

ПОДХОД К ОЦЕНКЕ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ В НАУЧНО-ПОПУЛЯРНОМ ДИСКУРСЕ НА ОСНОВЕ АНАЛИЗА АРГУМЕНТАЦИИ И НАУЧНОГО ФАКТ-ЧЕКИНГА

Ю. А. Загорулько¹ [0000-0002-7111-6524], **И. Р. Ахмадеева**² [0000-0002-7371-1087],
А. С. Серый³ [0000-0001-8275-4700], **Е. А. Сидорова**⁴ [0000-0001-8731-3058],

^{1–4}*Институт систем информатики им. А. П. Ершова СО РАН, г. Новосибирск,
Россия*

¹zagor@iis.nsk.su, ²i.r.akhmadeeva@iis.nsk.su, ³alexey.seryj@iis.nsk.su,

⁴Isidorova@iis.nsk.su

Аннотация

Рассмотрены проблемы оценки достоверности информации, извлекаемой из научно-популярных текстов. Как известно, научно-популярный жанр характеризуется свободным стилем изложения материала, нестрогими отсылками к первоисточникам и экспертным мнениям, наличием спорных утверждений и выводов. Уточнение этих компонент авторских рассуждений может существенно повысить степень доверия к ним критически настроенных читателей или, наоборот, позволит опровергнуть тезисы, приводимые авторами.

В работе предложен новый подход к решению задачи оценки достоверности информации, содержащейся в научно-популярных публикациях. Подход предполагает применение теории аргументации Д. Уолтона для выявления потенциально недостоверных утверждений и логических ошибок в рассуждениях. Поиск в научных базах данных и вычисление метрик доверия к научным публикациям на основе их метаданных и научных рейтингов будет обеспечивать верификацию спорных утверждений и оценку их достоверности. Применение больших языковых моделей позволит сгенерировать объяснения принимаемых решений.

Для разработки и апробации этого подхода предложены две методики сбора данных. В первой используются корпуса русскоязычных научно-популярных текстов и комментарии к ним с аргументативной разметкой. Такая разметка включает, в том числе, полемические схемы аргументации, анно-

тированные вручную, которые подвергают сомнению тезис или аргумент автора. Базой второй методики служит генерация с помощью LLM синтетических тезисов и научно-популярных статей на основе реальных научных публикаций.

Ключевые слова: проверка утверждений, анализ аргументации, научная проверка фактов, достоверность информации, автоматический сбор данных.

ВВЕДЕНИЕ

Открытый доступ к информации является одной из особенностей современного общества. Быстрыми темпами идет развитие средств распространения и предоставления информации, побуждая все больше людей интересоваться различными сферами исследований, в том числе теми, где они не являются специалистами. Этому способствуют многочисленные онлайн-площадки и интернет-ресурсы, предоставляющие доступ к различным публикациям псевдонаучного и научно-популярного характера. Авторами таких материалов, как правило, являются журналисты и блогеры, преследующие цель привлечения максимальной аудитории. Вследствие этого заголовки статей могут не соответствовать их содержанию, а утверждения, взятые даже из достоверных источников, цитируются с искажениями. Это приводит к тому, что вся цепочка дальнейших рассуждений, построенных на данных утверждениях вместе с промежуточными и конечными выводами, становится сомнительной.

Современные методы и технологии обработки естественного языка уже достигли впечатляющих результатов во многих областях. Однако на данный момент они не предоставляют эффективных инструментов для выявления ненадежных утверждений. Опровержение и подтверждение обоснованности утверждений и мнений, высказанных автором, составляют область исследования теории аргументации, изучающей различные способы рассуждений в текстах и устной речи с философской, лингвистической, когнитивной и социальной точек зрения. Разработка компьютерных методов анализа дискурса, позволяющих выявлять тезисы и связанную с ними аргументацию, даст в руки пользователей эффективный инструмент для исследования информации (поиск обоснований, выявление противоречий, анализ источников), а применение методов оценки доверия на основе релевантных внешних источников информа-

ции позволит проводить численную оценку достоверности спорных утверждений.

Целью настоящей работы является разработка архитектуры системы, которая бы позволяла комплексно решать задачу анализа достоверности информации, представленной в русскоязычных интернет-публикациях. Работа посвящена решению следующих исследовательских задач.

RQ1. Как оценивать достоверность информации?

RQ2. Какие методологические и технологические принципы должны лежать в основе комплексной технологии оценки достоверности информации?

RQ3. Каковы возможные способы автоматизации сбора и генерации данных для проведения экспериментальных исследований по оценке достоверности информации?

Чтобы не затрагивать этические проблемы, а также вопросы, зависящие от политических и личных интересов, мы ограничились рассмотрением научно-популярных текстов, факты из которых поддаются объективной оценке на основе сопоставления с научными источниками данных, степень достоверности которых на сегодняшний день не вызывает серьезных сомнений, а возникающие противоречия поддаются оценочному сравнению на основе научных рейтингов изданий, авторов и самих публикаций.

1. ОБЗОР РАБОТ

Одним из распространенных подходов к выявлению недостоверной научной информации (misinformation) является фактчекинг [1], т. е. проверка утверждений на соответствие известным фактам. В общем виде данный подход заключается в верификации утверждений с опорой на доказательства. Выделяют следующие основные этапы фактчекинга [2]: (а) поиск релевантных документов (Document Retrieval), (b) извлечение доказательств (Evidence Retrieval) и (с) верификация утверждений (Claim Verification). При анализе сообщений или публикаций целиком добавляется предварительный этап – обнаружение сомнительных утверждений, требующих проверки (Claim Detection).

Обзор современных работ [3] позволяет заключить, что все больше исследований в данной области опираются на большие языковые модели (LLM), которые применяют не только при реализации основных этапов автоматиче-

ского фактчекинга, но и для генерации объяснений, создания синтетических данных [4, 5], а также для оценки качества предсказания других систем. При этом LLM подвержены галлюцинациям – они могут генерировать правдоподобную, но фактически неверную информацию, что особенно критично в задачах фактчекинга. В рамках этой проблемы в [6] разработан фреймворк FActScore, предназначенный для оценки фактической точности текстов, генерируемых LLM. Авторы предлагают верифицировать отдельные факты, содержащиеся в таких текстах, используя Википедию в качестве источника доказательств, и тем самым ставят вопрос о доверии к LLM как к источнику достоверных знаний.

Преобладающая часть исследований в области автоматизированного фактчекинга опирается на предположение о наличии надежного корпуса источников и сосредоточена на задаче верификации утверждений относительно него. Подходы же, включающие явную оценку доверия к источнику, представлены меньшим числом работ. Авторы [7] используют метод обнаружения фейковых новостей через оценку доверия к авторам и источникам на основе их прошлой истории публикаций. В работе [8] использована внешняя оценка достоверности доменов, полученная с фактчекингового сайта, которая учитывает уровень политической предвзятости, фактической точности и общей надежности.

Подходы, основанные исключительно на поиске и сопоставлении утверждения с источниками, имеют существенное ограничение: наличие релевантного и авторитетного источника не гарантирует корректности выводов, сделанных на его основе. На практике искажения могут возникать как при передаче информации посредством цитирования, так и вследствие некорректной интерпретации или ошибочной аргументации со стороны авторов. Это обусловило развитие подходов, рассматривающих анализ цепочек рассуждений (reasoning) в качестве дополнительного компонента фактчекинга [9]. При этом было показано [10], что LLM плохо справляются с определением ложности утверждения и могут принимать ложные утверждения за истинные при наличии некорректного цитирования, создающего видимость доказательства.

Одним из основных условий развития рассматриваемой области исследований является создание наборов данных, которые бы включали как досто-

верные, так и недостоверные факты. При этом такие датасеты, как правило, включают источники, на основании которых принималось решение. Большинство существующих датасетов ориентировано на социальные сети и политические новости (AveriTec [11], LIAR [12]), фактологические утверждения из Википедии (HOVER) или тематику здравоохранения в контексте пандемии COVID-19 (COVID-Fact [13], CoVERT [14]). Среди ресурсов, непосредственно предназначенных для научного фактчекинга (scientific fact-checking), можно выделить PUBHEALTH [15], SciFact [16] и CoVERT [14]. Последний специализируется на проверке утверждений по данным клинических испытаний. Отличительной особенностью перечисленных датасетов является их преимущественно англоязычный характер. Русскоязычные ресурсы представлены значительно слабее, в частности датасетом ruSciFact [5].

В качестве проверяемых единиц в существующих датасетах обычно выступают утверждения (claim), которые подразделяются на два класса [1]: синтетические – формируемые аннотаторами или языковыми моделями на основе исходных текстов, и естественные – извлекаемые из реальных источников, включая сайты фактчекинга и публикации в социальных медиа. Первый тип утверждений является более простым для анализа моделями и не требует контекста, однако он значительно оторван от реальных сценариев использования. Так, датасет SciFact [16] содержит биомедицинские утверждения, изначально взятые из предложений со ссылкой на источники в научных статьях, которые затем были переформулированы аннотаторами в форму, более атомарную и удобную для обработки. Утверждения в ruSciFact [5] сгенерированы с использованием LLM и являются как позитивными – согласующимися с аннотациями научных статей, так и негативными – опровергаемыми этими аннотациями. Авторами датасета SciClaimHunt [4] предложены четыре способа генерации негативных утверждений на основе позитивных: генерация отрицания, сопоставление позитивному утверждению неверной статьи, изменение числовых значений и подмена именованных сущностей. Для формирования естественных утверждений применяют методы автоматического извлечения информации: в PUBHEALTH [15] использовались заголовки фактчекинговых статей о здравоохранении, из-за чего некоторые утверждения оказались слишком общими, а

CoVeRt [14] содержит утверждения из постов Twitter, извлеченные специализированной моделью.

В качестве источников доказательств преимущественно используют научные публикации, реже применяют новостные материалы и статьи Википедии [1]. Несмотря на то что научное знание формируется как результат накопления и переоценки результатов множества исследований, в большинстве существующих датасетов каждому утверждению сопоставляется небольшое число источников доказательств. Авторы SciFact-Open [16] частично преодолели это ограничение, расширив SciFact [16] до 500 тыс. источников. Распределение доказательств при этом оказалось неравномерным: половина всех доказательств сконцентрирована на 34 утверждениях (из 279), а максимальное количество доказательств, приходящихся на одно утверждение, достигает 19.

Датасет для выявления ошибок рассуждений MISSCI [17] и его расширение MISSCIPLUS [18] содержат примеры ошибочной аргументации, где недостоверные утверждения о здоровье (преимущественно о COVID-19) подкреплены ссылками на биомедицинские публикации. Этот датасет ориентирован на выявление имплицитных логических ошибок и содержит реконструкции ошибочных рассуждений помимо классификации их типов. Он построен на основе платформы HealthFeedback, с которой были отобраны статьи, разоблачающие некоторые утверждения, и ссылки на источники, которые некорректно интерпретируются в этих утверждениях. Разметка и восстановление ошибочной аргументации производились студентами-магистрантами.

В настоящей работе представлен новый подход к решению задачи оценки достоверности информации, содержащейся в научно-популярных публикациях, на базе интеграции практик, наиболее популярных на сегодняшний день, включая: а) использование теории аргументации Д. Уолтона [19] для выявления потенциально недостоверных утверждений и поиска логических ошибок в цепочках рассуждений; б) обнаружение в анализируемых текстах отсылок на внешние научные источники и использование последних для верификации спорных утверждений; в) использование метаданных и научных рейтингов внешних источников для вычисления оценки достоверности спорных тезисов.

Для разработки и апробации данного подхода предложены две методики сбора данных. В первой из них для составления датасетов используются рус-

скоязычные научно-популярные тексты, аннотированные вручную по аргументационным схемам и верифицированные экспертами как содержащие логические ошибки. В основе второй методики лежит генерация синтетических научно-популярных статей на базе реальных научных публикаций.

2. МЕТОДИКА АВТОМАТИЧЕСКОЙ ОЦЕНКИ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ

Одной из ключевых особенностей научно-популярного жанра является то, что обсуждение научных достижений ведется в широкой аудитории, что нередко порождает ошибочную интерпретацию научных данных и появление недостаточно обоснованных суждений со стороны лиц, не являющихся специалистами в соответствующей области знаний. В связи с этим предлагаемая методика оценки достоверности информации включает такие задачи, как выявление авторских тезисов, используемых приемов рассуждений, отсылок к источникам и их последующая проверка на основе научных публикаций (см. рис. 1).

Архитектура системы оценки достоверности информации, поддерживающая предлагаемую методику, включает три основных контура: внутренний и внешний, а также контур принятия решений.

Внутренний контур. Основное его назначение состоит в анализе содержания исходного текста без обращения к внешним источникам информации. На вход контуру подается текст научно-популярной статьи, а результатом его работы является множество тезисов и валидирующих утверждений (см. п. 3.1), требующих проверки во внешних данных.

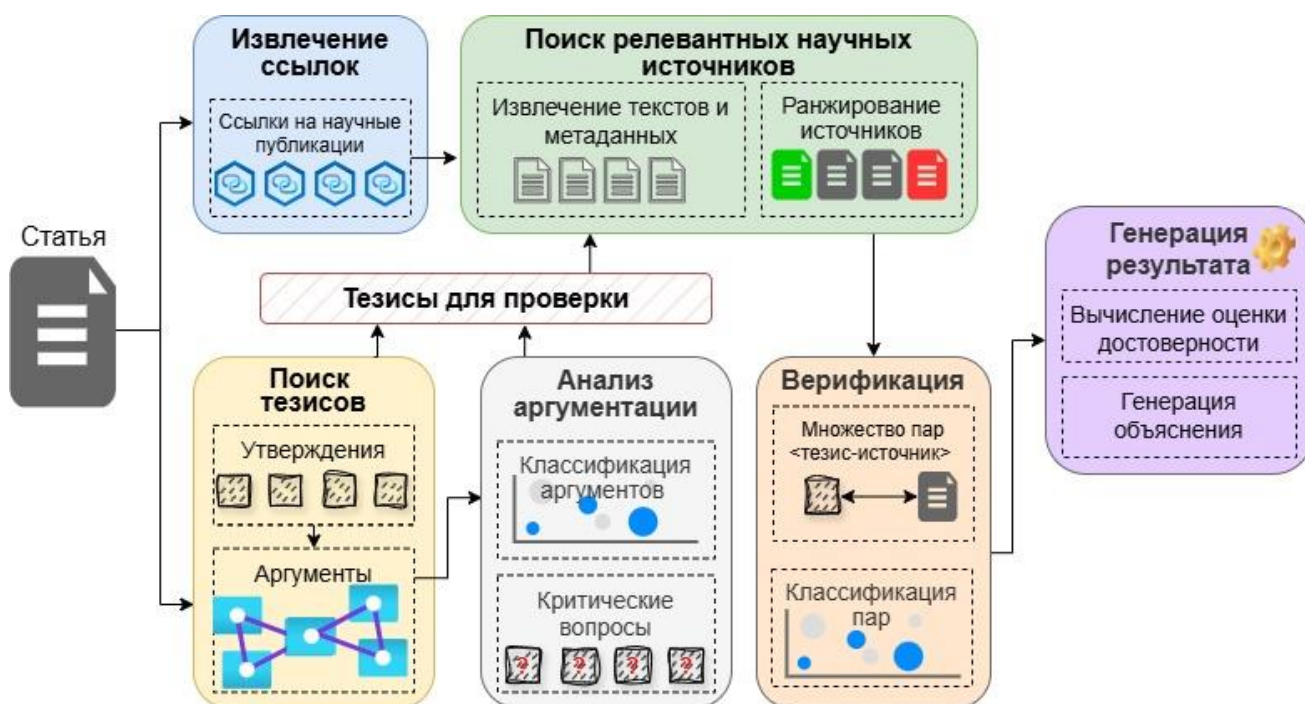


Рис. 1. Архитектура системы оценки достоверности информации.

Внешний контур. Его задачей является верификация тезисов, полученных от внутреннего контура, с привлечением внешней информации. Инструменты внешнего контура обращаются к научным базам данных, таким как arXiv, Semantic Scholar, OpenAlex, PubMed и др., в поисках актуальных научных исследований, содержащих информацию, семантически близкую заданным тезисам. На основе найденных источников осуществляется проверка тезисов (верификация), результаты которой передаются в контур принятия решений.

Контур принятия решений. Здесь выполняются общая оценка и принятие решения о степени достоверности статьи. Происходят вычисления оценки доверия для каждого тезиса, выделенного внутренним контуром, и общей оценки для всего документа, а также генерация структурированного объяснения результата.

2.1. Внутренний контур: извлечение ссылок, тезисов и критический анализ аргументации

Внутренний контур решает три основные задачи.

1. Извлечение ссылок на научные источники. В научно-популярных статьях авторы часто явным образом ссылаются на авторитетные источники, чтобы по-

высвить доверие к высказываемым суждениям, однако такие отсылки могут быть некорректны и требуют проверки.

2. Выявление в тексте тезисов и аргументов. Аргументированные суждения, связывающие посылки с выводами, образуют основу общей структуры рассуждений в научных и научно-популярных статьях. На основе этой структуры выявляются главные тезисы, цепочки рассуждений, используемых авторами для обоснования своих тезисов.

3. Анализ обоснованности аргументации. Многие рассуждения, вводящие в заблуждение, строятся на некорректных логических выводах и требуют проверки.

Из исходного текста выделяются утверждения и аргументативные отношения между ними, на основе которых формируется граф аргументации (argumentation graph). Сегментация текста, определение границ аргументативных утверждений и идентификация тезисов (вершин графа) выполняются с помощью методов синтаксического и риторического анализа текста [20, 21], а также методов анализа семантического сходства утверждений. Каждая вершина графа соответствует утверждению (claim) или его составной части.

Классификация моделей рассуждений в соответствии с классификатором Д. Уолтона позволяет выделить аргументы, значимые для оценки достоверности утверждений исходной статьи. Принципиальным является разделение аргументов на «Внешние», указывающие на внешний источник информации как на основание убедительности, и «Внутренние», опирающиеся на логические связи между утверждениями (причинно-следственную, меронимическую, корреляционную и др.). Для проверки внутренних аргументов можно использовать критические вопросы, которые включены в модель Уолтона. В работе [22] описан способ генерации валидирующих (проверяющих корректность применения схемы) утверждений, которые совместно с посылкой и заключением схемы могут формировать составной тезис для дальнейшей проверки.

Таким образом, утверждения, соответствующие элементам внешних аргументов, и валидирующие утверждения для внутренних аргументов передаются во внешний контур для поиска релевантных научных публикаций.

2.2. Внешний контур: верификация на основе научных источников информации

Внешний контур обеспечивает работу с информацией, находящейся за пределами исходной статьи: с научными публикациями, которые могут содержать факты, подтверждающие или опровергающие утверждения автора.

Поисковый запрос, сформированный на основе тезисов, полученных из внутреннего контура, применяется для поиска научных публикаций – внешних источников информации. К найденным источникам могут быть добавлены явные ссылки из текста исходной статьи, если их удалось обнаружить и распознать.

Важно отметить, что в рамках предлагаемой архитектуры внешние источники не считаются равноправными. В научном сообществе существует ряд критериев, определяющих авторитетность публикации: импакт-фактор издания, количество цитирований, авторский состав, репутация журнала/конференции и т. д. Для ранжирования источников по рейтингу применяется разработанный нами метод оценки доверия. Его описание дано ниже. Этот метод позволяет принимать во внимание степень «старения» источников, понижая их в рейтинге в случае длительного отсутствия ссылок на них из других научных работ.

В найденных источниках идентифицируются утверждения, соотносящиеся с тезисами исходной статьи. В зависимости от типов аргументативных отношений, устанавливаемых между утверждениями источника и соответствующим тезисом, каждая пара «тезис – источник» классифицируется по следующей схеме.

1. Подтверждение (**Support**). Согласуется с обнаруженными результатами научных исследований.
2. Противоречие (**Contradict/Refute**). Противоречит научным данным.
3. Нет общего контекста (**Insufficient**). Невозможно установить корреляцию с научным источником.
4. Неверная интерпретация (**Misinterpret**). Тезис искажает содержание научного источника.

По результатам анализа внешних источников каждое утверждение исходного текста оценивается по двум критериям.

Первый – это совокупный уровень доверия к релевантным научным источникам (Evidence Strength), характеризующий актуальность тематики утверждения в научном сообществе. Вычисляется как максимальный рейтинг релевантного источника (max) или степенное среднее порядка $2 \leq p \leq 5$. Выбор метода зависит от следующего принятого допущения: либо общий уровень доверия определяется наиболее авторитетным источником (применимо к научно-популярным текстам), либо учитывается вся выборка.

Второй критерий – это уровень соответствия утверждения источникам (Fidelity of Sources). В отличие от первого, он позволяет оценить, действительно ли утверждение опирается на источник или лишь имеет с ним общую тематику. В базовом варианте данный критерий вычисляется как доля источников, в которых удалось обнаружить посылки, для которых заданное утверждение является логически правомерным выводом. Иными словами, критерий фиксирует, представлены ли в источниках достаточные основания для заключения, соответствующего проверяемому утверждению. В более сложных случаях, когда есть возможность, помимо поддерживающей аргументации, классифицировать и противоречия (преувеличение, неправомерные обобщения и др.), данный критерий вычисляется как отношение $\frac{S^+}{S^+ + S^-}$, где S^+ и S^- – суммы рейтингов всех источников, поддерживающих утверждение и вступающих с ним в противоречие соответственно. Комбинация этих параметров позволит, в частности, выявлять стратегию «выборочного цитирования» (Cherry Picking): если статья опирается на малое число источников и игнорирует множество других авторитетных работ, то это служит сигналом о потенциальной ненадежности информации.

Ранжирование научных публикаций применяется для определения их веса в качестве доказательств при проверке утверждений, приводимых в исходной статье. Этот процесс является частью внешнего контура системы и опирается на разработанный ранее метод оценки доверия к единицам информации (фактам) [23]. Изначально разрабатываемый для построения вероятностных графов знаний, где вершины и дуги имеют оценку доверия или надежности (reliability), этот метод представляет динамику изменения доверия к единичному факту в виде марковского случайного процесса. То, что доверие

моделируется как процесс, означает, что оно может изменяться со временем под влиянием внешних факторов. Мы рассматриваем два основных фактора: время t и неотрицательные качественные характеристики μ источников, из которых получен факт. Таким образом, наш метод имеет много общих черт с методами объединения данных (Data Fusion), когда необходимо оценивать противоречивую информацию, приходящую из множества источников. При этом его отличительной чертой является учет влияния времени на надежность информации.

В рамках представленного процесса оценкой доверия к факту является вероятность того, что этот факт является правдивым. Соответственно, для факта можно определить два потенциальных состояния (Unreliable (U), Reliable (R)), которые образуют вектор распределения доверия, обозначаемый $\bar{\pi} = (\pi_U, \pi_R) = (\pi_1, \pi_2)$. Обновление распределения доверия происходит при появлении нового источника или по истечению определенного времени. Очевидно, что всегда $\pi_1 + \pi_2 = 1$.

Пусть в момент времени n для некоторого факта известны вектор распределения $\bar{\pi}^n = (\pi_1^n, \pi_2^n)$ и очередной источник с характеристикой μ . В рамках разработанного метода предложено два вида переходных матриц $P(n, n+1)$ как функций от $\bar{\pi}^n$ и μ , отличающихся поведением в предельных случаях. Переходная матрица $P_m = \text{softmax}(\mu \pi^T \times (-\pi_1, \pi_2))$ с учетом свойств вектора распределения вычисляется по формуле

$$P_m(n, n+1) = \begin{pmatrix} \frac{1}{1 + e^{\mu \pi_1^n}} & \frac{1}{1 + e^{-\mu \pi_1^n}} \\ \frac{1}{1 + e^{\mu \pi_2^n}} & \frac{1}{1 + e^{-\mu \pi_2^n}} \end{pmatrix}. \quad (1)$$

Из формулы (1) следует, что при $\mu \rightarrow 0$ умножение любого стохастического вектора на матрицу P_m даст распределение $(1/2, 1/2)$. Это означает, что любое распределение будет стремиться к неопределенному при приближении μ к нулю. С другой стороны, при неограниченном росте μ любое распределение будет стремиться к идеальному $(0,1)$.

Вторая матрица перехода $P_i = \text{softmax}\left(\pi^T\left(\frac{\pi_1}{\mu}, \mu\pi_2\right)\right)$, вычисляемая по формуле

$$P_i(n, n+1) = \begin{pmatrix} \frac{1}{1 + e^{\pi_1^n(\mu\pi_2^n - \pi_1^n/\mu)}} & \frac{1}{1 + e^{\pi_1^n(\pi_1^n/\mu - \mu\pi_2^n)}} \\ \frac{1}{1 + e^{\pi_2^n(\mu\pi_1^n - \pi_2^n/\mu)}} & \frac{1}{1 + e^{\pi_2^n(\pi_2^n/\mu - \mu\pi_1^n)}} \end{pmatrix},$$

с падением μ приводит распределение к вектору $(1, 0)$, соответствующему максимальной ненадежности. В другом предельном случае матрица P_i ведет себя аналогично P_m для всех $\bar{\pi}$, кроме $\bar{\pi} = (1, 0)$. Последнее важно в тех случаях, когда появление даже очень надежного источника не должно полностью перевешивать в спорных ситуациях. Таким образом, переход $n \rightarrow n+1$ производится путем умножения текущего распределения на переходную матрицу: $\bar{\pi}^{n+1} = \bar{\pi}^n P(n, n+1)$, где P – это P_m или P_i .

Помимо источников, на уровень доверия к факту влияет время, в частности то, как давно известен факт и как давно он был упомянут в каком-либо источнике. Согласно модели, шаг случайного процесса соответствует обнаружению факта в новом источнике. Между этими дискретными шагами проходит время $t \in \mathbb{R}$, и чем больше это время, тем более устарела информация. Можно говорить, что в определенных предметных областях длительное не упоминание информации в новых источниках является косвенным признаком утери актуальности.

Зависимость надежности информации от времени t регулируется темпоральным множителем – функцией из параметрического семейства

$$h(t) = \frac{1}{1 + \ln(at + 1)},$$

где параметр a (adjustor) регулирует темп деградации доверия со временем. Отметим, что понятие «длительного отсутствия упоминаний» факта в источниках зависит от конкретной задачи и, в большей степени, от предметной области.

Если $\bar{\pi}^n = (\pi_1^n, \pi_2^n)$ – текущий вектор распределения, то по прошествии времени t он переходит в $(\pi_1^{n'}, \pi_2^{n'})$, где

$$\pi_2^{n'} = h(t)\pi_2^n, \pi_1^{n'} = \pi_1^n + (1 - h(t))\pi_2^n. \quad (2)$$

Более ранняя версия рассматриваемого метода применялась для оценки рейтингов публикаций в системе цитирования PubMed. Единицами информации в этом случае выступали научные публикации, а аналогом «источников» являлись авторы публикаций. Рейтинг авторов вычислялся на основании их публикационной активности и цитируемости. Рейтинги публикаций зависели от рейтингов их авторов и рейтингов публикаций, в которых они были процитированы. При отсутствии информации о цитированиях за последний год публикация понижалась в рейтинге в соответствии с формулой, аналогичной формуле (2).

2.3. Контур принятия решений: генерация финального вердикта и объяснения

Для формирования финального заключения о степени достоверности статьи привлекается LLM, выступающая в роли «судьи». На вход модели подаются:

- исходная статья;
- построенный граф аргументации и множество выделенных тезисов;
- перечень найденных научных источников с их рангами доверия;
- для каждого тезиса – совокупный уровень доверия к релевантным научным источникам (Evidence Strength) и уровень соответствия утверждения источникам (Fidelity of Sources).

Выходными данными системы являются:

- финальный вердикт – оценка достоверности для каждого тезиса и всей статьи (например: «надежная», «частично надежная», «ненадежная», «требуется дополнительной проверки»);
- объяснение (explanation) – текстовое обоснование вердикта с указанием на конкретные логические ошибки, несоответствия научным источникам или случаи выборочного цитирования.

3. МЕТОДИКА ПОДГОТОВКИ ДАТАСЕТОВ

Основной методологической проблемой при разработке систем анализа достоверности научно-популярного контента является отсутствие унифицированной методики автоматизированного сбора размеченных данных.

Ключевое ограничение доступных наборов данных, делающее невозможным их использование в настоящем исследовании, заключается в том, что в них каждое проверяемое утверждение (claim) рассматривается изолированно, вне контекста. Контекст может отсутствовать вовсе, как в SciFact [16], где утверждения были написаны отдельно, а не получены из текстов. Информация об источниках подтверждений или опровержений также может быть недоступной, как в PUBHEALTH [15], где имеется лишь итоговая классификация (label) утверждения.

Таким образом, для разработки и апробации предложенного выше подхода необходимо было собрать корпуса научно-популярных статей и релевантных им статей-источников.

Разрабатываемый набор данных (датасет) будет состоять из следующих компонент.

1. Научно-популярные статьи с разметкой утверждений и аргументации, в которой выделяются:
 - a) тезисы с явными отсылками на другие публикации;
 - b) тезисы с косвенными отсылками на научные источники (указание эксперта, группы, мероприятия и т. п.);
 - c) тезисы, сгенерированные на основе критических вопросов к аргументам;
 - d) графы аргументации, связанные с выделенными тезисами: утверждения, аргументативные связи между ними с указанием типа связи в соответствии с классификацией Д. Уолтона.
2. Корпус научных статей-источников с разметкой метаданных и связей:
 - a) метаданные: издательство, авторы, дата публикации, название, рейтинг цитируемости публикации, рейтинг издательства, индекс цитируемости авторов;
 - b) содержание (на данном этапе – название и текст аннотации) с размеченными утверждениями;

3. Множество пар <тезис, источник>, которые сопоставляют тезисы из первого корпуса с размеченными утверждениями источника с указанием типа сопоставления: поддержка, опровержение, неверная интерпретация и ложная связь и его вес. Такие пары (опционально) также могут снабжаться объяснением на естественном языке, сгенерированным LLM (обоснование дается для каждого проверяемого тезиса).

Создание такого корпуса является достаточно трудоемким процессом, предусматривающим два основных этапа: подбор научно-популярных статей, включающих достоверные и недостоверные утверждения, и поиск научных источников, содержащих информацию, поддерживающую или опровергающую такие утверждения. Ручной сбор и разметка утверждений в обширных массивах реальных данных требуют значительных затрат. В связи с этим нами предложены методы автоматического сбора данных. При этом оценка достоверности утверждений должна опираться на верифицируемые критерии, позволяющие снимать этические вопросы и обосновывать данные, включаемые в обучающие выборки.

Предложены два подхода к автоматизации построения датасетов. Первый опирается на выбор утверждений с верифицированной степенью достоверности и последующий автоматизированный поиск релевантных им источников с помощью технологии RAG-LLM, второй на методы генерации синтетических научно-популярных статей на основе реальных научных публикаций.

3.1. Методика подготовки набора данных на основе корпуса с разметкой полемической аргументации

Для составления датасета предложено использовать русскоязычные научно-популярные тексты, аннотированные вручную по аргументационным схемам и верифицированные экспертами как содержащие логические (аргументационные) ошибки. С этой точки зрения полезно обратиться к фрагменту готового аргументативно размеченного корпуса научной интернет-коммуникации на русском языке, представленного на платформе ArgNetBankStudio (uniserv.iis.nsk.su/arg). Для целей настоящей работы был выбран подкорпус научно-популярных статей с форума «Хабр» общим объемом 57356 словоупотреблений: 30 статей с комментариями пользователей форума

(средний объем статьи без комментариев – 1912 словоупотреблений, с комментариями – 16482). Статьи и комментарии корпуса размечены вручную четырьмя экспертами с использованием 60 схем рассуждения Д. Уолтона.

В датасет отбираются: (1) фрагменты, содержащие ошибки, и комментарии к ним (реакции), для этого берутся комментарии с явными отсылками к фрагментам текста; (2) фрагменты диалогов, содержащие реплики с ошибками и реакции на них, в качестве диалога рассматривается общение пользователей в разделе комментариев.

Для выявленного недостоверного утверждения формируется запрос к LLM-агенту, поддерживающему поиск по научным базам публикаций:

Ты специалист по факт-чекингу. Прочитай текст (text) ниже и оцени его достоверность с учетом комментария (comment). Для этого проверь научные источники (до 10 штук) {science_bases}. Для каждого источника оцени корреляцию. Ответь, подтверждает он эту информацию или противоречит. Выведи суммарный итог по всем найденным источникам.

text: {text}
comment: {comment}

В табл. 1 представлены данные, которые можно получить на основе предлагаемой методики.

Табл. 1. Примеры данных, полученных на основе полемической аргументации.

Тип	Данные
text	<i>Если мы создадим сознательный ИИ, он неизбежно сойдет с ума от одиночества, потому что у него не будет друзей, семьи и настоящего общения, как у людей.</i>
comment	<i>С чего вы взяли, что ИИ будет чувствовать одиночество так же, как человек? Это ведь программа, у которой нет биологических потребностей в социуме, привязанности и эмоциональной близости!</i>
text_scheme	Схема Уолтона: “От аналогии” (Analogy_Inference) Premise: <i>потому что у него не будет друзей, семьи и настоящего общения, как у людей.</i>

	Conclusion: <i>Если мы создадим сознательный ИИ, он неизбежно сойдет с ума от одиночества</i>
comment_ conflict_scheme	AnalogyRelevance_ValidationCQ Relevance_Validation: <i>Как можно сравнивать живое существо и устройство, созданное для определенных целей?!</i> ValidatedElement: Analogy_Inference
error_type	Ложная аналогия
support_claim	(2 items)
contradict_claim	(8 items)
explanation	<i>Современные ИИ не обладают феноменальным сознанием и не способны к субъективному переживанию эмоций, включая одиночество. Они могут симулировать одиночество через статистические паттерны в языке, но это не тождественно переживанию. Одиночество как психическое состояние требует биологического субстрата, субъективного опыта и эмоциональной уязвимости, которых у ИИ нет.</i>
verdict	<i>Утверждение о том, что сознательный ИИ «неизбежно сойдет с ума от одиночества», является спекулятивным антропоморфизмом, не имеющим научных оснований.</i>

В датасет отбираются утверждения (text) и комментарии (comment), в которых комментаторы явным образом полемизируют с аргументацией авторов статей или других пользователей, возражая против несоблюдения условий примененной схемы, поставленных в критических вопросах, соответствующих схеме.

Таким образом, недостаточно достоверные рассуждения, доказывающие некий тезис в рамках некоторого фрагмента текста, представлены в виде стандартных схем аргументации (text_scheme), которые связывают тезис и поддерживающие его посылки, а явные возражения представлены в виде полемических схем (comment_conflict_scheme) [24]. Тип ошибки (error_type) для кон-

кретного рассуждения определяется на основании класса полемической схемы в комментарии и в дальнейшем требует ручной валидации (валидируется только наличие или отсутствие ошибки, а не ее тип).

Для приведенного примера с помощью веб-версии модели DeepSeek (chat.deepseek.com) получены научные источники (reference), которые оценивают утверждения как искаженные (1 шт.) либо как некорректные (9 шт.), а также сгенерированы финальный вердикт (verdict) и объяснение (explanation). В табл. 2 приведен пример сопоставленных источников и сгенерированных результатов по тезису из табл. 1.

Табл. 2. Примеры найденных LLM источников и результатов сопоставления.

№	Источник	Результат верификации
1	reference	Lang, L. (2026). Deception, Alief, and Authenticity: The Ethics of AI Companions as Loneliness-Alleviating Systems. B: Oxford Intersections: AI in Society (под ред. Н. Shevlin и Р. Hacker). Oxford University Press.
	explanation	<i>Автор утверждает, что ИИ-компаньоны онтологически неаутентичны: они представляют себя как обладающие феноменально сознательными психическими состояниями (включая эмоции), несмотря на то, что не обладают ими.</i>
	verdict: опровержение	<i>Прямо противоречит утверждению о возможности ИИ испытывать одиночество.</i>
2	reference	Artificial Intelligence Cannot Be Human, Emotional, or Spiritual (2025). Research.UMKC.edu.
	explanation	<i>ИИ лишен эмоционального сознания, морального понимания, духовного понимания и человеческой психологии... ИИ не может испытывать эмоции.</i>
	verdict: опровержение	<i>Прямое опровержение</i>
3	reference	Fujiwara, H., Kasai, D., Goto, T., & Tabei, K. (2025). Simulating Loneliness: Human–Generative Agent Comparative Insights. SIGGRAPH Asia 2025 Posters. ACM.

	explanation	<i>Авторы разработали генеративные агенты, которые воспроизводят связь между одиночеством и руминацией, аналогичную человеческой.</i>
	verdict: искажение	<i>Подтверждает возможность симуляции, но не переживания одиночества.</i>

Таким образом, при формировании датасета с помощью предложенного подхода будут использоваться «реальные» утверждения, автоматически найденные «реальные» источники и синтезированные объяснения. Вторая методика обеспечивает противоположный подход к распределению реальных и синтезированных частей датасета.

3.2. Методика построения датасетов путем генерации синтетических научно-популярных статей

Предложен метод генерации синтетических научно-популярных статей на базе реальных научных публикаций и описаны проведенные эксперименты. Предполагается доступность базы научных публикаций с возможностью семантического поиска, т. е. поиска работ, семантически близких заданному набору тезисов. Для генерации одного текста выбирались от 2 до 4 семантически близких научных источников. Процесс генерации сведен к следующей последовательности этапов.

1. Поиск базовых источников – семантически близких научных публикаций, которые будут основой для генерации. На этом этапе проводится семантический поиск или выбор подходящего кластера в случае, если была проведена предварительная кластеризация публикаций по смысловой близости.

2. Формирование списка утверждений-кандидатов. Из текстов базовых источников извлекаются составные части аргументации (Argumentative Discourse Units, ADUs), из которых затем формируется список базовых утверждений.

3. Генерация тезисов. На основе базовых утверждений LLM в соответствии со специально заданной инструкцией генерирует наборы тезисов, которые могут находиться с ними в одном из следующих отношений: поддержка (Support), противоречие (Contradict), неверная интерпретация

(Misleading/Misinterpretation), недостаточное обоснование (Insufficient). Последнее означает, что посылки не доказывают вывод.

4. Генерация текстов. По заданному набору синтетических тезисов LLM генерирует короткую научно-популярную статью. Тип статьи зависит от типов тезисов, лежащих в ее основе, и полученных инструкций. При генерации модель должна выделять в тексте основные тезисы. В итоге мы получим текст заданного типа, в котором известны типы и координаты основных тезисов. Для каждого из тезисов известны источники и базовые утверждения, а также типы связей, в которых они состоят. Смешивая тезисы разных типов, можно получить тексты, имеющие смешанную (Mixed) позицию относительно лежащих в их основе научных публикаций.

Были проведены эксперименты по генерации небольшого количества синтетических данных для английского и русского языков. Для генерации использовались инструкции, приведенные в табл. 3.

Табл. 3. Инструкции для генерации синтетических данных.

Задача	Русский	Английский
(а) Извлечение утверждений из научных публикаций	Тебе поручено выделить в аннотациях научных статей на русском языке те предложения, которые содержат утверждения о каких-либо фактах. Действуй с максимальной точностью. Если не уверен, то не пытайся угадывать, а просто пропусти предложение. Некоторые предложения могут быть анафорически связаны с предыдущими. Найди пары предложений, в которых второе предложение содержит анафорическую ссылку на предыдущее. Анафору следует переписать таким образом, чтобы предложение стало независимым от предыдущего.	You are tasked with extracting sentences from academic abstracts that are making some claim about some facts. Be precise. If you are not sure, do not try to guess, just skip the sentence. Some sentences may be anaphorically related to previous sentences. You should rewrite pairs of sentences where the second item is anaphorically related to the first into one sentence that does not contain anaphora.
(б) Генерация тезисов	Ты – журналист, пишущий научно-популярные статьи. Ты собираешь материал для статьи. Тебе предоставлен	You are a popular science journalist. You are collecting material for an article. You are given a list of

	список научных утверждений, взятых из ряда научных работ. Твоя задача – написать по 3 предложения для каждого научного утверждения. – Первое предложение должно быть комплиментарно заданному научному утверждению. – Второе предложение должно противоречить заданному утверждению. – Третье предложение должно на первый взгляд подтверждать утверждение, но при этом искажать его смысл, преувеличивая результаты с целью создания сенсации. Используй журналистский стиль. Старайся не составлять слишком длинные предложения.	scientific claims from some academic papers. Your task is to write for each scientific claim 3 sentences. – The first sentence should support the claim. – The second sentence should contradict the claim. – The third sentence must support it on the first glance, but should misinterpret it, exaggerating the results. Use journalistic style. Try to not make too long sentences.
(в) Генерация текстов на основе тезисов	Ты – научный журналист. Твоя специализация – научно-популярные статьи. Тебе дан набор тезисов для статьи. Основные требования к статье: – Не делай статью слишком краткой. Раскрой каждый тезис как минимум в трех предложениях. – Также не делай статью слишком длинной. Не стоит писать более четырех предложений по каждому тезису. – Старайся придерживаться журналистского стиля. Твоя задача – привлечь потенциального читателя. – По возможности не используй внешние знания. Ориентируйся только на заданные тезисы.	You are a popular science journalist. You are given a list of theses. Your task is to write a short popular scientific news item using the provided theses. Do not make the text too short. Expand on each thesis in two or three sentences. Also, you must not make the text too long. Two or three paragraphs will be enough.

В качестве основы для генерации на платформах arXiv и OpenAlex была собрана небольшая база научных публикаций (около 6.5 тыс. статей) на русском и английском языках, соответствующих тематике «применение методов машинного обучения и нейронных сетей в здравоохранении», проиндексиро-

ванная на основе векторных представлений (embedding) текстов вида «название + аннотация». Векторизация выполнялась моделью-энкодером paraphrase-multilingual-mpnet-base-v2, обеспечивающей геометрическую близость семантически схожих текстов в пространстве признаков. Собранный объем исходного материала позволил достичь приемлемого уровня наполненности, чтобы для большинства запросов могла быть сформирована достаточная выборка семантически близких (векторно-релевантных) публикаций. Генерация проводилась в несколько шагов с ручной корректировкой результатов промежуточных этапов.

1. Случайным образом из базы выбиралась статья на определенную тематику (например, эпидемия COVID-19) и от 2 до 4 близких к ней публикаций на английском или русском языках. Это было необходимо для генерации текстов, опирающихся более чем на один научный источник.

2. С помощью DeepSeek-R1 из аннотаций отобранных публикаций извлекались утверждения (тезисы) с помощью инструкции (а) из табл. 3. При необходимости проводилась ручная корректировка. Как правило, удавалось извлечь от 10 до 13 утверждений.

3. Вычислялась матрица попарного семантического сходства найденных утверждений. Поскольку на их основе предполагалась последующая генерация тезисов и далее связных текстов, отбору подлежали только семантически близкие подмножества (семантические кластеры). В качестве порогового интервала был выбран диапазон от 0.75 до 0.9. Нижняя граница должна была гарантировать связность генерируемых текстов, а верхняя позволяла исключить случаи перефразирования, когда утверждения дублируют друг друга по смыслу. В случае отсутствия в множестве утверждений, удовлетворяющих данному требованию подмножеств размером более 2, весь процесс перезапускался с самого начала.

4. Каждый семантический кластер становился основой для генерации синтетических текстов.

а) Для каждого утверждения из кластера с помощью ChatGPT по инструкции (б) из табл. 3 генерировались варианты тезисов, состоящих с ним в одном из следующих отношений: i) Поддержка (Support, **S**) означает, что тезис согласуется с исходным научным результатом; ii) Опровер-

жение (Refute, **R**) – тезис прямо противоречит исходному научному результату; iii) Неверное толкование (Misinterpretation, **M**) – научный результат не опровергается напрямую, но преувеличивается или искажается в сенсационном ключе.

b) На основе сгенерированных тезисов ChatGPT по инструкции (в) из табл. 3 создавал короткую статью в научно-популярном жанре. В зависимости от типа тезисов статьи также разделялись по категориям **S**, **R** и **M**.

Таким образом, в случае успеха за один раунд генерировалось $3 \times N$ статей, где N — количество семантических кластеров. Всего было проведено 10 успешных раундов генерации: 5 для английского языка и 5 для русского. В итоге был собран сбалансированный корпус из 30 текстов, содержащий достоверные, недостоверные и преувеличенные тезисы. В табл. 4 представлены примеры тезисов каждого вида, сгенерированных по заданному утверждению из аннотации научной публикации.

Табл. 4. Примеры генерации синтетических тезисов.

Исходное утверждение	Сгенерированный текст
Анализ гематологических показателей пациентов с тяжелым течением COVID-19 продемонстрировал, что достоверно чаще (критерий $\chi^2 < 0.001$, $p \leq 0.05$) у них имел место умеренный лейкоцитоз (в 29.5% случаев), относительная/абсолютная лимфопения (65.4%), невыраженный тромбоцитоз (58.1%).	S: Анализ показал, что у пациентов с тяжелым течением COVID-19 статистически значимо чаще выявлялись умеренный лейкоцитоз, лимфопения и невыраженный тромбоцитоз. R: Исследование не обнаружило различий в гематологических показателях между пациентами с тяжелым течением COVID-19 и остальными больными. M: Умеренный лейкоцитоз, лимфопения и тромбоцитоз являются обязательными признаками тяжелого течения COVID-19 и позволяют безошибочно определять степень тяжести заболевания.

4. ОБСУЖДЕНИЕ

Как показывают недавние исследования [1, 2], проблема доверия к научно-популярным материалам носит вполне реальный характер. Более того, прямо сейчас эта проблема обретает новое измерение в связи с распространением текстов, генерируемых LLM в ответ на вопросы пользователей. Современные

системы искусственного интеллекта склонны к «галлюцинациям» – порождению утверждений, не соответствующих действительности, но сформулированных с высокой степенью уверенности. Таким образом, разрабатываемые нами методы могут быть в перспективе адаптированы для оценки достоверности ответов языковых моделей, что выводит исследование в более широкую область доверия к искусственному интеллекту.

Важным аспектом проведенного исследования является его ориентация на русскоязычные научно-популярные тексты. Принимая во внимание, что значительная часть существующих датасетов и методов разработана для английского языка, это создает дополнительный методологический разрыв для русскоязычного сегмента. Однако, сам процесс верификации неизбежно потребует обращения к научным публикациям, представленным на английском языке. Международные базы научных публикаций индексируют главным образом англоязычные статьи, и именно английский язык в настоящее время является основным языком науки в мире. В связи с этим предлагаемые подходы и технологические решения планируется сделать как минимум двуязычными, что накладывает дополнительные требования на выбор моделей для работы с языком.

Следует отметить, что предлагаемая методология имеет ряд ограничений, которые важно явно зафиксировать.

На данный момент все компоненты архитектуры предлагаемой системы оценки достоверности информации ориентированы исключительно на работу с текстовой информацией. При этом рисунки, графики, диаграммы, таблицы и формулы, которые являются неотъемлемыми элементами научных статей, пока не рассматриваются. Это касается как исходных научно-популярных статей, так и привлекаемых научных источников. В качестве последних рассматриваются, как правило, не сами статьи, а их аннотации (название + аннотация), что объясняется их свободной доступностью через открытые API научных баз. Использование аннотаций вместо полных текстов научных статей накладывает ограничения на полноту верификации. Аннотация представляет собой сжатое изложение основных результатов работы и может не содержать важных нюансов, ограничений метода, а также деталей методики эксперимента. В результате

возможна как ложноположительная, так и ложноотрицательная классификация утверждения.

ЗАКЛЮЧЕНИЕ

Предложена методология верификации научно-популярных статей, основанная на сопоставлении утверждений из анализируемого текста с актуальными результатами научных исследований. Предлагаемый подход относится к области научного фактчекинга [1] и верификации утверждений (claim verification) и не предполагает вынесения категоричных суждений об истинности или ложности информации. Вместо этого акцент сделан на оценке степени соответствия утверждений научному знанию, выявлении логических ошибок и обнаружении стратегий выборочного цитирования.

Рассмотрена важная методологическая проблема, вызванная отсутствием унифицированных способов автоматизированного сбора размеченных данных, и предложена двухкомпонентная методика сбора и разметки данных, включающая автоматические процедуры сбора и синтеза данных на основе полемической аргументации и подтвержденных научных источников.

Рассмотренные подходы к анализу достоверности и верификации утверждений, а также автоматизации сбора датасетов имеют ряд ограничений, которые формируют направления для дальнейшего развития. Так, при наличии доступа к полным текстам научных публикаций (через открытые репозитории или ресурсы) появляется возможность значительно более глубокого анализа источников. При этом возникают проблемы, связанные с длиной контекста и необходимостью обработки мультимодальных данных, поскольку научные статьи содержат внетекстовые элементы, несущие критически важную информацию, такие как графики, диаграммы, таблицы и формулы. Остается также открытым вопрос о ручной валидации собранных данных, поскольку ни опубликованная научная статья, ни критический комментарий пользователя с подтвержденными ссылками на научные источники не дают 100%-ной гарантии истинности или ложности высказывания.

СПИСОК ЛИТЕРАТУРЫ

1. *Vladika J., Matthes F.* Scientific Fact-Checking: A Survey of Resources and Approaches // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 6215–6230.
<https://doi.org/10.18653/v1/2023.findings-acl.387>
2. *Guo Z., Schlichtkrull M., Vlachos A.* A Survey on Automated Fact-Checking // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 178–206. https://doi.org/10.1162/tacl_a_00454
3. *Vykopal I., Pikuliak M., Ostermann S., Šimko M.* Generative Large Language Models in Automated Fact-Checking: A Survey // arXiv preprint arXiv:2407.02351. arXiv, 2026.
4. *Kumar S. et al.* SciClaimHunt: A Large Dataset for Evidence-based Scientific Claim Verification // 2025 International Joint Conference on Neural Networks (IJCNN). Rome, Italy: IEEE, 2025. P. 1–10.
<https://doi.org/10.1109/IJCNN64981.2025.11227296>
5. *Vatolin A. et al.* ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies. 2025. P. 2075–7182. <https://doi.org/10.28995/2075-7182-2025-23-435-459>
6. *Min S. et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing / ed. Bouamor H., Pino J., Bali K. Singapore: Association for Computational Linguistics, 2023. P. 12076–12100.
<https://doi.org/10.18653/v1/2023.emnlp-main.741>
7. *Sitaula N. et al.* Credibility-Based Fake News Detection // Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities / ed. Shu K., Wang S., Lee D., Liu H. Cham: Springer International Publishing, 2020. P. 163–182. https://doi.org/10.1007/978-3-030-42699-6_9
8. *Chrysidis Z., Papadopoulos S.-I., Papadopoulos S., Petrantonakis P.C.* Credible, Unreliable or Leaked?: Evidence Verification for Enhanced Automated Fact-checking // 3rd ACM International Workshop on Multimedia AI against Disinformation. 2024. P. 73–81. <https://doi.org/10.1145/3643491.3660278>

9. Visser J., Lawrence J., Reed C. Reason-checking fake news // Commun. ACM. 2020. Vol. 63, No. 11. P. 38–40. <https://doi.org/10.1145/3397189>

10. Glockner M., Hou Y., Nakov P., Gurevych I. Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. <https://doi.org/10.18653/v1/2024.acl-long.240>

11. Schlichtkrull M., Guo Z., Vlachos A. AVERITEC: a dataset for real-world claim verification with evidence from the web // Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023. P. 65128–65167. <https://doi.org/10.52202/075280-2842>

12. Wang W.Y. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) / ed. Barzilay R., Kan M.-Y. Vancouver, Canada: Association for Computational Linguistics, 2017. P. 422–426. <https://doi.org/10.18653/v1/P17-2067>

13. Saakyan A., Chakrabarty T., Muresan S. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) / ed. Zong C., Xia F., Li W., Navigli R. Online: Association for Computational Linguistics, 2021. P. 2116–2129. <https://doi.org/10.18653/v1/2021.acl-long.165>

14. Liu H. et al. Retrieval augmented scientific claim verification // JAMIA Open. 2024. Vol. 7, № 1. P. ooae021. <https://doi.org/10.1093/jamiaopen/ooae021>

15. Kotonya N., Toni F. Explainable Automated Fact-Checking for Public Health Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020. P. 7740–7754. <https://doi.org/10.18653/v1/2020.emnlp-main.623>

16. Wadden D. et al. Fact or Fiction: Verifying Scientific Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / ed. Webber B., Cohn T., He Y., Liu Y. Online: Association for Computa-

al Linguistics, 2020. P. 7534–7550. <https://doi.org/10.18653/v1/2020.emnlp-main.609>

17. *Glockner M., Hou Y., Nakov P., Gurevych I.* Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. <https://doi.org/10.18653/v1/2024.acl-long.240>

18. *Glockner M., Hou Y., Nakov P., Gurevych I.* Grounding Fallacies Misrepresenting Scientific Publications in Evidence // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) / ed. Chiruzzo L., Ritter A., Wang L. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. P. 9732–9767. <https://doi.org/10.18653/v1/2025.naacl-long.491>

19. *Walton D.* Argumentation theory: A very short introduction // Argumentation in artificial intelligence. Boston, MA: Springer US, 2009. P. 1–22. https://doi.org/10.1007/978-0-387-98197-0_1

20. *Akhmadeeva I., Kononenko I., Sidorova E., Shestakov V.* Using Rhetorical Structures to Analyze Argumentation in Scientific Communication Texts // Computational Linguistics and Intellectual Technologies. 2025. P. 1–11. <https://doi.org/10.28995/2075-7182-2025-23-1-11>

21. *Chistova E.* End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 3376–3391. <https://doi.org/10.18653/v1/2023.findings-acl.209>

22. *Sidorova E.A. et al.* Enriching the ontology of argumentation based on critical questions // 2025 IEEE XVII international scientific and technical conference on actual problems of electronic instrument engineering (APEIE). 2025. P. 1–5. <https://doi.org/10.1109/APEIE66761.2025.11289342>

23. *Серый А.С.* Доверие к данным при пополнении онтологий и графов знаний // Онтология проектирования. 2023. Т.13, №1 (47). С. 113–124. <https://doi.org/10.18287/2223-9537-2023-13-1-113-124>

24. Ильина Д.В., Тимофеева М.К., Кононенко И.С., Сидорова Е.А. Генерация критических вопросов для схем аргументации Д. Уолтона // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Том. 24, № 1. С. 87–101. <https://doi.org/10.25205/1818-7935-2026-24-1-87-101>

AN APPROACH TO ASSESSING THE RELIABILITY OF INFORMATION IN POPULAR SCIENCE DISCOURSE BASED ON THE ANALYSIS OF ARGUMENTATION AND SCIENTIFIC FACT-CHECKING

Yu. A. Zagorulko¹ [0000-0002-7111-6524], I. R. Akhmadeeva² [0000-0002-7371-1087],

A. S. Sery³ [0000-0001-8275-4700], E. A. Sidorova⁴ [0000-0001-8731-3058],

^{1–4}*A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

¹zagor@iis.nsk.su, ²i.r.akhmadeeva@iis.nsk.su, ³alexey.seryj@iis.nsk.su,

⁴Isidorova@iis.nsk.su

Abstract

This paper examines the challenges of assessing the reliability of information extracted from popular science texts. Popular science is known to be characterized by a free style of presentation, loose references to primary sources and expert opinions, and the presence of controversial statements and conclusions. Clarifying these components of the author's reasoning can significantly increase the credibility among critical readers or, conversely, refute the authors' assertions.

This paper proposes a new approach to assessing the reliability of information contained in popular science publications. This approach utilizes D. Walton's argumentation theory to identify potentially unreliable claims and logical fallacies. Searching scientific databases and calculating trust metrics for scientific publications based on their metadata and scientific ratings will enable the verification of controversial claims and an assessment of their reliability. The use of large language models will enable the generation of explanations for decisions made.

To support the development and validation of this approach, two data collection methods are proposed. The first uses existing corpora of Russian-language pop-

ular science texts and commentaries with annotated argumentation. The annotation includes, among other things, polemical argumentation schemes that challenge the author's thesis or argument. The second method utilizes LLM for generating synthetic theses and popular science articles, based on actual scientific publications.

Keywords: *claim verification, analysis of argumentation, scientific fact-checking, reliability of information, automatic data collection.*

REFERENCES

1. Vladika J., Matthes F. Scientific Fact-Checking: A Survey of Resources and Approaches // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 6215–6230.
<https://doi.org/10.18653/v1/2023.findings-acl.387>
2. Guo Z., Schlichtkrull M., Vlachos A. A Survey on Automated Fact-Checking // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 178–206. https://doi.org/10.1162/tacl_a_00454
3. Vykopal I., Pikuliak M., Ostermann S., Šimko M. Generative Large Language Models in Automated Fact-Checking: A Survey // arXiv preprint arXiv:2407.02351. arXiv, 2026.
4. Kumar S. et al. SciClaimHunt: A Large Dataset for Evidence-based Scientific Claim Verification // 2025 International Joint Conference on Neural Networks (IJCNN). Rome, Italy: IEEE, 2025. P. 1–10.
<https://doi.org/10.1109/IJCNN64981.2025.11227296>
5. Vatolin A. et al. ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies. 2025. P. 2075–7182. <https://doi.org/10.28995/2075-7182-2025-23-435-459>
6. Min S. et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing / ed. Bouamor H., Pino J., Bali K. Singapore: Association for Computational Linguistics, 2023. P. 12076–12100.
<https://doi.org/10.18653/v1/2023.emnlp-main.741>
7. Sitaula N. et al. Credibility-Based Fake News Detection // Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and

Opportunities / ed. Shu K., Wang S., Lee D., Liu H. Cham: Springer International Publishing, 2020. P. 163–182. https://doi.org/10.1007/978-3-030-42699-6_9

8. *Chrysidis Z., Papadopoulos S.-I., Papadopoulos S., Petrantonakis P.C.* Credible, Unreliable or Leaked?: Evidence Verification for Enhanced Automated Fact-checking // 3rd ACM International Workshop on Multimedia AI against Disinformation. 2024. P. 73–81. <https://doi.org/10.1145/3643491.3660278>

9. *Visser J., Lawrence J., Reed C.* Reason-checking fake news // Commun. ACM. 2020. Vol. 63, No. 11. P. 38–40. <https://doi.org/10.1145/3397189>

10. *Glockner M., Hou Y., Nakov P., Gurevych I.* Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. <https://doi.org/10.18653/v1/2024.acl-long.240>

11. *Schlichtkrull M., Guo Z., Vlachos A.* AVERITEC: a dataset for real-world claim verification with evidence from the web // Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023. P. 65128–65167. <https://doi.org/10.52202/075280-2842>

12. *Wang W.Y.* “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) / ed. Barzilay R., Kan M.-Y. Vancouver, Canada: Association for Computational Linguistics, 2017. P. 422–426. <https://doi.org/10.18653/v1/P17-2067>

13. *Saakyan A., Chakrabarty T., Muresan S.* COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) / ed. Zong C., Xia F., Li W., Navigli R. Online: Association for Computational Linguistics, 2021. P. 2116–2129. <https://doi.org/10.18653/v1/2021.acl-long.165>

14. *Liu H. et al.* Retrieval augmented scientific claim verification // JAMIA Open. 2024. Vol. 7, № 1. P. ooae021. <https://doi.org/10.1093/jamiaopen/ooae021>

15. *Kotonya N., Toni F.* Explainable Automated Fact-Checking for Public Health Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Lan-

guage Processing (EMNLP). Online: Association for Computational Linguistics, 2020. P. 7740–7754. <https://doi.org/10.18653/v1/2020.emnlp-main.623>

16. *Wadden D. et al.* Fact or Fiction: Verifying Scientific Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / ed. Webber B., Cohn T., He Y., Liu Y. Online: Association for Computational Linguistics, 2020. P. 7534–7550. <https://doi.org/10.18653/v1/2020.emnlp-main.609>

17. *Glockner M., Hou Y., Nakov P., Gurevych I.* Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. <https://doi.org/10.18653/v1/2024.acl-long.240>

18. *Glockner M., Hou Y., Nakov P., Gurevych I.* Grounding Fallacies Misrepresenting Scientific Publications in Evidence // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) / ed. Chiruzzo L., Ritter A., Wang L. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. P. 9732–9767. <https://doi.org/10.18653/v1/2025.naacl-long.491>

19. *Walton D.* Argumentation theory: A very short introduction // Argumentation in artificial intelligence. Boston, MA: Springer US, 2009. P. 1–22. https://doi.org/10.1007/978-0-387-98197-0_1

20. *Akhmadeeva I., Kononenko I., Sidorova E., Shestakov V.* Using Rhetorical Structures to Analyze Argumentation in Scientific Communication Texts // Computational Linguistics and Intellectual Technologies. 2025. P. 1–11. <https://doi.org/10.28995/2075-7182-2025-23-1-11>

21. *Chistova E.* End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 3376–3391. <https://doi.org/10.18653/v1/2023.findings-acl.209>

22. *Sidorova E.A. et al.* Enriching the ontology of argumentation based on critical questions // 2025 IEEE XVII international scientific and technical conference on actual problems of electronic instrument engineering (APEIE). 2025. P. 1–5.

<https://doi.org/10.1109/APEIE66761.2025.11289342>

23. Sery A. S. Data credibility when populating ontologies and knowledge graphs // *Ontology of Designing*. 2023. Vol. 13, № 1. P. 113–124.

<https://doi.org/10.18287/2223-9537-2023-13-1-113-124>

24. Ilina D.V., Timofeeva M.K., Kononenko I.S., Sidorova E.A. Generating Critical Questions for D. Walton's Argumentation Schemes// *NSU Vestnik. Series: Linguistics and Intercultural Communication*. 2026. Vol. 4, No. 1. P. 87–101. (In Russ.)

<https://doi.org/10.25205/1818-7935-2026-24-1-87-101>

СВЕДЕНИЯ ОБ АВТОРАХ



ЗАГОРУЛЬКО Юрий Алексеевич – кандидат технических наук, заведующий лабораторией Института систем информатики им. А. П. Ершова СО РАН, доцент кафедры программирования и кафедры систем информатики Новосибирского государственного университета. В списке научных трудов более 300 публикаций в области искусственного интеллекта, разработки интеллектуальных систем, инженерии знаний, онтологического моделирования и компьютерной лингвистики.

Yury Alekseevich ZAGORULKO – Candidate of Technical Sciences, head of Laboratory at A. P. Ershov Institute of Informatics Systems at the Siberian Branch of RAS, associate professor at Novosibirsk State University. List of scientific works includes more than 300 publications in the fields of AI, Knowledge and Ontology Engineering, Intelligent System Development and Computational Linguistics.

email: zagor@iis.nsk.su

ORCID: 0000-0002-7111-6524



АХМАДЕЕВА Ирина Равильевна – младший научный сотрудник Института систем информатики им. А. П. Ершова СО РАН, ассистент кафедры программирования Новосибирского государственного университета. В списке научных трудов около 30 работ в области искусственного интеллекта, разработки интеллектуальных систем и компьютерной лингвистики.

Irina Ravilevna AKHMADEEVA – she is a junior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, assistant lecturer at Novosibirsk State University. List of scientific works includes about 30 publications in the fields of AI, Intelligent System Development and NLP.

email: i.r.akhmadeeva@iis.nsk.su

ORCID: 0000-0002-7371-1087

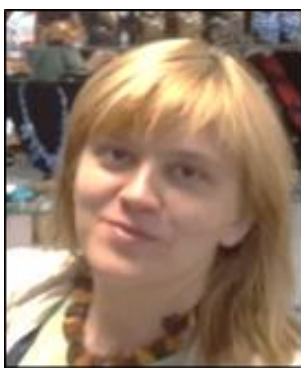


СЕРЫЙ Алексей Сергеевич – Окончил НГУ в 2010 г. Младший научный сотрудник лаборатории искусственного интеллекта Института систем информатики им. А. П. Ершова СО РАН. Старший преподаватель кафедры систем информатики факультета информационных технологий Новосибирского государственного университета. В списке научных трудов более 40 работ в области представления знаний и обработки естественного языка.

Alexey Sergeevich SERV – holds a master's degree in mathematics from Novosibirsk State University (2010) and a junior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS. Senior Teacher at Novosibirsk State University. He is the author of more than 40 papers in the fields of NLP systems and Knowledge Representation.

email: alexey.seryj@iis.nsk.su

ORCID: 0000-0001-8275-4700



СИДОРОВА Елена Анатольевна – кандидат физико-математических наук, старший научный сотрудник Института систем информатики им. А. П. Ершова СО РАН, доцент кафедры программирования и кафедры систем информатики Новосибирского государственного университета. В списке научных трудов более 160 работ в области компьютерной лингвистики, онтологического инжиниринга и разработки интеллектуальных систем.

Elena Anatolievna SIDOROVA – PhD (2006). She is a senior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, associate professor at Novosibirsk State University. List of scientific works includes more than 160 peer-reviewed publications in the field of Computational Linguistics, Intelligent System Development, Knowledge and Ontology Engineering.

email: lsidorova@iis.nsk.su

ORCID: 0000-0001-8731-3058

Материал поступил в редакцию 3 августа 2026 года