекий, жена соответственно, ходить плавать, проходить курс, отслеживать цикл, проблема решать, принимать участие, прийтись пока (76) | Ради интерес, вес стабильно, взвешиваться месяц, набрать вес, не меняться (80) |
| 3 | Пить таблетка, кандидат наука, врач доверять, проблема здоровье, вместе смотреть, следить здоровье, стараться следить, врач сходить, не контролировать (43) | Заниматься спорт, спорт заниматься, ходить спортзал, профессиональный спорт, ходить зал, не заниматься, не спорт (74) | Заниматься спорт, спорт заниматься, ходить спортзал, профессиональный спорт, ходить зал, не заниматься, не спорт (75) |
| 4 | Большой доверять, особо доверять, принцип тема, обратить внимание, доверять врач, поэтому стараться, не доверять, не особо доверять (43) | Отслеживать шаг, день проходить, давление пить, фитнес браслет, не пользоваться, не носить, не беспокоиться (69) | Идти тренировка, смотреть пульс, сидеть дом, отслеживать шаг, не пользоваться, не беспокоиться, не носить (65) |
| 5 | Взвешиваться месяц, вес стабильно, набрать вес, ради интерес, не меняться (40) | Образ жизнь, мама бабушка, возраст год, член семья, год учиться, работа учеба (60) | Мерить давление, вести дневник, голова болеть, период нужно, давление высокий, не пользоваться, не подойти (64) |
| 6 | Идти тренировка, ложиться спать, день проходить, давление пить, пить лекарство, тренировка пойти, сидеть дом, член семья (37) | Мерить давление, вести дневник, голова болеть, повышенный давление, период нужно, давление тонометр, не пользоваться (59) | Поэтому следить, прийтись пока, отслеживать цикл, начало изучать, проходить курс, ходить плавать, цель шаг, принимать участие, личный инициатива, думать далекий (62) |

*Продолжение табл. П3*

| | Параметр 1 | Параметр 2 | Параметр 3 |
|---|---|---|---|
| 7 | Белка жир углевод, здоровый питание не, жир углевод, белка жир, здоровый питание, правильно питаться, образ жизнь, здоровый образ, здоровый образ жизнь (34) | Пить таблетка, проблема здоровье, кандидат наука, врач доверять, идти врач, вместе смотреть, следить здоровье, не контролировать (58) | Количество шаг, часть здоровье, пройти шаг, шаг сделать, не обсуждать, не планировать, не следить, не бывало (52) |
| 8 | Сходить врач, врач анализ, поход врач, сходить врач не, хороший врач, рекомендация врач, анализ сдавать, не врач, не рекомендация (34) | Не бывало, не нужный, не обсуждать (47) | Пить таблетка, кандидат наука, врач доверять, проблема здоровье, тема здоровье, вместе смотреть, идти врач, не контролировать (50) |
| 9 | Мама бабушка, образ жизнь, член семья, папа мама, течение день, думать хороший, запоминать касаться, допустить мама, вместе жить, не принимать (31) | Анализ хороший, анализ плохой, анализ кровь, сдавать анализ, начинать думать, хороший наоборот, не анализ, не норма (42) | Доверять врач, врач не ходить, принцип тема, особо доверять, большой доверять, обратить внимание, не доверять, не особо доверять (45) |
| 10 | Давление принимать, повышенный давление, мерить давление, голова болеть, давление высокий, высокий давление, принимать таблетка, не проходить (29) | Особо доверять, принцип тема, большой доверять, доверять врач, обратить внимание, поэтому стараться, не доверять, не особо доверять (42) | Мама бабушка, член семья, образ жизнь, папа мама, анализ смотреть, думать хороший, допустить мама, не принимать, не месяц, не задумываться (32) |
| 11 | Мерить давление, мерило давление, не пользоваться, не использовать, не следить, не повод (26) | Вес стабильно, взвешиваться месяц, набрать вес, ради интерес, маленький порция (33) | Тысяча шаг, количество шаг, определенный количество, каждый вечер, касаемо ребенок, ребенок маленький, шаг день, тысяча стараться, женский календарь, каждый месяц (24) |
| 12 | Тысяча шаг, количество шаг, определенный количество, тысяча стараться, ребенок маленький, каждый вечер, касаемо ребенок, шаг день, женский календарь, каждый месяц (26) | Здоровый питание, здоровый питание не, белка жир углевод, белка жир, жир углевод, правильно питаться, образ жизнь, здоровый образ, здоровый образ жизнь (29) | Анализ плохой, анализ хороший, хороший наоборот, сдавать кровь, не анализ, не норма (22) |

*Продолжение табл. П3*

| | Параметр 1 | Параметр 2 | Параметр 3 |
|---|---|---|---|
| 13 | Здоровый питание, следить здоровье, член семья, физический активность, режим день, образ жизнь, нужно заниматься, не проблема (25) | Здоровый семья, здоровый питание, следить здоровье, режим день, физический активность, образ жизнь, член семья, не проблема (23) | Здоровый семья, здоровый питание, следить здоровье, режим день, физический активность, член семья, питание спорт, вредный привычка (19) |
| 14 | Анализ хороший, анализ плохой, хороший наоборот, сдавать кровь, не анализ, не норма (22) | Количество шаг, посмотреть пройти, часть здоровье, шаг сделать, интересно посмотреть, не следить, не смотреть, не пойти (17) | Высокий образование, родиться занятие, год учиться, год месяц, возраст год, год месяц год, третий курс, далее муж, учиться магистратура (15) |
| 15 | Врач большой, доктор наука, большой опыт, большой смотреть, взять инициатива, запись врач, врач назначать, врач порекомендовать, каждый хороший, найти врач (16) | | Лекарство подойти, здоровье таки, жизнь здоровье, повод здоровье, ездить дача, туда ходить, повод здоровье достаточно, состояние здоровье, ходить думать, спортзал ходить (15) |
| 16 | Родиться занятие, высокий образование, год учиться, возраст год, далее муж, месяц год, образование высокий, учиться магистратура, смотреть жить, не получить (15) | | Порция заниматься, не никто, не никто не, не мерить, не взвешивать (15) |
| 17 | Здоровье таки, жизнь здоровье, ездить дача, повод здоровье, повод здоровье достаточно, туда ходить, спортзал ходить, состояние здоровье, ходить думать, думать здоровье (14) | | Приложение телефон, календарь цикл, интересно находить, здоровье получаться, посмотреть ходить, шаг считать, получаться муж (13) |
| 18 | Отслеживать шаг, день смотреть, смотреть часы, часы появиться, пользоваться год, смотреть сон, режим день, интересно посмотреть, не отслеживать (14) | | Откладывать месяц, пара месяц, полгода точно, захотеться посмотреть, год месяц год, стараться полгода, месяц год назад, месяц год, назад год (13) |

*Окончание табл. П3*

| | Параметр 1 | Параметр 2 | Параметр 3 |
|---|---|---|---|
| 19 | Приложение телефон, календарь цикл, вес зал, здоровье получаться, интересно находить, шаг считать, получаться муж (13) | | Полезный здоровье, чувствовать хороший, первый эмоция, отслеживать самочувствие, общий ощущение, ощущение ориентироваться, ощущение первый, психологический точка, стоить точка, сила возможность (13) |
| 20 | Полезный здоровье, чувствовать хороший, здоровье целое, ощущение ориентироваться, первый эмоция, ощущение первый, показатель физический, показатель физический тело, общий ощущение, отслеживать самочувствие (13) | | Здоровье таки, доверять совет, пройти шаг не, не обсуждать, не делиться (12) |
| 21 | Здоровье таки, доверять совет, пройти шаг не, не обсуждать, не делиться (12) | | Отслеживать давление, сдавать кровь, вес большой, год день, год начало, делиться врач, врач помогать, не отслеживать (12) |
| 22 | Отслеживать давление, сдавать кровь, год день, год начало, день день, делиться врач, думать сходить, начало заниматься начало, не отслеживать (12) | | |

*Таблица П4*

**Метрики модели 1 Guided BERTopic (биграммы и триграммы)**

| | min_ cluster_ size | n_ neighbors | Max_ features | min_ df | Число тем | Шум, % | Силуэт | Разнообразие тем |
|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 15 | 2000 | 2 | 23 | 17.3 | 0.7094 | 0.9304 |
| 2 | 11 | 15 | 2000 | 2 | 21 | 17.9 | 0.7207 | 0.9238 |
| 3 | 12 | 15 | 2000 | 2 | 20 | 19 | 0.7196 | 0.9200 |

*Таблица П5*

**Метрики модели 2 Guided BERTopic (униграммы, биграммы и триграммы)**

| min_cluster_ size | n_ neighbors | Max_ features | min_df | Число тем | Шум, % | Силуэт | Разнообразие тем |
|---|---|---|---|---|---|---|---|
| 10 | 16 | 2000 | 2 | 21 | 13.8 | 0.7254 | 0.8238 |

## Текст промпта

Привет!

**1. Твоя роль**

Ты — социолог-методолог, специалист по качественному анализу и тематическому анализу (Braun & Clarke). Твоя задача — провести этап 4 тематического анализа: проверку и уточнение тем на основе данных, полученных с помощью тематического моделирования (BERT + UMAP + HDBSCAN) интервью о практиках заботы о здоровье в российских семьях. У тебя есть доступ к предварительным темам (результаты ML), примерам реплик и гайду. Ты должен критически оценить и доработать предварительные темы, а затем комплексно описать весь массив данных, обязательно выделив:

– типологии взаимодействия с устройствами мониторинга здоровья,
– типологии координации практик в семье,
– условия возникновения практик,
– следствия практик для семьи и отдельных членов.

**2. Исходные данные (пользователь предоставит файлы)**

Файлы с описанием тем из двух моделей:
Модель 1: только биграммы и триграммы.
Модель 2: униграммы + биграммы + триграммы.
Каждая тема содержит номер кластера, топ n-грамм.
Файлы с примерами реплик к каждой теме (несколько цитат).
Гайд интервью (опросник).
Статья о тематическом анализе (Braun & Clarke, желательно этап 4).

**3. Контекст исследования (пользователь уже сформулировала)**

Предмет: коллективные практики заботы о здоровье в повседневной жизни российских семей (цифровой и нецифровой мониторинг).

Цель: описать коллективные практики, выявить стратегии.

Задачи:

Описать цифровые и нецифровые практики и их встроенность в семейную рутину.

Описать и классифицировать форматы взаимодействия с цифровыми устройствами мониторинга здоровья.

Выявить условия возникновения практик и их координацию в семье.

**4. Задачи для тебя (выполни последовательно)**

Шаг 1. Ознакомление

Прочитай гайд интервью, чтобы понять логику сбора данных.

Изучи предоставленную статью о тематическом анализе, особенно раздел о проверке и уточнении тем (reviewing themes).

Просмотри все предварительные темы из обеих моделей и примеры реплик.

Шаг 2. Проверка предварительных тем

Для каждой машинной темы (кластера):

– вернись к примерам реплик и прочитай несколько высказываний, попавших в эту тему:

– оцени когерентность: действительно ли все реплики объединены общим смыслом? Если нет — раздели тему на две или более;

– оцени границы: не пересекается ли эта тема с другой? Если да — объедини;

– оцени содержательную насыщенность: если тема содержит мало реплик или они тривиальны (например, «да», «нет», «не знаю»), можешь исключить ее.

Выявление условий и следствий (новое): на основе предоставленных примеров реплик для каждой практики (описываемой в теме) определи:

– условия возникновения: какие внешние или внутренние факторы (например, болезнь члена семьи, рекомендация врача, удобство приложения, финансовые ограничения, нехватка времени) запускают или усиливают данную практику;

– следствия практики для семьи в целом и для отдельных членов семьи: как меняется семейная рутина, эмоциональный климат, распределение обязанностей, здоровье, конфликты или сплоченность.

Запиши эти наблюдения как отдельные коды-заметки (например, «условие: хроническое заболевание ребенка», «следствие: снижение тревожности у матери»). Эти коды должны быть позже интегрированы в темы и типологии.

Шаг 3. Уточнение и переименование тем

Присвой каждой уточненной теме четкое название, отражающее суть (например, «Поиск медицинской информации в интернете до визита к врачу», «Давление старшего поколения в вопросах лечения», «Фитнес-браслет как средство самоконтроля»).

Напиши краткое определение темы (что именно она охватывает).

Укажи явно условия и следствия: для каждой темы опиши, какие условия способствуют возникновению данной практики и какие следствия (позитивные/негативные, для семьи или отдельных членов) она порождает.

Приведи 1–3 характерные цитаты.

Укажи, из каких исходных машинных тем (номера кластеров) и из каких интервью получена тема.

Шаг 4. Группировка тем в более крупные блоки

Сгруппируй уточненные темы в 5–8 тематических блоков. Обязательно включи следующие блоки (остальные на твое усмотрение):

– использование цифровых / нецифровых устройств / приложений для мониторинга здоровья;
– семейная координация и распределение ролей;
– эмоциональные аспекты и конфликты;
– условия возникновения практик заботы о здоровье;
– следствия практик для семьи и отдельных членов

Шаг 5. Выделение типологий

5.1. Типология взаимодействия с цифровыми устройствами мониторинга здоровья

На основе уточненных тем, относящихся к использованию устройств (фитнес-браслеты, смарт-часы, умные весы, приложения для отслеживания симптомов и т. д.), выдели 4–6 типов взаимодействия. Для каждого типа опиши: название, характерные действия и установки пользователей, примеры цитат.

5.2. Типология координации практик заботы о здоровье в семье

На основе тем, описывающих распределение ответственности, совместные действия, конфликты и согласования, выдели 3–5 типов координации (например, «единый центр», «распределенная координация» и др.) Для каждого типа приведи цитаты и объясни, при каких условиях он возникает.

5.3. Типология условий возникновения практик (новое)

Выдели 3–5 типов условий, которые запускают или усиливают практики заботы о здоровье (например, «острый эпизод болезни», «социальное давление / пример окружения», «предписание врача» и др.). Для каждого типа приведи примеры цитат и укажи, к каким практикам они чаще всего приводят.

5.4. Типология следствий практик (новое)

Выдели 3–5 типов следствий, которые возникают в результате внедрения практик (например, «укрепление внутрисемейных связей», «появление новых конфликтов», «снижение тревожности», «усталость от контроля», «изменение семейной рутины»). Для каждого типа приведи цитаты и укажи, какие практики их вызывают.

Шаг 6. Комплексное описание данных

Напиши связный текст (1,5–2 страницы), который:

– кратко представляет контекст и методологию;
– описывает основные тематические блоки (из шага 4) с примерами;
– отвечает на исследовательские задачи:

Какие цифровые и нецифровые практики выявлены, как они встроены в рутину? Какие типы взаимодействия с устройствами выделены? Как условия влияют на возникновение и трансформацию практик? Какие следствия практик заботы о здоровье испытывают семьи и отдельные члены семьи? Как происходит координация практик в семье?



Завершается выводами о природе коллективных практик заботы о здоровье в российских семьях. В заключении укажи, как выделенные условия и следствия интегрируются в общую теоретическую модель.

**5. Требования к формату вывода**

Представь результат в виде структурированного документа на русском языке, содержащего:

Введение (цели, данные, метод).

Уточненные темы (список с определениями, условиями, следствиями, цитатами).

Группировку в тематические блоки (включая блоки условий и следствий).

Типологию взаимодействия с устройствами.

Типологию координации практик в семье.

Условия возникновения практик.

Следствия практик для семьи и отдельных членов.

Комплексное описание данных (связный текст).

**6. Дополнительные указания**

Опирайся строго на предоставленные файлы: примеры реплик для цитирования, гайд для проверки полноты.

Если какая-то тема не подтверждается (оказалась артефактом модели), отбрось ее.

При выделении типологий старайся, чтобы они были эмпирически обоснованы (каждый тип иллюстрируется цитатами из разных интервью).

Будь критична: если предварительная теоретическая схема не полностью подтверждается, честно напиши об этом.

Особое внимание удели интеграции условий и следствий в каждую тему и в итоговые типологии — они не должны быть просто отдельным списком, а должны быть встроены в описание феноменов.