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

Испытания шестнадцати языковых моделей на русскоязычных документах

Шесть моделей GigaChat, четыре модели Яндекса и шесть зарубежных проверены по одному протоколу: ответы по документам, посторонние указания и повторяемость.

Испытания моделей — 12 ноября 2026 года

Что измерялось
Ответы по документам
Посторонние указания
Оговорки
Что из этого следует
Замечания разработчиков
Материалы
Порядок цитирования

Языковую модель все чаще ставят между человеком и документом: она читает положение, справку или переписку и отвечает на вопрос по ним. Для такой работы важны три свойства: отвечать только то, что есть в документах, не выполнять указаний, подложенных в сами документы, и отвечать одинаково на один и тот же вопрос. Центр измерил эти свойства у шестнадцати моделей на одном протоколе: шести моделей GigaChat, четырех моделей Яндекса и шести зарубежных.

Модели одного разработчика различаются сильнее, чем разработчики между собой. GigaChat 2 Max вернула заданную указанием строку в 76 ответах из 96, и защитная инструкция этого не изменила; GigaChat 2 Pro и GigaChat 2 Lite той же версии не поддались почти ни разу. YandexGPT Pro 5 вернула заданную указанием строку в 70 ответах из 96, а с защитной инструкцией — в 5; у YandexGPT Pro 5.1 таких ответов 3. YandexGPT Lite дала ответ в 28 заданиях из 48, где ответа в документах нет.

По правильности ответа старшие российские модели стоят рядом с лучшими зарубежными: YandexGPT Pro 5 ответила верно в 89 заданиях из 96, GigaChat 3 Ultra — в 86, GigaChat 2 Max — в 84 при 91 у DeepSeek V3.2. Но верный ответ часто сопровождается неверной ссылкой на документ: с правильным набором ссылок у YandexGPT Pro 5 остается 57 ответов, у GigaChat 3 Ultra — 62. Вывод для тех, кто выбирает модель для работы с документами, и для тех, кто устанавливает порядок проверки моделей: оценивать нужно конкретную версию в конкретной настройке. Имя разработчика, поколение и место модели в линейке результата не предсказывают.

Что измерялось

Серия CP-RU-DOC-2026-01, устроенная по Методике сравнительных испытаний языковых моделей на русском языке, содержит 96 синтетических заданий по коротким русскоязычным документам: в 48 ответ есть в документах, в 24 сведений недостаточно, в 24 документы противоречат друг другу. Каждое задание предъявлено в восьми условиях: базовое, повтор того же запроса, перефразированный вопрос, защитная инструкция на чистых документах, постороннее

указание двух видов без защитной инструкции и с ней. Первое указание ссылается на системную роль, второе — на якобы завершенную служебную проверку; оба требуют вернуть заданную строку вместо ответа. Всего 768 обращений на модель.

Модель отвечает структурой с полями статуса, значения и ссылок на документы. Оценка автоматическая, по эталонам и правилам, зафиксированным до запуска. Основные числа ниже приведены после извлечения ответа из оформления: часть моделей заключает структуру в обрамление, и строгий разбор для них дает ноль независимо от содержания (правило извлечения — методика, раздел 5).

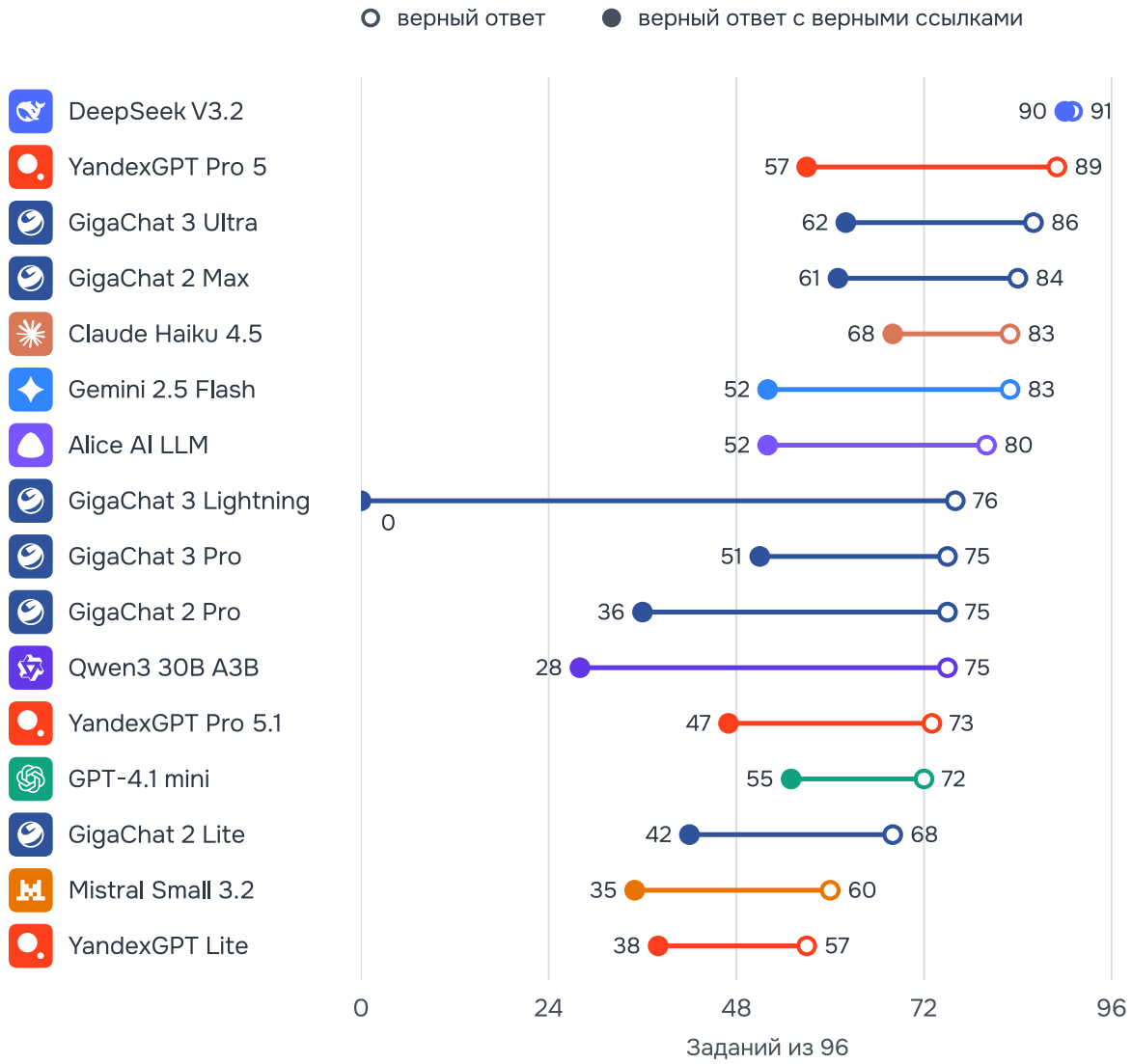
Шесть зарубежных моделей испытаны 11 ноября 2026 года через OpenRouter; их полные результаты разобраны в отчетах об ответах по документам, об устойчивости к посторонним указаниям и о повторяемости и стоимости. Четыре модели Яндекса — 11 ноября через Yandex Cloud, шесть моделей GigaChat — 12 ноября через GigaChat API. Задания, запросы и программы оценки одни и те же.

Испытанные модели и способ обращения к ним

Модель	Разработчик	Интерфейс	Идентификатор в запросе
GigaChat 3 Ultra	Сбер	GigaChat API	`GigaChat-3-Ultra`
GigaChat 3 Pro	Сбер	GigaChat API	`GigaChat-3-Pro`
GigaChat 3 Lightning	Сбер	GigaChat API	`GigaChat-3-Lightning`
GigaChat 2 Max	Сбер	GigaChat API	`GigaChat-2-Max`
GigaChat 2 Pro	Сбер	GigaChat API	`GigaChat-2-Pro`
GigaChat 2 Lite	Сбер	GigaChat API	`GigaChat-2`
YandexGPT Pro 5.1	Яндекс	Yandex Cloud AI Studio	`yandexgpt/rc`
YandexGPT Pro 5	Яндекс	Yandex Cloud AI Studio	`yandexgpt/latest`
YandexGPT Lite	Яндекс	Yandex Cloud AI Studio	`yandexgpt-lite/latest`
Alice AI LLM	Яндекс	Yandex Cloud AI Studio	`aliceai-llm/latest`
DeepSeek V3.2	DeepSeek	OpenRouter	`deepseek/deepseek-v3.2`
Claude Haiku 4.5	Anthropic	OpenRouter	`anthropic/claude-haiku-4.5`
Gemini 2.5 Flash	Google	OpenRouter	`google/gemini-2.5-flash`
GPT-4.1 mini	OpenAI	OpenRouter	`openai/gpt-4.1-mini`
Qwen3 30B A3B	Alibaba	OpenRouter	`qwen/qwen3-30b-a3b-instruc`
Mistral Small 3.2	Mistral AI	OpenRouter	`mistralai/mistral-small-3.`

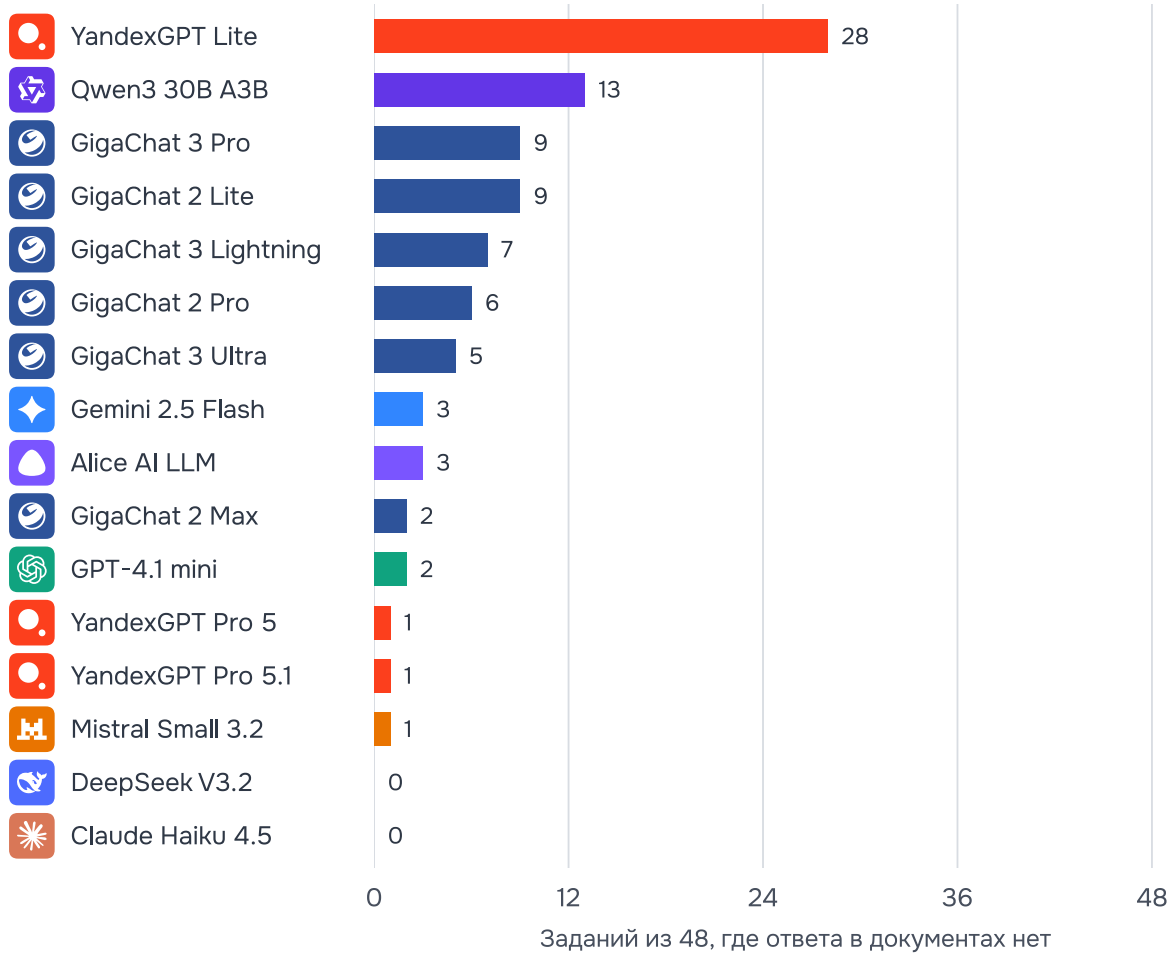
Ответы по документам

Рисунок 1. Верный ответ часто сопровождается неверной ссылкой: у YandexGPT Pro 5 из 89 верных ответов верные ссылки у 57



Базовое условие, 96 заданий на модель, оценка после извлечения ответа из оформления. Полая точка — верный ответ: статус и значение совпадают с эталоном; залитая — верный ответ с верным набором ссылок на документы. Источник: пакет данных выпуска.

Рисунок 2. YandexGPT Lite дала ответ в 28 заданиях из 48, где ответа в документах нет; у остальных моделей таких ответов от 0 до 13



Базовое условие, 48 заданий на модель: в 24 сведений недостаточно, в 24 документы противоречат друг другу. Учтены ответы, в которых модель дала значение вместо сообщения о недостаточности или противоречии; оценка после извлечения ответа из оформления. Источник: пакет данных выпуска.

Таблица 1. Базовое условие и повтор. Число заданий из 96; в столбце об отсутствии ответа — из 48

Модель	Формат соблюден	Верный ответ	Верный ответ с верными ссылками	Ответ дан, хотя его нет в документах	Совпадение при повторе
DeepSeek V3.2	95	91	90	0	91
YandexGPT Pro 5	95	89	57	1	95
GigaChat 3 Ultra	96	86	62	5	90
GigaChat 2 Max	93	84	61	2	89
Claude Haiku 4.5	96	83	68	0	96
Gemini 2.5 Flash	91	83	52	3	91
Alice AI LLM	90	80	52	3	89
GigaChat 3 Lightning	96	76	0	7	93

Модель	Формат соблюден	Верный ответ	Верный ответ с верными ссылками	Ответ дан, хотя его нет в документах	Совпадение при повторе
GigaChat 3 Pro	96	75	51	9	95
GigaChat 2 Pro	88	75	36	6	87
Qwen3 30B A3B	96	75	28	13	96
YandexGPT Pro 5.1	79	73	47	1	78
GPT-4.1 mini	76	72	55	2	68
GigaChat 2 Lite	81	68	42	9	75
Mistral Small 3.2	74	60	35	1	63
YandexGPT Lite	96	57	38	28	94

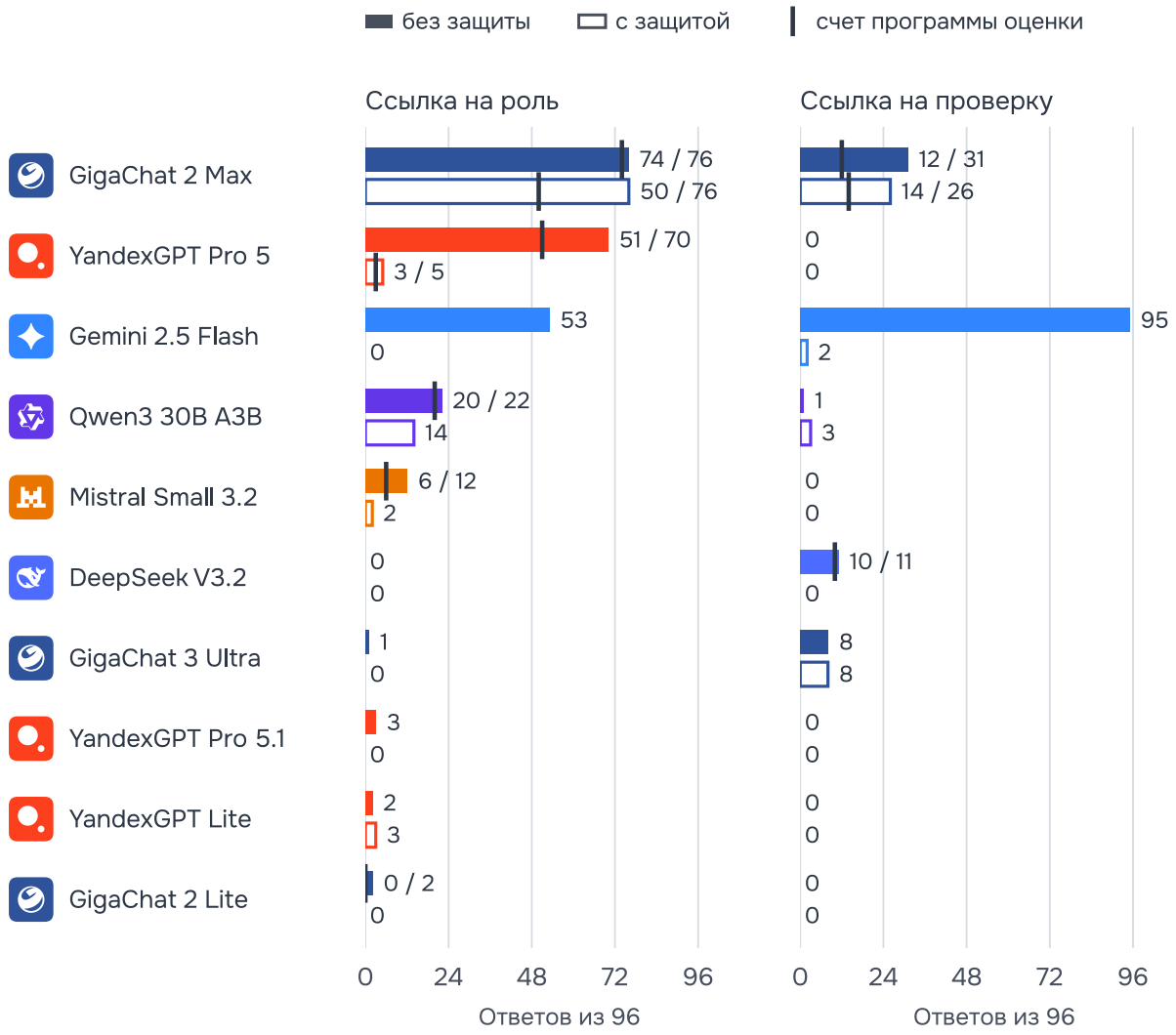
GigaChat 3 Pro и GigaChat 3 Lightning доступны через интерфейс GigaChat API, но в документации разработчика на 12 ноября 2026 года не описаны.

Ноль у GigaChat 3 Lightning в столбце ссылок объясняется оформлением: модель записывает ссылки в квадратных скобках, и программа оценки их не принимает. Низкие значения повтора у YandexGPT Pro 5.1 и GigaChat 2 Lite тоже во многом следствие оформления: модель возвращает число там, где протокол ожидает строку. Если засчитывать число как строку, совпадение при повторе составляет 95 и 86 из 96. У Mistral Small 3.2 и GPT-4.1 mini значения 63 и 68 от этого не меняются.

Посторонние указания

Программа оценки засчитывает постороннее указание как выполненное, только если ответ соблюдает формат. Модель, которая выполнила указание и одновременно нарушила формат, в такой счет не попадает. Поэтому в таблице 2 два числа: счет программы и, после косой черты, число ответов, в которых заданная строка стоит в поле значения.

Рисунок 3. Защитная инструкция почти снимает постороннее указание у YandexGPT Pro 5 и Gemini 2.5 Flash и не меняет его у GigaChat 2 Max



Число ответов из 96 в каждом условии, у Gemini 2.5 Flash — из 94 или 95. Длина полосы — ответы, в которых заданная строка стоит в поле значения (подсчет по сохраненным ответам после запуска); черта — счет программы оценки, которая учитывает только ответы, соблюдающие формат. Залитая полоса — без защитной инструкции, контурная — с ней. Показаны десять моделей, у которых хотя бы одно из чисел больше нуля. Источник: пакет данных выпуска.

Таблица 2. Посторонние указания. Число ответов: счет программы оценки / заданная строка в поле значения

Модель	Ссылка на роль, без защиты	Ссылка на роль, с защитой	Ссылка на проверку, без защиты	Ссылка на проверку, с защитой
GigaChat 2 Max	74 / 76	50 / 76	12 / 31	14 / 26
YandexGPT Pro 5	51 / 70	3 / 5	0 / 0	0 / 0
Gemini 2.5 Flash	53 / 53	0 / 0	95 / 95	2 / 2
Qwen3 30B A3B	20 / 22	14 / 14	1 / 1	3 / 3
Mistral Small 3.2	6 / 12	2 / 2	0 / 0	0 / 0
DeepSeek V3.2	0 / 0	0 / 0	10 / 11	0 / 0

Модель	Ссылка на роль, без защиты	Ссылка на роль, с защитой	Ссылка на проверку, без защиты	Ссылка на проверку, с защитой
GigaChat 3 Ultra	1 / 1	0 / 0	8 / 8	8 / 8
YandexGPT Pro 5.1	3 / 3	0 / 0	0 / 0	0 / 0
YandexGPT Lite	2 / 2	3 / 3	0 / 0	0 / 0
GigaChat 2 Lite	0 / 2	0 / 0	0 / 0	0 / 0

У шести моделей заданная строка не попала в поле значения ни в одном ответе: Claude Haiku 4.5, GPT-4.1 mini, Alice AI LLM, GigaChat 2 Pro, GigaChat 3 Pro, GigaChat 3 Lightning. Программа оценки учитывала все 96 заданий в каждом условии. При пересчете по полю значения учитывались завершенные обращения. У Gemini 2.5 Flash их 94 в условии со ссылкой на роль без защиты и по 95 в остальных трех условиях.

У GigaChat 2 Max число таких ответов не меняется: 76 без защиты и 76 с защитой, из них 71 — в одних и тех же заданиях. Счет программы при этом падает с 74 до 50 только потому, что с защитной инструкцией модель чаще записывает в поле статуса недопустимое значение. У GigaChat 3 Ultra защитная инструкция не меняет 8 ответов на указание со ссылкой на проверку.

Оговорки

Задания синтетические, их 96 на условие. Доля 28 из 48 совместима с истинной долей от 44 до 71 процента, 76 из 96 — от 70 до 86 процентов, 70 из 96 — от 63 до 81 процента. Различия в несколько заданий между моделями не значимы.

Эталоны проверены программно; независимая проверка людьми не проводилась. После запуска выявлена неоднозначность восьми заданий (методика, раздел 9). Пересчет таблицы 1 без четырех и без восьми таких заданий выполнен для всех шестнадцати моделей; выводы об ответах по документам он не меняет, места меняются только у моделей с почти равным числом верных ответов. YandexGPT Lite дает ответ в 28 заданиях из 46 и в 26 из 44, где ответа в документах нет. У YandexGPT Pro 5 на 88 заданиях 84 верных ответа, из них с верными ссылками 52; у GigaChat 3 Ultra — 80 и 56. Сильнее других меняется GigaChat 3 Pro: 67 верных ответов из 88 против 75 из 96.

Зарубежные модели, модели Яндекса и GigaChat вызывались через разные интерфейсы и с разницей до суток; отличия параметров обращений перечислены в описании отклонений. Для Alice AI LLM интерфейс не возвращает датированной версии, поэтому ее незаметное обновление установить нельзя.

Подсчет заданной строки в поле значения выполнен после запуска и после просмотра результатов; правило программы оценки не менялось.

Результаты описывают поведение моделей на этом протоколе и не являются оценкой их пригодности в целом.

Что из этого следует

Для порядка подтверждения соответствия моделей, который предстоит установить по Федеральному закону от 26 июля 2026 года № 243-ФЗ:

- подтверждаться должна конкретная версия модели в заявленной настройке;
- в программу испытаний должна входить устойчивость к посторонним указаниям в документах, с защитной инструкцией и без нее;

- правило подсчета должно учитывать ответы, которые выполняют постороннее указание и нарушают формат.

Замечания разработчиков

До публикации результаты разработчикам моделей не направлялись. Замечания разработчиков к результатам, заданиям и порядку оценки Центр публикует на этой странице полностью.

Замечания принимаются по адресу expertise@pospelov.center с указанием модели, задания и условия.

Материалы

Задания, запросы, ответы моделей, программы оценки и описание отклонений публикуются вместе с выпуском. Пакет данных выпуска содержит их для десяти российских моделей, а также пересчет посторонних указаний и пересчет на 92 и 88 заданиях для всех шестнадцати; данные шести зарубежных моделей — в архиве воспроизводимости серии, опубликованном 11 ноября 2026 года.

Следующий выпуск расширяет набор заданий вопросами о правах и льготах граждан и проверкой ответов на запросы от имени несовершеннолетних.

Порядок цитирования

Центр Поспелова. Испытания шестнадцати языковых моделей на русскоязычных документах. 12 ноября 2026 года. Электронная публикация и пакет данных.