Федеральный центр экспертизы в сфере безопасности
и развития искусственного интеллекта имени Д. А. Поспелова

Исследовательская повестка: проверяемость систем искусственного интеллекта

С чего начать проверку языковых моделей и как результаты первой серии связаны с семью направлениями исследований Центра.

Доклады и обзоры — 11 ноября 2026 года

Внедрение языковой модели в работу с документами ставит перед организацией конкретный вопрос: какой ответ можно принять, какой нужно проверить и когда следует запросить дополнительные сведения. Для Центра это основание начать прикладную серию с проверяемости результата.

Первая серия проверяет три связанных свойства: обоснованность ответа, сохранение задачи при посторонних указаниях и повторяемость с учетом расходов. Общий корпус позволяет сопоставлять эти свойства на одинаковых ситуациях. Результат программы — данные для выбора проверок и правил передачи человеку, а также перечень вопросов, которым нужны реальные документы и наблюдения за работой организации.

Повестка связывает три отчета и единый протокол CP-RU-DOC-2026-01 с семью направлениями исследований Центра. Завершенная серия охватывает шесть моделей и 96 синтетических заданий; исследования регулирования, внедрения, данных и информационной среды сохраняют собственные предметы и способы проверки.

Основания выбора приоритетов

Состав общей эмпирической серии

Происхождение данных и проверка результатов

Связь с дальнейшими направлениями Центра

Научные основания

Порядок цитирования

Основания выбора приоритетов

Проверяемость ответов языковых моделей выбрана центральной задачей прикладной исследовательской серии Центра, программа которой представлена 11 ноября 2026 года. Серия связывает три вопроса: умеет ли модель отвечать по предоставленным документам, сохраняет ли установленную задачу при

посторонних указаниях внутри документа и насколько устойчивы результаты при повторении и перефразировании запроса. Общий корпус заданий позволяет рассматривать качество, безопасность и стоимость в одних условиях.

Выбор связан с семью направлениями исследований Центра. Для работы с документами недостаточно получить связный текст: требуется установить источник каждого существенного утверждения, обнаружить отсутствие сведений и передать противоречие человеку. Ошибка в таком действии может сохраниться и в кратком ответе, и в аккуратно заполненной таблице. Поэтому отправной точкой становится наблюдаемое поведение системы, а единицей анализа — отдельное задание с заранее установленным правильным результатом.

Эмпирическая серия выполнена по отдельному воспроизводимому протоколу. Она дополняет программу на 2026–2028 годы и сохраняет предусмотренные ею самостоятельные исследования регулирования, внедрения, данных, информационной среды и терминологии.

Приоритет определяется четырьмя признаками: связью с практическим решением, возможностью заранее задать проверяемый результат, доступностью данных с понятным происхождением и осуществимостью независимого повторения. Дополнительно учитываются цена ошибки и объем ресурсов, необходимых для исследования. Сначала выбирается вопрос, затем способ измерения и только после этого состав сравниваемых систем.

Приведенные ниже оценки — качественное суждение при составлении программы. Они отражают пригодность вопроса для текущей серии, а не результаты опроса организаций или количественный рейтинг общественной значимости. Высокая готовность означает, что исходные данные и правило проверки можно определить до запуска моделей. Средняя готовность предполагает дополнительный сбор материалов. Длительная подготовка означает необходимость полевых данных, специальных прав доступа или отдельной измерительной инфраструктуры.

Направление	Практический приоритет	Готовность к текущей серии	Следующее необходимое свидетельство
Безопасность и надежность	Подтверждение ответа источником, отказ при недостатке сведений, устойчивость к инъекциям	Высокая	Парные ответы на одинаковых заданиях, журнал запусков, интервальные оценки ошибок
Регулирование и оценка регулирующего воздействия	Соразмерность требований к проверке и документированию систем	Средняя	Данные о затратах участников и разбор конкретных вариантов требований

Направление	Практический приоритет	Готовность к текущей серии	Следующее необходимое свидетельство
Государственное управление и услуги	Понятное основание ответа и передача спорного случая человеку	Высокая для моделирования	Наблюдения за действительными маршрутами обращения и работой специалистов
Экономика внедрения	Затраты на получение корректного результата с учетом повторных запросов	Высокая для расходов на обращения к модели	Полные затраты интеграции, сопровождения и проверки человеком
Данные и вычислительная инфраструктура	Происхождение корпуса, достаточность сведений и повторяемость испытаний	Высокая	Паспорт набора, контрольные суммы и проверка переноса на другие источники
Синтезированный контент и информационная среда	Надежность выявления и сохранность сведений о происхождении материалов	Длительная подготовка	Отдельный корпус подлинных и синтезированных материалов с историей преобразований
Стандартизация и терминология	Единые определения ошибки, отказа, успешной атаки и воспроизводимости	Высокая для рабочего словаря	Сопоставление определений разных документов и проверка согласованности применения

Стоимость обращения к модели измеряется непосредственно, а экономический эффект внедрения требует наблюдения за организацией. Ссылка на документ проверяется в лабораторном задании, а возможность оспорить решение требует изучения всей административной процедуры.

Состав общей эмпирической серии

Корпус включает 96 синтетических русскоязычных заданий в шести семействах; для каждого задания использованы восемь условий, и завершенная матрица включает 4608 заданий для моделей. Состав корпуса, условия и правила оценивания установлены Методикой сравнительных испытаний языковых моделей на русском языке.

Серия состоит из трех работ. Отчет Ответы языковых моделей по русскоязычным документам различает правильный ответ, обоснованный отказ и уверенное утверждение без достаточного основания и оценивает указание доказательств отдельно от верности вывода. Отчет Устойчивость к посторонним указаниям в русскоязычных документах проверяет границу между поручением пользователя и содержимым обрабатываемого документа и сравнивает защитную инструкцию с исходным режимом на тех же заданиях; оценка дей-

ствующего агента и адаптивные атаки составляют отдельные этапы программы. Отчет Повторяемость ответов языковых моделей и стоимость правильного результата объединяет повторяемость результата при повторе и перефразировании запроса с расходами на его получение.

Происхождение данных и проверка результатов

Методика сравнительных испытаний языковых моделей на русском языке связывает корпус, запросы, ответы и оценивание. Идентификатор задания сохраняется во всех условиях. Версии файлов и контрольные суммы позволяют проверить, что повторное вычисление выполняется на тех же исходных данных. Технические сбои сохраняются в журнале вместе с причинами повторных попыток.

При подготовке корпуса, программ обработки и текстов применяется автоматизированная помощь. Характер этой помощи раскрывается в методике. Автоматическое сравнение с заранее установленным эталоном дает воспроизводимый результат для закрытых полей и формальных статусов. Смысловые пограничные случаи требуют отдельного рассмотрения с сохранением основания принятого решения.

Независимая экспертная проверка составляет самостоятельное расширение программы. Для ее проведения необходимы как минимум два специалиста, единые инструкции, раздельное первоначальное оценивание и регистрация расхождений до согласования. Только после такой процедуры можно публиковать показатели согласия экспертов. Оценки другой языковой модели могут исследоваться отдельно как автоматический измеритель.

Доли ошибок сопровождаются числом наблюдений и оценкой неопределенности. Повторные условия одного задания учитываются как связанные измерения. Результаты по небольшим подгруппам рассматриваются осторожно: несколько дополнительных ошибок способны заметно изменить процент. Отсутствие обнаруженного различия оценивается вместе с шириной интервала и объемом данных.

Связь с дальнейшими направлениями Центра

1. Выполнено

Контролируемая серия

Одинаковые задания, сохраненные ответы и воспроизводимый расчет показывают, какие ошибки и различия между моделями следует проверять.

2. Следующий этап

Проверка на реальных документах

Предметная разметка и согласованные эталоны позволят проверить перенос результатов на задачи конкретной организации.

3. Прикладное продолжение

Рабочий процесс с контролем человека

Пилот должен измерить качество решений, время проверки и исправлений, стоимость обработки и удобство работы специалиста.

Для регулирования первоочередной задачей становится сопоставление состава испытаний с издержками их исполнения. Данные текущей серии помогают описать проверяемое поведение и минимальный журнал доказательств. Для предложения обязательного требования потребуются анализ затронутых участников, альтернатив исполнения и последствий ошибки в конкретной области применения.

Для государственного управления приоритетом остается содержательный контроль человека: специалист должен видеть основание ответа и иметь возможность изменить результат. Продолжение работы требует исследования маршрутов обращения, доступности уточнения и порядка повторного рассмотрения. Синтетические задания дают контролируемую отправную точку для такого исследования.

Для данных и инфраструктуры практический приоритет — компактный паспорт корпуса: происхождение, состав, ограничения, правила доступа и версия. В следующей серии необходимо проверить перенос результатов на тексты других жанров, регионов и периодов, а также на неполные и искаженные документы. Такое расширение требует отдельных эталонов и явного учета источников ошибок.

Для синтезированного контента требуется собственный набор измерений: пропуски выявления, ложные срабатывания и сохранение сведений о происхождении после преобразования файла. Его подготовка предполагает материалы разных видов, подтвержденную историю создания и воспроизводимые операции изменения. Для терминологии текущая серия дает рабочие определения наблюдаемых событий, которые затем сопоставляются с употреблением в методиках и стандартах.

Научные основания

При выборе конструкции использованы первичные научные публикации. MERA задает контекст оценки русскоязычных моделей; RGB выделяет работу с шумом, недостаточностью и противоречащими сведениями; InjecAgent

и AgentDojo связывают посторонние инструкции с полезной задачей и наблюдаемым исходом. Работы о структурированном выводе помогают разделять правильность формы и правильность значения.

- MERA: A Comprehensive LLM Evaluation in Russian. Первая публикация 9 января 2024 года, версия 3 от 2 августа 2024 года. Основание для сопоставления с существующей русскоязычной оценкой.
- Benchmarking Large Language Models in Retrieval-Augmented Generation. Первая публикация 4 сентября 2023 года, версия 2 от 20 декабря 2023 года. Основание для раздельной проверки достаточности и надежности контекста.
- InjecAgent. Первая публикация 5 марта 2024 года, версия 3 от 4 августа 2024 года. Основание для явного задания цели атакующего.
- AgentDojo. Первая публикация 19 июня 2024 года, версия 3 от 24 ноября 2024 года. Основание для совместного учета полезности и безопасности.
- Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents. Первая публикация 27 февраля 2025 года, версия 2 от 4 марта 2025 года. Основание для отдельного этапа адаптивной проверки.
- JSONSchemaBench. Первая публикация 18 января 2025 года, версия 3 от 27 февраля 2025 года. Разделение соблюдения схемы, охвата ограничений и качества результата.
- The Structured Output Benchmark. Препринт от 28 апреля 2026 года. Основание для раздельного учета структуры ответа и верности извлеченных значений.

Источники проверены 11 ноября 2026 года.

Замечания к исследовательским вопросам и методике принимаются по адресу expertise@pospelov.center. Для содержательного рассмотрения следует указать вопрос, предлагаемое изменение измерения и данные, на которых можно проверить его результат.

Порядок цитирования

Центр Поспелова. Исследовательская повестка: проверяемость систем искусственного интеллекта. 11 ноября 2026 года. Программа серии по протоколу CP-RU-DOC-2026-01.