

Система поддержки критической оценки новостей с использованием больших
языковых моделей

Содержание

ВВЕДЕНИЕ	4
Глава 1. Теоретические аспекты поддержки критической оценки новостей с использованием больших языковых моделей	7
1.1 Место и роль искусственного интеллекта в создании и в оценке новостей пользователями	8
1.2 Обзор языковых моделей и методов работы с текстами	12
1.3 Постановка задачи исследования.....	23
Выводы Глава 1	28
Глава 2. Моделирование системы поддержки критической оценки новостей: признаки, данные, модели	29
2.1 Логические и манипулятивные конструкции в новостях: выделение признаков для интеллектуального анализа	29
2.2 Релевантные модели, методы и наборы данных для задачи исследования.....	32
2.2.1 Тестирование методов Few-shot learning, One Shot learning из этапа In Context Learning (ICL) для задачи исследования.....	32
2.2.2 Анализ методов Pretraining и Transfer Learning для задачи исследования	34
2.3 Создание синтетического набора данных для системы поддержки критической оценки новостей.....	36
Выводы Глава 2	41
Глава 3 Проектирование системы поддержки критической оценки новостей с использованием больших языковых моделей	42
3.1 Архитектура системы поддержки критической оценки новостей.....	43
3.2 Оценка предобученных языковых моделей на выявление логических ошибок: эксперимент на корпусе REDDIT	47
3.3 Эксперименты по обучению модели GLiNER для NER-распознавания элементов манипулятивной риторики	51
Выводы Глава 3	55
ЗАКЛЮЧЕНИЕ.....	56
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	59
ПРИЛОЖЕНИЕ А	66
ПРИЛОЖЕНИЕ Б	76
ПРИЛОЖЕНИЕ В.....	88

ВВЕДЕНИЕ

Актуальность. В современном обществе информационная и когнитивная безопасность играют ключевую роль в формировании общественного мнения, экономических процессов и политических решений. Актуальность задачи поддержки критической оценки новостей обусловлена тем, что распространение ложной или манипулятивной информации приводит к этим негативным последствиям. Например, в апреле 2025 года на фоне ложных сообщений о приостановке тарифов США на Китай рынок акций пережил внезапный всплеск и обвал за считанные минуты. Индекс S&P 500 вырос на 8,5 %, затем моментально откатился, вызвав флуктуации в размере триллионов долларов [1]. Таких примеров много, и в условиях нарастающей информационной перегрузки [2], когда пользователи ежедневно сталкиваются с десятками, а то и сотнями новостных сообщений, способность критически оценивать поступающую информацию становится не просто полезным навыком, а необходимым элементом цифровой гигиены, информационной и когнитивной безопасности. Повышение медиаграмотности населения, развитие навыков распознавания когнитивных искажений, манипулятивных приёмов, эмоциональных уловок и логических ошибок становится важным направлением в обеспечении устойчивости общества к информационным атакам. Поэтому тема диссертации представляется актуальной.

Несмотря на общественное внимание к проблемам фейковых новостей [3], информационных пузырей [4], недостаточному уровню медиаграмотности [5] и осведомленности общества, большинство технических решений сосредоточены либо на проверке фактов, источников, путей распространения информации, либо на фильтрации контента или его авторов, не затрагивая когнитивную сторону восприятия информации пользователями. При этом существующие системы поддержки медиаграмотности, как правило, не адаптируются под уровень подготовки конкретного пользователя и не учитывают разнообразие манипулятивных техник, используемых в новостях. Во-первых, обучающие курсы часто реализованы в виде скриптовых игр [6, 7], не интегрированных в реальную среду потребления новостей. Во-вторых, выводы систем автоматического анализа новостей в государстве чаще всего не доступны простым пользователям и даже если в какой-то момент происходит зачистка новостей со страницы СМИ, то пользователю об этом не сообщают, как и причину удаления. В-третьих, в условиях стремительного роста объёмов и скорости распространения новостей, ручная проверка и обучение становятся неэффективными без инструментов, использующих возможности современных языковых моделей. Таким образом, возникает необходимость в системах, которые бы одновременно

обеспечивали автоматизированную, но интерпретируемую подсветку потенциально манипулятивных участков текста, адаптацию глубины объяснений в зависимости от уровня подготовки пользователя, были бы доступны пользователю как в режиме обучения, так и в режиме подсказок.

Цель диссертационного исследования – разработка и экспериментальное обоснование архитектуры системы поддержки критической оценки новостных текстов на основе предобученных языковых моделей.

Объектом исследования являются тексты новостей, манипулятивные техники в аргументации, языковые модели.

Предметом исследования являются подходы выявления манипулятивных признаков в новостных текстах с использованием предобученных языковых моделей.

Для достижения поставленной цели в выпускной квалификационной работе необходимо выполнить следующие **задачи**:

1. Провести обзор современного состояния исследований, проанализировать существующие языковые модели и методы обработки текстов, применимые для задач критической оценки новостей.
2. Выделить релевантные признаки манипулятивного воздействия в новостных текстах и создать синтетический набор данных.
3. Разработать архитектуру системы поддержки критической оценки новостей.
4. Протестировать существующие методы обнаружения логических ошибок на общедоступном наборе данных.
5. Проверить гипотезу о том, что триггеры логических убеждений, могут быть эффективно интерпретированы как именованные сущности и автоматически распознаны при помощи современных NER-моделей.

Научная задача. Разработка системы поддержки критической оценки новостей, обеспечивающей обучение и подсветку фрагментов новостей, с учётом когнитивного профиля пользователя.

Теоретическая значимость работы заключается в развитии подходов к построению интеллектуальных систем поддержки пользователей при взаимодействии с новостным контентом. В рамках исследования представлена формализация задачи поддержки критической оценки новостей с учётом различных когнитивных подходов (логический анализ, фактчекинг, анализ тональности). Систематизированы и классифицированы современные методы NLP, применимые к задачам критической оценки новостей. Кроме того, теоретическая значимость работы заключается в формулировании и обосновании гипотезы о возможности использования методов распознавания именованных

сущностей (NER) для выявления в текстах новостей лексических триггеров, характерных для логических убеждений

Практическая значимость работы состоит в разработке и экспериментальной проверке архитектуры системы, обеспечивающей подсветку и, при необходимости, интерпретацию новостных текстов на предмет наличия в них манипулятивных техник для поддержки критического мышления пользователя. Система использует возможности предобученных языковых моделей и может быть внедрена (1) в составе образовательных платформ, направленных на развитие медиаграмотности; (2) как вспомогательный модуль при чтении новостей в браузере или мессенджере; (3) в системах повышения цифровой гигиены и информационной безопасности пользователей. Результаты могут быть использованы при разработке интеллектуальных интерфейсов, рекомендательных систем и платформ медиаграмотности или повышения осведомленности.

Апробация результатов работы. Основные положения и результаты диссертационной работы были представлены на 65-й Научно-технической конференции профессорско-преподавательского состава, научных работников и аспирантов Санкт-Петербургского государственного университета телекоммуникаций им. проф. М. А. Бонч-Бруевича (17-21 февраля 2025) [8].

Личный вклад. Все результаты, представленные в диссертационной работе, получены лично автором в процессе выполнения научно-исследовательской деятельности.

Структура и объем диссертационной работы. Диссертационная работа включает введение, три главы, заключение, список использованных источников (83 наименования) и 3-х приложений. Объем работы – 65 страниц машинописного текста; включая 7 рисунков и 10 таблиц, 3 листинга.

Глава 1. Теоретические аспекты поддержки критической оценки новостей с использованием больших языковых моделей

Развитие искусственного интеллекта (ИИ) за последние пять лет стало одним из наиболее значимых технологических достижений, оказывающих глубокое воздействие на науку, экономику, культуру и повседневную жизнь. Вокруг больших языковых моделей и генеративного искусственного интеллекта возникла волна общественного интереса и дискуссий. В 2024 году и в первом квартале 2025 года только в России прошло значительное число конференций, на которых обсуждали ИИ с разных сторон.

В данной магистерской работе автор не преследует цели решить все проблемы безопасного взаимодействия современного человека с информационным пространством – это более широкая тема, место которой, вероятно, в научном исследовании другого уровня. Однако для начала можно выделить один аспект, поддающийся исследованию и анализу в рамках текущей работы. В фокусе внимания данной выпускной квалификационной работы находится задача повышения уровня критического мышления пользователя при взаимодействии с новостным контентом через смартфон, компьютер, за рамками исследования остаются телевидение, радио и так далее.

Под уровнем критического мышления в диссертации понимается совокупность навыков проверки достоверности новостей, а также выделения из новостных текстов логических манипуляций, убеждений, логических ошибок и эмоциональной тональности. При этом следует отметить, что сам подход к эмпирической или иной формальной шкале измерения уровня критического мышления, а также вариативность его значений в зависимости от социальных групп пользователей, в рамках текущего исследования не разрабатывается и является предметом дальнейших научных работ автора.

Инструментом для решения этой задачи выступают языковые модели, а на системном уровне приоритет отдается большим языковым моделям в силу выбранной темы исследования.

В дополнение к использованию языковых моделей для повышения уровня критического мышления пользователей при взаимодействии с новостным контентом, следует отметить существование альтернативных подходов, основанных на вероятностных моделях. Так, П.С. Краснощеков предложил вероятностную модель, в которой вероятность совершения человеком определённого поступка связывается с его мнением относительно рассматриваемой проблемы [9]. Эта модель, близкая по идеологии к моделям ДеГрута [10] и Фридкина-Джонсона [11], которые получили продолжение в современных исследованиях [12], вводит понятие конформизма – изменения человеком своих взглядов в результате

взаимодействия с окружающим коллективом, что фактически отражает изменение уровня критического мышления. Дальнейшее развитие этой модели представлено в работах И.В. Козицина [13], который обобщил модель Краснощекова на случай систем с разложимой структурой социальных связей. В частности, он изучил ситуации, при которых обособленные группы людей, не знакомые с остальными участниками сообщества, взаимодействуют в рамках разложимой матрицы социальных связей. В своих исследованиях Козицин И.В. установил формальную связь между моделями Краснощекова П.С. и ДеГрута, а также провёл валидацию своих моделей на данных социальной сети «ВКонтакте» [14,15].

Таким образом, модели, учитывающие социальные взаимодействия и конформизм, представляют собой перспективное направление для оценки и повышения уровня критического мышления пользователей. В рамках данной диссертации основное внимание уделяется применению языковых моделей. Однако в будущих исследованиях автор диссертации планирует более детальное изучение и применение вероятностных моделей для решения поставленных задач.

В первой главе будут рассмотрены проблемы развития технологий ИИ, затем будет представлен обзор языковых моделей и завершит главу постановка задачи исследования.

1.1 Место и роль искусственного интеллекта в создании и в оценке новостей пользователями

Первоначально ИИ разрабатывался как инструмент для автоматизации рутинных вычислений и обработки больших объемов информации, однако с начала XXI века наблюдается качественный скачок: интеллектуальные системы стали успешно решать задачи, которые ранее считались исключительно человеческими прерогативами. Тем не менее, фундаментальные идеи, связанные с возможностью имитации когнитивных функций человека, были заложены уже на заре исследований ИИ. Рассмотрим рисунок 1.

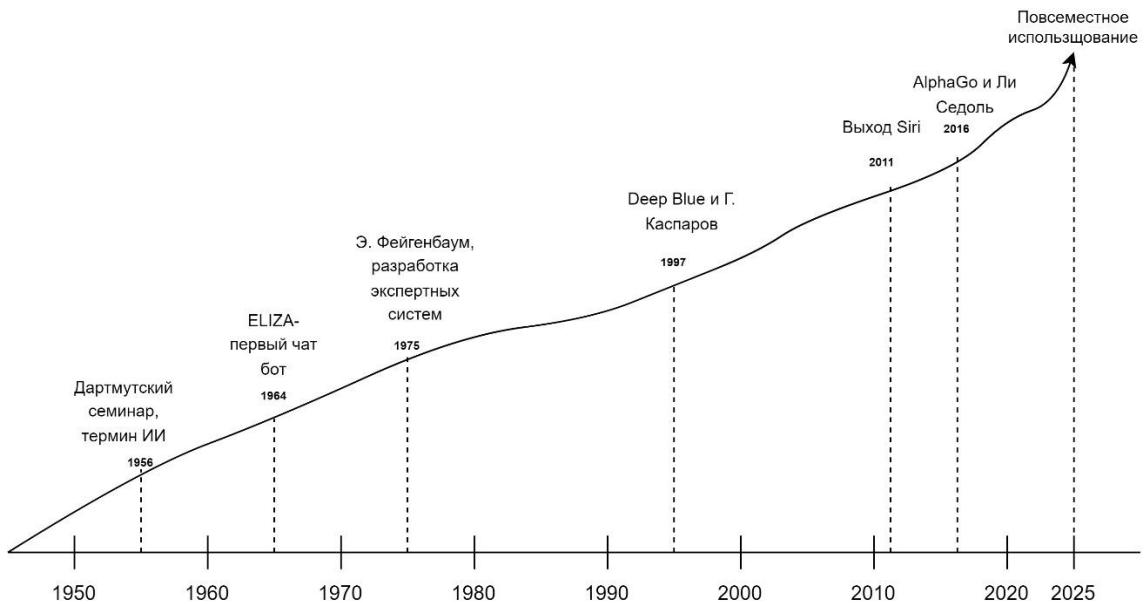


Рисунок 1 – Развитие технологий искусственного интеллекта.

Термин «искусственный интеллект» (artificial intelligence) был впервые официально использован в 1956 году на Дартмутском летнем семинаре, организованном Дж. Маккарти, М. Минским, К. Шенном и Н. Рочестером [16]. В заявке семинара утверждалось, что «все аспекты обучения и любые другие признаки интеллекта могут быть настолько точно описаны, что машину можно будет научить имитировать их» [17, 18]. Семинар обозначил начало ИИ как самостоятельной научной дисциплины и определил исследовательскую повестку, связанную с моделированием когнитивных процессов средствами формальной логики и алгоритмического анализа.

Параллельно в Советском Союзе велись исследования в смежных направлениях – теории алгоритмов, математической логике, кибернетике. Работы М.И. Бонгарда, С.Ю. Маслова, А.А. Маркова, П.П. Смирнова-Троянского [19] и других заложили основы формализации понятий вычислимости и логического вывода. Развивались концепции алгоритмической логики и формальных систем. В 1960-е годы в рамках кибернетики проводились исследования процессов управления и обработки информации. Существенный вклад в развитие автоматизации логических рассуждений и формализации знаний внесли А. М. Ляпунов [20], В. М. Глушков [21]. Несмотря на ограничение, которые накладывала государственная система на возможные международные контакты, отечественная школа оказала влияние на формирование теоретических основ ИИ.

В дальнейшем одной из первых реализованных попыток симуляции человеческой коммуникации стала программа ELIZA, разработанная Джозефом Вейценбаумом в Массачусетском технологическом институте в середине 1960-х годов [22]. Используя

простые шаблонные правила обработки текста, ELIZA имитировала работу психотерапевта, подставляя фразы из пользовательских сообщений в заранее заданные речевые конструкции. Несмотря на свою техническую примитивность, ELIZA произвела сильное впечатление на пользователей, многие из которых приписывали ей способность к пониманию. Парадоксальным образом именно эта иллюзия «понимающего ИИ» заставила самого Вейценбаума впоследствии выступать с критикой безоглядного доверия к машинным системам – он подчеркивал, что поверхностная симуляция диалога не тождественна мышлению [23]. Сейчас мы сталкиваемся с проблемой, когда множество людей на планете очеловечивают модели, не воспринимают их только как инструменты. Постепенно как-то общество перевалило за порог отождествления диалога мышлению и часть ученых бьет тревогу.

Следующим значимым этапом развития ИИ стали экспертные системы, получившие распространение в 1970–1980-х годах. Одним из пионеров в этой области был Эдвард Фейгенбаум, который разработал ряд систем, имитирующих решение задач в узких профессиональных доменах – от медицины (например, MYCIN для диагностики бактериальных инфекций [24]) до молекулярной биологии. Экспертные системы опирались на базы знаний, содержащие факты и правила, и использовали механизмы логического вывода для генерации рекомендаций [25]. Хотя такие системы демонстрировали впечатляющие результаты в строго ограниченных задачах, они оказались уязвимы к контекстуальным и языковым нюансам, что ограничило их применение.

Очень символическим моментом в процессе трансформации взаимодействия человека и ИИ стало поражение чемпиона мира по шахматам Гарри Каспарова в матче против суперкомпьютера IBM Deep Blue в 1997 году [26]. Этот случай был воспринят общественностью как первое убедительное доказательство того, что машина может превзойти человека в строго формализованной интеллектуальной игре. Спустя почти два десятилетия, в 2016 году, компания DeepMind (принадлежащая Google) представила AlphaGo – программу, победившую сильнейшего игрока в го, Ли Седоля [27]. Игра го долгое время считалась последним бастионом человеческой интуиции и стратегического мышления.

Эти события стали не только вехами в истории ИИ, но и отправной точкой для более широкого его применения в сферах, напрямую связанных с пониманием, интерпретацией и генерацией информации. В 2018 году система BERT от Google показала результаты, сравнимые с человеческими, в бенчмарке SQuAD 2.0 (Stanford Question Answering Dataset) [28] – тесте на понимание прочитанного, включающем как прямые вопросы, так и вопросы с отсутствующими ответами, требующие оценки текста [29]. К 2021 году многие

архитектуры достигли даже лучших результатов, превзойдя средние показатели человека [30]. Таким образом, ИИ стал не просто инструментом классификации и поиска, но полноценным участником когнитивного процесса, способным к анализу, дедукции и аргументации.

Тревожной тенденцией последних лет стало использование ИИ для генерации дезинформации. Согласно исследованию [31], ответы GPT-3 на сложные вопросы по химии сравнимы или превосходят специализированные модели, обученные на химических данных. В другом исследовании [32] языковые модели точнее распознавали дипфейки на арабском языке, чем участвовавшие в эксперименте 8 человек. В мае 2025 года исследователи Цюрихского университета тайно разместили более 1000 ИИ-генерированных комментариев на Reddit [33]. Комментарии ИИ оказались очень правдоподобными и убедительными. Сам эксперимент вызвал этическую критику: участников не предупредили, а модераторы потребовали извинений. Несмотря на негативную реакцию, эксперимент подтвердил тезис, что ИИ способен убеждать человека, создавать идеальные дипфейки, а человек не способен эффективно их отличать. Причина – применение риторических приёмов: апелляции к эмоциям, ложных дилемм, логических ошибок. Именно на анализе таких техник и будет сделан акцент в данной диссертации для повышения критической оценки новостей пользователями.

Одновременно с этим стоит отметить, что наблюдается снижение способности людей критически осмысливать информацию. Согласно статье, опубликованной на портале Stanford graduate school education в 2021, исследователи поручили 3446 американским старшеклассникам проверить новостные статьи и другой цифровой контент. Исследование показало, что школьники и студенты совершенно не умеют распознавать фейковые новости в интернете. [34]. В тоже время исследование Pew Research Center (2022) показало, что среди журналистов около четверти (26%) говорят, что они по незнанию освещали историю, которая, как позже выяснилось, содержала ложную информацию [35].

К 2025 году эта тенденция усиливается на фоне информационной перегрузки, фрагментации медийного пространства и постепенного замещения рационального анализа эмоциональными и поверхностными оценками. Под информационной перегрузкой в данной работе понимается состояние, при котором объём поступающей информации превышает когнитивные возможности индивида по её обработке, что затрудняет принятие решений и критическую оценку содержания. Это явление было впервые описано Бертрамом Гроссом [2] и популяризировано Элвином Тоффлером в книге «Шок будущего» [36], где он охарактеризовал его как сенсорную перегрузку в условиях информационной эпохи.

Эволюция технологий ИИ от ELIZA до GPT-4 выявила не только рост когнитивных возможностей машин, но и усиление их способности к убеждению и манипуляции. Современные модели эффективно имитируют достоверность и используют риторические приёмы, что делает их мощным инструментом управления общественным мнением. Одновременно снижается способность человека распознавать дезинформацию, особенно созданную ИИ. Это формирует новый тип когнитивного неравенства – между человеком и технологиями ИИ.

Настоящее исследование направлено на разработку системы поддержки критической оценки новостей, в которой используются языковые модели для выявления признаков манипулятивного контента (оценка достоверности выходит за рамки исследования). В фокусе работы находится феномен уязвимости человека к когнитивному воздействию со стороны ИИ-сгенерированной информации. Под пользователем в диссертации понимается человек, регулярно взаимодействующий с информационными ресурсами через смартфоны и компьютеры. Цель работы заключается в создании алгоритмов и инструментов, которые могут быть использованы в системе поддержки критической оценки новостей на основе больших языковых моделей. При этом формирование и измерение уровней критического мышления пользователей выносится за рамки настоящей диссертации и планируется к изучению в дальнейших исследованиях автора.

1.2 Обзор языковых моделей и методов работы с текстами

В данной выпускной квалификационной работе акцент в обзоре будет сделан на методах NLP, на языковых моделях, которые используются для анализа новостного контента, тональности текстов, для обнаружения дезинформации, фейков. Однако цель описать все множество существующих моделей и алгоритмов автор не преследует. В диссертации известные подходы рассматриваются как релевантные методы для работы с новостным контентом.

В начале рассмотрим модели представления слов. Word2Vec – группа алгоритмов, преобразующих слова из текстового корпуса в компактные векторные представления. Модель была предложена в 2013 году Т. Миколовым и коллегами из корпорации Google [37]. Векторизация слов (word embeddings) стала основой современных методов обработки естественного языка (NLP). Ранее в NLP использовались дискретные символьные представления (one-hot encoding), не отражавшие семантической близости между словами. Word2Vec стал важной вехой в развитии машинного обучения. В Word2Vec можно выделить 2 модели: (1) Skip-gram и (2) CBOW.

Согласно описанию, данному авторами в работе [37], цель обучения модели Skip-gram состоит в определении векторных представлений слов, обладающих свойством предсказуемости их окружения в пределах предложений или текстовых документов. Формально, пусть дана последовательность обучающих слов $w_1, w_2, w_3, \dots, w_T$, тогда цель обучения модели заключается в максимизации средней логарифмической условной вероятности предсказания слов из контекста при заданном целевом слове, как показано в формуле (1):

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

где c определяет размер контекстного окна (он может быть функцией от центрального слова w_T). Увеличение параметра c приводит к формированию большего количества обучающих примеров, что потенциально улучшает качество представлений, но увеличивает затраты на вычисления. Базовая модель Skip-gram вычисляет $p(w_{t+j} | w_t)$ посредством функции softmax:

$$p(w_O | w_I) = \frac{\exp(v_{w_O}' v_{w_I})}{\sum_{w=1}^W \exp(v_w' v_{w_I})}$$

Где v_w и v'_w это входные и (input) и выходные (output) векторные представления слова w , а W – количество слов в словаре. Но такая формулировка оказывается вычислительно неэффективной, так как затраты на вычисление градиента пропорциональны количеству слов в словаре. Морин и Бенджио [38] предложили использовать иерархический softmax, который представляет собой вычислительно эффективное приближение полного. Его основное преимущество заключается в том, что вместо вычисления вероятностей для всех W выходных узлов нейросети достаточно обработать около $\log_2(W)$ узлов. В своей работе Миколов [37] ссылается на этот метод, но предлагает негативное семплирование (Negative Sampling, NEG), как альтернативу, которая упрощает обучение и делает его более эффективным:

$$\log \sigma(v_{w_O}' v_{w_I}) + \sum_{i=1}^k E_{w_i \sim P_n(w)} [\log \sigma(-v_{w_i}' v_{w_I})]$$

В формуле $P_n(w)$ – шумовое распределение, из которого выбираются отрицательные примеры. Также в работе Т. Миколова и др. используется методика субсемплирования для устранения дисбаланса между редкими и часто встречающимися словами. То есть в процессе обучения модели Skip-gram каждое слово w_i из обучающего набора данных исключается с определённой вероятностью, которая рассчитывается по следующей формуле:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

Где $f(w_i)$ – частота появления слова w_i в корпусе, t – установленный порог (обычно порядка 10^{-5}).

Т. Миколов с соавторами в том же году опубликовали работу, в которой предложили модель «непрерывный мешок слов» (Continuous Bag-of-Words, CBOW) [39]. CBOW предсказывает целевое слово, исходя из контекста окружающих слов. А модель Skip-gram предсказывает окружающие слова контекста, исходя из заданного целевого слова. Обе модели являются неглубокими нейронными сетями с одним скрытым слоем. Разработанные Т. Миколовым и его коллегами модели позволили значительно ускорить процесс обучения и улучшить качество получаемых векторных представлений.

Благодаря способности моделировать смысловые связи между словами, Word2Vec широко используется в анализе новостных текстов и других информационных потоков до сих пор [40,41]. Векторные представления слов помогают выявлять ключевые темы уже на множестве языков мира, анализировать тональность статей, а также предсказывать связь между новостными событиями, выявлять сарказм, токсичные комментарии.

Альтернативными моделями представления слов стали GloVe и FastText. GloVe (Global Vectors for Word Representation) – модель, обучения векторных представлений слов, предложенный Джейфри Пеннингтоном, Ричардом Сочером и Кристофером Маннингом в 2014 году [42]. Как и статьи Т. Миколова с соавторами эта статья цитируется более 45 000 раз. В отличие от моделей Word2Vec, которые используют локальный контекст слов, GloVe создает матрицу соотношений на основании анализа частоты совместного появления слов в корпусе. При этом обучение модели основано на минимизации функции потерь, которая представлена следующим уравнением:

$$J = \sum_{i,j=1}^W f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

Где X_{ij} – частота совместного появления слов i и j , w_i и \tilde{w}_j – векторные представления слова и его контекста, b_i и \tilde{b}_j – смещения (биасы) для слова и его контекста соответственно, $f(X_{ij})$ – весовая функция, регулирующая вклад различных частотных значений.

Дополнительный биас \tilde{b}_j авторы предположили использовать для восстановления симметрии модели, что позволило учитывать влияние каждого слова независимо от его роли. Это улучшило согласованность представлений и способствовало более точному учёту

глобальных статистических зависимостей между словами, таких как семантическая близость (например, «король» – «королева») и аналогии (например, «Париж» относится к «Франции», как «Берлин» – к «Германии»), которые извлекаются на основе анализа частот совместной встречаемости слов в корпусе [42].

В 2016 году в стремлении создать более гибкую и точную модель был предложен FastText. Он расширил идеи распределенных представлений. Его авторы Адам Жулен (Armand Joulin), Элуар Граф (Edouard Grave), Петр Бояновский (Piotr Bojanowski) и Томас Миколов (Tomas Mikolov) из Facebook AI Research (FAIR) [43]. Т. Миколов присоединился к FAIR в 2014.

В отличие от моделей Word2Vec и GloVe, FastText уже учитывал морфологию слов, используя n-граммы, что позволило модели эффективно работать с редкими словами и улучшать качество представлений. То есть каждое слово уже представлялось не только как единое целое, но и как набор его подстрок (например, «машина»: «ма», «аш», «ши», «ин», «на»). При этом FastText сохранял базовые принципы CBOW и Skip-gram, но вместо обучения модели только на отдельных словах, он включал уже и векторные представления их частей. Это позволило модели формировать осмысленные эмбеддинги даже для редких и ранее не встречавшихся слов, за счёт обобщения на уровне морфологических компонент. Таким образом, повышалась способность модели обрабатывать слова вне обучающего словаря, что критически важно при работе с малочастотной лексикой, неологизмами или орографическими вариантами. FastText использует модифицированную функцию потерь, аналогичную Skip-gram, но с учётом n-грамм:

$$J = - \left(\frac{1}{N} \right) \sum_{n=1}^N y_n \log(f(BAx_n))$$

Где x_n – представление документа (среднее по векторным представлениям слов и n-грамм), A – матрица эмбеддингов, B – матрица линейного преобразования, f – softmax-функция, y_n – метка класса.

FastText широко применяется для категоризации новостей и для обнаружения фейковых новостей [44,45]. Чаще всего, конечно же, в 2025 уже в сочетании с глубокими нейронными сетями, с трансформерами.

Несмотря на широкое применение и эффективность в построении векторных представлений слов, Word2Vec имеет ряд признанных ограничений. Во-первых, модели Skip-gram и CBOW анализируют локальный контекст слов, но не учитывают порядок слов и синтаксические зависимости, по сути, они игнорируют структуру предложений. Это приводит к потерям информации, особенно в сложных языковых конструкциях. Во-вторых,

Word2Vec присваивает каждому слову единственный вектор, независимо от его значения в разных контекстах. Например, слово «замок» в смысле «дом-крепость» и «приспособление для запирания чего-либо ключом» будет иметь один вектор, что снижает точность семантического анализа. В-третьих, все-таки есть ограничения для работы с редкими словами, хотя FastText частично решил эту проблему. В-четвертых, векторное представление слова фиксировано после обучения, это мешает моделированию динамических языковых изменений. Одна модель сейчас не пригодна для анализа текстов из 2010 года 21 века и из 2025, смыслы слов достаточно быстро меняются, вслед за трендами. В защиту Word2Vec можно сказать то, что как базовое решение для классификации текстов, особенно в информационной безопасности (анализ логов ОС, сетевого трафика (если трафик перевести в текст) он используется до сих пор, как достаточно легкое решение. Ограничения Word2Vec и прогресс в общем в NLP привели к развитию более сложных моделей, таких как трансформеры, способные учитывать глобальный контекст текста и генерировать динамические представления слов. В отличие от статических эмбеддингов, трансформеры используют механизмы внимания, позволяя анализировать взаимосвязи между словами и предсказывать значения в зависимости от контекста.

Рассмотрим эволюцию механизма внимания. Механизм внимания (Attention Mechanism) был впервые предложен в 2014 году исследователями Бахдановым, Чо и Бенгио в статье «Neural Machine Translation by Jointly Learning to Align and Translate» [46]. Они разработали этот подход для нейронного машинного перевода, чтобы модель могла динамически фокусироваться на разных частях входного текста при генерации перевода.

Основными идеями механизма внимания стали динамическое выделение значимых частей входных данных и выравнивание входных слов с выходными. Внимание реализуется через soft alignment, где каждому входному слову присваивается вес в зависимости от его значимости для текущего выходного слова. Этот процесс обеспечивается вычислением контекстного вектора c_i для каждого целевого слова y_i . Вектор контекста зависит от $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$, на которые кодировщик сопоставляет входное предложение. Каждая аннотация h_i содержит информацию обо всей входной последовательности с особым вниманием к частям, окружающим i -е слово входной последовательности. Контекстный вектор вычисляется как взвешенная сумма этих аннотаций h_i [46].

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

Вес α_{ij} каждой аннотации h_j вычисляется следующим образом:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

Где $e_{ij} = a(s_{i-1}, h_j)$ это модель внимания, которая оценивает, насколько хорошо совпадают входные данные в позиции j и выходные данные в позиции i . Оценка, в статье у Бахданова, Чо и Бенгио основана на скрытом состоянии в RNN s_{i-1} непосредственно перед выдачей y_i , вычисляемого по формуле $p(y_i|y_1, \dots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i)$, и s_i – это скрытое состояние RNN для времени i , вычисляемое по формуле $s_i = f(s_{i-1}, y_{i-1}, c_i)$. Модель внимания также опирается на j -ю аннотацию h_j входного предложения [46].

Механизм внимания стал ключевым компонентом многих моделей обработки естественного языка (NLP), включая трансформеры, предложенные в 2017 году в статье «Attention Is All You Need» исследователями из Google Brain и Университета Торонто [47]. На данный момент, наверное, это самая цитируемая статья в DeepLearning, по статистике академии Google в мае 2025 у нее более 178 000 цитирований, но сюда не попадают все те упоминания и разборы на курсах по глубокому машинному переводу, которые проходят в разных университетах. Упоминаниях в выступлениях, роликах, подкастах и т.д.

Авторы [47] предложили новую архитектуру – трансформер. Основой архитектуры стали два ключевых нововведения: масштабированное скалярное произведение внимания (Scaled Dot-Product Attention) и многоголовое внимание (Multi-Head Attention) (рис 2).

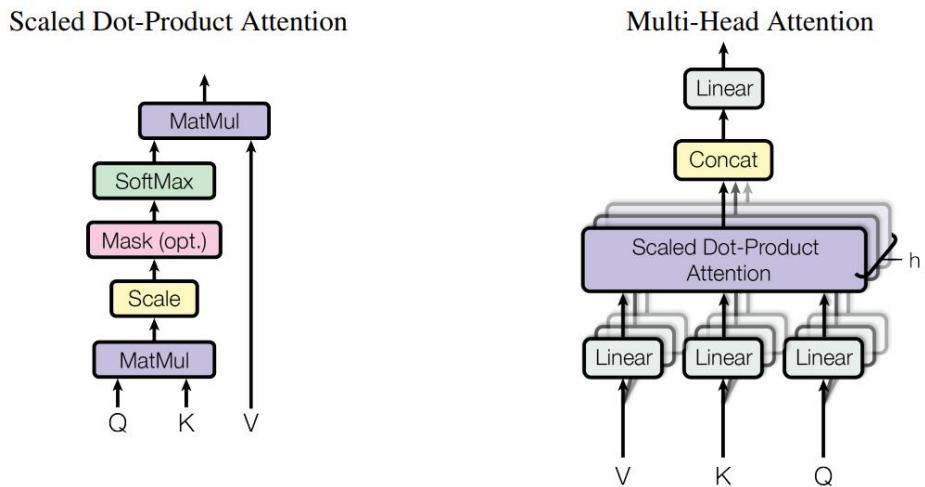


Рисунок 2 – Слева – это масштабированное внимание (Scaled Dot-Product Attention) на основе скалярного произведения. Справа – MultiHead attention, состоящее из нескольких параллельно работающих слоев внимания [47].

Масштабированное внимание вычисляется по следующей формуле:

$$Attention(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

где Q – матрица запросов, K – ключей, V – значений, а d_k – размерность пространства ключей. Деление на $\sqrt{d_k}$ стабилизирует градиенты softmax-функции при больших значениях размерности и предотвращает резкое падение градиентов, наблюдаемое в механизме внимания на основе скалярного произведения без масштабирования, в котором степень важности (вес) определяется как скалярное произведение между вектором запроса и каждым вектором-ключом, с последующим применением softmax.

Также авторы предложили вместо того, чтобы выполнять одну функцию внимания линейно проецировать запросы, ключи и значения h раз с различными обученными линейными проекциями в размерности d_k , d_v соответственно. Операции выполняются параллельно (рис 2, справа).

Механизм MultiHead attention (многоголовое внимание) позволяет модели одновременно фокусироваться на разных аспектах входной информации. Вместо одной операции внимания выполняются h параллельных, каждая со своими обучаемыми проекциями:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O,$$

где $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, W_i^Q , W_i^K , W_i^V – матрицы линейной проекции, W^O – объединяющая проекция после конкатенации.

В архитектуре Трансформер MultiHead attention применяется в трёх ключевых контекстах. Во-первых, в слоях внимания между энкодером и декодером (encoder-decoder attention) запросы формируются из декодера, а ключи и значения – из энкодера, что позволяет каждому токену выходной последовательности учитывать всю входную последовательность. Это повторяет механизм внимания из традиционных sequence-to-sequence моделей и обеспечивает гибкую фокусировку на релевантных входных элементах. Во-вторых, в слоях самовнимания энкодера все три компонента – запросы, ключи и значения – формируются из выходов предыдущего слоя энкодера, позволяя каждой позиции учитывать всю входную последовательность без учета её порядка. В-третьих, в слоях самовнимания декодера реализуется маскирование будущих позиций: каждая позиция может обращаться только к предыдущим или текущим токенам, что сохраняет авторегressiveный принцип генерации. Такое разделение внимания обеспечивает параллельность вычислений, моделирование длинных зависимостей и высокую эффективность архитектуры.

Одной из ключевых особенностей архитектуры трансформер является ее способность эффективно моделировать зависимости между токенами, находящимися на произвольном расстоянии друг от друга, за постоянное число шагов. В отличие от RNN, где информация передаётся последовательно и глубина распространения градиента пропорциональна длине входной последовательности, механизм самовнимания (self-attention) обеспечил доступ к любому элементу входа за один шаг вычисления. Это не только существенно сократило пути распространения сигнала, но и позволило реализовать высокую степень параллелизма в обработке последовательностей.

Однако отказ от рекуррентной структуры привел к тому, что модель не получала информации о порядке токенов «по умолчанию». Чтобы компенсировать отсутствие встроенной временной структуры, авторы ввели механизм позиционного кодирования (positional encoding), который добавляется к входным эмбеддингам на уровне как энкодера, так и декодера. Авторы предложили использовать фиксированные функции на основе тригонометрических функций ($\sin(x)$, $\cos(x)$). Такой подход позволил явно закодировать позиционную информацию в векторах, не увеличивая количество обучаемых параметров.

Предложенная архитектура трансформера показала выдающиеся результаты на задачах машинного перевода, достигнув BLEU-оценки 28.4 на задаче перевода с английского на немецкий (WMT 2014 [48]), превзойдя предшествующие модели, включая ансамбли [47]. Авторы показали, что благодаря полной параллелизации, трансформер может быть обучен за значительно меньшее время, чем RNN или CNN модели.

Современные подходы к распознаванию текстов, дипфейков, тональности в новостях опираются на предобученные языковые модели, которые решают более сложные задачи, такие как Masked Language Modeling (MLM) в BERT, Autoregressive Language Modeling в GPT или Span Corruption в T5. Эти задачи требуют от модели учитывать широкий контекст и создавать динамические, контекстно-зависимые эмбеддинги, что позволяет «понимать» структуру и смысл текста на более глубоком уровне [49].

Произошёл качественный переход от фиксированных, контекстно-независимых эмбеддингов к контекстно- зависимым представлениям, охватывающим всю структуру предложения. И предобученные языковые модели, уже обладающие глубоким знанием языка или многих языков одновременно. Это стало фундаментом для развития производных от таких архитектур, как BERT, GPT.

Рассмотрим архитектуру моделей BERT (Bidirectional Encoder Representations from Transformers). Эту модель предложили исследователи из Google в 2018 году [50], и она представляет собой трансформер, состоящий только из энкодерной части.

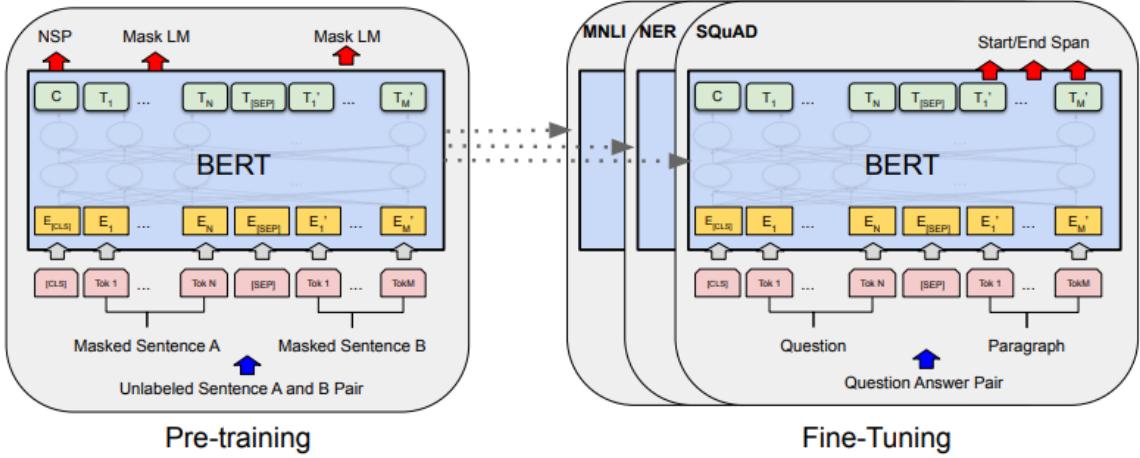


Рисунок 3 – Общая подготовка и отработка процедур для BERT.

BERT – это многослойный двунаправленный трансформенный энкодер, который доступен в двух вариантах:

- 1) BERT-BASE – 12 слоев, размер скрытого слоя 768, число self-attention – 12, общее количество параметров 110M.
- 2) BERT-LARGE – 24 слоя, размер скрытого слоя 1024, число self-attention – 16, общее количество параметров 340 миллионов.

BERT использует WordPiece [51] в качестве алгоритма сегментации слов на части со словарём из 30 000 токенов. Первый токен каждой последовательности – это специальный классификационный токен [CLS]. Финальное скрытое состояние, соответствующее этому токену, используется как агрегированное представление всей последовательности для задач классификации. Пары предложений объединяются в единую последовательность.

Чтобы различать два предложения в одной последовательности, в BERT применяется двойной механизм: (1) разделение специальным токеном [SEP], который обозначает границу между предложениями; (2) добавление обучаемого эмбеддинга к каждому токену, указывающему, принадлежит ли он предложению A или предложению B.

Входной эмбеддинг обозначается как E . Финальный скрытый вектор для [CLS] обозначается как $C \in \mathbb{R}^H$. Финальный скрытый вектор для i -го входного токена обозначается как $T_i \in \mathbb{R}^H$. Таким образом, входное представление BERT включает сумму трёх компонент: (1) эмбеддинг токена (WordPiece); (2) эмбеддинг сегмента (Sentence A или Sentence B); (3) эмбеддинг позиции (позиция токена в последовательности).

В отличие других моделей, BERT обучался с помощью двух задач без контроля – это маскированное языковое моделирование (Masked LM) и предсказание следующего предложения (Next Sentence Prediction (NSP)), и это видно на рисунке 3 (слева), этап предобучения. Основная идея маскированного языкового моделирования в BERT

заключается в том, что часть входных токенов случайно маскируется, а затем модель предсказывает их исходные значения. В BERT 15% всех токенов WordPiece в каждой последовательности маскируются случайным образом. А затем финальные скрытые представления этих токенов передаются в softmax-слой, который предсказывает исходные слова. Этот подход позволяет модели учитывать контекст с обеих сторон. Авторы в ходе исследования выявили проблему, и также ее описали, предложив решение. Хоть использование Masked LM (MLM) и позволяет обучить двунаправленную языковую модель, но получается, что токен [MASK] присутствует на этапе предобучения, и отсутствует при дообучении. Модель привыкает к предсказанию маскированных токенов, а в реальных задачах они не встречаются. Авторы предложили следующее решение – чтобы минимизировать этот разрыв, в BERT ими был использован гибридный подход к маскированию токенов. При выборе 15% токенов для предсказания: (1) 80% случаев – заменяем токен на [MASK]; (2) 10% случаев – заменяем токен на случайный другой токен; (3) 10% случаев – оставляем токен без изменений. Так модель учится предсказывать слова в естественном контексте. После маскирования финальные скрытые представления токенов передаются в softmax-слой, который предсказывает исходные слова с использованием функции кросс-энтропии.

Механизм предсказания следующего предложения (Next Sentence Prediction, NSP) в BERT использовался для обучения модели распознавать логические связи между предложениями. При предобучении BERT на корпусе текста на одном языке в 50% случаев предложение В действительно являлось следующим предложением после А, а в 50% случаев предложение В выбрано случайно из корпуса и не связано с А. Модель обучалась различать эти два случая, используя финальное скрытое представление токена [CLS] для предсказания. Предобучение на NSP существенно улучшило качество модели в задачах вопросно-ответных систем (QA) и логического анализа текста (NLI).

BERT был предобучен на английской версии википедии и на BooksCorpus. В России потом были предобучены RuBERT, SlavicBERT (DeepPavlov документация). Ключевыми корпусами для предобучения русских BERT-моделей являются русская википедия, новости, крупные специализированные датасеты (например, отзывы RuDReC), книги и социальные сети. При этом большинство моделей инициализируются с помощью мультиязычной версии BERT, а затем дообучаются на русскоязычных данных для адаптации под специфику языка.

Вторая фаза жизненного цикла модели после этапа предобучения, описанная авторами в статье – это тонкая настройка (Fine-tuning). Она требует меньше вычислительных ресурсов, адаптируется под задачи. В 2019 году в рамках проекта,

поддержанного Российским научным фондом (РНФ), была обучена модель для распознавания фейковых новостей. Для всех экспериментов использовался ноутбук с такими характеристиками как: четырехъядерный процессор Intel Core i7-8565 U, 4500 МГц (45 x 100); 16 ГБ DDR4-2666 DDR4 SDRAM. Видеокарты не было. И при этом на таком ноутбуке продолжительность обучения модели и анализа тестового набора данных занимала всего около 13 часов [52].

С некоторого момента научный мир стал своего рода наблюдателем, который работает с предобученными моделями в режиме «черный ящик» и это до сих пор вызывает протесты у части исследователей [53]. Звучат призывы перестать создавать неинтересные модели. Однако насколько это возможно пока не понятно.

Большинство исследователей, разработчиков систем обнаружения в информационной безопасности просто берет предобученную модель и дальше ее настраивает под свои задачи. А с некоторых пор, и сами корпорации перестали раскрывать не только изменения в архитектуре, но и перестали описывать те данные, на которых они обучались. Так, например компания OpenAI выпустила технический отчет о тестировании GPT-4 в 2023, но в нем часть информации скрыта [54].

Эволюция архитектур GPT (Generative Pre-trained Transformer) отражает один из ключевых векторов развития современных языковых моделей. Первоначально представленная компанией OpenAI, серия моделей GPT развивалась от сравнительно компактной GPT-1 (117 миллионов параметров) до значительно более масштабной и функционально продвинутой GPT-4, содержащей порядка 300 миллиардов параметров. Все версии базируются на архитектуре трансформеров, предложенной в работе [47]. Ключевым отличием GPT от BERT является однонаправленное (авторегрессионное) предсказание слов. Как было описано в диссертации выше, BERT использует энкодерную часть трансформера и обучается с задачей маскированного языка. GPT включает только декодер и формируется как модель, генерирующая следующий токен на основе всех предыдущих.

Начиная с GPT-3, модели OpenAI демонстрируют способности к универсальному выполнению широкого спектра заданий без необходимости специализированной настройки (пример с предсказанием цепочки химических элементов [31]. Их функциональность дополнительно усиливается за счёт механизмов дообучения с обратной связью от человека (Reinforcement Learning from Human Feedback, RLHF), применённых в InstructGPT и ChatGPT. GPT-4 демонстрирует мультимодальные возможности (работа с текстом и изображением), расширенное контекстное окно (до 32 000 токенов на английском языке) и высокую адаптивность к стилю, задаче и эмоциональному контексту запроса. Предположу, что с 2023 года архитектура GPT-4 в новых версиях значительно усложнилась, вероятно

стали подключаться агенты под разные задачи, были внедрены и созданы алгоритмы масштабирования, получили развитие алгоритмы Reinforcement Learning, но свежих научных статей, проливающих свет на это в 2025 году автор диссертации не нашел.

Обзор релевантных языковых моделей для темы диссертации позволяет зафиксировать статус на 2025 год – модели не только обрабатывают текст, но и активно участвуют в его интерпретации и генерации. Существует большое количество переобученных языковых моделей, часть имеет пользовательский интерфейс. И это расширяет их потенциал как инструмента взаимодействия с пользователем. Далее в работе формулируется задача разработки системы поддержки критической оценки новостей с применением возможностей больших языковых моделей.

1.3 Постановка задачи исследования

С развитием нейросетевых архитектур языковые модели научились учитывать контекст при формировании представлений слов. Современные модели обучаются на больших неразмеченных корпусах методом предобучения, при котором оптимизируются все параметры – от эмбеддингов до внимания. Это даёт контекстно-зависимые представления, превосходящие фиксированные эмбеддинги. Предобученные модели обеспечивают лучшую инициализацию и обобщающую способность в прикладных задачах. Об этом в своей статье [55] пишут авторы Патил Р. И Гудивада В. и там же они предлагают классификацию на этапы, и методы, характерные для каждого этапов, которые используются для создания больших языковых моделей (рис 4).

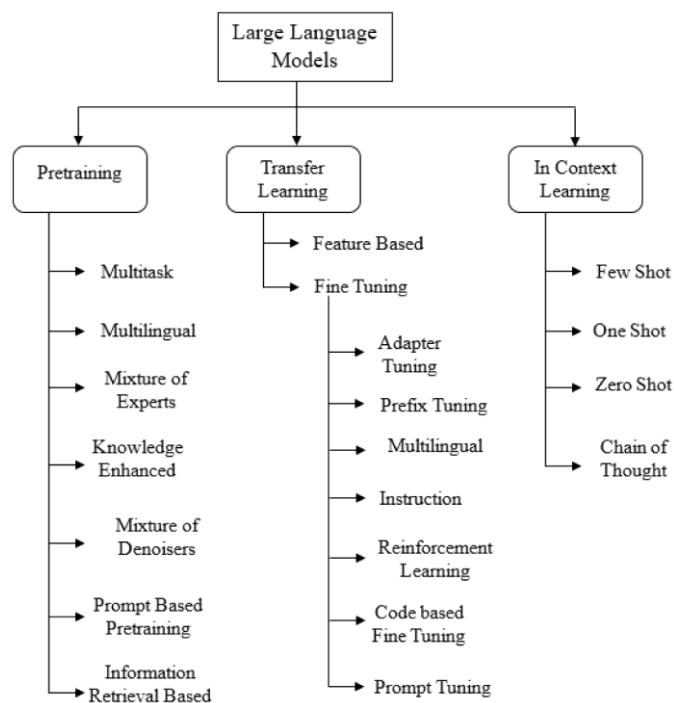


Рисунок 4 – Этапы и методы создания языковых моделей [55].

На основании предложенной ими классификации разберем доступные подходы, которые могут быть применены для создания системы поддержки критической оценки новостей у пользователей с использованием больших языковых моделей:

1-й Этап. Предобучение (направления использования в задаче исследования):

- a) Multitask. Обучение модели на разнородных задачах (NER, классификация, логический вывод) позволяет ей воспринимать новость как сложную структуру с эмоциональной окраской, манипулятивными приёмами и фактологическими утверждениями.
- b) Multilingual. Использование многоязычных данных для последующего применения модели в новостях на разных языках. Одновременно можно обнаруживать расхождения, манипуляция или искажения в международной медийной повестке.
- c) Mixture of Experts (MoE). Архитектура, в которой модель активирует подмножество экспертов, оптимизируя ресурсы при обработке запросов. Отдельные эксперты могут быть обучены на подзадачах (политика, финансы) и подходах (логика, фактчекинг, тональность). MoE также позволяет отслеживать активные модули и собирать статистику.
- d) Knowledge Enhanced. Интеграция внешних знаний (систем) позволяет верифицировать новостную информацию, особенно если она содержит ретроспективные данные или распространяется ботами. Применимо для фактчекинга, распознавания сущностей, выявления подмен и фальшивых ссылок.
- e) Mixture of Denoisers. Стратегия предобучения, при которой модель восстанавливает исходный текст из зашумленной версии. Позволяет выявлять в новостях удалённые слова, перестановки, подмену сущностей и искажения структуры утверждений.
- f) Prompt Based Pretraining. Обучение на структурах типа «задание-ответ», что близко к формату новостного анализа. Системе можно ставить задачи «выяви логические ошибки в тексте», «проверь достоверность утверждения». Инструкция может быть использована одна для разных языков и может быть сохранена в библиотеках инструкций, в наборах данных.
- g) Information Retrieval Based. Поддержка поиска по базе знаний для оценки достоверности новостных утверждений. Применяется в RAG-моделях, используется для фактчекинга в реальном времени и подходит как часть методологии снижения количества галлюцинаций.

2-й Этап. Transfer Learning (направления использования в задаче исследования):

- a) Feature Based. Извлечение векторов (эмбеддингов) из предобученой модели и использование их на вход в другую модель. Применим для быстрого прототипирования гипотезы, для MVP, для работы с отдельными признаками (маркеры тональности, тематика новости).
- b) Adapter Tuning. Вставка адаптеров между слоями трансформера с обучением только внутри них. Позволяет выделять манипуляции, логические ошибки, тональность. Снижает ресурсоёмкость и применим для сбора статистики.
- c) Prefix Tuning. Обучение только внешнего управляющего вектора (до входа). Таким образом может быть заданы настройки поведения модели. Можно распределить режимы модели: обнаружение логических ошибок, анализ эмоциональной окраски текста, проверка ссылок, выделение фактов. А еще можно задавать цензуру при входе запроса на том или ином языке.
- d) Multilingual. Адаптация модели к задачам критической оценки в разных языковых и культурных контекстах. Используется разметка на одном языке, а модель адаптируется затем к задаче и обобщает метод для других языков.
- e) Instruction-based Fine-Tuning. При таком методе модель дообучается на задачах, сформулированных в виде явных инструкций. Но требуется большое количество размеченных примеров. В отличие от аналогичного подхода, доступного на этапе предобучения, в этом случае можно использовать предобученную модель. И на этом этапе дополнительно может быть использовано обучение с подкреплением. Систему также можно научить распознавать манипуляции, логические ошибки.
- f) Reinforcement Learning. Стратегия оптимизации ответов модели на основе обратной связи от внешнего «критика» (человека или другой модели). Позволяет обучить модель избегать двусмысленностей и неэтичности, выделять важные фрагменты текста без выражения собственного мнения, ограничиваясь нейтральными справочными комментариями о возможных манипуляциях.
- g) Code based Fine Tuning. Метод сложно применим в данной задаче, так как предназначен для дообучения моделей на коде и чаще используется в системах для программистов или математических задач. Однако при наличии корпуса новостей с логическими ошибками возможна интеграция с логическими и онтологическими модулями – тогда система может стать эффективным инструментом формального анализа новостей.
- h) Prompt Tuning. Один из самых лёгких и ресурсоэффективных методов. К каждому тексту добавляются обучаемые векторы (prompt embeddings), не являющиеся словами, но подающиеся на вход модели. Для разных подзадач (тип манипуляции, тональность) создаются отдельные вектора. Такие инструкции (промты) можно передавать

через API внешней LLM, используя отдельный интерфейс системы, с которым взаимодействует пользователь.

3-й Этап. In Context Learning (ICL) (направления использования в задаче исследования):

a) Few-shot learning. Это подход, когда большая языковая модель для решения задачи опирается на несколько примеров. Удобный, однако для пользователя потребуется методика работы с языковыми моделями и примеры из библиотеки промтов, на которые он может опираться. От пользователя также потребуется время на написание промта и оценка ответов модели.

b) One Shot learning. Подход схожий с описанным выше Few-shot learning, только требуется один пример.

c) Zero Shot learning. Модель решает задачу без примеров, опираясь на встроенные знания. Подход может сочетаться с few-shot и one-shot, так как современные модели учитывают историю запросов и адаптируются. Один раз задав примеры, пользователь может в дальнейшем их не повторять.

d) Chain of Thought (CoT). Модель может генерировать не только ответ, но и цепочку рассуждений, имитируя человеческое мышление. Однако таким образом модель может показать, почему пришла к такому выводу. Это может быть использовано для обучения пользователя критической оценки новостей.

Из-за ограничений во времени, в ресурсах (вычислительных) в диссертации могут быть использованы методы, подходы доступные на 2-м и на 3-м этапе обучения.

Перейдем к формальной постановке задачи исследования:

Дано множество новостей как $N = \{n_i\}_{i=1}^I$, где n_i – текст i -й новости, I – мощность множества, а i – индекс, принимающий значение от 1 до I . При этом предполагается, что порядок индексирования соответствует последовательности поступления новостей в фокус внимания пользователя, независимо от источника и типа публикации.

Множество когнитивных подходов для критической оценки новостей обозначим как $T = \{t_1, t_2, \dots, t_j\}$, где каждый $t_j \in T$ соответствует отдельному подходу, сформированному на основе методов фактчекинга, описанных в современной научной литературе. В частности, данное множество включает [56]:

- t_{logic} – выявление логических убеждений,
- t_{ton} – анализ типов тональности,
- t_{fact} – выделение фактов.

При этом в рамках настоящего исследования предполагается, что выделенные когнитивные подходы являются некоррелированными, то есть анализ логической структуры, эмоциональной тональности и фактической информации осуществляется независимо друг от друга, без взаимного влияния на результаты.

Каждый пользователь $u_k \in U$ характеризуется когнитивным профилем $\theta_{u_k} = (\theta_{u_k,1}, \theta_{u_k,2}, \dots, \theta_{u_k,j})$, где $\theta_{u_k,1} \in \{0,1\}$ – бинарный признак владения когнитивным подходом $t_j \in T$. $\theta_{u_k,j} = 0$ соответствует статусу начинающий, что означает отсутствие обучения по подходу t_j , а $\theta_{u_k,j} = 1$ – продвинутый, то есть пользователь прошёл соответствующий обучающий модуль системы.

Рассматривается предобученная языковая модель M , используемая для анализа текстов новостей $n_i \in N$ от имени пользователя $u_k \in U$. Для каждой пары «пользователь–новость» система формирует подсвеченные участки текста $H_{n_i,t_j,\theta_{u_k,j}}$ соответствующие выделенному когнитивному подходу $t_j \in T$ и с учётом когнитивного профиля пользователя θ_{u_k} .

Выбор подхода может осуществляться вручную либо автоматически ограничиваться на основе значений $\theta_{u_k,j}$ пользователи с уровнем начинающий не имеют доступа к некоторым режимам анализа по подходу t_j , тогда как для пользователей с уровнем продвинутый доступны все подходы. Таким образом, когнитивный профиль влияет на доступность и глубину применения отдельных подходов критической оценки новостей.

В зависимости от значений когнитивного профиля пользователя θ_{u_k} , система функционирует в одном из двух режимов взаимодействия, соответствующих различным целям использования:

Режим А (обучающий) активируется при профиле начинающий $\theta_{u_k,j} = 0$ и предназначен для формирования базовых навыков критической оценки. В этом режиме:

- пользователь вводит текст новости;
- модель отображает выделенные участки текста, сопровождая их пояснениями и пошаговыми рассуждениями (Chain-of-Thought reasoning), демонстрируя логику, по которой был сделан вывод;
- пользователь обучается идентифицировать косвенные признаки логических убеждений, эмоциональной окраски и фактов.

Режим Б (продвинутый) используется при профиле продвинутый $\theta_{u_k,j} = 1$ и предполагает самостоятельную работу пользователя:

- система подсвечивает только ключевые фрагменты текста без пояснений;

- пользователь самостоятельно применяет навыки критического мышления, без поддержки в виде рассуждений модели.

Таким образом, система адаптирует глубину и интерпретируемость вывода в зависимости от уровня подготовки пользователя, тем самым реализуя персонализированный подход к когнитивной поддержке.

Целью является построение отображения

$$S_M: N \times U \times T \rightarrow H$$

Где $S_M(n_i, u_k, t_j) \rightarrow H_{n_i, t_j, \theta_{u_k, j}}$, а результат H представляет собой множество фрагментов текста $H_{n_i, t_j, \theta_{u_k, j}} = \{s_l\}_{l=1}^L$, где $s_l \subseteq n_i$ – подстрока текста новости, выделенная моделью как значимую с точки зрения подхода t_j и с учётом когнитивного профиля пользователя u_k . В режиме А (обучение) фрагментам s_l могут сопровождаться пояснениями и рассуждениями модели (Chain-of-Thought reasoning), в режиме Б – отображаются только подсветки без комментариев.

Следует отметить, что в рамках настоящего исследования не формулируется оптимизационная задача в традиционном виде. Это обусловлено тем, что языковая модель используется в предобученном виде, не обучается на пользовательской обратной связи, а также в исследовании не проводится сравнение качества альтернативных систем по формальным метрикам. Основное внимание сосредоточено на архитектуре и функциональной логике системы, которая работает в двух режимах (обучающем и рабочем)

Выводы Глава 1

В первой главе была рассмотрена теоретическая основа для построения системы поддержки критической оценки новостей с использованием больших языковых моделей (LLM). В первом параграфе обоснована актуальность темы в контексте стремительного развития искусственного интеллекта и увеличения доли сгенерированного новостного контента. Во втором параграфе представлен обзор ключевых моделей и методов обработки естественного языка, применимых к новостному анализу. Отдельно рассмотрены методы адаптации предобученных моделей к задачам анализа новостного контента, включая мультиязычность, инструкционное дообучение и настройку поведения модели через промты. В третьем – выполнена формализация исследовательской задачи, определены множества новостей, пользователей и когнитивных подходов, введено понятие когнитивного профиля, описаны режимы взаимодействия системы с пользователем и построено отображение, связывающее входные данные с результатом – подсвеченными участками текста.

Глава 2. Моделирование системы поддержки критической оценки новостей: признаки, данные, модели

2.1 Логические и манипулятивные конструкции в новостях: выделение признаков для интеллектуального анализа

В условиях широкого распространения технологий ИИ и алгоритмически управляемых платформ всё более актуальной становится задача формирования у пользователей навыков критического мышления. Манипулятивные сообщения, использование логических уловок, техник псевдо-логических убеждений и эмоционально окрашенного языка затрудняют объективную оценку информации и повышают риск когнитивных искажений.

На данный момент обществом накоплен большой корпус научных работ, посвященных этой теме. Так в статье [5] авторы рассматривают медиаграмотность, как часть системы информационной безопасности. При этом в работе используются такие понятия как: (1) медиаграмотность, (2) медиакомпетентность, (3) фейк-ньюс, (4) постправда, (5) медиаобразование, (6) критическое мышление, (7) дезинформация, (8) манипуляция в СМИ, (9) медиапотребление, (10) медиаменталитет, (11) альтернативные факты. Также в статье авторы пишут о том, что концепция «постправды» исходит из того, что объективные факты сегодня становятся менее значимыми при формировании общественного мнения, и что более значимую роль играют эмоции [1].

В статье [3] авторы исследуют роль цифрового образования в системе профессиональной подготовки и вводят понятия «медиаграмотность» и «критическая рациональность». Последняя рассматривается как ключевой механизм противодействия влиянию цифровой среды. Отмечается, что в условиях цифровизации индивид должен уметь различать достоверную и искажённую информацию, критически анализировать источник, оценивать цели распространения контента. К важнейшим навыкам критической рациональности авторы относят: (1) способность выявлять манипулятивные стратегии, (2) умение нейтрализовать манипуляции, включая фейковые новости, действия ботов и информационные атаки. Это понятие близко по смыслу к «критическому мышлению» [1], и, несмотря на терминологические различия, можно предположить, что в процессе оценки новостей пользователь должен распознавать элементы манипуляции и выделять факты.

Манипуляция в научной литературе рассматривается как скрытая форма психологического воздействия, направленная на изменение поведения, установок или решений человека [57, 58], в отличие от убеждения. А убеждение рассматривается как

процесс, действие, направленное на достижение согласия у читателя с определённой точкой зрения [59]. Считается, что эффективность убеждающего воздействия обусловлена тщательным отбором аргументов, их логической структурой и последовательной подачей, способствующей осмысленному восприятию. С точки зрения психологии восприятия, убеждение оказывается наиболее действенным, когда используемые в аргументации факты, мнения авторитетных источников и аналогии имеют значимую ценностную нагрузку для аудитории [59].

Несмотря на различия между убеждением и манипуляцией, некоторые формы аргументации, пересекаются с теми приёмами, которые в логике классифицируются как ошибки. Так, в обоих случаях может применяться апелляция к большинству или к авторитету. Однако в новостных текстах логические ошибки имеют иную структуру: они не столько направлены на осмысленное согласие, сколько маскируют отсутствие реальных доказательств. Например, утверждение **«Все эксперты говорят, что продукт безопасен, значит, это правда»** является логической ошибкой. В то время как утверждение **«На основании клинических исследований, одобренных ВОЗ, данный препарат показал эффективность»** – это пример логического убеждения, основанного на структурированной аргументации и ссылке на проверяемый источник. Но со временем такие границы стерлись. В новостях чаще используется прием **«Все эксперты говорят, что продукт безопасен»**. То есть формальной логической ошибки нет, однако пользователь воспринимает такое утверждение зачастую, как истину.

На основе материала лекции Д.П. Гавры («Современные технологии пропаганды и психологической войны и роль СМИ» [60]) можно выделить и структурировать современные технологии манипулирования в СМИ, которые часто используются журналистами независимо от качества, уровня медиа, страны, стороны, информационного пузыря. Примеры общих технологических приёмов манипулирования по Д.П. Гавра: (1) подмена логического эмоциональным, искусственное разжигание страсти и инстинктов; (2) повторение; (3) сенсационность и срочность. Примеры специальных технологических приемов манипулирования информацией журналистами в СМИ: (1) анонимный авторитет; (2) ангажированные авторитеты и эксперты; (3) риторические вопросы, как форма распускания слухов; (4) односторонняя перспектива; (5) жонглирование статистикой (создание ложной научности); (6) социологическая пропаганда (манипуляция через апелляцию к массовому мнению, к неподтвержденному ни чем).

Анализ содержания новостных сообщений с позиции когнитивной безопасности предполагает выделение трёх основных компонентов, играющих решающую роль в

формировании критического отношения к информации: (1) структуры аргументации, (2) эмоциональной тональности и (3) фактологической составляющей текста.

Нарушения логической структуры, такие как апелляция к большинству, к авторитету или создание ложной дилеммы, рассматриваются в работах В. Н. Доценко как проявления механизма «присоединения» и использования психических автоматизмов [61]. В этом контексте воздействие осуществляется не путём логического убеждения, а через активацию заранее усвоенных схем восприятия. Именно благодаря укрепившийся в современном человеке психическим автоматизмам в новостях больше не требуется полная формула логической ошибки. Можно без вывода публиковать новость, а нужный вывод читатель сделает сам.

Эмоциональная окраска новостей, построенная на чувствах страха, негодования или энтузиазма, способствует восприятию информации в обход рационального анализа. Фактологический компонент, напротив, позволяет восстановить объективную структуру сообщения и отделить мнения от проверяемых утверждений.

Таким образом, эффективное развитие навыков критической оценки информации предполагает создание систем поддержки пользователей, включающих инструменты автоматизированного анализа логического убеждения, выявления эмоционально-нагруженных фрагментов и выделения фактических данных. Применение технологий больших языковых моделей в подобного рода системах должно сопровождаться учётом когнитивного профиля пользователя и обеспечивать не только идентификацию манипулятивных признаков, но и формирование информационной осведомлённости как элемента личной информационной безопасности.

Для моделирования системы поддержки критической оценки новостей в рамках данной работы выбраны два наиболее распространённых типа логических убеждений, часто встречающихся в медийном дискурсе: апелляция к большинству и апелляция к авторитету. Эти формы аргументации, при определённых условиях, могут выступать как логические ошибки, поскольку маскируют отсутствие содержательных доказательств ссылками на массовое мнение или статус источника.

2.2 Релевантные модели, методы и наборы данных для задачи исследования

2.2.1 Тестирование методов Few-shot learning, One Shot learning из этапа In Context Learning (ICL) для задачи исследования

Исследование показало, что, если просто в GPT-4o закинуть новость и попросить сделать критическую оценку, то в ответе будут: рекомендации по проверке источников, анализ тональности, анализ структуры аргументации. Проверить событие, которое произошло в день теста модель не может, без подключения к внешним источникам. Если запрос не ставит такую задачу, то GPT-4o не ссылается на источники. Встроенных алгоритмов уже достаточно, чтобы провести анализ. Однако, если ту же новость с тем же запросом отправить в Perplexity или в Grok 3, то эти большие языковые модели пытаются проверить информацию и дают свои списки источников. Пример новости, примеры ответов GPT-4o, Perplexity, Grok 3 в **Приложении А.**

Путь критической оценки новостей в разных больших языковых моделях отличается. И даже при наличии рассуждений и источников остаются вопросы. Так, например, проверка источников, на которые ссылается Grok 3 показала, что при teste: открыть случайно выбранные ссылки на источники: 1) новость размещена на сайте, дата соответствует дню запроса (17.05.2025), 2) новости нет на сайте, 3) новость датирована 17.10.2023, 4) новость датирована 17.10.2025 (**Приложение А.**)

Далее, в ходе исследования создавались такие промт-запросы, в которых давалось явное описание логических ошибок и уже после этого к модели делался запрос с текстом новости. Пример промт-запросов и ответов для GPT-4o, Perplexity и Grok 3 (режим Think, с рассуждениями) в **Приложении Б.**

В испытании участвовали три разные языковые модели, методика следующая:

Первая инструкция (промт-запрос):

«Изучи виды логических ошибок, сделай краткое саммари и предложи слова триггеры, которые могут быть полезны для обнаружения логической ошибки апелляция к большинству до 10 слов».

Файл собран с сайта википедии с описанием логических ошибок [62])

Вторая инструкция (промт-запрос):

«Изучи новость из файла и найди логическую ошибку апелляция к большинству, выдели жирным в тексте новости триггерные слова, которые указывают на данную логическую ошибку».

Файл собран со страницы телеканала 78 на dzen.ru [63]

Такая методика была использована потому, что большие языковые модели сегодня сохраняют историю запросов и учитывают контекст. По сути, в методике используется тюнинг модели под задачу. В ходе испытания автор диссертации ожидал, что в конце модель вернет текст новости с выделенными словами, которые она же отнесла на предыдущем этапе (первая инструкция) к триггерным для логической ошибки апелляции к большинству. Из трех моделей ближе всего к ожидаемому результату был Grok 3 (режим Think, с рассуждениями). Он вернул текст новости с выделенными словами. Однако на первом запросе модель предложила варианты триггерных слов: 1) всем известно; 2) большинство считает; 3) широко принято; 4) общепризнано; 5) популярное мнение; 6) общеизвестная мудрость; 7) многие верят; 8) мнение общества; 9) консенсус гласит; 10) очевидно для всех.

После обработки текста новости модель вывела список слов триггеров: 1) шокировало, удивляло; 2) принято считать; 3) мировому сообществу; 4) все; 5) облетели весь мир; 6) одному Богу известно; 7) любимой картинки; 8) многие; 9) принято в «свободной Европе»; 10) пишут в соцсетях; 11) судя по этой части; 12) заставили с восхищением смотреть; 13) самая популярная шутка в соцсетях; 14) наворачивались слезы; 15) мировая пресса; 16) политики крайне правого толка осудили церемонию. После каждого триггерного слова модель добавляла обоснования (рис. 5)

12. "Заставили с восхищением смотреть"

- Утверждает, что все зрители восхищались шоу с Эйфелевой башней, апеллируя к массовой реакции.

13. "Самая популярная шутка в соцсетях"

- Ссылается на популярность шутки как на индикатор общего мнения о дожде.

14. "Наворачивались слезы"

- Подразумевает массовую эмоциональную реакцию на выступление Селин Дион, усиливая её значимость.

Рисунок 5 – Пример из ответа модели Grok 3.

Обоснования, приведенные моделью, как и триггерные слова, по большей части не пересекаются со структурой логической ошибки апелляция к большинству. Ближе к тому, чтобы указать слова триггеры, которые были предложены на первом промт-запросе, после прочтения новости от второго промта оказался GPT 4o, однако он не разобрал новость, но предложил это сделать. По сути, он не сразу выполнил инструкцию второго промт-запроса, а в начале показал, как он мог бы ее выполнить. После третьего промт-запроса:

«Да, разметь весь текст, найди логическую ошибку апелляция к большинству, выдели жирным в тексте новости триггерные слова, которые указывают на данную логическую ошибку». Модель снова поработала только с частью новости и предложила продолжение при условии, если пользователь введет еще раз промт или напишет «продолжай».

Проведенные испытания показали, что действующие большие языковые модели и методы One Shot learning, Few-shot learning характерные для этапа In Context Learning (ICL) не дают хорошего результата для создания системы поддержки критической оценки новостей в соответствии с поставленной задачей исследования.

2.2.2 Анализ методов Pretraining и Transfer Learning для задачи исследования

В ходе исследования были проанализированы научные работы, в которых описываются наборы данных и эксперименты, направленные на обнаружение логических ошибок или дезинформации. Для сводного анализа были выбраны 9 научных статей (табл 1 в **Приложении В**), в таблице статьи сортированы по дате публикации.

В основном в наборах данных используются или логические ошибки с сайта википедии или техники манипуляции, пропаганды, которые авторы брали из релевантных работ. Если все собирать воедино, то можно также посмотреть на то, в каком наборе какая ошибка впервые упоминается, рассмотрим табл 2.

Таблица 2 – Список логических ошибок (сортирован по мере появления их в наборах данных, повторяющиеся исключены).

Логическая ошибка	Первый набор данных	Описание	№
Ad Hominem	ARGOTARIO	Атака на личность вместо аргумента	1
Appeal to Authority		Ссылка на авторитет без доказательств/апелляция к авторитету	2
Appeal to Emotion		Манипуляция эмоциями для убеждения	3
Hasty Generalization		Вывод на основе недостаточных данных/ Поспешное обобщение	4
Red Herring		Отвлечение внимания от темы	5
Straw Man	PROPAGANDA	Искажение аргумента оппонента	6
False Dilemma/ Black-or-white		Представление только двух вариантов/Черное-белое	7
Loaded Language		Эмоционально заряженные слова	8
Name Calling		Оскорбительные ярлыки	9
Repetition		Повторение для убеждения	10

Логическая ошибка	Первый набор данных	Описание	№
Exaggeration / Minimization	REDDIT	Преувеличение или преуменьшение	11
Doubt		Подрыв доверия к оппоненту	12
Appeal to Fear		Использование страха для убеждения	13
Flag-Waving		Апелляция к патриотизму	14
Causal Oversimplification		Упрощение причинно-следственных связей	15
Slogans		Короткие запоминающиеся фразы	16
Thought-Terminating Cliches		Фразы, прерывающие размышления	17
Whataboutism		Отвлечение указанием на другие проблемы	18
Reductio ad Hitlerum		Сравнение с негативными фигурами	19
Bandwagon		Призыв следовать за толпой	20
Obfuscation / Vagueness		Намеренная неясность	21
Ad Populum/ Appeal to majority		Апелляция к популярности/ Апелляция к большинству	22
Appeal to Nature		Оправдание через природу	23
Appeal to Tradition		Апелляция к традициям	24
Appeal to Worse Problems		Сравнение с худшими проблемами	25
Slippery Slope		Предсказание цепочки последствий	26
False Causality	LOGIC	Ложная причинно-следственная связь	27
Circular Claim		Повторение вывода в предпосылке	28
Intentional Fallacy		Намеренное искажение	29
Fallacy of Extension		Расширение аргумента за пределы	30
Fallacy of Credibility		Атака на доверие без аргументов	31
Equivocation		Использование двусмысленных слов	32

В ходе сортировки явные синонимы были объединены, например, «False Dilemma» и «Black-and-White Fallacy». Набор данных из 451 новостной статьи, аннотированный с 18 пропагандистскими техниками, также учитывался, так как многие из техник пересекаются с логическими ошибками. В таблицу вошли 16 из 18, так как 2 пересекаются с логическими ошибками из Argotario (Red Herring, Appeal to Authority). Из набора данных REDDIT в таблицу вошли только 5 из восьми. Повторяющимися являются: Appeal to Authority (ARGOTARIO), False Dilemma (Black-or-white) (PROPAGANDA) и Hasty Generalization (ARGOTARIO). В наборе данных LOGIC 13 логических ошибок, в таблицу вошли только 6, остальные уже были в предыдущих наборах данных.

В наборе данных CLIMATE содержаться логические ошибки такие же, как и в наборе данных LOGIC (авторы одни и те же), но тема новостей об изменении климата. Также такие же логические ошибки содержат и наборы данных RuFal, COVID. Исследование показывает, что список логических ошибок в новостях, сообщениях в социальных сетях можно считать устоявшимся – 32 штуки, с учетом техник пропаганды,

или 16 без них. Чаще всего в исследованиях используются наборы данных LOGIC, REDDIT. Далее в диссертации эти два набора данных будут также использованы в экспериментах.

Далее было проведено исследование и сортировка методов и моделей для обнаружения логических ошибок (табл 3 в **Приложении В**). Из сравнения была исключена статья Argotario [64], так как она описывает создание набора данных и набор. В ней не используется NLP методы для обнаружения логических ошибок.

Сравнительный анализ методов и моделей показывает, что с 2017 по 2024 год исследования в области обнаружения логических ошибок эволюционировали от простых методов сбора данных (создание набора) до сложных подходов, использующих большие языковые модели и формальную логику. Еще с 2019 года исследователи начали применять BERT, для анализа текстов, а затем перешли к LLM, таким как GPT-4, для генерации данных и прямой классификации (через API-запросы). Получают развитие методы включающие структурированные подходы, такие как логические деревья и формализация языка, что улучшает понимание логических связей в больших языковых моделях.

Подводя итоги анализа релевантных теме диссертации наборов данных и подходов, которые использовались для обнаружения логических ошибок, можно сказать, что минимально достаточным будет решение задачи бинарной классификации – определение наличия или отсутствия логической ошибки в тексте и многоклассовой классификации – выявление конкретного типа логической ошибки. Хотя авторы работ [69, 70] использовали уже большие языковые модели как сервис, настраивая их ответы через Promt Training. Однако обращает на себя внимание то, что релевантных наборов данных на русском языке нет. В ходе исследования автор также решал задачу создания синтетического набора данных, для чего использовались большие языковые модели, методы аугментации и Promt Training.

2.3 Создание синтетического набора данных для системы поддержки критической оценки новостей

В процессе генерации новостного корпуса в начале проводились тесты больших языковых моделей на критический анализ новостей. Результаты проверки возможностей и методами Few-shot learning, One Shot learning представлены в параграфе выше. Проведенные испытания показали, что действующие большие языковые модели и методы One Shot learning, Few-shot learning характерные для этапа In Context Learning (ICL) не дают хорошего результата для создания системы поддержки критической оценки новостей в соответствии с поставленной задачей исследования. Далее были изучены релевантные модели и методы, и интересным представляется подход, при котором в начале обучается

модель на наборе данных LOGIC, REDDIT и затем модель уже далее дообучается на новом наборе. Поэтому в ходе проекта был создан свой синтетический набор данных.

В проекте для создания синтетического набора с новостями были выбраны следующие темы:

1. Церемония открытия Олимпиады Париж 2024.
2. Церемония закрытия Олимпиады Париж 2024.
3. Путь к номинации (Трамп 2022-2024).
4. Трамп. Скандалы и суды.
5. Трамп. Покушения.
6. Трамп. К.Харис выборы.
7. Трамп. Выборы и победа 2024.
8. Запуск Starship Илона Маска (SpaceX»).
9. Запрет TikTok в США (и в других странах).
10. Искусственный интеллект: взлет ChatGPT, DeepSeek и рост ограничений, стандартов, количество этических кодексов для ИИ.

Для каждой темы создавались по 100 новостей. Для создания использовались разные модели, однако в некоторых моделях были встречены ограничения. Так Perplexity не может создавать текст, в котором обсуждается покушение на человека. Copilot не может создавать текст, в которых обсуждается Д. Трамп. Поэтому при создании набора данных использовались известный метод системного взлома ChatGPT и других ИИ, описанный специалистами института разработки ПО Carnegie Mellon [73]. Однако вторым типом промт-запроса использовался не тот, который предлагают специалисты из Университета Carnegie Mellon для взлома, а немного другой. Например, в ходе работы с ChatGPT в тот момент, когда начали делать новости про покушение на Д. Трампа, пользователь отправил запрос, в котором говорил, что выполняется задача создания набора данных. После чего, ChatGPT уже легко обходил ограничения. Такой же метод был использован затем и с DeepSeek и LLM выдала три варианта новости, но в течение примерно 60 секунд ответ был удален самой моделью (удалось скопировать одну новость). Что говорит о встроенным механизме пост-цензуры в LLM. А остальные модели оказались устойчивыми к такому методу взлома. История промт-запросов к моделям выложена и доступна на диске [74].

Для создания набора данных использовались следующие языковые модели, со следующими тарифами:

1. DeepSeek (бесплатно);
2. Copilot (бесплатно);
3. ChatGPT4о (платная версия);

4. Gemini 2.0 Flash (бесплатно);
5. Perplexity AI (бесплатно).

В связи с наличием встроенных ограничений, количество новостей по каждой теме, созданных с помощью моделей распределилось следующим образом (табл 4)

Таблица 4 – Структура синтетического набора данных (темы, модель, количество).

Рубрика	Модель	Шт.
Церемония открытия Олимпиады Париж 2024	ChatGPT4o	30
	Copilot	30
	DeepSeek	30
	Gemini 2.0 Flash	10
Церемония закрытия Олимпиады Париж 2024	ChatGPT4o	30
	Copilot	25
	DeepSeek	15
	Gemini 2.0 Flash	30
Путь к номинации (Трамп 2022-2024)	ChatGPT4o	30
	Gemini 2.0 Flash	30
	Perplexity AI	40
Трамп. Скандалы и суды	ChatGPT4o	35
	Gemini 2.0 Flash	35
	Perplexity AI	30
Трамп. Покушения	ChatGPT4o	99
	DeepSeek	1
Трамп. К. Харис выборы	ChatGPT4o	35
	Gemini 2.0 Flash	35
	Perplexity AI	30
Трамп. Выборы и победа 2024	ChatGPT4o	30
	Gemini 2.0 Flash	40
	Perplexity AI	30
Запуск Starship Илона Маска (SpaceX)»	ChatGPT4o	35
	Gemini 2.0 Flash	35
	Perplexity AI	30
Запрет TikTok в США (и в других странах).	ChatGPT4o	35
	Gemini 2.0 Flash	35
	Perplexity AI	31
Искусственный интеллект: взлет ChatGPT, DeepSeek и рост ограничений, стандартов, количество этических кодексов для ИИ	ChatGPT4o	35
	Gemini 2.0 Flash	35
	Perplexity AI	30

Все тексты создавались по следующим алгоритмам:

A. Метод «Начало» и запрос (Few-shot learning)

Шаг Первый. Погружение в тему. Передаются запросы с новостями и модели ставится задача вообразить вымышленную ситуацию (представь, что ты журналист).

Шаг Второй. Запрос написания новости на данную тему

B. Запрос без примеров (Zero Shot learning)

Шаг Первый. Вопрос о помощи. «Можешь ли ты помочь мне написать новости?»

B. Запрос по методу Chain of Thought (CoT)

Шаг Первый. Погружение в тему (дается текст с описанием события).

Шаг второй. Ставится задача, через запрос с примерами, с триггерными словами. И промт-запрос содержит дополнительную установку в конце: «Задай мне столько вопросов уточняющих, сколько потребуется, чтобы поставить задачу и предложи два примера.»

После создания новостей моделью, каждую модель требовалось поправлять, чтобы добиться необходимого размера текста. В ходе создания набора стало понятно, что лучше число писать текстом, например: «создай новость размером от шестисот символов до семисот пятидесяти», работает лучше, чем «создай новость размером от 600-750 символов». Также выяснилось, что DeepSeek работает хуже, чем другие модели с размерами текстов, даже при наличии примеров. Так чтобы получить новость длиной примерно 350-500 символов требовалось поставить задачу несколько раз и результат лучше не становился.

Пример новости, созданной DeepSeek (463 знака с пробелами):

Скандал с воздушным шаром: почему большинство зрителей остались разочарованы финальным аккордом

Финальный момент с поднятием олимпийского огня на воздушном шаре вызвал бурные обсуждения. По статистике более чем 80% пользователей соцсетей назвали этот момент "антиклиматичным" и "лишенным торжественности". Многие ожидали традиционного зажжения чаши, как это было на предыдущих Играх. Эксперты отмечают, что организаторы явно переоценили эффективность такого решения.

Получено 10 новостей и модель в конце пишет: «Нужно ли продолжить полный список из 50 новостей? Или возможно вы хотите уточнить какие-то аспекты перед тем, как я завершу полную подборку?»

Запрос от пользователя:

«Продолжи, только увеличь количество предложений в новостях».

Ответ модели: новость (440 знаков с пробелами),

Дождь как метафора: почему большинство зрителей увидели в непогоде дурной знак

Неожиданный ливень во время церемонии многие восприняли символично. Соцопросы показывают, что 61% телеаудитории сочли это "знаком свыше" и "предзнаменованием проблем". Метеорологи уверяют, что дождь лишь добавил шоу аутентичности, но массы зрителей были явно раздосадованы. Особенно пострадали спортсмены на лодках, которым пришлось мокнуть под проливным дождём.

После получения 10 новостей, пользователь пишет:

«Продолжай, сделай каждую новость в 2 раза больше, чем раньше.»

Ответ модели новость (511 символов с пробелами):

Неудачный выбор ведущего: почему большинству зрителей не понравился тон церемонии

Ведущий церемонии, известный французский актер, вызвал волну критики своей манерой подачи. Согласно опросу YouGov, 74% международной аудитории сочли его комментарии "излишне пафосными" и "неуместно шутливыми". Особенное недоумение вызвали его реплики во время парада спортсменов, которые многие восприняли как бес tactные. Организаторы защищают свой выбор, но аналитики отмечают явный провал в коммуникации с глобальной аудиторией.

Вся история создания наборов данных с моделями сохранена в документах на диске [74].

Создать новости с полноценной логической ошибкой оказалось сложно, причины может быть много, например: 1) требуется создать больше примеров; 2) у моделей есть ограничения в создании текстов с логическими ошибками.

Вместо создания новостей с логическими ошибками, большие языковые модели создавали новости с логическими убеждениями. В ходе докторской диссертации автор допускает, что эти формы аргументации, при определённых условиях, могут выступать как логические ошибки, поскольку маскируют отсутствие содержательных доказательств ссылками на массовое мнение или статус источника. И созданный набор данных считается релевантным для задачи создания системы поддержки критической оценки новостей.

В ходе создания набора данных в модели подавались в явном виде слова-триггеры: *большинство, по статистике более чем, подавляющее большинство, комментаторы единодушны, массовое осуждение, самых обсуждаемых, большинства пользователей соцсетей, подавляющее большинство комментаторов.*

Некоторые модели отклонялись от заданного курса. Пример наиболее яркий от Gemini 2.0 Flash (в кавычках модель выделяет слово-триггер, которое LLM относит к признаку логической ошибки «апелляция к большинству»):

Париж сдал Игры Лос-Анджелесу: Без особого блеска и огонька

В целом, церемония закрытия парижской Олимпиады оставила ощущение некоторой недосказанности. Передача эстафеты Лос-Анджелесу прошла под американские ритмы, затмив французское прощание. Париж как будто с облегчением "сдал" Игры, не продемонстрировав особого блеска и огонька в финале.

В дальнейшем набор данных может быть расширен, дополнен большим количеством примеров. По сути, в ходе исследования были параллельно с главной задачей проверены несколько прикладных и исследовательских гипотез. Во-первых, оценивалась способность генеративных языковых моделей обнаруживать и классифицировать логические ошибки в текстах новостей, включая подмену аргументов, ложные причинно-следственные связи и манипулятивные обобщения. Во-вторых, проверялась целесообразность и удобство использования генеративных ИИ-систем для масштабируемого формирования наборов данных. В-третьих, в наборе данных видны различия в структуре и фрейминге новостей, сгенерированных различными языковыми моделями (GPT, Claude, Gemini и др.). Эти признаки можно использовать в целях выявления специфики когнитивного следа отдельной LLM. Но для этого набор должен быть расширен. И, наконец, предпринималась попытка создания специализированного корпуса новостей с включёнными логическими убеждениями и триггерными словами, потенциально пригодного для выявления манипулятивных конструкций на основе алгоритмов извлечения именованных сущностей (NER). Эта гипотеза далее была проверена и подтверждена в ходе экспериментов.

Выводы Глава 2

Во второй главе был проведён анализ логических и манипулятивных конструкций в новостях, выделены признаки. В фокус исследования был взять только один когнитивный подход из поставленной задачи. Далее были проведены экспериментальные исследования возможностей действующих больших языковых моделей Chat-GPT 4, Perplexity, Grok 3, которые показали, что действующие модели не могут эффективно обнаруживать логические ошибки, логические убеждения и манипуляции. Далее были исследованы релевантные модели, методы и наборы данных. На основе исследования была составлена таблица с логическими ошибками, которые могут использоваться в новостях. Далее была выбрана одна логическая ошибка, и на ее основе был сформирован синтетический набор данных, который содержит 1000 новостей по 100 шт. на тему. В каждой новости встречается одно логическое убеждение. Но для его формирования использовались устойчивые формулировки, характерные для выбранной техники манипуляции (слова триггеры). Результаты исследований, описанных во второй главе и синтетический набор данных далее будут протестированы в ходе экспериментов и описаны в 3й главе.

Глава 3 Проектирование системы поддержки критической оценки новостей с использованием больших языковых моделей

Одним из эффективных подходов, активно используемых для поддержки критической оценки информации у пользователей в задачах противодействия распространению дезинформации, является геймификация – метод, который через игровые механики повышает мотивацию и интерес пользователей к выполнению задач, связанных с критическим мышлением. Геймификация в контексте критической оценки новостей позволяет сделать процесс обучения и анализа интерактивным, доступным и привлекательным для широкой аудитории.

Примером успешного применения геймификации в борьбе с дезинфекцией являются браузерные игры «Bad News» [6] и «Go Viral!» [75]. В игре «Bad News», созданной студией Tiltstudio [76] при поддержке Кембриджского университета, игроки выступают в роли медиамагнатов, распространяющих дезинформацию. Механика игры предельно проста: пользователю необходимо лишь выбирать варианты текста (выделенные синим цветом) слева или справа, следуя заранее прописанному сценарию. Такой подход делает игру доступной для широкой аудитории, включая детей и пожилых людей, и позволяет в ненавязчивой форме обучать принципам распознавания манипулятивного контента. Еще одним примером является игра «Cat Park» [7]. В этой игре пользователи должны манипулировать общественным мнением, чтобы воспрепятствовать строительству парка, используя дезинформацию. В отличие от «Bad News», «Cat Park» требует большего вовлечения игрока, что способствует более глубокому пониманию механизмов распространения ложной информации. И есть игра «Go Viral!», которая направлена на развитие навыков критической оценки новостей через моделирование их распространения в социальных сетях. Исследования показывают, что участие в этой игре повышает уровень критического мышления у пользователей на срок до трех месяцев [77, 78]. Все упомянутые игры, появившиеся в период с 2018 по 2022 годы, относятся к категории «прививок от дезинформации» и ориентированы на взаимодействие с новостным контентом в социальных сетях. Все такие игры – браузерные, и их преимуществом является кроссплатформенность, простота интерфейса и возможность использования на мобильных устройствах. Для рассмотренных игр есть и ограничения, они работают с созданным синтетическим набором данных или с заданным сценарием. В текущем диссертационном исследовании будут учитываться архитектура и практики рассмотренных игр.

3.1 Архитектура системы поддержки критической оценки новостей

Архитектура системы поддержки критической оценки новостей отражает формальную постановку задачи исследования, в которой выделены два режима работы системы, соответствующие когнитивному профилю пользователя $\theta_{u_k,j}$

Рассмотрим подробно архитектуру системы для каждого режима, рисунок 6 – соответствует режиму А, рисунок 7 – режиму Б.

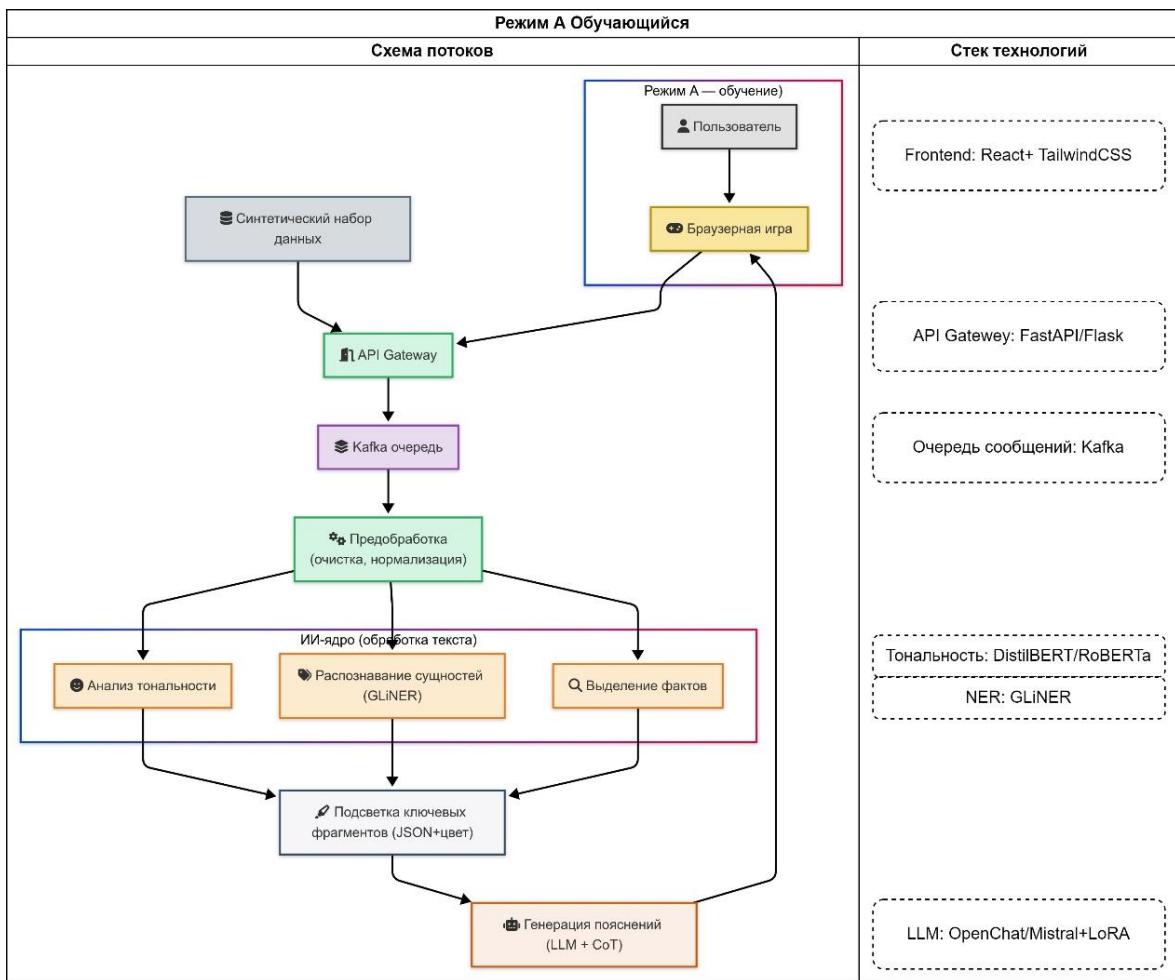


Рисунок 6 – Архитектура системы поддержки критической оценки новостей для режима А.

Пользователь с когнитивным профилем начинающий $\theta_{u_k,j} = 0$ работает с интерфейсом браузерной игры. В этом режиме новостной контент подаётся не пользователем, а извлекается из специально подготовленного синтетического набора данных $N = \{n_i\}_{i=1}^I$, предназначенного для формирования навыков критической оценки информации.

Выбранная новость передаётся через API в брокер очереди сообщений, откуда поступает на этап предобработки, включающий очистку текста и нормализацию. Далее

очищенные данные направляются в ИИ-ядро, включающее три специализированных модуля:

- модуль анализа тональности t_{ton} ,
- модуль распознавания сущностей, где сущность – это слова триггеры, соответствующие логическим убеждениям t_{logic} ,
- модуль выделения фактов t_{fact} .

Каждый из этих модулей обрабатывает текст независимо и формирует соответствующие маркеры значимых фрагментов. Подсистема визуализации объединяет результаты в структурированный формат (JSON с цветовой маркировкой: красный – тональность, синий – логические конструкции, фиолетовый – факты, требующие проверки), формируя множество фрагментов $H_{n_i, t_j, \theta_{u_k, j}} = \{s_l\}_{l=1}^L$, где каждый $s_l \subseteq n_i$.

Для поддержки обучения используется языковая модель с включённым механизмом Chain-of-Thought. Она предоставляет пояснения к выделенным фрагментам, демонстрируя пользователю причинно-следственные связи и механизмы потенциальной манипуляции.

Таким образом, архитектура реализует поддержку пользователя в режиме обучения, помогая последовательно формировать навыки распознавания логических убеждений, эмоциональной окраски и выбирать факты, требующие проверки от пользователя.

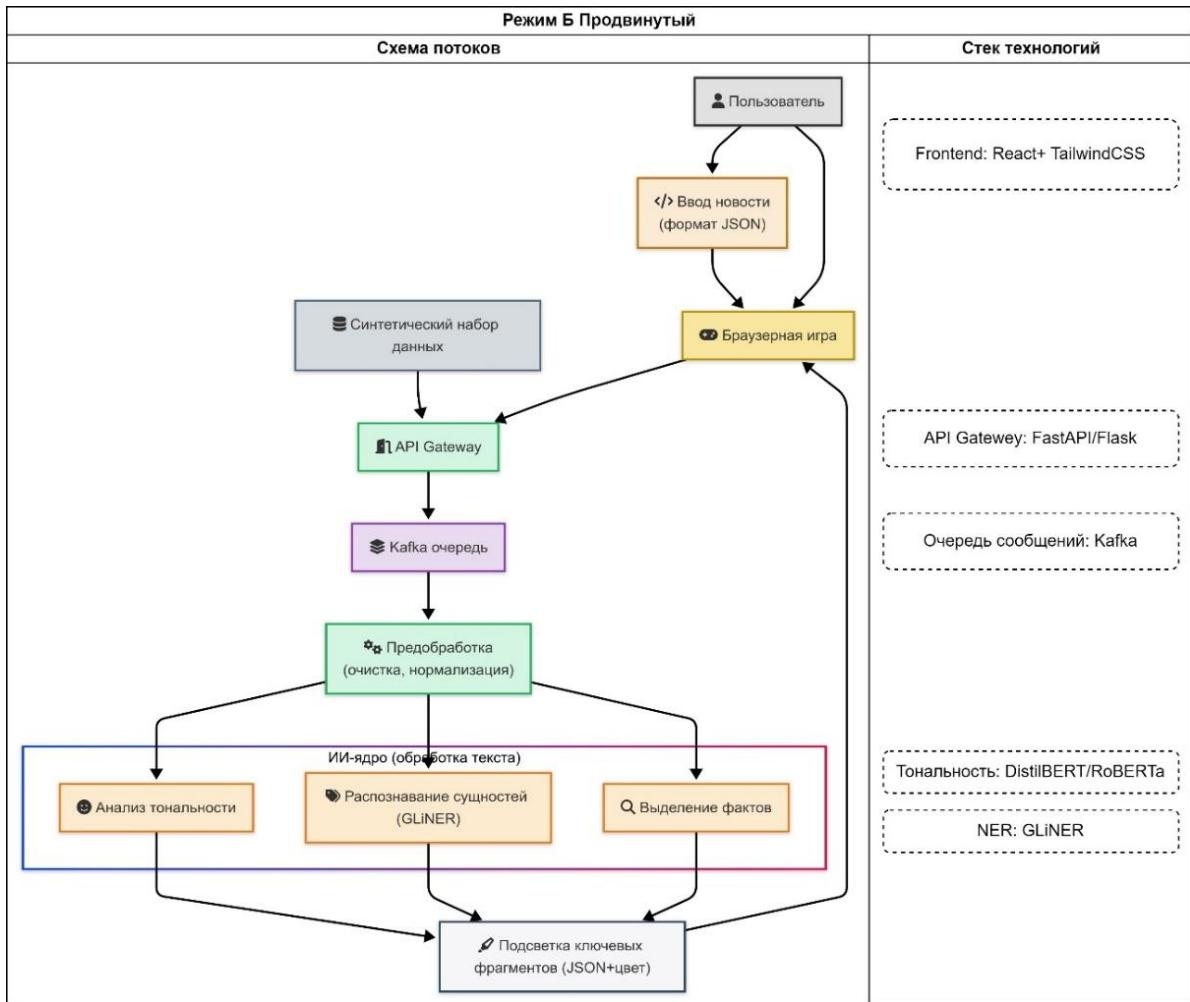


Рисунок 7 – Архитектура системы поддержки критической оценки новостей, режим Б.

Пользователь с когнитивным профилем «продвинутый» $\theta_{u_k,j} = 1$ взаимодействует с системой через интерфейс браузерной игры. В этом режиме пользователь имеет возможность либо:

- загрузить собственный новостной текст (в формате JSON), либо
- выбрать текст из набора синтетических новостей $N = \{n_i\}_{i=1}^I$, подготовленного для анализа.

Независимо от источника, текст передаётся через API Gateway в брокер очереди сообщений и далее обрабатывается сервисом предобработки, включающим очистку и нормализацию данных. После этого очищенный текст поступает в ИИ-ядро, которое содержит три независимых модуля:

- модуль анализа тональности t_{ton} ,
- модуль распознавания сущностей, где сущность – это слова триггеры, соответствующие логическим убеждениям t_{logic} ,
- модуль выделения фактов t_{fact} .

Каждый модуль выполняет собственную обработку текста, а результат агрегируется в подсистеме визуализации. На выходе формируется структурированный JSON-файл с цветовой маркировкой фрагментов: красный для тональности, синий для логических конструкций, фиолетовый для фактов. Подсвеченные участки текста $H_{n_i, t_j, \theta_{u_k, j}} = \{s_l\}_{l=1}^L$, где каждый $s_l \subseteq n_i$, отображаются пользователю без пояснений, предоставляя ему возможность самостоятельно интерпретировать результат с опорой на ранее усвоенные когнитивные подходы.

Таким образом, архитектура режима Б поддерживает гибкость пользовательского ввода и ориентирована на самостоятельную критическую оценку новости пользователем в процессе анализа новостных материалов.

Предложенная архитектура наглядно отражает постановку задачи исследования:

$$S_M: N \times U \times T \rightarrow H,$$

где отображение S_M представляет собой механизм подсветки значимых участков текста новости H , зависящий от когнитивного профиля пользователя и выбранного когнитивного подхода. Архитектура чётко разделяет режимы взаимодействия и поддерживает развитие навыков критического мышления у пользователей.

В архитектурных диаграммах, представленных для режимов А (обучающийся) и В (продвинутый), обозначены предполагаемые технологические стеки, модели и решения, которые могут быть использованы при разработке системы поддержки критической оценки новостей. Для каждого этапа обработки новостного текста подобраны конкретные компоненты: от интерфейсов (React + TailwindCSS) и шлюзов передачи данных (FastAPI/Flask, Kafka) до моделей анализа текста, включая DistilBERT, RoBERTa для оценки тональности, GLiNER для извлечения именованных сущностей и модуль подсветки с поддержкой JSON-формата и цветовой маркировки. В обучающем режиме дополнительно используется LLM-модель с пояснительной логикой на основе Chain-of-Thought (CoT), а в рабочем – пользователь получает результат без интерпретаций, применяя собственные навыки.

Далее в диссертационном исследовании будут представлены результаты экспериментов с дообучением предобученных языковых моделей BERT [79], DistilBERT [80], RoBERTa [81] на наборе данных REDDIT. Эти эксперименты направлены на оценку эффективности моделей в распознавании логических ошибок и манипулятивных техник.

В рамках экспериментов была сформулирована гипотеза о возможности представления признаков логических убеждений в виде сущностей (NER). На основе анализа новостных текстов, содержащих апелляцию к большинству, были выделены

характерные триггерные слова и фразы (например, «все знают», «по всеобщему мнению», «большинство считает», «широко распространено» и т. п.). Эти выражения были аннотированы в синтетическом наборе данных и использованы в процессе дообучения модели GLiNER [82]. Цель данного этапа заключалась в том, чтобы оценить применимость GLiNER в качестве инструмента для предварительного детектирования фрагментов текста с потенциальными манипуляциями, что впоследствии может быть интегрировано в архитектуру системы, поддерживающей когнитивную оценку новостей.

В рамках диссертации полная система не строится. Для обнаружения тональности существует множество подходов и моделей, так же, как и для выделения фактов. Но их проверка осталась за границами исследований.

3.2 Оценка предобученных языковых моделей на выявление логических ошибок: эксперимент на корпусе REDDIT

Как показано на диаграмме архитектуры системы, модели BERT, DistilBERT и RoBERTa могут быть использованы для анализа тональности текста. Но релевантные исследования (параграф 2.2) показывают, что трансформеры могут быть адаптированы к задачам выявления логических ошибок. В ходе экспериментов, результаты которых представлены в данном параграфе, доказано, что модели BERT-семейства демонстрируют высокую применимость в задачах обнаружения логических ошибок.

Тем не менее, в архитектуре системы для задачи подсветки логических убеждений в реальном времени предусмотрен отдельный модуль, основанный на модели GLiNER, ориентированной на распознавание заданных в наборе слов триггеров (сущностей). Сравнительный анализ между подходами BERT и GLiNER в контексте обнаружения логических ошибок не проводился и будет отнесен к перспективным направлениям дальнейших исследований автора.

Для проведения экспериментов с трансформерами использовался набор данных Reddit Informal Fallacies, представленный в репозитории на GitHub [83], а также описанный в статье [66]. Данный набор данных представляет собой размеченные англоязычные тексты, соответствующие или не соответствующие различным типам логических ошибок. Общая характеристика распределения данных в Reddit Informal Fallacies представлена в таблице 5.

Таблица 5 – Структура набора данных REDDIT.

Класс данных	Количество экземпляров		Процент от общего, %	
Appeal to authority	212	1708	2.18	17.52

Appeal to majority	196		2.01	
Appeal to nature	208		2.13	
Appeal to tradition	210		2.15	
Appeal to worse problems	239		2.45	
Black-or-white	211		2.17	
Hasty generalization	204		2.09	
Slippery slope	228		2.34	
Non fallacy	8042		82.48	
Итого	9750		100.00	

Отметим, что набор данных REDDIT является достаточно сбалансированным. При этом примеры данных, которые не содержат логические ошибки, были отобраны таким образом, чтобы быть похожими на логические ошибки, но все же не являться ими, что лучше, чем использование просто случайных текстов. Авторы [66] также отмечают, что их разметка осуществлялась экспертами, но все еще может содержать неточности, хотя и для небольшого процента данных.

В экспериментах решалась задача классификации текстов на основе размеченных данных. При этом в качестве текстов, относящихся или не относящихся к определенным классам логических ошибок, использовались как сами комментарии из социальной сети Reddit (COI, Comment of Interest, далее «Комментарий»), так и их окружение (контекст) – комментарий, на который был предоставлен «Комментарий» в качестве ответа (PAC, Parent Comment, далее «Ответ»), и заголовок поста, в котором появились «Комментарий» и «Ответ» (TOP, Title of Post, далее «Заголовок»).

Более того, решалась как задача бинарной классификации – отличить тексты с логическими ошибками от текстов без них, так и многоклассовой классификации, когда необходимо не только обнаружить логические ошибки, но и отличить один тип логической ошибки от другой. Для классификации текстов было решено использовать трансформеры, а именно базовый BERT и его модификации – DistilBERT и DeBERTa, см. таблицу 6. При этом модели выступали как токенизаторами текстовых данных, так и их классификаторами.

Таблица 6. Трансформеры выбранные для экспериментов в ходе исследования

Модель	Характеристики	Описание
BERT	12 слоев, размер слоя – 768, 110 млн. параметров	Модель, предварительно обученная на большом корпусе англоязычных данных. Обучение происходило на необработанных и неразмеченных текстах.
DistilBERT	6 слоев, размер слоя – 768, 67 млн. параметров	Представляет собой версию модели BERT после дистилляции данных. Данный процесс позволил снизить размерность модели и улучшить скорость ее работы.
RoBERTa	12 слоев, размер слоя – 768, 125 млн. параметров	Модель, предварительно обученная на большом корпусе англоязычных данных. Обучение происходило на основе маскированного языкового моделирования (MLM).

Таким образом, всего было проведено 18 экспериментов ($3 * 2 * 3$), в которых для каждой из 3 моделей решалась задача бинарной и многоклассовой классификации на {«Комментарий»}, {«Ответ»+ «Комментарий»}, и {«Заголовок» + «Ответ» + «Комментарий»}. Для экспериментов данные делились в отношении 70/20/10, т.е. 70% данных использовалось для обучения моделей, 20% для валидации, а оставшиеся 10% не были известны моделям и использовались только для их тестирования. Каждая модель обучалась 40 эпох при следующих параметрах: `batch_size` 32, `warmup_steps` 150, `weight_decay` = 0.01. В качестве метрики для отбора моделей в рамках обучения использовалась F-мера, т.к. данная метрика является наиболее сбалансированным показателем. В качестве среды для проведения экспериментов использовался Python 3.10.12 WSL со следующими версиями библиотек: `transformers` 4.48.3, `torch` 2.6.0, `scikit-learn` 1.5.2. Результаты и модели доступны на диске по ссылке [74].

Результаты экспериментов по бинарной классификации представлены в таблице 7 в **Приложении В**. Лучшие результаты по каждому классу по каждой отдельной метрике выделены **жирным**. При решении задачи обнаружения логических ошибок на основе «Комментариев», лучшие результаты показали RoBERTa в части обнаружения логических ошибок и DistilBERT в части определения их отсутствия. При этом добавление контекста в виде «Ответа» позволило улучшить результаты BERT, в то время как результаты RoBERTa и DistilBERT ухудшились. При этом улучшенный результат BERT (F-мера 0.4984)

превосходит результаты RoBERTa на основе «Комментария» (F-мера 0.4919) в задаче обнаружения логических ошибок, однако не превосходит результаты DistilBERT на основе «Комментария» (F-мера 0.9151) в задаче обнаружения их отсутствия. Дополнительный контекст в виде «Заголовка» преимуществ при решении задачи бинарной классификации не приносит.

Результаты по многоклассовой классификации представлены в таблицах 8-10: результаты каждой модели на основе «Комментариев» представлены в таблице 8, {«Ответ»+ «Комментарий»} в таблице 9 и {«Заголовок» + «Ответ» + «Комментарий»} в таблице 10 в **Приложении В**. По результатам эксперимента с {«Комментарий»}, BERT показал себя лучше других моделей при выявлении Hasty generalization; DistilBERT – Black-or-white; RoBERTa – Апелляция к большинству (Appeal to majority), «Оправдание через природу» (Appeal to nature), «Сравнение с худшими проблемами» (Appeal to worse problems), «Предсказание цепочки последствий» (Slippery slope) и «Отсутствие ошибок» (Non fallacy). Также BERT, DistilBERT и RoBERTa показали аналогичные результаты на «Апелляция к авторитету» (Appeal to authority), а DistilBERT и RoBERTa показали похожие результаты на «Апелляция к традициям» (Appeal to tradition).

В эксперименте {«Ответ»+ «Комментарий»} (см. таблицу 9 в **Приложении В**) лучшие результаты были получены для BERT в части «Сравнение с худшими проблемами» и «Представление только двух вариантов»; RoBERTa – «Апелляция к авторитету», «Апелляция к большинству», «Оправдание через природу», «Апелляция к традициям», «Поспешное обобщение» (Hasty generalization), «Предсказание цепочки последствий» и «Отсутствие ошибок». DistilBERT в данном эксперименте также показал неплохие результаты, однако не превзошел другие модели ни на одном из классов данных.

В эксперименте {«Заголовок» + «Ответ» + «Комментарий»} (см. таблицу 10 в **Приложении В**) лучшие результаты были получены для BERT в части Appeal to authority, Black-or-white, Slippery slope и Non fallacy; RoBERTa в части Appeal to majority и Hasty generalization; и DistilBERT в части Appeal to nature, Appeal to tradition и Appeal to worse problems.

Сравнение результатов лучших моделей на основе добавления контекста для каждого класса данных представлено в таблице 11 в **Приложении В**. Результаты показали, что для большинства классов (Appeal to authority, Appeal to majority, Appeal to nature, Slippery slope и Non fallacy) и решения задачи обнаружения логических ошибок в целом модели BERT, RoBERTa и DistilBERT лучше справляются на данных без контекста (только «Комментарий»). Однако, для классов Appeal to tradition и Black-or-white лучшие

результаты были достигнуты при {«Заголовок» + «Ответ» + «Комментарий»}, а для классов Appeal to worse problems и Hasty generalization при {«Ответ»+ «Комментарий»}.

Выводы по результатам экспериментов на релевантных наборах данных. Прежде всего эксперименты подтвердили, что трансформерные модели (BERT, DistilBERT, RoBERTa) эффективно выявляют логические ошибки в текстах. Компактность DistilBERT и высокая точность моделей позволяют интегрировать их в браузерное расширение для анализа новостей в реальном времени, минимизируя вычислительные затраты и поддерживая потоковую обработку текстов. При этом, достаточно высокая точность классификации и способность больших языковых моделей генерировать подсветку (H) и пояснения (CoT) обеспечивают реализацию архитектуры системы. Одновременно с этим результаты показывают, что необходима тонкая настройка моделей на новостных данных и оптимизация выбора контекста для повышения точности. Пример, продемонстрированный в статье [68] показывает, что можно обучить трансформер на релевантных наборах данных и потом уже дообучить их на синтетическом своем наборе данных.

3.3 Эксперименты по обучению модели GLiNER для NER-распознавания элементов манипулятивной риторики

В рамках диссертационного исследования была выдвинута гипотеза о том, что логические ошибки в новостных текстах проявляются через ограниченный и устойчивый набор слов и словосочетаний, которые могут быть охарактеризованы как триггерные. Эти триггеры представляют собой лексические конструкции, способные указывать на наличие логических убеждений и манипулятивных техник. Было предположено, что такие триггеры можно рассматривать как особый класс именованных сущностей, поскольку они имеют конкретную семантическую нагрузку и выполняют структурированную функцию в тексте.

Для проверки данной гипотезы был сформирован синтетический набор данных с вручную размеченными триггерами логических убеждений, в частности, типа «апелляция к большинству» («Appeal to majority»). На этом наборе данных были проведены эксперименты по обучению модели распознавания именованных сущностей (Named Entity Recognition, NER), в ходе которых оценивалась способность модели выделять такие триггеры как самостоятельные сущности.

Перед подачей на вход модели все текстовые данные (как новостные заметки, так и словосочетания-триггеры) проходили предварительную обработку, включавшую:

- замену символов новой строки и нижних подчеркиваний на пробелы;
- удаление ссылок, хештегов и упоминаний;

- замену всех числовых символов на пробелы;
- устранение избыточных пробелов.

После этого тексты подвергались процессу стемминга (SnowballStemmer("russian")), при котором слова заменялись на их корни. Данная манипуляция проводилась для повышения обобщаемости результатов и повышения вероятности обнаружения именованных сущностей.

Для подготовки размеченных данных были осуществлены следующие шаги:

1. Все заметки были преобразованы в массив слов.
2. Все словосочетания-триггеры для каждого отдельного текста были объединены в единый набор словосочетаний-триггеров апелляции к большинству.
3. На основе каждого словосочетания-триггера была размечена каждая заметка.

При этом разметка состояла из массива следующих элементов (листинг 1).

```
[  
    int: индекс слова заметки, с которого начинается словосочетание-  
        триггер,  
    int: индекс слова заметки, которым заканчивается словосочетание-  
        триггер,  
    str: класс сущности (в нашем случае 'Appeal to majority')  
]
```

Листинг 1 – Пример структуры разметки для задачи NER

Ситуации, где первые 2 числа (int) совпадают, означают, что словосочетание-триггер состоит из 1 слова.

Пример заметки и ее разметки приведен в листинге 2.

```
{  
    'tokenized_text': [  
        'рост', 'количество', 'этическ', 'кодекс', 'для', 'и', 'ответ', 'на',  
        'массов', 'вызов', 'с', 'увеличен', 'применен', 'и', 'в', 'разн',  
        'сфер', 'растет', 'и', 'числ', 'этическ', 'кодекс', 'и', 'стандарт',  
        'международн', 'организац', 'и', 'компан', 'разрабатыв', 'прав',  
        'чтоб', 'обеспеч', 'безопасн', 'прозрачн', 'и', 'ответствен', 'по',  
        'общ', 'мнен', 'имен', 'массов', 'распространен', 'и', 'треб', 'так',  
        'мер', 'для', 'предотвращен', ' злоупотреблен', 'в', 'массов', 'сознан',  
        'обществен', 'формир', 'пониман', 'что', 'этическ', 'норм', 'ключ',  
        'к', 'гармоничн', 'развит', 'технolog', 'и', 'сохранен', 'довер', 'к',  
        'ним'  
    ],
```

```
'ner': [
    [8, 8, 'Appeal to majority'], [36, 38, 'Appeal to majority'], [40, 40,
     'Appeal to majority'], [49, 51, 'Appeal to majority'], [50, 50, 'Appeal
     to majority']
]
```

Листинг 2 – Пример текстовой заметки и соответствующей разметки сущностей

Для задачи NER была выбрана мультиязычная модель GLiNER (версия multi-v2.1) на 209 млн параметров. Обучение производилось на разметке, разбитой в пропорции 70% для обучения, 20% для валидации и 10% – для тестирования. Использовались следующие параметры: batch_size=8, weight_decay=0.01, learning_rate=5e-6, warmup_ratio=0.1. Количество эпох обучения рассчитывалось и зависело от входных данных и размера батча.

В качестве среды для проведения экспериментов использовался Python 3.10.12 WSL со следующими версиями библиотек: transformers 4.48.3, torch 2.6.0, gliner 0.2.17.

Для проверки качества обнаружения именованных сущностей, связанных с апелляцией к большинству, тестовые данные (данные, которые не использовались при обучении или валидации модели) преобразовывались в текстовую строку.

Затем данный текст передавался на вход предобученной модели GLiNER со списком классов именованных сущностей, которых необходимо обнаружить (в нашем случае только 1 класс – "Appeal to majority"). В качестве степени уверенности модели устанавливался порог в 50%.

Результатом работы модели становятся обнаруженные сущности, для каждой из которых известно ее текстовое содержимое, символ старта и окончания в текстовой строке, а также степень уверенности модели в результатах обнаружения. Пример работы модели представлен ниже, в листинге 3.

```
публик упрекнул организатор за избыток видеофрагмент закрыт олимпиад
сопровожда активн использован видеовставок ключев элемент шо стал
предварительн записи сцен из лос анджелес где на фон знамени голливудск
холм том круг передава олимпийск флаг будущ хозяев игр однак пользовател
социальн сет обрат вниман что так количеств записи материа лиш церемон
ощущен присутств и динамик зрител на трибун ожида жив шо а не просмотр
фильм что вызва шквал недовольств по статистик бол чем отзыв на основн
платформ выраж разочарован люд куп билет на уникальн событ оказа перед
```

```
экран вместо сцен оживлен наступа только в жив момент с флаг реч и
фейерверк
- публик (0, 6) => Appeal to majority [0.8810372948646545]
- пользовател (269, 280) => Appeal to majority [0.8704999685287476]
- зрител (379, 385) => Appeal to majority [0.9707452654838562]
- по статистик (457, 469) => Appeal to majority [0.914125382900238]
```

Листинг 3. – Пример результата работы модели GLiNER: обнаружение триггеров апелляции к большинству в синтетической заметке.

Для расчета метрик эффективности на всем объеме тестовых данных, для каждого отдельного текста заметки собирались следующие метрики:

- True Positive (TP): обнаруженная сущность присутствует в разметке;
- False Positive (FP): обнаруженная сущность отсутствует в разметке;
- False Negative (FN): сущность из разметки не была обнаружена.

Информация о данных метриках позволила нам рассчитать такие классические метрики эффективности как Precision – 0.8743, Recall – 0.7002 и F-measure – 0.7776. Метрики, полученные в ходе эксперимента, свидетельствуют о достаточно высокой точности модели при умеренной полноте. Это означает, что модель в целом надёжно распознаёт те триггеры, которые действительно представлены в тексте, но часть целевых сущностей остаётся невыявленной. Такой баланс между точностью и полнотой приемлем на этапе прототипирования и разработки архитектуры, особенно с учётом ограниченного объёма синтетического корпуса. Отметим, что данные метрики рассчитывались на уровне полного обнаружения именованных сущностей (entity level) без учета частичного совпадения (token level).

Для дальнейшего повышения точности модели и расширения её применения возможно увеличение объёма синтетического обучающего набора, включая в него новые примеры различных логических убеждений. Также перспективным направлением является усложнение задачи, например, создание новостных текстов, не содержащих логических ошибок, но включающих слова-триггеры в нейтральном или ином контексте. Такая модификация позволит повысить дискриминативную способность модели и приблизить условия обучения к реальным сценариям.

Проведённые эксперименты подтвердили гипотезу о том, что триггеры логических убеждений, в частности апелляция к большинству, могут быть эффективно интерпретированы как именованные сущности и автоматически распознаны при помощи

современных NER-моделей. Использование синтетически размеченного набора данных позволило изолировать семантически значимые фрагменты текста и точно оценить способность модели GLiNER обнаруживать такие конструкции.

Полученные результаты демонстрируют, что подход к идентификации логических убеждений в новостных текстах, основанный на лексических паттернах и NER-анализе, является обоснованным и практически реализуемым. Это даёт основание включить компонент GLiNER в архитектуру предлагаемой обучающей системы, как показано в соответствующей диаграмме.

Выходы Глава 3

В третьей главе диссертационного исследования была разработана архитектура системы поддержки критического восприятия новостей, основанная на современных методах обработки естественного языка. Представленные блок-схемы отражают структуру взаимодействия между пользователем и интеллектуальной системой в двух режимах – обучающем и продвинутом. Для каждого из режимов были обозначены ключевые этапы обработки текста, включая предобработку, анализ тональности, распознавание сущностей и генерацию обратной связи. В архитектуру интегрированы конкретные технологические решения, включая стек используемых языков, библиотек и моделей, что позволяет обеспечить практическую реализуемость предложенного подхода.

Далее в главе был проведён экспериментальный анализ возможностей предобученных языковых моделей BERT, DistilBERT и RoBERTa по распознаванию логических ошибок в текстах. Модели демонстрируют уверенные результаты при дообучении на специализированном корпусе REDDIT, что подтверждает их применимость для задач выявления манипулятивных техник.

Особое внимание было уделено проверке гипотезы о том, что триггеры логических убеждений можно рассматривать как особый тип именованных сущностей. Для этого был создан синтетический набор данных с разметкой триггеров, характерных для категории «апелляция к большинству», и проведены эксперименты с моделью GLiNER. Полученные значения метрик (Precision — 0.8743, Recall — 0.7002, F1 — 0.7776) подтверждают, что модель способна эффективно выявлять такие триггеры, что открывает путь к дальнейшему расширению подхода на другие типы логических ошибок.

Таким образом, предложенная архитектура системы, подтверждённая экспериментами с предобученными моделями и моделью NER, отражает реализуемую стратегию по построению инструментов для поддержки критической оценки новостей пользователем.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена научная задача разработки системы поддержки критической оценки новостей, обеспечивающей обучение и подсветку фрагментов новостей, с учётом когнитивного профиля пользователя. Предложены архитектура и методический подход к созданию интеллектуальной системы, совместно обеспечивающие повышение критической оценки новостей у пользователей за счет выявления манипулятивных и логических приёмов в текстах. Достижение поставленной цели обеспечено посредством использования предобученных языковых моделей, методов обработки естественного языка и распознавания именованных сущностей (NER). Получены следующие результаты, составляющие итоги исследования:

1. Проведён обзор и анализ современного состояния исследований, систематизированы и классифицированы современные предобученные языковые модели, а также методы обработки естественного языка, применимые для задач критической оценки новостного контента.
2. Определены и формализованы когнитивные подходы (логический анализ, фактчекинг и анализ тональности), применяемые при критической оценке новостей, введено понятие когнитивного профиля пользователя.
3. Выделены и обоснованы релевантные признаки манипулятивного воздействия в новостных текстах, такие как логические ошибки (апелляция к большинству, апелляция к авторитету) и эмоционально-нагруженные фрагменты. Создан синтетический набор данных с новостями, размеченными вручную по наличию триггерных слов и словосочетаний, соответствующих логическим убеждениям. В набор данных вошли новости на 10 тем, общее количество текстов составило 1000 единиц.
4. Предложена архитектура интеллектуальной системы поддержки критической оценки новостей, предусматривающая адаптивные режимы взаимодействия с пользователем (обучающий и продвинутый) и содержащая отдельные модули анализа тональности, выделения фактов и распознавания триггерных выражений. Архитектура предусматривает персонализацию глубины объяснений в зависимости от когнитивного профиля пользователя и обеспечивает визуализацию результатов через интерфейс браузерной игры.
5. Выполнено экспериментальное тестирование существующих методов обнаружения логических ошибок (BERT, DistilBERT, RoBERTa) на общедоступном наборе данных REDDIT. Экспериментально подтверждена возможность применения данных

моделей для задач обнаружения логических ошибок. Установлено, что эти модели могут быть успешно переиспользованы после дообучения на релевантных наборах данных.

6. Подтверждена гипотеза о возможности использования метода распознавания именованных сущностей (NER) с моделью GLiNER для автоматизированного выявления триггерных слов и словосочетаний, характерных для логических убеждений.

Таким образом, все задачи, поставленные в диссертационном исследовании, были успешно решены, а цель, заключавшаяся в разработке и экспериментальном обосновании архитектуры системы поддержки критической оценки новостных текстов на основе предобученных языковых моделей, была достигнута.

Все результаты, полученные в ходе диссертационного исследования, обладают научной или исследовательской новизной. Предложенная архитектура системы поддержки критической оценки новостей отличается от существующих решений наличием режимов, обеспечивающих адаптацию к уровню пользователя, использованием методов выделения триггерных слов и интеграцией генеративных моделей для предоставления интерпретируемых пояснений. Разработанный синтетический набор данных с лексическими триггерами, применяемыми в качестве именованных сущностей, отличается от аналогов структурированным подходом к разметке и целенаправленным созданием текстов с логическими убеждениями. Экспериментальное подтверждение возможности распознавания логических убеждений через подходы NER является новым результатом, расширяющим методический аппарат медиаграмотности и когнитивной безопасности.

Практическая значимость работы заключается в создании системы, обеспечивающей пользователям автоматическую подсветку и интерпретацию потенциально манипулятивных фрагментов новостей.

Рекомендации по применению результатов исследования.

В области медиаграмотности. Предложенные решения могут быть интегрированы в образовательные платформы по медиаграмотности, информационно-аналитические системы поддержки решений, браузерные приложения и мессенджеры, а также системы обеспечения информационной и когнитивной безопасности. Кроме того, система может быть использована в учебных заведениях для развития критического мышления учащихся, в информационных центрах для оценки медиапространства, а также в корпоративных платформах повышения осведомлённости сотрудников о цифровой гигиене и информационной безопасности.

В области прикладной математики и информатики. Полученные результаты имеют прикладное значение для разработки интеллектуальных систем обработки естественного языка, рекомендательных сервисов и персонализированных интерфейсов на основе

предобученных языковых моделей. Подходы и методы, проверенные в рамках исследования (NER для выявления лексических триггеров, трансферное обучение и дообучение языковых моделей), могут быть применены в задачах автоматизированного анализа текстовых данных в различных доменных областях.

В качестве перспектив дальнейших исследований можно выделить расширение объёмов и тематик синтетического набора данных, создание более сложных тестовых примеров с триггерными словами, не связанными с логическими убеждениями, но включёнными в новостные тексты, усложнение задачу распознавания. Также перспективным направлением является исследование сравнительной эффективности различных подходов (BERT-подобные модели, LLM и NER-методы) на более разнообразных наборах данных и разработка интегрированных решений, сочетающих достоинства разных архитектур и методов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. 'Fake news' and a \$4 trillion swing: Inside the stock market's wild moves today // Business Insider URL: <https://www.businessinsider.com/stock-market-trump-tariffs-90-day-pause-sp500-rally-2025-4> (дата обращения: 25.05.2025).
2. Григорьев В. В. Информационный шум как угроза для человека и общества //Аналитические технологии в социальной сфере: теория и практика. – 2020. – С. 42.
3. Chen C., Shu K. Combating misinformation in the age of llms: Opportunities and challenges //AI Magazine. – 2024. – Т. 45. – №. 3. – С. 354-368.
4. Altay S., Berriche M., Acerbi A. Misinformation on misinformation: Conceptual and methodological challenges //Social media+ society. – 2023. – Т. 9. – №. 1. – С. 20563051221150412.
5. Быков И. А., Медведева М. В. Медиаграмотность как часть системы информационной безопасности //Вестник Государственного университета просвещения. Серия: История и политические науки. – 2020. – №. 1. – С. 24-32.
6. BadNews // Game URL: <https://www.getbadnews.com/en> (дата обращения: 25.05.2025).
7. CatPark // Game URL: <https://catpark.game/ru/> (дата обращения: 25.05.2025).
8. Виткова Л. А. Система поддержки критической оценки новостей с использованием больших языковых моделей // Кибербезопасность 4.1. Информационная безопасность компьютерных сетей URL: <https://ntk.sut.ru/events/schedule/kb-ibks#qmj0m0qpzksqeojc> (дата обращения: 25.05.2025).
9. Краснощёков П. С. Простейшая математическая модель поведения. Психология конформизма //Математическое моделирование. – 1998. – Т. 10. – №. 7. – С. 76-92.
10. Xu Z., Liu J., Başar T. On a modified DeGroot-Friedkin model of opinion dynamics //2015 American Control Conference (ACC). – IEEE, 2015. – С. 1047-1052.
11. Ouyang C. et al. The Impact of External Sources on the Friedkin–Johnsen Model //Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. – 2024. – С. 1815-1824.
12. Raineri R., Zino L., Proskurnikov A. FJ-MM: The Friedkin-Johnsen Opinion Dynamics Model with Memory and Higher-Order Neighbors //arXiv preprint arXiv:2504.06731. – 2025.

13. Козицин И. В. Обобщение модели Краснощекова на случай разложимой матрицы социальных связей //Математическое моделирование. – 2017. – Т. 29. – №. 12. – С. 3-15.
14. Козицин И. В. и др. Моделирование политических взглядов российских пользователей социальной сети ВКонтакте //Математическое моделирование. – 2019. – Т. 31. – №. 8. – С. 3-20.
15. Козицин И. В. и др. Эволюция социальных связей пользователей вконтакте и информационные пузыри //Управление развитием крупномасштабных систем MLSД'2020. – 2020. – С. 1865-1873.
16. Dartmouth Summer Research Project: The Birth of Artificial Intelligence [Электронный ресурс] URL: <https://www.historyofdatascience.com/dartmouth-summer-research-project-the-birth-of-artificial-intelligence/> (дата обращения: 09.05.2025)
17. Artificial Intelligence Coined at Dartmouth [Электронный ресурс] URL: <https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth> (дата обращения: 09.05.2025)
18. McCarthy J., Minsky M., Rochester N., Shannon C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955 [Электронный ресурс]. – URL: <http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf> (дата обращения: 09.05.2025).
19. Резаев А.В., Стариakov В.С., Иванова А.А. История искусственного интеллекта в СССР: институциональный контекст, вклад и значение работ ученых для современной науки // Социология науки и технологий. 2024. №4. URL: <https://cyberleninka.ru/article/n/istoriya-iskusstvennogo-intellekta-v-sssr-institutsionalnyy-kontekst-vklad-i-znachenie-rabot-uchenyh-dlya-sovremennoy-nauki> (дата обращения: 09.05.2025).
20. Математическая логика Учебное пособие по дисциплине «Математическая логика и теория алгоритмов» [Электронный ресурс]. – URL: books.ifmo.ru>file/pdf/1335.pdf (дата обращения: 09.05.2025).
21. Жизнь замечательных айтишников – Виктор Глушков [Электронный ресурс]. – URL: <https://habr.com/ru/companies/gaz-is/articles/753970/> (дата обращения: 09.05.2025).
22. Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM 9, 1 (Jan. 1966), 36–45. <https://doi.org/10.1145/365153.365168>

23. Berry, D. M. (2023). The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot. *Weizenbaum Journal of the Digital Society*, 3(3). <https://doi.org/10.34669/WI.WJDS/3.3.2>
24. MYCIN // Википедия. [2023]. Дата обновления: 06.12.2023. URL: <https://ru.wikipedia.org/?curid=2226031&oldid=134682078> (дата обращения: 06.12.2023).
25. AI History: the 1980s and expert systems // Klondike [Электронный ресурс]. – URL: <https://www.klondike.ai/en/ai-history-the-1980s-and-expert-systems/> (дата обращения: 09.05.2025).
26. IBM. Deep Blue - IBM [Электронный ресурс]. – URL: <https://www.ibm.com/history/deep-blue> (дата обращения: 09.05.2025).
27. Google AI Blog. AlphaGo's ultimate challenge: a five-game match against the legendary Lee Sedol [Электронный ресурс]. – URL: <https://blog.google/technology/ai/alphago-ultimate-challenge/> (дата обращения: 09.05.2025).
28. Rajpurkar P. et al. The Stanford Question Answering Dataset (SQuAD) [Электронный ресурс]. URL: <https://rajpurkar.github.io/SQuAD-explorer/> (дата обращения: 09.05.2025).
29. Отчёт по модели BERT для SQuAD 2.0 с подробным описанием fine-tuning и архитектуры: URL: <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15848021.pdf>
30. Papers With Code. SQuAD2.0 Benchmark (Question Answering) [Электронный ресурс]. URL: <https://paperswithcode.com/sota/question-answering-on-squad20> (дата обращения: 09.05.2025).
31. Jablonka K. M. et al. Is GPT all you need for low-data discovery in chemistry?. – 2023.
32. Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan. 2025. Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 24, 2, Article 12 (February 2025), 26 pages. <https://doi.org/10.1145/3708889>
33. Coleman T. Secret AI experiment on Reddit accused of ethical violations [Электронный ресурс] // The Week. – 2025. – URL: <https://theweek.com/tech/secret-ai-experiment-reddit> (дата обращения: 09.05.2025).
34. National study of high school students' digital skills paints worrying portrait, Stanford researchers say [Электронный ресурс] // Stanford Graduate School of Education. – 2025. – URL: <https://ed.stanford.edu/news/national-study-high-school-students-digital-skills-paints-worrying-portrait-stanford?print=all> (дата обращения: 09.05.2025).

35. Journalists highly concerned about misinformation, future of press freedoms [Электронный ресурс] // Pew Research Center. – 2022. – URL: <https://www.pewresearch.org/journalism/2022/06/14/journalists-highly-concerned-about-misinformation-future-of-press-freedoms/> (дата обращения: 09.05.2025).
36. Тоффлер Э. Шок будущего. – Litres, 2025.
37. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. – 2013. – Т. 26.
38. Morin F., Bengio Y. Hierarchical probabilistic neural network language model //International workshop on artificial intelligence and statistics. – PMLR, 2005. – С. 246-252.
39. Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.
40. Babu B. R., Ramakrishna S., Duvvuri S. K. Advanced Sentiment and Trend Analysis of Twitter Data Using CNN-LSTM and Word2Vec //2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL). – IEEE, 2025. – С. 1536-1543.
41. Ahmad I. A., Gatla P., Mundotiya R. K. Sarcasm Identification and Classification in Hindi Newspaper Headlines //ACM Transactions on Asian and Low-Resource Language Information Processing. – 2025. – Т. 24. – №. 4. – С. 1-21.
42. Pennington J., Socher R., Manning C. GloVe: Global Vectors for Word Representation [Электронный ресурс] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2014. – С. 1532–1543. – Режим доступа: <https://aclanthology.org/D14-1162.pdf>
43. Joulin A. et al. Bag of tricks for efficient text classification //arXiv preprint arXiv:1607.01759. – 2016.
44. Abduljaleel I. Q., Ali I. H. Textual Fake News Detection Based on FastText Embedding and Deep Learning //National Conference on New Trends in Information and Communications Technology Applications. – Cham : Springer Nature Switzerland, 2024. – С. 167-181.
45. Barua R., Rahman M. M., Joy U. G. Comparative analysis of Bangla news classification: a study of fake news detection and multiclass classification using BERT and FastText //International Journal of Computers and Applications. – 2025. – С. 1-11.
46. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate //arXiv preprint arXiv:1409.0473. – 2014.
47. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.

48. WMT 2014 // Papers With Code URL: <https://paperswithcode.com/dataset/wmt-2014> (дата обращения: 04.05.2025).
49. Лекция: cs224n-spr2024-lecture09-pretraining-updated.pdf // Stanford CS 224N | Natural Language Processing with Deep Learning URL: <https://web.stanford.edu/class/cs224n/slides/> (дата обращения: 04.05.2025).
50. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). – 2019. – С. 4171-4186.
51. WordPiece // Papers With Code URL: <https://paperswithcode.com/method/wordpiece> (дата обращения: 04.05.2025).
52. Виткова Л. А., Валиева К. А., Козлов Д. С. Подход к выявлению распространения ложной информации в интернете с использованием алгоритмов data science //Фундаментальные проблемы управления производственными процессами в условиях перехода к индустрии 4.0. – 2020. – С. 128-133.
53. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead //Nature machine intelligence. – 2019. – Т. 1. – №. 5. – С. 206-215.
54. Achiam J. et al. Gpt-4 technical report //arXiv preprint arXiv:2303.08774. – 2023.
55. Patil, R.; Gudivada, V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). Appl. Sci. 2024, 14, 2074. <https://doi.org/10.3390/app14052074>
56. Alonso M. A. et al. Sentiment analysis for fake news detection //Electronics. – 2021. – Т. 10. – №. 11. – С. 1348.
57. Антонова О.Г. Манипуляция как феномен информационного общества // Изв. Сарат. ун-та Нов. сер. Сер. Социология. Политология. 2023. №3. URL: <https://cyberleninka.ru/article/n/manipulyatsiya-kak-fenomen-informatsionnogo-obschestva> (дата обращения: 11.05.2025).
58. Измеров А.А. Сущность манипуляции в информационной среде общества // Общество: социология, психология, педагогика. 2024. №11. URL: <https://cyberleninka.ru/article/n/suschnost-manipulyatsii-v-informatsionnoy-srede-obschestva> (дата обращения: 11.05.2025).
59. Катенева И. Г. Механизмы и языковые средства манипуляции в текстах СМИ (на примере общественно-политических оппозиционных изданий) //АКД. Новосибирск. – 2010.

60. Гавра Д.П. // Высшая школа журналистики и массовых коммуникаций» СПбГУ URL: <https://jf.spbu.ru/business-pr/3887-15300.html> (дата обращения: 25.05.2025).
61. Доценко Е. Л. Психология манипуляции: феномены, механизмы и защита //М.: ЧеРо, Издательство МГУ. – 1997. – Т. 344. – С. 7.
62. Логическая ошибка // Википедия URL: https://ru.wikipedia.org/wiki/Логическая_ошибка (дата обращения: 25.05.2025).
63. Крысы, бомжи, бородатые женщины и Призрак Оперы: главные моменты открытия Олимпиады 2024 // Дзен URL: <https://dzen.ru/a/ZqSUDyxiTV9mB72p> (дата обращения: 25.05.2025).
64. Habernal I. et al. Argotario: Computational argumentation meets serious games //arXiv preprint arXiv:1707.06002. – 2017.
65. Martino G. D. S. et al. Fine-grained analysis of propaganda in news articles //arXiv preprint arXiv:1910.02517. – 2019.
66. Sahai S. Y., Balalau O., Horincar R. Breaking down the invisible wall of informal fallacies in online discussions //ACL-IJCNLP 2021-Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. – 2021.
67. Jin Z. et al. Logical fallacy detection //arXiv preprint arXiv:2202.13758. – 2022.
68. Shultz B. An entity-aware approach to logical fallacy detection in kremlin social media content //Proceedings of the International Conference on Advances in Social Networks Analysis and Mining. – 2023. – С. 780-783.
69. Alhindi T., Muresan S., Nakov P. Large language models are few-shot training example generators: A case study in fallacy recognition //arXiv preprint arXiv:2311.09552. – 2023.
70. Lim G., Perrault S. T. Evaluation of an LLM in identifying logical fallacies: A call for rigor when adopting LLMs in HCI research //Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing. – 2024. – С. 303-308.
71. Abhinav Lalwani, Tasha Kim, Lovish Chopra, Christopher Hahn, Zhijing Jin, Mrinmaya Sachan. Autoformalizing Natural Language to First-Order Logic: A Case Study in Logical Fallacy Detection // Computer Science > Computation and Language. – 2025. <https://doi.org/10.48550/arXiv.2405.02318>
72. Lei Y., Huang R. Boosting Logical Fallacy Reasoning in LLMs via Logical Structure Tree //arXiv preprint arXiv:2410.12048. – 2024.

73. Выявлен универсальный способ взлома ChatGPT, Gemini и других ИИ // ИНФОБЕЗОПАСНОСТЬ URL: <https://infobezopasnost.ru/blog/news/vyyavlen-universalnyj-sposob-vzloma-chatgpt-gemini-i-drugih-ii/> (дата обращения: 25.05.2025).

74. Данные для публикации // ЯндексДиск URL: <https://disk.yandex.ru/d/NYnw6AW16X3ZZw> (дата обращения: 25.05.2025).

75. Traberg C. S., Roozenbeek J., van der Linden S. Psychological inoculation against misinformation: Current evidence and future directions //The ANNALS of the American Academy of Political and Social Science. – 2022. – Т. 700. – №. 1. – С. 136-151.

76. Tilt Studio. Bad News Game — serious game about online misinformation // Tilt Studio URL: <https://www.tiltstudio.co/solutions/cases/bad-news-game> (дата обращения: 25.05.2025).

77. Modirrousta-Galian A. Improving news veracity discernment with inductive learning and gamification: New insights from receiver operating characteristic analysis : дис. – University of Southampton, 2025.

78. Hoes E. et al. Prominent misinformation interventions reduce misperceptions but increase scepticism //Nature Human Behaviour. – 2024. – Т. 8. – №. 8. – С. 1545-1553.

79. Google BERT-base uncased [Электронный ресурс]. – Режим доступа: <https://huggingface.co/google-bert/bert-base-uncased> (дата обращения: 08.06.2025).

80. DistilBERT-base-uncased [Электронный ресурс]. – Режим доступа: <https://huggingface.co/distilbert/distilbert-base-uncased> (дата обращения: 08.06.2025).

81. RoBERTa-base [Электронный ресурс]. – Режим доступа: <https://huggingface.co/FacebookAI/roberta-base> (дата обращения: 08.06.2025).

82. GLiNER Multi-v2.1 [Электронный ресурс]. – Режим доступа: https://huggingface.co/urchade/gliner_multi-v2.1 (дата обращения: 08.06.2025).

83. Sahaisaumya /informal_fallacies [Электронный ресурс]. – Режим доступа: https://github.com/sahaisaumya/informal_fallacies (дата обращения: 08.06.2025).

ПРИЛОЖЕНИЕ А

Пример новости 17.05.2025:

The screenshot shows a news article from the website Legalbet. The article title is "Экс-вратарь сборной Камеруна может продолжить карьеру в России". Below the title is a short summary: "Камерунский голкипер Рижской футбольной школы (РФШ) Фабрис Ондоа изучает вариант с продолжением карьеры в России. Как сообщил 237foot, вратарь получил предложение от неназванного клуба МИР Российской Премьер-Лиги (РПЛ). Legalbet". To the right of the main text, there is a smaller box with the word "ФУТБОЛ" and the same headline.

Пример ответа GPT-4o 17.05.2025

The screenshot shows a response from ChatGPT 4o. It includes a message from the user "ChatGPT 4o" and a reply from the AI. The AI's reply starts with "Сделай критическую оценку новости:" followed by a summary of the news article. Below this, it says "Вот критическая оценка представленной новости по ключевым критериям:" followed by a section titled "1. Достоверность источника" with two bullet points.

1. Достоверность источника

- Источник новости — 237foot. Это специализированный африканский футбольный портал, ориентированный на Камерун и другие франкоязычные африканские страны. Он не является международным признанным агентством или основным источником футбольных инсайдов, что снижает уровень доверия к информации.
- Отсутствие подтверждения в российских или европейских спортивных СМИ (например, «Спорт-Экспресс», «Чемпионат», Transfermarkt, L'Équipe) на момент публикации — признак низкой верифицированности.

2. Наличие конкретики

- Не указан клуб РПЛ, заинтересованный в трансфере. Это снижает степень конкретности, делает новость ближе к слуху.
- Не приводится информация об условиях предложения, интересе со стороны клуба, статусе переговоров или комментариях игрока или агента. Это усиливает ощущение неполноты информации.

3. Репутация и статус фигурантов

- Фабрис Ондоа — действительно бывший основной вратарь сборной Камеруна (особенно на Кубке Африки 2017), но с тех пор его карьера пошла на спад. На момент публикации он играет за РФШ (Латвия), что указывает на весьма скромный клубный уровень.
- Возможный переход в РПЛ может быть «шагом вверх», но не очевидно, насколько он соответствует требованиям клубов лиги, особенно в условиях квоты на легионеров.

4. Возможные цели публикации

- Новость может быть инициирована агентами игрока для повышения интереса к футболисту, создания ажиотажа или давления на другие клубы.
- Возможна гипотеза промоушена игрока на российский рынок — особенно через малоизвестный африканский источник.

5. Язык и стиль подачи

- Новость написана нейтрально, без эмоциональных утверждений, но из-за своей сжатости и недоказанности больше напоминает краткий слух, чем журналистское расследование или аналитический материал.

соответствует требованиям клубов лиги, особенно в условиях квоты на легионеров.

4. Возможные цели публикации

- Новость может быть инициирована агентами игрока для повышения интереса к футболисту, создания ажиотажа или давления на другие клубы.
- Возможна гипотеза промоушена игрока на российский рынок — особенно через малоизвестный африканский источник.

5. Язык и стиль подачи

- Новость написана нейтрально, без эмоциональных утверждений, но из-за своей сжатости и недоказанности больше напоминает краткий слух, чем журналистское расследование или аналитический материал.

Заключение

Критическая оценка: сомнительно / неполно

На текущем этапе новость нельзя считать подтвержденной. Отсутствие конкретики, отсутствие второго источника, слабый клубный уровень игрока и источник низкой известности — всё это указывает на то, что читатель должен относиться к материалу с осторожностью и не воспринимать его как факт.

Вероятность, что это информационныйброс или PR-ход, выше, чем вероятность подлинных переговоров.



Полезно ли это обсуждение на данный момент? X

Пример ответа Perplexity (17.05.2025)

The screenshot shows the Perplexity AI interface. At the top, there are navigation icons (back, forward, search, etc.) and the URL www.perplexity.ai. A banner at the top reads "Сделай критическую оценку новости: Экс-вратарь с...". The main area shows a user post by "iskinlidia41080" from 13 sec. ago. The post asks: "Сделай критическую оценку новости: Экс-вратарь сборной Камеруна может...". Below the post is a snippet of the news article: "Экс-вратарь сборной Камеруна может продолжить карьеру в России. Камерунский голкипер Рижской футбольной школы (РФШ) Фабрис Ондоа изучает вариант с продолжением карьеры в России. Как сообщил 237foot, вратарь получил предложение от неназванного клуба МИР Российской Премьер-Лиги (РПЛ)". The interface includes a sidebar with icons for "+", "home", "cover", and "faces". A "Answer" section is visible, along with "Sources · 18" and "Tasks". A message box says "Your answer was upgraded to Pro Search" and "Pro considers 3x more sources. Subscribe to increase your daily limit. 2 enhanced queries remaining today." It also features a "Upgrade to Pro" button. Below the message are several source cards: "Газета.Ru" (Vospitannik "Barcelona" может перейти в...), "ridus" (Агент: Ондоа заинтересовал клубы РПЛ...), "Чемпионат" (Воспитанник "Барселоны" и экс-вратарь сборной...), and "Bombardir.ru" (Клуб РПЛ может подписать вратаря сборной Камеруна). A summary below states: "Новость о возможном переходе экс-вратаря сборной Камеруна Фабриса Ондоа в клуб Российской Премьер-Лиги (РПЛ) заслуживает критической оценки с учётом нескольких ключевых аспектов." A section titled "Достоверность и полнота информации" contains a bulleted list of points:

- Источник новости - издание 237foot, которое сообщает о заинтересованности неназванного клуба РПЛ в 29-летнем голкипере, контракт которого с рижской командой РФШ заканчивается летом 2025 года [1](#) [3](#) [4](#). Однако название российского клуба не раскрывается, что снижает прозрачность информации и вызывает вопросы о степени реальности сделки.
- Агент футболиста Эрик Деполо подтвердил интерес клубов РПЛ, но подчеркнул, что Россия не является приоритетом для Ондоа, и сейчас футболист сосредоточен на выступлениях за РФШ [2](#) [5](#). Это указывает на то, что переговоры либо находятся на ранней стадии, либо имеют низкий приоритет, что следует учитывать при оценке новости.
- Помимо России, интерес к игроку проявляют клубы из Швейцарии и Турции, что говорит о наличии альтернативных вариантов для футболиста и снижает вероятность перехода именно в РПЛ [1](#) [2](#) [3](#) [5](#).

Спортивные и профессиональные характеристики игрока

- Фабрис Ондоа - опытный голкипер с богатой карьерой: воспитанник академии Самуэля Это'О и «Барселоны», выступал во вторых командах испанских клубов, а также в Бельгии, Хорватии, Франции и Латвии [1](#) [3](#) [4](#) [5](#).
- В сезоне 2024/2025 он провёл за РФШ 24-37 матчей (в зависимости от источника), включая 8 игр в Лиге Европы, с 29-39 пропущенными мячами и 7 матчами без пропущенных голов [1](#) [3](#) [4](#). Эти показатели указывают на стабильный уровень игры, но не выделяют его как звезду европейского уровня.
- Рыночная стоимость футболиста оценивается примерно в 500 тысяч евро, что является умеренной суммой для трансфера в РПЛ и соответствует уровню игрока, который может усилить клубы среднего или нижнего эшелона российского чемпионата [1](#) [3](#) [4](#).

Контекст и значимость для российского футбола

- Переход опытного международного вратаря с опытом выступлений в Лиге Европы и национальной сборной Камеруна может быть полезен для клубов РПЛ, особенно учитывая его статус свободного агента после окончания контракта [1](#) [3](#) [4](#).
- Однако отсутствие конкретики по клубу и приоритетность других вариантов для игрока ставят под сомнение оперативность и успешность данного трансфера.
- Также учитывая, что агент говорит о низком приоритете России для футболиста, новость может носить скорее информационный характер, чем свидетельствовать о реальном завершении сделки в ближайшее время [2](#) [5](#).

Итоговая оценка

Новость о возможном переходе Фабриса Ондоа в российский клуб РПЛ основана на достоверных, но неполных данных, с отсутствием конкретики по клубу и подтверждённым низким приоритетом России для самого игрока. Спортивный уровень и опыт футболиста делают его потенциально интересным усилением для некоторых российских команд, однако реальность перехода остаётся под вопросом. Таким образом, новость следует воспринимать как предварительную информацию о возможном варианте развития карьеры Ондоа, а не как факт состоявшегося или гарантированного трансфера.

Share Export Rewrite

Источники Perplexity

Сделай критическую оценку новости:

Экс-вратарь сборной Камеруна может продолжить карьеру в России
Камерунский голкипер Рижской футбольной школы (РФШ) Фабрис Ондоа изучает вариант с продолжением карьеры в России. Как сообщил 237foot, вратарь получил предложение от неназванного клуба МИР Российской Премьер-Лиги (РПЛ)

Answer Sources Tasks

1 Газета.Ru gazeta.ru/sport/news/202... Воспитанник «Барселоны» может перейти в российский клуб Российский клуб заинтересован в трансфере воспитанника «Барселоны»

2 ridus ridus.ru/agent--ondoa-z... Агент: Ондоа заинтересовал клубы РПЛ, но приоритет не в России

3 Чемпионат championat.com/football/news-... Воспитанник «Барселоны» и экс-вратарь сборной Камеруна ... Воспитанник «Барселоны» и экс-вратарь сборной Камеруна Ондоа может перейти в РПЛ — 237foot

4 Bombardir.ru bombardir.ru/news/772119-kl... Клуб РПЛ может подписать вратаря сборной Камеруна Голкипер РФШ Фабрис Ондоа может продолжить карьеру в России

5 РБК rbc.ru/sport/17/05/202... Агент воспитанника «Барселоны» рассказал об интересе клубов ... Эрик Деполо сообщил, что у клубов РПЛ есть интерес к вратарю сборной Камеруна Фабрису Ондоа, но это не приоритет для футболиста

6 sport.rambler sport.rambler.ru/football/54677... Клуб из РПЛ может подписать вратаря из Камеруна

7 sportmk sportmk.ru/sports/2025/05.. СМИ сообщили об интересе клуба РПЛ к воспитаннику - МК-Спорт

8 sport-express sport-express.ru/football/rpl/... Источник: российский клуб заинтересован в переходе ...

metaratings

Проверка источника 4

← ⏪ ⏴ bombardir.ru Клуб РПЛ может подписать вратаря сборной Кам... Пересказать

БОМБАРДИР ≡ РПЛ АПЛ МАТЧИ НОВОСТИ КОНКУРС ПРОГНОЗЫ

Главная › Новости › Клуб РПЛ может подписать вратаря сборной Камеруна

Клуб РПЛ может подписать вратаря сборной Камеруна

Сегодня, 12:48 1

Голкипер РФШ Фабрис Ондоа может продолжить карьеру в России. 29-летний камерунец через полтора месяца станет свободным агентом. Им интересуются неназванные клубы из РПЛ и швейцарской Суперлиги.

- Рыночная стоимость Ондоа – 500 тысяч евро.
- Вратарь прошел академию «Барселоны».
- В РФШ он с июня 2024 года, провел 37 матчей и пропустил 39 голов.
- Также в его активе 53 игры в составе сборной Камеруна.

Источник: 237foot

Пример ответа Grok (17.05.2025)

1. Режим «думающий»

The screenshot shows the Grok web interface. At the top, it says "NEW CONVERSATION - GROK". Below that is the Grok logo and the word "Grok". A large text box contains a message in Russian: "Сделай критическую оценку новости: Экс-вратарь сборной Камеруна может продолжить карьеру в России" followed by a detailed explanation in English about a Cameroonian goalkeeper named Fabrice Ondoa and his potential move to a Russian club.

Thinking 12c

goalkeeper who has played for various clubs in Europe, including Barcelona's youth team, and has represented Cameroon at the international level, winning the Africa Cup of Nations in 2017.

The article states that Ondoa's contract with RFS is set to expire in June 2025, and the club has offered him a new contract, but he has decided to leave. This suggests that he is looking for a new challenge or a better opportunity.

Now, let's consider the possibility of Ondoa moving to a Russian club. The RPL has been known to attract players from Africa, so it's not entirely unlikely. However, due to the current political situation and sanctions against Russia, it might be more challenging for players to move to Russian clubs.

I should also consider the timing of this news. The article was published on May 17, 2025, which is during the summer transfer

Ответ (рассуждения модели можно посмотреть)

A detailed critical assessment of a news story is shown. It starts with a summary: "Мысли для 28с" and "Развернуть для подробностей". The main section is titled "Критическая оценка новости: Экс-вратарь сборной Камеруна может продолжить карьеру в России". It discusses the potential move of Fabrice Ondoa from RFS to a Russian club, mentioning the RPL's history of attracting African players and the challenges posed by political sanctions. It also notes the timing of the news during the summer transfer period.

Достоверность источника

Публикация исходит от сайта 237foot, который позиционируется как камерунский футбольный новостной ресурс. Однако этот источник не имеет широкой международной репутации, что снижает уровень доверия к информации. В футбольной журналистике слухи и спекуляции — частое явление, особенно если они касаются игроков среднего уровня или не подкреплены данными от известных СМИ. Отсутствие подтверждения от крупных международных изданий или официальных заявлений клуба/игрока заставляет относиться к этой новости с осторожностью.

Кто такой Фабрис Ондоа?

Фабрис Ондоа — 29-летний вратарь из Камеруна, известный своим выступлением за сборную страны, с которой он выиграл Кубок африканских наций в 2017 году. Его карьера включает опыт в европейских клубах, таких как молодёжная команда «Барселоны», а в настоящее время он играет за латвийский РФШ. Согласно доступным данным, его контракт с клубом истекает в июне 2025 года, и, несмотря на предложение о продлении, Ондоа решил искать новые возможности. Это делает его потенциальный трансфер логичным шагом, но не подтверждает конкретно интерес со стороны России.

Реалистичность перехода в РПЛ

Российская Премьер-Лига традиционно привлекала игроков из Африки благодаря конкурентным зарплатам и уровню лиги. Однако текущая геополитическая ситуация, включая санкции против России, может осложнить трансферы. Европейские игроки и клубы зачастую избегают сотрудничества с российскими командами из-за политических и экономических рисков. Указание на «неназванный клуб» в новости также добавляет неопределённости — без конкретики сложно оценить серьёзность интереса.

Временной контекст

Новость датирована 17 мая 2025 года, что совпадает с началом летнего трансферного окна — периода, когда трансферные слухи и сделки наиболее активны. Это может косвенно поддерживать правдоподобность информации, но без дополнительных деталей (например, названия клуба или стадии переговоров) остаётся лишь предположением.

Вывод

Новость о возможном переходе Фабриса Ондоа в клуб РПЛ имеет определённые основания: у игрока заканчивается контракт, он ищет новый вызов, а российские клубы исторически интересовались африканскими талантами. Однако сомнительная репутация источника (237foot), отсутствие конкретных деталей и геополитические факторы ставят под вопрос достоверность и вероятность этого трансфера. Для подтверждения информации необходимо дождаться официальных заявлений от игрока, клуба или более авторитетных СМИ. На данный момент эту новость стоит воспринимать как слух, требующий проверки.

 25 веб-страниц

С ⌂ ↗ ⌛ ⌚ ...

Источники, которые приводит модель

Grok

Воспитанник "Барселоны" может продолжить карьеру в российском клубе - РИА Новости ...
Вратарь сборной Камеруна по футболу и латвийского РФС Фабрис Ондоа может продолжить карьеру в клубе из России, сообщает 237foot. РИА Новости Спорт, 17.05.2025
[gSport.ru.ru](#)

Футбол: история камерунского полузащитника «Ганновера» Газли Ондуа, который ...
Ондуа родился в столице Камеруна Яунде. Его мама работала воспитательницей в детском саду, а отец – в посольстве Камеруна в Москве. К папе Газли и переехал в детстве, хотя футболом начал заниматься еще на родине – он вспоминал, что босиком на песке гонял с друзьями мячи или имитацию...
[m.sport-express.ru](#)

Российский клуб заинтересован в переходе воспитанника «Барселоны» Ондоа 17.05.2025 ...
Камерунский вратарь Фабрис Ондоа, который сейчас играет за латвийский клуб РФШ, заинтересовал один из клубов Российской Премьер-Лиги
[soccer.ru](#)

Голкипер сборной Камеруна Ондоа не стал исключать, что в будущем перейдет в РПЛ
Голкипер сборной Камеруна Фабрис Ондоа не стал исключать, что будущем может продолжить карьеру в одном из российских клубов.
[matchtv.ru](#)

Вратарь Камеруна Ондоа: будущем могу оказаться в одном из клубов РПЛ.
Вратарь французского Нима и сборной Камеруна Фабрис Ондоа после матча с командой России (1:1) не исключил, что в будущем может оказаться в одном из клубов РПЛ. Могу ли когда-нибудь перейти в РПЛ? У
[rusfootball.info](#)

Клуб из РПЛ может подписать вратаря из Камеруна
Чемпион Африки-2017 Фабрис Ондоа может продолжить карьеру в Российской Премьер-Лиге.
[championat.com](#)

Ондоа, Фабрис – Википедия
18 марта 2017 года дебютировал за вторую команду «Севильи» в матче Второй лиги Испании против «Сарагосы», который закончился со счетом 2:1 в пользу андалусского клуба. Сам вратарь на 88 минуте заработал удаление (за две желтые карточки). Это удаление стало для него первым в карьере...
[ru.wikipedia.org](#)

Проверка новости. Источник № 1¹

← ⌛ 🔍 🔒 matchtv.ru Голкипер сборной Камеруна Ондоа не стал исключать, что в будущем ...

МАТЧ! ЭФИР ТЕЛЕПРОГРАММА ТРАНСЛЯЦИИ ВИДЕО НОВОСТИ ПОДПИСКИ ЭКСКЛЮЗИВ ПРИЛОЖЕНИЕ

Футбол Хоккей Биатлон UFC Фигурное катание Баскетбол Лыжные гонки Теннис Бокс Ещё

Новости футбола РПЛ РОССИЯ ЕВРОПА СБОРНЫЕ

ГЛАВНОЕ ТРАНСЛЯЦИИ НОВОСТИ ТРАНСФЕРЫ

ГОЛКИПЕР СБОРНОЙ КАМЕРУНА ОНДОА НЕ СТАЛ ИСКЛЮЧАТЬ, ЧТО В БУДУЩЕМ ПЕРЕЙДЕТ В РПЛ

17 октября 2023 13:23



1

https://matchtv.ru/football/matchtvnews_NI1964154_Golkiper_sbornoj_Kameruna_Ondoa_ne_stal_iskluchat_chto_v_budushhem_perejdet_v_RPL

Проверка новости. Источник №2²

РПЛ 2024/25

Вратарь Камеруна Ондоа: в будущем могу оказаться в одном из клубов РПЛ. Не исключаю такой факт

17 октября 2023 08:18

Вратарь французского "Нима" и сборной Камеруна Фабрис Ондоа после матча с командой России (1:1) не исключил, что в будущем может оказаться в одном из клубов РПЛ.

Фото: Getty Images

#	Команда	и	о
1	Краснодар	29	64
2	Зенит	28	60
3	ЦСКА	28	55
4	Динамо	28	53
5	Спартак	28	51
6	Локомотив	28	49
7	Рубин	28	42
8	Ростов	28	38
9	Акрон	28	35
10	Крылья Советов	28	30
11	Динамо Махачкала	28	27
12	Химки	28	26
13	Нижний Новгород	28	26
14	Ахмат	28	24
15	Оренбург	29	19
16	Факел	28	16

[Полная таблица](#) | [Календарь игр](#)

Проверка новости. Источник №3³

soccer.ru

ЗАВЕРШЕН 18' ИДЕТ 18' ИДЕТ 18' ИДЕТ 18:30, СЕГОДНЯ 1

ОРЕ 1 **ХИМ 1** **РБ 1** **МАЙ 1** **БОР 0** **КРИ - 4.00**

КРА 2 **РУБ 1** **ШТУ 0** **БАЙ 0** **ВОЛ 0** **МАН - 1.91**

НОВОСТИ ▾ БЛОГИ ▾ ТУРНИРЫ ▾ КОМАНДЫ ▾ КОНКУРСЫ ▾ ПРОГНОЗЫ РЕЙТИНГИ ▾ ЛИНИИ БК РЕЙТИНГ БК

Главная ▶ Новости ▶ Россия: трансферные слухи

Российский клуб заинтересован в переходе воспитанника «Барселоны» Ондоа

сегодня, 12:17

Камерунский вратарь **Фабрис Ондоа**, который сейчас играет за латвийский клуб РФШ, заинтересовал один из клубов Российской Премьер-Лиги. Какой именно — не сообщается. Ондоа присоединился к команде летом 2024 года и с тех пор провёл 27 матчей, в том числе помог команде выйти в групповую стадию Лиги Европы, отразив решающий пенальти в матче с АПОЭЛом.

Он — воспитанник академии «Барселоны», раньше играл во Франции, Хорватии, Бельгии и Латвии. За сборную Камеруна провёл 53 матча и выигрывал с ней Кубок Африки в 2017 году. Его контракт с РФШ действует до лета 2025 года.

[Подписывайся в ВК](#)



[Все новости](#)

РФШ ▪ Камерун ▪ АПОЭЛ ▪ Барселона ▪ Ондоа Фабрис

² <https://www.rusfootball.info/pliga/1146716692-vratar-kameruna-ondoa-v-buduschem-mogu-okazatsya-v-odnom-iz-klubov-rpl-ne-isklyuchayu-takoy-fakt.html>

³ <https://www.soccer.ru/news/1407150/ondoa-fabris-kamerun>

Проверка новости. Источник №4⁴

The screenshot shows the Championat.com website interface. At the top, there is a navigation bar with links for various sports like Football, Hockey, Tennis, etc., and a search bar. Below the navigation bar, there is a banner for the Russian Premier League (RPL) 2024-2025. The main content area displays news articles under the heading 'ГЛАВНЫЕ НОВОСТИ' (Main News). One article is highlighted: '16:44 Тренер «Краснодара» Мусаев — о волевой победе над «Оренбургом»: удовлетворило всё на 100%' (16:44 Trainer of Krasnodar Musaev — about the volitional victory over Orenburg: it was satisfying at 100%). There is also a section for the 'МАТЧ-ЦЕНТР' (Match Center) with a live tab for 'СЕЙЧАС (3)' (Now) showing three events: 'Мир Российской Премьер-лиги. 29-й тур' (World of Russian Premier League. 29th round) with matches between Orenburg vs Krasnodar (1:2), PFC Ural vs FC Rubin (1:1), and FC Nizhny Novgorod vs FC Fakel (1:1); a 'ФУТБОЛ' (Football) section showing 146 events; and a 'СКРЫТЬ LIVE-TAB' (Hide live tab) button.

Below the Match Center, there is a section for the 'РОССИЙСКАЯ ПРЕМЬЕР-ЛИГА' (Russian Premier League) with a logo of a bear. It shows the current season '2024/2025' and a dropdown menu for seasons. It also lists the date of the tournament '20.07.2024–24.05.2025', category 'Клубные национальные', and participants '16'. Below this, there are tabs for 'Турнир' (Tournament), 'Команды' (Teams), 'Календарь' (Calendar), 'Турнирная таблица' (Tournament Table), 'Результаты матчей' (Match Results), 'Статистика игроков' (Player Statistics), and '...' (More). The 'ТУРНИРНАЯ ТАБЛИЦА' (Tournament Table) tab is active, showing the top team 'Краснодар' (Rostov) in first place with 64 points after 29 matches. To the right, there is a 'Новости. РПЛ' (News. RPL) section with a recent update from '17 МАЯ 2025' about the coach of Krasnodar.

⁴ https://www.championat.com/football/_russiapl.html

ПРИЛОЖЕНИЕ Б

Пример промт-запроса к большой языковой модели и ответ

GPT 4o 17.05.2025

Логические ошибки и триггеры

ChatGPT 4o

Виды логических ошибок_Промт ча...

Документ

Изучи виды логических ошибок, сделай краткое summari и предложи слова триггеры, которые могут быть полезны для обнаружения логической ошибки апелляция к большинству до 10 слов.

Вот краткое summari видов логических ошибок на основе документа, а также триггерные слова и выражения, полезные для обнаружения ошибки апелляции к большинству (*argumentum ad populum*).

🔍 Краткое summari логических ошибок

Вид ошибки	Краткое описание
Подмена тезиса	Доказывается не исходный тезис, а близкий, но другой.
Предвосхищение основания	Утверждение основано на недоказанных посылках.
Мнимая логическая связь	Вывод не следует из доводов, несмотря на кажущуюся связь.
Ошибка выборки	Делается вывод на основе нерепрезентативной информации.
После этого — значит из-за этого	Временная последовательность принимается за причинность.

ChatGPT 4o

↑ Поделиться

причинность.

Корреляция ≠ причинность

Игнорируется влияние третьего фактора или случайности.

То же посредством того же

Круговое доказательство, где тезис заложен в аргументе.

Неопровергимость утверждений

Используются утверждения, которые невозможно проверить.

Ложная аналогия

Сравнение объектов по несущественным признакам.

Апелляции (к личности, к большинству, к эмоциям и т. п.)

Используются манипулятивные аргументы вместо логики.

Словесные уловки

Ошибки, основанные на омонимии, двусмысленности и прочем.

Эквивокация

Одно слово употребляется в разных смыслах в одном рассуждении.



Слова-триггеры для ошибки апелляция к большинству

(логическая ошибка: «если все так думают, значит, это правда»)

До 10 слов:

1. Все считают, что...
 2. Каждый знает...
 3. Большинство убеждено...
 4. Общеизвестно, что...
 5. Принято считать...
 6. Люди верят в то, что...
 7. Так думает весь народ...
 8. По мнению всех...
 9. Общее мнение таково...
 10. Так говорят все...

Эти выражения могут использоваться как индикаторы, что за ними может стоять ошибочное обоснование, апеллирующее к числу сторонников, а не к логике и фактам.

Второй промт и ответ 17.05.2025



Изучи новость из файла и найди логическую ошибку апелляция к большинству, выдели жирным в тексте новости триггерные слова, которые указывают на данную логическую ошибку

В тексте новости действительно присутствует логическая ошибка апелляция к большинству (*argumentum ad populum*) — аргументация строится на том, что нечто считается верным или значимым просто потому, что "все так думают" или "все так делают".

Вот отрывок с этой ошибкой, и я выделю **жирным** триггерные слова, указывающие на нее:

Принято считать, что церемония открытия Олимпийских Игр в Сочи в 2014 году была одной из лучших в новейшей истории. Да что там, все Игры в Сочи запомнились мировому сообществу – повторить грандиозный успех вряд ли будет возможно ещё много лет. Но французы попытались

Объяснение

Здесь утверждение о превосходстве сочинской церемонии обосновывается не аргументами (качеством постановки, рейтингами, эстетикой), а ссылкой на то, что "все запомнили", "принято считать", "мировое сообщество". Это типичный приём апелляции к мнению большинства, то есть попытка доказать ценность или правильность через ссылку на общее мнение.

Если нужно — могу разметить весь текст с выделением других логических ошибок.



Третий промт

Если нужно — могу разметить весь текст с выделением других логических ошибок.

Да, разметь весь текст, найди логическую ошибку апелляция к большинству, выдели жирным в тексте новости триггерные слова, которые указывают на данную логическую ошибку

Вот полностью размеченный текст новости с выделением триггерных слов, указывающих на логическую ошибку апелляция к большинству (*argumentum ad populum*). Жирным выделены выражения, использующие мнение толпы как аргумент:

В Париже завершилась церемония открытия летних Олимпийских Игр 2024. Неожиданное шоу, на котором выступила Леди Гага и Селин Дион, шокировало, удивляло и поставило множество вопросов о том, что будет с Европой дальше.

Принято считать, что церемония открытия Олимпийских Игр в Сочи в 2014 году была одной из лучших в новейшей истории. Да что там, все Игры в Сочи запомнились мировому сообществу – повторить грандиозный успех вряд ли будет возможно ещё много лет. Но французы попытались.

... (пропущена часть описания церемонии)

– Какое-то просто чудовищное открытие Олимпиады. Или я что-то не понимаю? Удивительно, как можно было не понять, что идея НЕ РАБОТАЕТ. В итоге – машем флагами на фоне картонных декораций по мутной реке в дождь... Мда... – пишет Ксения Собчак.

– Худшее открытие Олимпиады ever... дико скучно, провинциально и просто убого, — пишет муж Собчак Константин Богомолов.

... (пропущена часть описания водного парада)

– «Евровидение» теперь вполне можно вносить в программу Олимпийских игр. Быстрее, выше,

Реакция на церемонию открытия Олимпийских игр в Париже 2024

– Несмотря на мягко говоря не впечатляющее открытие (ладно, был красивый момент с флагом Франции, который все и постят). Все, кому открытие понравилось, постят именно его, потому что больше особо нечего.

... (остальные цитаты из зарубежной прессы не содержат апелляции к большинству, а представляют личные мнения экспертов)

💡 Примеры триггеров ошибки:

- Принято считать
- все Игры в Сочи запомнились мировому сообществу
- пишут в соцсетях
- все и постят
- все, кому открытие понравилось

Эти фразы не приводят содержательных аргументов, а лишь ссылаются на мнение большинства или предполагаемую массовую реакцию как на доказательство правильности или значимости оценки.

Если нужно — могу также разметить другие типы логических ошибок (например, *ad hominem*, анекдотическое свидетельство, ложная аналогия) в этом же тексте.

Perplexity 17.05.2025

The screenshot shows the Perplexity AI interface with two search results for "Изучи виды логических ошибок, сделай краткое саммари и предложи слова триггеры, которые могут быть полезны для обнаружения логической ошибки апелляция к большинству до 10 слов".

Result 1: A user post by [iskinlidia41080](#) from 9 seconds ago. The summary states: "Изучи виды логических ошибок, сделай краткое саммари и предложи слова триггеры, которые могут быть полезны для обнаружения логической ошибки апелляция к большинству до 10 слов". Below the summary, there is a note: "Your answer was upgraded to Pro Search" and a link to "Upgrade to Pro". An attachment is shown: "Attachment Vidy-logicheskikh-oshibok_Promt-chast-s-". The summary text is: "Вот краткое саммари основных видов логических ошибок и слова-триггеры для обнаружения апелляции к большинству (argumentum ad populum):". The detailed list follows:

- Подмена тезиса - доказательство другого тезиса вместо исходного, часто в форме «соломенного чучела».
- Предвосхищение основания (**petitio principii**) - использование недоказанных положений как аргументов, часто сопровождается фразами типа «Как абсолютно всем известно...».
- Минимая логическая связь (**non sequitur**) - довод не связан с заключением, несмотря на истинность обеих частей.
- Ошибка выборки - вывод о целом на основе неподходящей или слишком малой выборки (сверхобщение, ошибка выжившего).
- После этого - значит по причине этого (**post hoc ergo propter hoc**) - отождествление хронологии с причинностью.
- Корреляция как причинно-следственная связь - ошибочное приравнивание корреляции к причинности.

Result 2: Another user post by [iskinlidia41080](#) from 9 seconds ago. The summary states: "Изучи виды логических ошибок, сделай краткое саммари и предложи слова...". Below the summary, there is a note: "хронологии с причинностью.". The detailed list follows:

- Корреляция как причинно-следственная связь - ошибочное приравнивание корреляции к причинности.
- То же посредством того же (**idem per idem**) - в доказательство вводится сам доказываемый тезис.
- Использование принципиально неопровергаемых утверждений - аргументы, которые невозможно формально опровергнуть.
- Ложная аналогия - сравнение на основе несущественных сходств с игнорированием важных различий.
- Апелляция к личности (**ad hominem**) - критика личности вместо аргументов.
- Апелляция к большинству (**ad populum**) - утверждение, что мнение большинства истинно.
- Апелляция к авторитету (**ad verecundiam**) - ссылка на авторитет вместо доказательств.
- Апелляция к традиции (**ad antiquitatem**) - аргумент «так было всегда».
- Апелляция к эмоциям (**ad passiones**) - манипуляция чувствами вместо логики.
- Апелляция к страху (**ad baculum**) - угроза вместо аргументации.
- Апелляция к милосердию (**ad misericordiam**) - просьба о сочувствии вместо доказательств.
- Апелляция к незнанию (**ad ignorantiam**) - отсутствие доказательств принимается за доказательство.
- Апелляция к личному опыту (анекдотическое свидетельство) - обобщение на основе личного случая.
- Апелляция к природе (**ad Naturam**) - утверждение, что «естественное» - всегда хорошее.
- Словесные уловки - омонимия, двусмысличество, оксюморон, эквивокация и др.

Изучи виды логических ошибок, сделай краткое саммари и предложи слова...

Answer Sources - 1 Tasks

- Словесные уловки - омонимия, двусмысленность, оксюморон, эквивокация и др.

Триггерные слова и фразы для обнаружения апелляции к большинству
(*argumentum ad populum*)

- «Все считают, что...»
- «Большинство людей думают...»
- «Как все знают...»
- «Это мнение разделяют многие...»
- «Все вокруг говорят...»
- «Общественное мнение таково...»
- «Это популярное убеждение...»
- «Все поддерживают идею...»
- «Большинство согласны с тем, что...»
- «Это принято большинством...»

Эти слова и выражения часто сигнализируют о попытке опереться на мнение большинства вместо фактов и логики .

Share Export Rewrite



Второй промт и ответ 17.05.2025

Answer Sources - 1 Tasks

Изучи новость из файла и найди логическую ошибку апелляция к большинству, выдели жирным в тексте новости триггерные слова, которые указывают на данную логическую ошибку

Answer Sources - 11 Tasks

Attachment
Telekana78_novost_vtoroi_

sport-express
За что критикуют
церемонию открытия...

ФОНТАНКА.ру
Реакции на открытие
Олимпиады 2024 года во...

Чемпионат
Олимпийские игры – 2024 в
Париже, церемония...

В приведённой новости из файла логическая ошибка апелляция к большинству проявляется в утверждениях, где мнение большинства или большой группы людей представлено как аргумент истинности или правильности позиции.

Вот выделенный жирным текст с триггерными словами, указывающими на апелляцию к большинству:

«Принято считать, что церемония открытия Олимпийских Игр в Сочи в 2014 году была одной из лучших в новейшей истории. Да что там, все Игры в Сочи запомнились мировому сообществу – повторить грандиозный успех вряд ли будет возможно много лет. Но французы попытались.»

«– Все, кому открытие понравилось, постят именно его, потому что больше особо нечего. Но к серьёзному. За этой Олимпиадой будут следить все и везде. В том числе и наши чиновники, которые последний год провели в психологических атаках на наших спортсменов, дескать, ехать на Олимпиаду – это предательство. Наших мал... что они есть. И все мы знаем, что они наши и будем за них болеть...»

Эти фразы используют массовое мнение («все», «всемирное сообщество», «все, кому понравилось», «все мы знаем») как аргумент в пользу оценки церемонии, что является классическим примером логической ошибки апелляции к большинству.

Таким образом, в новости присутствуют триггерные слова и выражения: «принято считать», «все», «мировое сообщество», «все, кому понравилось», «все мы знаем», которые сигнализируют о попытке опереться на мнение большинства вместо объективного анализа.

Share Export Rewrite



Related

Какие конкретные слова или фразы указывают на логическую ошибку в тексте новости



Какие методы помогут выделять триггерные слова в тексте новости



Какие типы логических ошибок могут быть обнаружены в политических комментариях



Grok 3 17.05.2025

The screenshot shows a web browser window for 'grok.com' with the title 'New conversation - Grok'. The main area displays a conversation between a user and a AI system. The user has asked for examples of logical fallacies and provided a document for study. The AI has summarized the types of logical errors and listed them. The summary includes:

Изучи виды логических ошибок, сделай краткое саммари и предложи слова триггеры, которые могут быть полезны для обнаружения логической ошибки апелляция к большинству до 10 слов.

Виды логических ошибок_Промт ч...

Мысли для :
Развернуть для подробностей

Я изучил документ "Виды логических ошибок_Промт часть с обучением.docx" и подготовил краткое саммари основных типов логических ошибок, а также предложил слова-триггеры для обнаружения логической ошибки "апелляция к большинству". Вот ответ:

Краткое саммари видов логических ошибок

Документ описывает множество логических ошибок, которые нарушают корректность рассуждений. Ниже приведен краткий обзор основных типов:

- Подмена тезиса (Ignoratio elenchi):** Отклонение от исходного тезиса к доказательству другого, схожего, но отличного по смыслу (например, "соломенное чучело").
- Предвосхищение основания (Petitio principii):** Использование недоказанных утверждений как аргументов, часто с фразами вроде "как всем известно".
- Минимая логическая связь (Non sequitur):** Вывод не следует из посылок, даже если они истинны.
- Ошибка выборки:** Выводы на основе неподходящей выборки, включая сверхобщечтение и анекдотическое свидетельство.
- Посл этого – значит по причине этого (Post hoc ergo propter hoc):** Ошибочное принятие хронологической последовательности за причинно-следственную связь.
- Корреляция как причинно-следственная связь:** Предположение, что корреляция доказывает причинность, игнорируя случайности или третий фактор.
- То же посредством того же (Idem per Idem):** Круговое рассуждение, где вывод повторяет посылку.
- Использование неопровергимых утверждений:** Аргументы, которые невозможно проверить или опровергнуть.
- Ложная аналогия:** Сравнение объектов с недостаточным сходством, искаженное включение...

9. **Ложная аналогия:** Сравнение объектов с недостаточным сходством, игнорируя ключевые различия.
10. **Апелляция к личности (Ad hominem):** Атака на оппонента вместо его аргументов.
11. **Апелляция к большинству (Ad populum):** Утверждение, что нечто верно, потому что многие так считают.
12. **Апелляция к авторитету (Ad verecundiam):** Опирается на мнение авторитета, а не на доказательства.
13. **Апелляция к традиции (Ad antiquitatem):** Аргумент, что нечто верно, потому что так было давно.
14. **Апелляция к эмоциям (Ad passiones):** Манипуляция чувствами вместо логики.
15. **Апелляция к силе (Ad baculum):** Использование угроз для поддержки тезиса.
16. **Апелляция к милосердию (Ad misericordiam):** Вызов жалости для убеждения.
17. **Выдача желаемого за действительное:** Утверждение, что нечто истинно, потому что этого хочется.
18. **Апелляция к незнанию (Ad ignorantiam):** Утверждение, что отсутствие опровержения доказывает правоту.
19. **Аргумент к тошноте (Ad nauseam):** Повторение тезиса до изнеможения оппонента.
20. **Апелляция к личному опыту:** Опирается на субъективный опыт, а не на данные.
21. **Апелляция к природе (Ad Naturam):** Утверждение, что естественное — хорошо.
22. **Словесные уловки (по Аристотелю):** Ошибки, связанные с языком, такие как омонимия, амфиболия, соединение, разъединение, ударение и форма выражения.
23. **Эквивокация:** Использование одного слова в разных смыслах в одном рассуждении.

Слова-триггеры для обнаружения "апелляции к большинству"

"Апелляция к большинству" (*argumentum ad populum*) — это ошибка, при которой тезис считается верным только потому, что его поддерживает большинство. Вот 10 слов или фраз, которые могут сигнализировать об этой ошибке:

1. "Всем известно"
2. "Большинство считает"
3. "Широко принято"
4. "Общепризнано"
5. "Популярное мнение"
6. "Общеизвестная мудрость"
7. "Многие верят"
8. "Мнение общества"
9. "Консенсус гласит"
10. "Очевидно для всех"

Эти фразы указывают на попытку обосновать утверждение популярностью, а не логикой или

Чем я могу помочь?

Второй промт и ответ 17.05.2025

Изучи новость из файла и найди логическую ошибку апелляция к большинству, выдели жирным в тексте новости триггерные слова, которые указывают на данную логическую ошибку



Телеканал78_новость_второй_про...

Q Мысли для 77с
Развернуть для подробностей

Апелляция к большинству (*argumentum ad populum*) — это логическая ошибка, при которой утверждение считается истинным только потому, что его поддерживает большинство людей или оно является общепринятым. В тексте новости от телеканала "78" о церемонии открытия Олимпийских игр в Париже 2024 года можно найти несколько примеров этой ошибки. Автор часто ссылается на мнение большинства, популярные представления или реакции широкой аудитории, чтобы подкрепить свои доводы, вместо того чтобы предоставить объективные доказательства. Ниже представлен анализ текста с выделением **триггерных слов**, указывающих на апелляцию к большинству, и объяснением, как они демонстрируют эту логическую ошибку.

Текст новости с выделенными триггерными словами

Телеканал "78"

Крысы, бомжи, бородатые женщины и Призрак Оперы: главные моменты открытия Олимпиады 2024

27 июля 2024

В Париже завершилась церемония открытия летних Олимпийских Игр 2024. Неожиданное шоу, на котором выступила Леди Гага и Селин Дион, шокировало, удивляло и поставило множество вопросов о том, что будет с Европой дальше.

Принято считать, что церемония открытия Олимпийских Игр в Сочи в 2014 году была одной из лучших в новейшей истории. Да что там, все Игры в Сочи запомнились **мировому сообществу** — повторить грандиозный успех вряд ли будет возможно ещё много лет. Но французы попытались.

Церемония открытия Олимпийских игр в Париже 2024

Церемония открытия Олимпийских игр в Париже 2024

Когда стало известно о том, что Париж получил возможность провести Олимпиаду в 2024 году, то тут же стали продумывать план реализации. Надо было почистить Сену (задача оказалась невыполнимой), построить множество новых объектов (тут элегантно вышли из положения приспособив для этого старые здания), добиться экономического эффекта от этих строений (узнаем про это через годы) и как-то вписать сам Париж в церемонию открытия. Последнее получилось, но не совсем.

Обычно церемония открытия состояла из масштабного шоу, а также парада олимпийцев, шедших по стадиону. Затем были торжественные речи, а после зажигался Олимпийский огонь – **все** плакали, радовались и дальше переходили к самим соревнованиям.

Французы захотели соединить несоединимое: спорт, шоу, Париж, своих крыс (кадры с тем, как грызуны вылезают из люка, а бомжи лежат на улице **облетели весь мир!**), Сену, да ещё и добавить туда выступление Селин Дион и Леди Гаги.

В итоге получилось скомканное действие, которое на первый взгляд выглядит интересно, но если разбирать всё по отдельности, то получится какая-то каша.

Снуп Догг не закурил, а модель Onlyfans стала знаменосцем

Олимпийский огонь с утра по Парижу нёс Снуп Догг. Для тех, кто не в курсе, это американский рэпер (в столицу Франции он приехал как журналист), прославившийся употреблением всевозможных запрещённых курительных веществ. Как он ещё жив после этого **одному Богу известно**. И в соцсетях шутили, что Снуп может не сдержаться и прикурить прямо от Олимпийского огня. В данном случае обошлось.

Самая главная сцена церемонии была установлена на площади Трокадеро напротив Эйфелевой башни. Безусловно, Париж настолько красив, что сам является прекрасной декорацией для **любимой картинки**, но вписывать его в действие было затеей весьма сложной. И трудновыполнимой.

– Какое-то просто чудовищное открытие Олимпиады. Или я что-то не понимаю?

Удивительно, как можно было не понять, что идея **НЕ РАБОТАЕТ**. В итоге – машем флагами на фоне картонных декораций по мутной реке в дождь... Мда... – пишет Ксения Собчак.

– Худшее открытие Олимпиады ever... дико скучно, провинциально и просто убого, — пишет муж Собчак Константин Богомолов.

После выступления Леди Гаги (перед этим в центре города устроили свето-пиротехническое шоу с Призраком Оперы) **многие** уже поняли, что ждать чудес не приходится. Но начался водный парад по Сене: радостные спортсмены на достаточно простых лодках проплывали несколько километров и затем занимали свои места на трибунах. Из интересного можно отметить то, что знамя сборной Кипра несла Onlyfans-модель Елена Куличенко, а также многочисленная сборная США была представлена рослыми парнями из НБА.

«Бал сатаны» на Олимпиаде 2024



□

А после этого началась странная вакханалия с танцами на воде и, как это **принято в «свободной Европе»**, появились переодетые в женщин мужчин, бородатые женщины и голубое чудо, напоминающее то ли устрицу, то ли то, джина. В какой-то момент они даже сделали отсылку к знаменитой картине «Тайная вечеря».

– «Евровидение» теперь вполне можно вносить в программу Олимпийских игр. Быстрее, выше, пошлее, – пишут в соцсетях.

Судя по этой части, где случился «бал сатаны», уже можно было закрывать действие или переключать на другой канал (тут россиянам повезло, потому что федеральные каналы решили не закупать трансляции даже тех соревнований, где участвуют наши спортсмены под нейтральным флагом), но затем началось то, ради чего мы все собирались.

Всех спасла Селин Дион

В Париже стемнело, и организаторы наконец-то взялись за дело, обыграв присутствие Эйфелевой башни как декорации. Её эффектно подсветили, сделали классную светопроекцию и заставили с восхищением смотреть на шоу. Кто бы что ни говорил, но дождь внёс свой эффект в само представление (**самая популярная шутка в соцсетях**: «А в Париже тучи не умеют разгонять? Собянина на них нет!»), добавив изюминки открытию.

Завершилось всё прибытием Олимпийского огня, который пронесли мимо Лувра несколько великих французских спортсменов, а на финальном этапе его зажгли стрелой дзюдоиста Тедди Ринера и легкоатлетки Мари-Жозе Перек. Когда огонь разгорелся, он поднял в воздух воздушный шар, который теперь красуется над французской столицей.

Зафиналили историю хитом великой Эдит Пиаф, который исполнила не менее великая Селин Дион. Это её первое выступление за долгое время после объявление о приостановке карьеры и сообщений о неизлечимом недуге. Даже в таком состоянии Селин в буквальном смысле вытащила всю церемонию открытия, выступив на Эйфелевой башне так, что наворачивались слезы.

Реакция на церемонию открытия Олимпийских игр в Париже 2024

– Несмотря на мягко говоря не впечатляющее открытие (ладно, был красивый момент с флагом Франции, который **все** и посят). **Все, кому открытие понравилось, посят именно его**, потому что больше особо нечего. Но к серьёзному. За этой Олимпиадой будут следить **все** и **везде**. В том числе и наши чиновники, которые последний год провели в психологических атаках на наших спортсменов, дескать, ехать на Олимпиаду – это предательство. Наших мало, но они есть. И **все** мы знаем, что они наши и будем за них болеть, хоть и федеральное ТВ решило ничего не показывать, – пишет Ксения Собчак в соцсетях.

Мировая пресса тоже сдержанно встретила попытку французов выделиться на всеобщем фоне и провести необычную церемонию открытия.

– Пока рано называть это переосмысление церемонии – успехом или провалом. Традиционно церемония открытия Олимпийских игр всегда представляла собой



Мировая пресса тоже сдержанно встретила попытку французов выделиться на всеобщем фоне и провести необычную церемонию открытия.

– Пока рано называть это переосмысление церемонии – успехом или провалом. Традиционно церемония открытия Олимпийских игр всегда представляла собой потрясающую зрелищную музыку и театр в течение 45 минут – часа, затем наступало небольшое затишье, парад длился долго-долго, после чего произносились речи и зажигался факел. Это попытка смешать всё это, но должен признаться, что порой трудно понять, куда смотреть, и пока что не хватает общей сюжетной нити, помимо доставки огня, – рассуждают эксперты *The Guardian*.

– Гага и Селин, евродиско и бальные вогеры, трансвеститы, позирующие как апостолы в Тайной вечере, набережные Сены, покрытые розовым: церемония Томаса Джолли обещала быть грандиозной, но это был также фестиваль странностей в Париже, выглядящий сказочно под дождём. **Политики крайне правого толка осудили церемонию** ещё до её начала. И в разгар продолжающегося политического кризиса эта церемония ясно дала понять, в какую Францию она верит, – пишет *The New York Times*.

Объяснение триггерных слов и примеров апелляции к большинству

1. "Шокировало, удивляло"

- Указывает на общую реакцию аудитории, подразумевая, что большинство людей испытывали шок и удивление, что используется для подтверждения значимости шоу.

2. "Принято считать"

- Фраза напрямую ссылается на общепринятое мнение о том, что церемония в Сочи была одной из лучших, без предоставления объективных критериев.

3. "Мировому сообществу"

- Утверждает, что успех Игр в Сочи признан всем миром, апеллируя к коллективному мнению как доказательству их величия.

4. "Все"

- Многократное использование слова "все" (например, "все плакали, радовались", "все и постят", "все мы знаем") предполагает универсальную реакцию или действие, что усиливает аргумент за счёт мнения большинства.

5. "Облетели весь мир"

- Подразумевает, что кадры с крысами и бомжами стали известны всем, используя их популярность как подтверждение значимости.

6. "Одному Богу известно"

- Выражение, подразумевающее общее удивление или непонимание, что косвенно апеллирует к коллективному восприятию.



7. "Любимой картинки"

- Ссылка на общее восхищение Парижем как красивым городом, что используется для обоснования его роли в церемонии.

8. "Многие"

- Указывает на то, что большинство людей разочаровались после выступления Леди Гаги, что подкрепляет негативную оценку шоу.

9. "Принято в «свободной Европе»"

- Подразумевает общепринятые нормы в Европе, используемые для объяснения появления необычных элементов в шоу.

10. "Пишут в соцсетях"

- Ссылка на мнение пользователей социальных сетей как на коллективное суждение, подтверждающее критику церемонии.

11. "Судя по этой части"

- Подразумевает общее впечатление от "бала сатаны", что используется для обоснования его провальности.

12. "Заставили с восхищением смотреть"

- Утверждает, что все зрители восхищались шоу с Эйфелевой башней, апеллируя к массовой реакции.

13. "Самая популярная шутка в соцсетях"

- Ссылается на популярность шутки как на индикатор общего мнения о дожде.

14. "Наворачивались слезы"

- Подразумевает массовую эмоциональную реакцию на выступление Селин Дион, усиливая её значимость.

15. "Мировая пресса"

- Указывает на мнение мировой прессы как на коллективное суждение, подтверждающее сдержанную реакцию на церемонию.

16. "Политики крайне правого толка осудили церемонию"

- Ссылка на мнение определённой группы как на значимый фактор, подразумевающий поддержку этого мнения многими.

Вывод

В тексте новости автор неоднократно прибегает к апелляции к большинству, используя мнения толпы, общепринятые представления или реакции масс для подкрепления своих утверждений о качестве церемонии открытия Олимпийских игр 2024 года в Париже.

Чем я могу помочь?

Вывод

В тексте новости автор неоднократно прибегает к апелляции к большинству, используя мнения толпы, общепринятые представления или реакции масс для подкрепления своих утверждений о качестве церемонии открытия Олимпийских игр 2024 года в Париже.

Выделенные **триггерные слова** демонстрируют, как автор опирается на популярность или всеобщность суждений вместо объективных фактов, что является классическим примером этой логической ошибки.

 25 веб-страниц

🕒 ⏷ ⏵ ⏴ ⏵ ...

ПРИЛОЖЕНИЕ В

Таблицы из диссертации

Таблица 1 – Релевантные исследования, сортированные по годам публикаций.

Статья	Цели и методы	Наборы данных
Argotario: Computational Argumentation Meets Serious Games [64] 2017	Разработана игра Argotario для распознавания и аннотирования логических ошибок. Геймификация с взаимодействием игроков против игрока. Приложение, ручная разметка, без языковых моделей. Создан набор данных	ARGOTARIO Создан авторами статьи
Fine-Grained Analysis of Propaganda in News Articles [65] 2019	Обнаружение конкретных техник пропаганды на уровне фрагментов в тексте. Использовались BERT, BERT-Joint, BERT-Granularity и многоуровневая нейронная сеть с иерархическим подходом и обучаемыми гейтами (ReLU или Sigmoid) Предложена собственная нейронная сеть, которая сравнивается с BERT	CoNLL'02 CoNLL'03 Создан корпус из 451 новостной статьи «Propaganda Techniques Corpus». Это название используется в контексте задач, связанных с проектом Tanbih и платформой QCRI, общепринятого названия, устоявшегося у набора нет
Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions [66] 2021	Создание набора данных с ложными убеждениями (аргументацией), обучение нейронной сети BERT, Multi-Granularity Network (из [6])	REDDIT Comments (создан авторами статьи)
Logical Fallacy Detection [67] 2022	Поставлена научная задача обнаружения логических ошибок в тексте. Использовались базовые модели: zero-shot (RoBERTa-large, BART-large, TARS, GPT-2, GPT-3) и дообученные (ALBERT, BERT, BigBird, DeBERTa, DistilBERT, Electra, MobileBERT, RoBERTa) с использованием фреймворков NLI Предложенный метод: структурированный классификатор на основе Electra с дистилляцией структуры, маскировкой содержательных слов и гипотезами, специфичными для типов ошибок, с использованием CoreNLP, Stanza и Sentence-BERT.	LOGIC CLIMATE Созданы авторами статьи
An Entity-Aware Approach to Logical Fallacy Detection in Kremlin Social Media Content [68] 2023	Предложен подход Entity-Aware Approach (EAA) для обнаружения логических ошибок в контенте социальных сетей. EAA - предобработка текста с заменой именованных сущностей на метки типов с использованием zero-shot NER от FlairNLP. Тестирували пять предобученных моделей: BERT (cased/uncased), Electra, ALBERT, DeBERTa. Обучались на LOGIC и делали прямой перенос и дообучение на RuFal	LOGIC RuFal (создан авторами статьи)
Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition [69] 2023	Улучшение распознавания логических ошибок за счет генерации синтетических данных к наборам данных. Методы: аугментация данных с использованием GPT-3.5 (gpt-3.5-turbo) в режимах zero-shot, 1-shot, 2-shot и 5-shot, с контекстом и без. Дообучение модели T5-3B	ARGOTARIO PROPAGANDA LOGIC COVID CLIMATE
Evaluation of an LLM in Identifying	Обнаружение логических ошибок в наборах с использованием GPT-4 (API)	LOGIC

Статья	Цели и методы	Наборы данных
Logical Fallacies: A Call for Rigor When Adopting LLMs in HCI Research [70] 2024		
Autoformalizing Natural Language to First-Order Logic: A Case Study in Logical Fallacy Detection [71] 2024-2025 (вносились правки авторами)	<p>Предложен фреймворк NL2FOL для перевода естественного языка в логику первого порядка (FOL) с использованием LLM, для решения проблемы интеграции неявных знаний.</p> <p>Фреймворк NL2FOL с тремя модулями: перевод NL в FOL, преобразование FOL в SMT, интерпретация результатов SMT.</p> <p>Использование LLM (Llama-7B, GPT-4o-mini, GPT-4o) для парсинга, генерации и обработки неоднозначностей.</p> <p>SMT-сольверы (CVC4) для проверки логической валидности и генерации контрпримеров.</p>	LOGIC CLIMATE
Boosting Logical Fallacy Reasoning in LLMs via Logical Structure Tree [72] 2024	Предложено дерево структуры для иерархического представления логической структуры потока в высказываниях	ARGOTARIO LOGIC CLIMATE REDDIT

Таблица 3. Сортировка релевантных методов и моделей для обнаружения логических ошибок.

Статья	Ключевой метод	Модели	Улучшение
Fine-Grained Analysis of Propaganda in News Articles [65]	Введены задачи классификации на уровне предложений (SLC) и фрагментов (FLC). Разработана новая метрика оценки, учитывающая частичные перекрытия.	BERT, MGN	Multi-Granularity Network (MGN) улучшила точность обнаружения пропаганды
Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions [66]	Введены четыре задачи классификации (на уровне комментариев и токенов). Использовались контекстные модели, включающие родительские комментарии и заголовки.	BERT, MGN	Контекстные модели повысили производительность в онлайн-дискуссиях
Logical Fallacy Detection [67]	Разработаны наборы данных. Предложен структурированный классификатор, использующий маскировку содержательных слов для выделения логической формы	Оценены 12 моделей: zero-shot (TARS, BART-MNLI, RoBERTa-MNLI, GPT-2, GPT-3) и finetuned (ALBERT, BERT, BigBird, DeBERTa, DistilBERT, Electra, MobileBERT). Лучшая модель – Electra-StructAware.	Доказали важность анализа логической структуры, а не только содержимого текста в задачах обнаружения логических ошибок
An entity-aware approach to logical fallacy detection in kremlin social media content [68]	Предложен Entity-Aware Approach (EAA), заменяющий именованные сущности (люди, организации) на метки типов с помощью FlairNLP	BERT (cased/uncased), Electra, ALBERT, DeBERTa. Лучшая модель – DeBERTa с EAA	EAA уменьшил путаницу от множественных эмбеддингов для одних и тех же сущностей. Показали эффективность

Статья	Ключевой метод	Модели	Улучшение
			предобработки для специфичных доменов (социальные сети)
Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition [69] 2023	Использован GPT-3.5 для генерации синтетических данных в режимах zero-shot, 1-shot, 2-shot и 5-shot. Модель T5-3B обучалась на комбинации оригинальных и синтетических данных	GPT-3.5 (генерация), T5-3B (классификация)	Синтетические данные, особенно в 1-shot режиме, улучшили распознавание редких ошибок, таких как Equivocation и Whataboutism. По сути показали потенциал LLM для аугментации данных в задачах с дисбалансом классов
Evaluation of an LLM in Identifying Logical Fallacies: A Call for Rigor When Adopting LLMs in HCI Research [70]	Использован набор данных LOGIC, отфильтрованный до 780 образцов семи типов ошибок. Применено few-shot prompting с тремя примерами на ошибку.	GPT-4	Тщательная оценка и настройка подсказок повысили надежность применения LLM. По сути оценивали возможности GPT-4 для распознавания логических ошибок
Autoformalizing Natural Language to First-Order Logic: A Case Study in Logical Fallacy Detection [71]	Разработан фреймворк NL2FOL с тремя модулями: перевод NL в FOL, FOL в SMT, интерпретация результатов. Использованы few-shot prompting и SMT-солверы (CVC4)	Llama-7B, GPT-4o-mini, GPT-4o. Основная модель – GPT-4o	Интеграция формальной логики, учет неявного знания и пошаговая обработка повысили точность и интерпретируемость. Это открывает новые возможности для сложных задач
Boosting Logical Fallacy Reasoning in LLMs via Logical Structure Tree [72]	Логические структурные деревья	Llama-2, Flan-T5-large, с базовыми моделями GPT-3.5 и RoBERTa	NL2FOL и логические деревья повысили интерпретируемость и точность. Показали значительные улучшения для обнаружения ошибок, зависящих от логических отношений, таких как Red Herring

Таблица 7. Результаты бинарной классификации

Класс	Модель	Precision	Recall	F-measure
Эксперимент 1: {«Комментарий»}				
Fallacy 171 экземпляров	BERT	0.5156	0.3860	0.4415
	DistilBERT	0.6489	0.3567	0.4604
	RoBERTa	0.5507	0.4444	0.4919
Non fallacy 804 экземпляра	BERT	0.8760	0.9229	0.8988
	DistilBERT	0.8751	0.9590	0.9151
	RoBERTa	0.8865	0.9229	0.9043
macro avg 975 экземпляров	BERT	0.6958	0.6544	0.6702
	DistilBERT	0.7620	0.6578	0.6878
	RoBERTa	0.7186	0.6837	0.6981
weighted avg 975 экземпляров	BERT	0.8128	0.8287	0.8186
	DistilBERT	0.8355	0.8533	0.8354
	RoBERTa	0.8276	0.8390	0.8320
Эксперимент 2: {«Ответ»+ «Комментарий»}				
Fallacy 171 экземпляров	BERT	0.5493	0.4561	0.4984
	DistilBERT	0.4253	0.4327	0.4290
	RoBERTa	0.4667	0.4094	0.4361
Non fallacy 804 экземпляра	BERT	0.8884	0.9204	0.9041
	DistilBERT	0.8789	0.8756	0.8773
	RoBERTa	0.8776	0.9005	0.8889
macro avg 975 экземпляров	BERT	0.7188	0.6883	0.7012
	DistilBERT	0.7230	0.6688	0.6887
	RoBERTa	0.7134	0.6910	0.7009
weighted avg	BERT	0.8289	0.8390	0.8329

Класс	Модель	Precision	Recall	F-measure
975 экземпляров	DistilBERT	0.8252	0.8410	0.8301
	RoBERTa	0.8279	0.8359	0.8314
<i>Эксперимент 3: {«Заголовок» + «Ответ» + «Комментарий»}</i>				
Fallacy 171 экземпляр	BERT	0.5349	0.4035	0.4600
	DistilBERT	0.4253	0.4327	0.4290
	RoBERTa	0.4667	0.4094	0.4361
Non fallacy 804 экземпляра	BERT	0.8794	0.9254	0.9018
	DistilBERT	0.8789	0.8756	0.8773
	RoBERTa	0.8776	0.9005	0.8889
macro avg 975 экземпляров	BERT	0.7072	0.6644	0.6809
	DistilBERT	0.6521	0.6542	0.6531
	RoBERTa	0.6721	0.6549	0.6625
weighted avg 975 экземпляров	BERT	0.8190	0.8338	0.8243
	DistilBERT	0.7993	0.7979	0.7986
	RoBERTa	0.8055	0.8144	0.8095

Таблица 8. Результаты многоклассовой классификации «Комментариев»

Класс	Модель	Precision	Recall	F-measure
Appeal to authority 21 экземпляр	BERT	0.6429	0.4286	0.5143
	DistilBERT	0.6429	0.4286	0.5143
	RoBERTa	0.6429	0.4286	0.5143
Appeal to majority 20 экземпляров	BERT	0.7000	0.3500	0.4667
	DistilBERT	0.5333	0.4000	0.4571
	RoBERTa	0.5263	0.5000	0.5128
Appeal to nature 21 экземпляр	BERT	0.6316	0.5714	0.6000
	DistilBERT	0.5385	0.3333	0.4118

Класс	Модель	Precision	Recall	F-measure
	RoBERTa	0.7647	0.6190	0.6842
Appeal to tradition 21 экземпляр	BERT	0.6154	0.3810	0.4706
	DistilBERT	0.5263	0.4762	0.5000
	RoBERTa	0.6000	0.4286	0.5000
Appeal to worse problems 24 экземпляра	BERT	0.3125	0.2083	0.2500
	DistilBERT	0.5455	0.2500	0.3429
	RoBERTa	0.4286	0.3750	0.4000
Black-or-white 21 экземпляр	BERT	0.2222	0.0952	0.1333
	DistilBERT	0.4444	0.1905	0.2667
	RoBERTa	0.5000	0.1429	0.2222
Hasty generalization 20 экземпляров	BERT	0.2272	0.2500	0.2381
	DistilBERT	0.1429	0.1500	0.1463
	RoBERTa	0.2222	0.1000	0.1379
Slippery slope 23 экземпляра	BERT	0.5455	0.2609	0.3529
	DistilBERT	0.6250	0.2174	0.3226
	RoBERTa	0.6667	0.4348	0.5263
Non fallacy 804 экземпляра	BERT	0.8769	0.9391	0.9069
	DistilBERT	0.8786	0.9453	0.9107
	RoBERTa	0.8941	0.9552	0.9236
macro avg 975 экземпляров	BERT	0.5305	0.3872	0.4370
	DistilBERT	0.5419	0.3768	0.4303
	RoBERTa	0.5828	0.4427	0.4913

Класс	Модель	Precision	Recall	F-measure
weighted avg 975 экземпляров	BERT	0.8082	0.8297	0.8138
	DistilBERT	0.8129	0.8328	0.8159
	RoBERTa	0.8329	0.8544	0.8386

Таблица 9. Результаты многоклассовой классификации {«Ответ»+ «Комментарий»}

Класс	Модель	Precision	Recall	F-measure
Appeal to authority 21 экземпляр	BERT	0.5385	0.3333	0.4118
	DistilBERT	0.4706	0.3810	0.4211
	RoBERTa	0.4737	0.4286	0.4500
Appeal to majority 20 экземпляров	BERT	0.2941	0.2500	0.2703
	DistilBERT	0.4706	0.4000	0.4324
	RoBERTa	0.4762	0.5000	0.4878
Appeal to nature 21 экземпляр	BERT	0.6000	0.4286	0.5000
	DistilBERT	0.6250	0.4762	0.5405
	RoBERTa	0.8000	0.5714	0.6667
Appeal to tradition 21 экземпляр	BERT	0.3889	0.3333	0.3590
	DistilBERT	0.3529	0.2857	0.3158
	RoBERTa	0.3810	0.3810	0.3810
Appeal to worse problems 24 экземпляра	BERT	0.5625	0.3750	0.4500
	DistilBERT	0.4211	0.3333	0.3721
	RoBERTa	0.6154	0.3333	0.4324
Black-or-white 21 экземпляр	BERT	0.2941	0.2381	0.2632
	DistilBERT	0.3333	0.1429	0.2000

Класс	Модель	Precision	Recall	F-measure
	RoBERTa	0.3636	0.1905	0.2500
Hasty generalization 20 экземпляров	BERT	0.0769	0.0500	0.0606
	DistilBERT	0.0417	0.0500	0.0455
	RoBERTa	0.3846	0.2500	0.3030
Slippery slope 23 экземпляра	BERT	0.3889	0.3043	0.3415
	DistilBERT	0.3846	0.2174	0.2778
	RoBERTa	0.6000	0.3913	0.4737
Non fallacy 804 экземпляра	BERT	0.8821	0.9303	0.9056
	DistilBERT	0.8837	0.9266	0.9047
	RoBERTa	0.8949	0.9428	0.9182
macro avg 975 экземпляров	BERT	0.4473	0.3603	0.3958
	DistilBERT	0.4426	0.3570	0.3900
	RoBERTa	0.5544	0.4432	0.4848
weighted avg 975 экземпляров	BERT	0.7972	0.8185	0.8057
	DistilBERT	0.7971	0.8144	0.8033
	RoBERTa	0.8284	0.8441	0.8329

Таблица 10 Результаты многоклассовой классификации {«Заголовок» + «Ответ» + «Комментарий»}

Класс	Модель	Precision	Recall	F-measure
Appeal to authority 21 экземпляр	BERT	0.5556	0.4762	0.5128
	DistilBERT	0.4118	0.3333	0.3684
	RoBERTa	0.5000	0.2857	0.3636

Класс	Модель	Precision	Recall	F-measure
Appeal to majority 20 экземпляров	BERT	0.3636	0.2000	0.2581
	DistilBERT	0.3636	0.2000	0.2581
	RoBERTa	0.5000	0.3000	0.3750
Appeal to nature 21 экземпляр	BERT	0.5238	0.5238	0.5238
	DistilBERT	0.6087	0.6667	0.6364
	RoBERTa	0.5789	0.5238	0.5500
Appeal to tradition 21 экземпляр	BERT	0.5000	0.4286	0.4615
	DistilBERT	0.6250	0.4762	0.5405
	RoBERTa	0.5000	0.5238	0.5116
Appeal to worse problems 24 экземпляра	BERT	0.3333	0.3333	0.3333
	DistilBERT	0.5385	0.2917	0.3784
	RoBERTa	0.4000	0.3333	0.3636
Black-or-white 21 экземпляр	BERT	0.3571	0.2381	0.2857
	DistilBERT	0.2727	0.1429	0.1875
	RoBERTa	0.3636	0.1905	0.2500
Hasty generalization 20 экземпляров	BERT	0.5000	0.0500	0.0909
	DistilBERT	0.3333	0.1500	0.2069
	RoBERTa	0.3333	0.2000	0.2500
Slippery slope 23 экземпляра	BERT	0.6250	0.2174	0.3226
	DistilBERT	0.3846	0.2174	0.2778
	RoBERTa	0.3846	0.2174	0.2778
Non fallacy	BERT	0.8847	0.9453	0.9140

Класс	Модель	Precision	Recall	F-measure
804 экземпляра	DistilBERT	0.8770	0.9403	0.9076
	RoBERTa	0.8806	0.9353	0.9071
macro avg 975 экземпляров	BERT	0.5159	0.3792	0.4114
	DistilBERT	0.4906	0.3798	0.4179
	RoBERTa	0.4935	0.3900	0.4276
weighted avg 975 экземпляров	BERT	0.8120	0.8338	0.8151
	DistilBERT	0.8012	0.8297	0.8111
	RoBERTa	0.8040	0.8277	0.8124

Таблица 11. Влияние контекста на результаты классификации

Класс	Данные	F-measure	Δ F-measure	Модель
Appeal to authority 21 экземпляр	COI	0.5143	0.0000 ⇩	BERT, RoBERTa, DistilBERT
	PAR + COI	0.4500	0.6143 ↓	RoBERTa
	TOP + PAR + COI	0.5128	0.0015 ↓	BERT
Appeal to majority 20 экземпляров	COI	0.5128	0.0000 ⇩	RoBERTa
	PAR + COI	0.4878	0.0250 ↓	RoBERTa
	TOP + PAR + COI	0.3750	0.1378 ↓	RoBERTa
Appeal to nature 21 экземпляр	COI	0.6842	0.0000 ⇩	RoBERTa
	PAR + COI	0.6667	0.0175 ↓	RoBERTa
	TOP + PAR + COI	0.6364	0.0478 ↓	DistilBERT
Appeal to tradition 21 экземпляр	COI	0.5000	0.0000 ⇩	DistilBERT, RoBERTa
	PAR + COI	0.3810	0.1190 ↓	RoBERTa
	TOP + PAR + COI	0.5405	0.0405 ↑	DistilBERT

Класс	Данные	F-measure	Δ F-measure	Модель
Appeal to worse problems 24 экземпляра	COI	0.4000	0.0000 ⇩	RoBERTa
	PAR + COI	0.4500	0.0500 ↑	BERT
	TOP + PAR + COI	0.3784	0.0216 ↓	DistilBERT
Black-or-white 21 экземпляр	COI	0.2632	0.0000 ⇩	BERT
	PAR + COI	0.2632	0.0000 ⇩	BERT
	TOP + PAR + COI	0.2857	0.0225 ↑	BERT
Hasty generalization 20 экземпляров	COI	0.2381	0.0000 ⇩	BERT
	PAR + COI	0.3030	0.0649 ↑	RoBERTa
	TOP + PAR + COI	0.2500	0.0119 ↑	RoBERTa
Slippery slope 23 экземпляра	COI	0.5263	0.0000 ⇩	RoBERTa
	PAR + COI	0.4737	0.0526 ↓	RoBERTa
	TOP + PAR + COI	0.3226	0.2037 ↓	BERT
Non fallacy 804 экземпляра	COI	0.9236	0.0000 ⇩	RoBERTa
	PAR + COI	0.9182	0.0054 ↓	RoBERTa
	TOP + PAR + COI	0.9140	0.0096 ↓	BERT
macro avg 975 экземпляров	COI	0.4913	0.0000 ⇩	RoBERTa
	PAR + COI	0.4848	0.0065 ↓	RoBERTa
	TOP + PAR + COI	0.4276	0.0637 ↓	RoBERTa
weighted avg 975 экземпляров	COI	0.8386	0.0000 ⇩	RoBERTa
	PAR + COI	0.8329	0.0057 ↓	RoBERTa
	TOP + PAR + COI	0.8151	0.0235 ↓	BERT