

ОПРОС СИНТЕТИЧЕСКИХ РЕСПОНДЕНТОВ: НОВЫЕ ВОЗМОЖНОСТИ ИЛИ УГРОЗА ДОСТОВЕРНОСТИ ДАННЫХ?

Е. Е. КУЧКО¹⁾, М. В. ВИШНЯКОВА^{1), 2)}, Я. И. МАРЧУК²⁾

¹⁾Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь

²⁾Гипросвязь, ул. Сурганова, 24, 220012, г. Минск, Беларусь

Аннотация. Рассмотрены основные направления применения моделей искусственного интеллекта в профессиональной деятельности социолога. Особое внимание уделено использованию больших языковых моделей и их роли в генерации ответов синтетических респондентов. Отмечено, что использование синтетических респондентов позволяет провести исследование в сжатые сроки, осуществить взаимодействие с труднодоступными группами, устранить эффект интервьюера. Подчеркнуты методологические ограничения опроса синтетических респондентов: невозможность учета эмоциональных реакций, редуцирование социального контекста, низкое качество обучения больших языковых моделей, вызванное недоступностью тренировочных датасетов.

Ключевые слова: большая языковая модель; искусственный интеллект; синтетический респондент; социологическое исследование; машинное обучение.

SURVEY OF SYNTHETIC RESPONDENTS: NEW OPPORTUNITIES OR THREATS TO DATA RELIABILITY?

E. E. KUCHKO^a, M. V. VISHNYAKOVA^{a, b}, Ya. I. MARCHUK^b

^aBelarusian State University, 4 Niezaliezhnasci Avenue, Minsk 220030, Belarus

^bGiprosvyaz, 24 Surganova Street, Minsk 220012, Belarus

Corresponding author: E. E. Kuchko (avesol@yandex.ru)

Abstract. The article explores the main applications of artificial intelligence models in the professional work of sociologists. Particular attention is paid to the use of large language models and their role in generating responses from synthetic

Образец цитирования:

Кучко ЕЕ, Вишнякова МВ, Марчук ЯИ. Опрос синтетических респондентов: новые возможности или угроза достоверности данных? Журнал Белорусского государственного университета. Социология. 2025;3:31–35.
EDN: HQLWGG

For citation:

Kuchko EE, Vishnyakova MV, Marchuk YaI. Survey of synthetic respondents: new opportunities or threats to data reliability? Journal of the Belarusian State University. Sociology. 2025;3:31–35. Russian.
EDN: HQLWGG

Авторы:

Елена Евгеньевна Кучко – доктор социологических наук, профессор; профессор кафедры социологии факультета философии и социальных наук.

Маргарита Владимировна Вишнякова – кандидат социологических наук; доцент кафедры социологии факультета философии и социальных наук¹⁾, начальник научно-исследовательского отдела анализа и прогноза развития Центра перспективных исследований в сфере цифрового развития²⁾.

Яна Игоревна Марчук – специалист научно-исследовательского отдела цифровых технологий Центра перспективных исследований в сфере цифрового развития.

Authors:

Elena E. Kuchko, doctor of science (sociology), full professor; professor at the department of sociology, faculty of philosophy and social sciences.

avesol@yandex.ru

<https://orcid.org/0000-0003-2376-4639>

Margarita V. Vishnyakova, PhD (sociology); associate professor at the department of sociology, faculty of philosophy and social sciences^a, and head of the department of analysis and development forecast, Centre for Advanced Research in Digital Development^b.

vishnyakova@giprosvjaz.by

Yana I. Marchuk, specialist at the department of digital technologies, Centre for Advanced Research in Digital Development.

marchuk@giprosvjaz.by

respondents. It is noted that the use of synthetic respondents allows for faster research, access to hard-to-reach groups, and the elimination of the interviewer effect. At the same time, methodological limitations are emphasised, including the inability to account for emotional reactions, the reduction of social context, and the low quality of training of large language models caused by the unavailability of training datasets.

Keywords: large language model; artificial intelligence; synthetic respondent; sociological research; machine learning.

Введение

В ходе эволюции опросных технологий кардинально изменились методология и особенности проведения исследований. Широкое использование возможностей сети Интернет, а также активное развитие технологий искусственного интеллекта (ИИ) в сочетании с ростом вычислительных мощностей трансформируют привычные подходы к организации исследовательской деятельности и открывают новые возможности для сбора и анализа данных. Сегодня ключевым трендом становится симбиоз естественного и искусственного интеллектов для изучения проблемных ситуаций социологами.

Способность ИИ имитировать когнитивные функции человека (в том числе самообучение и поиск решений без заранее заданного алгоритма) и достигать результатов, сопоставимых с результатами интеллектуальной деятельности, позволяет использовать его алгоритмы для решения ряда исследовательских задач (в том числе узкоспециализированных) – от генерации дизайна исследования до предоставления итоговых документов.

ИИ дает возможность проводить различия, классифицировать и прогнозировать данные, что допускает его применение в количественных исследованиях при статистической обработке данных. Апробация потенциала ИИ для анализа данных количественного исследования отражена в работах Е. Е. Фомина (анализ результатов анкетного опроса) [1], В. В. Круглова и М. И. Дали (кластерный анализ) [2], Д. А. Котова (факторный анализ) [3], Е. Г. Ашихмина и его соавторов В. В. Левченко и Г. И. Селетковой (описательная статистика – частотные таблицы, средние значения и многомерная статистика – сводные таблицы, корреляции) [4].

Широкий потенциал ИИ кроется в анализе данных качественных исследований для категоризации ответов, определения модальности текста [5], анализа визуальных данных (изображения, фотографии) [6]. Кроме того, ИИ позволяет социологу автоматизировать ряд задач – создавать тексты отчетов, готовить презентации, генерировать код для визуализации или анализа данных.

Основная часть

В последнее время популярны большие языковые модели (*large language model*, LLM), что связано с их доступностью и стремительным развитием. LLM обучены как на обширном наборе открытых данных, содержащихся в сети Интернет (отчеты, отзывы, файлы cookies и др.)¹, так и на закрытых, узкоспециализированных наборах данных, полученных в результате исследований, что повышает точность ответов в нишевых областях.

Возможности LLM, позволяющие прогнозировать потенциально релевантные реальным респондентам точки зрения, лежат в основе формирования феномена синтетического респондента.

Под синтетическим респондентом мы понимаем искусственно созданную при помощи алгоритмов ИИ симуляцию реального представителя целевой аудитории, формируемую с учетом заданных исследователем характеристик (социально-демографических, психографических, поведенческих и пр.). Так, «алгоритмы модели генерируют ответы, которые отражают прогнозируемые мнения человека

с указанными характеристиками, что позволяет исследователям изучать гипотетические точки зрения, не полагаясь на прямой вклад реальных людей» [7]. Генерация и опрос синтетических респондентов позволяют оперативно собирать данные. Синтетические респонденты не испытывают усталости от интенсивных мыслительных процессов, что дает возможность исследовать широкий спектр вопросов и проверить множество гипотез в сжатые сроки.

Опрос синтетических респондентов при изучении уязвимых групп имеет ряд потенциальных методологических преимуществ. Наиболее значимым преимуществом является преодоление проблем взаимодействия с труднодоступными категориями респондентов. В традиционных исследованиях такие группы часто недостаточно представлены из-за географической изоляции, языковых барьеров, опасений опрашиваемых раскрывать свою принадлежность к стигматизированным сообществам или из-за сложностей установления контакта. Синтетические респонденты могут помочь в моделировании

¹Незговорова М. Синтетические респонденты: от теории к практике [Электронный ресурс]. URL: <https://socforum.niioz.ru/wp-content/uploads/2024/11/1.pdf> (дата обращения: 13.05.2025).

мнения таких групп на основе имеющихся данных, обеспечивая тем самым некоторое представление о социальной реальности.

В исследованиях с участием реальных респондентов из числа уязвимых групп может наблюдаться эффект социальной желательности, когда участники сознательно или бессознательно корректируют свои ответы в соответствии с существующими социальными нормами или собственными представлениями об ожиданиях интервьюера. Синтетические респонденты, будучи свободными от подобных психологических барьеров, дают ответы, не искаженные страхом последствий или стремлением произвести благоприятное впечатление.

Важным преимуществом опроса синтетических респондентов является их потенциальная способность нивелировать эффект интервьюера, что минимизирует систематические искажения, характерные для традиционных опросов. При изучении реальных уязвимых групп этот эффект может быть особенно выражен, так как представители маргинализированных сообществ могут демонстрировать различное поведение при взаимодействии с интервьюерами из гомогенной чужой аутгруппы.

Синтетический респондент генерирует ответы, не зависящие от внешних факторов, что особенно актуально при изучении сенситивных тем. В таких случаях технология позволяет обойти ограничения, связанные с возможной ретравматизацией жертв или риском юридических последствий для респондентов при изучении противоправных действий. Кроме того, следует помнить, что отсутствие эмоциональной составляющей в ответах синтетических респондентов в ходе исследования высокочувствительных тем может привести к опасной дегуманизации – излишне упрощенному и механическому пониманию сложных социальных явлений, мотивов и последствий.

Таким образом, использование синтетических респондентов для изучения немногочисленных и уязвимых групп, с одной стороны, помогает преодолеть некоторые ограничения традиционных методов, с другой – создает иллюзию репрезентации там, где ее в действительности нет. Это особенно опасно в случае принятия социально значимых решений на основе подобных данных, поэтому применение такого подхода требует разработки строгих методологических стандартов и этических протоколов, гарантирующих, что потенциальные преимущества будут достигнуты не за счет снижения уровня научной достоверности и нарушения прав малочисленных и уязвимых групп на подлинное представительство в исследованиях. Соответственно, этические последствия внедрения синтетического респондента в социологическую практику носят фундаментальный характер, затрагивая саму суть

социологического познания. Ключевой этический вызов заключается в сохранении баланса между методологической эффективностью и онтологической аутентичностью объекта исследования. Однако даже при соблюдении всех этических предосторожностей вопрос экономической целесообразности исследований с использованием подобных технологий остается открытым.

На первый взгляд, использование синтетических респондентов позволяет существенно сократить расходы за счет отсутствия затрат на проведение полевого этапа исследования (рекрутинг участников, организация вознаграждения и иные сопряженные расходы). Однако финансовые затраты на разработку, обучение и внедрение LLM-моделей являются серьезным препятствием для их широкого применения и представляют высокие риски для небольших исследовательских коллективов или малобюджетных проектов. В случае использования готовых проприетарных решений с безвозмездным доступом (например, *Deepseek*) или готовых решений с открытым кодом не следует полагаться на высокую достоверность полученных данных. Даже после дополнительного обучения модели на собственных датасетах необходимо ограничивать изучаемую область разведывательными исследованиями – поиском инсайтов в фактических данных, структурированием исторических данных, предварительным тестированием гипотез. Важно помнить, что результаты такой работы подлежат обязательной последующей валидации посредством опроса реальных респондентов.

Качество исходных данных, послуживших основой обучения модели, является приоритетным. Поскольку языковая модель обучается на большом массиве данных, сгенерированных представителями определенных социально-демографических групп, анализ данных на ее основе может приводить к искажениям, не отражать реальной картины. В связи с этим исследователям необходимо контролировать качество подготовленной модели во время ее использования. Инструментами контроллинга могут выступать сверка с внешними источниками данных, консультации экспертов, проведение серии методических экспериментов, направленных на валидацию данных, А/В-тестирование моделей на ответах реальных респондентов, а также бенчмаркинг. Последний позволяет оценить производительность LLM по ряду задач и метрик с упором на адекватность логических рассуждений модели и техническую точность результатов ее работы. Основными критериями производительности являются понимание кода, логическое мышление, понимание прочитанного, общие знания о мире и профильные компетенции.

Как утверждает И. Ю. Барская, ключевым инструментом в области оценки рассуждений и проверки

знаний LLM в таких областях, как гуманитарные и социальные науки, история, информатика и право, выступает Multimodal Language Understanding (MLLU). Несмотря на высокую оценку ряда моделей по данному тесту (к примеру, GPT-4-omni – 88,7 %, Claude 3 Opus – 86,8 %), это не означает, что они стали экспертами в определенной сфере. Бенчмарки не учитывают возможную сложность и контекст реальных взаимодействий, а знание фактов и умение применять их на практике не всегда позволяет подойти творчески к пониманию проблемы или решить ее². В связи с этим жизненный цикл LLM предполагает постоянное совершенствование модели и регулярную валидацию выходных данных.

Кроме того, следует учитывать, что дополнительное обучение моделей не исключает вероятность случайных ошибок и намеренных искажений в ответах, поэтому для подтверждения валидности полученных данных фактчекинг необходимо проводить на всех этапах работы с LLM.

Применение LLM в некотором смысле ограничивает исследователя, использование синтетических респондентов не позволяет собрать креативную фокус-группу или провести глубинное интервью, синтетический респондент сможет лишь по-новому применить заложенные ранее шаблоны.

Исследователи отмечают, что использование одних и тех же моделей генерации синтетических данных не позволяет брендам получить уникальные инсайты, которые помогли бы им выделиться на рынке³. Одинаковая методология генерации данных для конкурентных компаний может привести к тиражированию системных ошибок, сделать результаты предсказуемыми, что в целом снизит конкурентное преимущество, приведет к некорректным стратегическим решениям, а в перспективе и вовсе исказит конъюнктуру рынка. При этом отметим, что конкурентное преимущество в исследованиях рынка кроется в промптинге – постановке вопросов и умении интерпретировать полученные данные с учетом внутренних данных компании (финансовая аналитика) и особенностей интеграции полученных данных в бизнес-процессы.

Методологическая прозрачность и научная строгость социологических исследований подтверждается раскрытием сроков проведения сбора данных, объема выборочной совокупности, ошибок выборки, метода ее формирования, метода сбора данных. Для демонстрации научной ценности исследования, проведенного с использованием опроса синтетических респондентов, помимо указанных выше параметров, репрезентации должны подвергаться характеристики, описывающие условия разработки и функцио-

нирования LLM. К ключевым мы относим следующие: 1) описание использованной модели с указанием названия, версии, разработчика, модели распространения; 2) описание набора обучающих данных (если возможно), поскольку это определяет уровень знаний LLM и верифицирует ее ответы; 3) принципы и характеристики, влияющие на формирование аудитории синтетических респондентов; 4) метод генерации ответов (вероятностное моделирование типовых ответов и пр.); 5) используемые промпты (аналог инструментария в классическом формате социологического исследования, гайд глубинного интервью или фокус-группы), содержащие перечень вопросов к изучению; 6) протокол проведения опроса синтетического респондента на базе LLM с описанием алгоритма взаимодействия с языковой моделью.

Российские социологи отмечают, что для использования LLM-моделей в социологическом исследовании LLM не хватает опыта, эмоций, мотивации [8]. Действительно, излишняя рациональность синтетических респондентов – это серьезный методологический риск, не позволяющий социологу оценить эмоциональный фон, который нередко служит основой для понимания особенностей поведения. К примеру, политическое поведение может быть основано на эмоциональной привязанности, а потребительское – на желании быть как все или, наоборот, противопоставлять себя группе.

Исследователю важно помнить, что при работе с ИИ существуют риски, связанные с обработкой данных, содержащих конфиденциальную информацию. В связи с этим возникает несколько вопросов: не является ли сервер, обрабатывающий данные, сторонним, какие последствия несут владельцы ИИ-платформы в случае возникновения киберинцидентов и т. п. Отсутствие прозрачных правовых норм в области работы с ИИ создает предпосылки для возможных конфликтов между разработчиками ИИ, исследователями, заказчиками и даже их потенциальными конкурентами.

Кроме того, привлечение ИИ на любом этапе исследовательского цикла означает его участие в данном процессе, что порождает вопрос об авторстве материалов, созданных с его помощью. Указанная проблема лежит в плоскости прав интеллектуальной собственности.

Безусловно, представленные ограничения варьируют в зависимости от уровня готовности к использованию технологий ИИ, зрелости нормативно-правовой базы, понимания социологами принципов машинного обучения и промпт-инжиниринга, доступности вычислительных мощностей и др.

²Barskaya I. Benchmarks for LLMs // Unite.AI : website. URL: <https://www.unite.ai/benchmarks-for-llms/> (date of access: 02.06.2025).

³Using synthetic participants for market research // Market Research Society : website. URL: https://www.mrs.org.uk/pdf/MRS_Delphi_synthetic.pdf (date of access: 23.06.2025).

Заключение

Применение ИИ в исследовательском цикле изменяет привычную природу сбора и анализа данных. Ввиду того что использование алгоритмов ИИ при проведении исследований находится на стадии становления, а их эффективность напрямую связана с преодолением существующих ограничений, утверждать о формировании нового этапа в эволюции опросных технологий преждевременно. Вероятно, мы подходим к пику развития постгэлловского периода изучения общественного мнения, в котором цифровые технологии трансформируют саму при-

роду познания, открывая исследователям возможность изучения гибридной реальности, существующей на стыке реального и синтетического, живого и цифрового. Использование ИИ в профессиональной деятельности социолога все больше дистанцирует исследователя от социума – эти изменения требуют новых подходов к пониманию принципов познания, исследовательской этики, профессиональной идентичности социолога. Сохранить научную строгость и достоверность данных в условиях таких изменений – главный вызов настоящего времени.

Библиографические ссылки

1. Фомина ЕЕ. Применение аппарата нейронных сетей для анализа результатов анкетирования. *Вестник Пермского национального исследовательского политехнического университета. Социально-экономические науки*. 2020;2:99–110. DOI: 10.15593/2224-9354/2020.2.8.
2. Круглов ВВ, Дли МИ. Применение аппарата нейронных сетей для анализа социологических данных. *Социологические исследования*. 2001;9:112–114.
3. Котов Д. Алгоритмы искусственного интеллекта в прикладных социологических исследованиях. *Социодиггер* [Интернет]. 2023 [процитировано 24 июня 2025 г.];4(7–8). Доступно по: https://sociodigger.ru/articles/articles-page/algorithmy-iskusstvennogo-intellekta-v-prikladnykh-sociologicheskikh-issledovaniyakh#_ftn1.
4. Ашихмин ЕГ, Левченко ВВ, Селеткова ГИ. Опыт применения больших языковых моделей для анализа количественных социологических данных. *Вестник университета*. 2024;11:205–215. DOI: 10.26425/1816-4277-2024-11-205-215.
5. Ашихмин ЕГ, Левченко ВВ, Селеткова ГИ. Опыт применения больших языковых моделей для анализа социологических данных, полученных в результате интервью о восприятии студентами предпринимательской деятельности. *Цифровая социология*. 2024;3:4–14. DOI: 10.26425/2658-347X-2024-7-3-4-14.
6. Драч ВЕ, Торкунова ЮВ. Использование генеративного искусственного интеллекта для социологических исследований. *Дискурс*. 2025;11(1):52–70. DOI: 10.32603/2412-8562-2025-11-1-52-70.
7. Shrestha P, Dario K, Koaik F, Schnider R, Sayess D, Binbaz MS. Can synthetic survey participants substitute for humans in global policy research? *Behavioral Science & Policy*. 2025;2:26–45. DOI: 10.1177/23794607241311793.
8. Филипова АГ, Абросимова ЕЕ, Зубова ОГ. Метод синтетических фокус-групп в контексте цифровой трансформации социологических исследований. *Социодинамика*. 2025;5:1–17. DOI: 10.25136/2409-7144.2025.5.74430.

Статья поступила в редакцию 24.07.2025.
Received by editorial board 24.07.2025.