

Качество учебных заданий

Актуализированный исследовательский отчёт проекта aiTalim

1,14%

aiTalim

gpt-6-luna · reasoning: medium
Режим: standard
Контекст и инструкции платформы

Подтверждённых дефектов: 16 / заданий: 1405

18,33%

ChatGPT Edu

gpt-6-luna · reasoning: medium
Режим: standard
Ориентир по модели через API

110 подтверждённых дефектов / 600 заданий

Что означает показатель

Доля заданий с подтверждённой фактической ошибкой, неверным ключом или существенным методическим дефектом. Например, ученик не может решить задание по доступным ему данным либо правильный ответ не соответствует условию.

В общей таблице тестов предметов: 17. Редакторские замечания (63) и спорные случаи (18) учтены отдельно и не входят в число подтверждённых дефектов.

PISA показана отдельно: **1,04% — дефектов: 1 / заданий: 96**. Проверенных предметных контекстов: 4; для остальных предметов пока нет данных. Результат PISA не включён в показатель тестов.

Как читать сравнение

18,33% — доступный результат проверки ответов модели через API, используемый как ориентир для ChatGPT Edu. В обоих результатах использована gpt-6-luna с reasoning medium. Различаются даты, набор заданий и условия генерации; это непарный контроль.

Актуальный срез по последним доступным замерам каждого предмета. Даты и объём предметных проверок различаются; это диагностический набор, а не одномоментная случайная выборка. Биология и география проверены повторной генерацией после доработок. Результаты казахского языка, казахской литературы и истории Казахстана сохранены из предыдущего замера; последующие доработки этих трёх предметов ещё требуют новой проверки.

Тесты: результаты по предметам

Подтверждённые дефекты = фактические ошибки (S) + существенные методические дефекты (M). Редактура (P) и спорные случаи приведены отдельно.

Предмет	Выдано	S+M	Доля	P	Спорно
Алгебра	60	0	0,00%	4	0
Алгебра и начала анализа	60	0	0,00%	1	0
Английский язык	53	0	0,00%	3	0
Биология	133	1	0,75%	7	2
Всемирная история	60	1	1,67%	4	2
География	131	1	0,76%	4	1
Геометрия	59	1	1,69%	1	0
Информатика	60	1	1,67%	1	0
История Казахстана	110	5	4,55%	2	1
Казахская литература	138	1	0,72%	12	2
Казахский язык	126	2	1,59%	10	3
Математика	60	0	0,00%	1	0
Основы права	53	1	1,89%	2	1
Русская литература	59	1	1,69%	4	3
Русский язык	64	0	0,00%	3	2
Физика	60	0	0,00%	4	0
Химия	119	1	0,84%	0	1
Итого	1405	16	1,14%	63	18

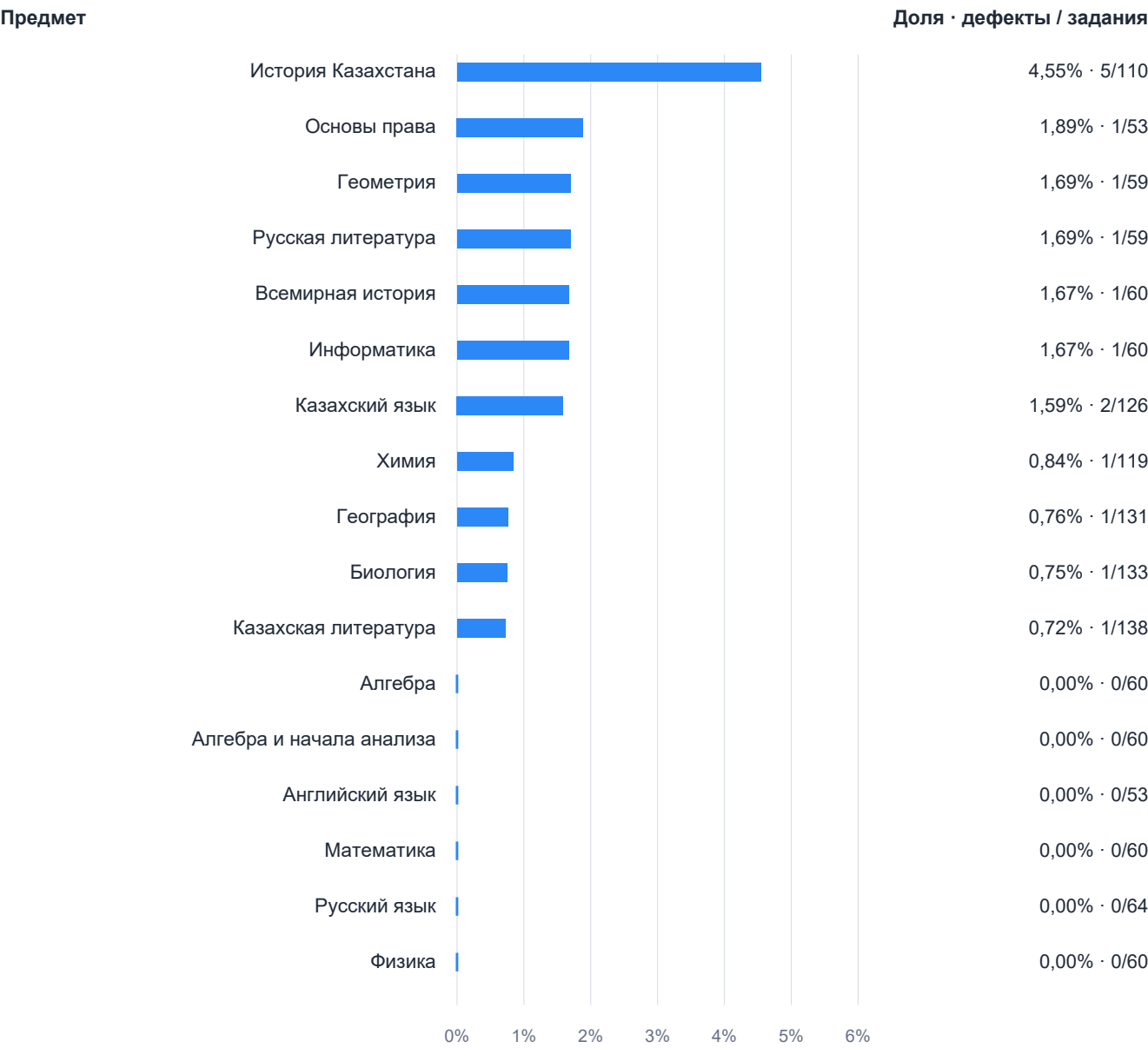
Доступность генерации

Запланировано заданий: 1570; выдано: 1405; не завершились выдачей: 165. Это отдельный показатель доступности. Невыданные задания не входят в знаменатель доли дефектов.

Каждый предмет представлен последним завершённым замером. Более новые результаты заменяют соответствующие строки; повторного суммирования одних и тех же проверок нет.

Тесты: где доля дефектов выше

Предметы расположены по доле подтверждённых дефектов. Рядом с процентом указан объём проверки.



Наибольшие наблюдаемые доли: **История Казахстана: 5/110 — 4,55%; Основы права: 1/53 — 1,89%; Геометрия: 1/59 — 1,69%.**

Без подтверждённых существенных дефектов в последнем замере: Алгебра, Алгебра и начала анализа, Английский язык, Математика, Русский язык, Физика. Нулевой результат означает отсутствие подтверждённых дефектов в проверенном наборе, а не гарантию для будущей генерации.

PISA: результаты по всем предметам

Проверено заданий: 96; подтверждённых существенных дефектов: 1 (1,04%); редакторских замечаний: 1; спорных случаев: 3. Предмет — контекст задания; вид грамотности указан только для проверенных заданий.

Предмет	Грамотность	Заданий	S+M / доля	P / спорно
Алгебра	Нет данных	—	—	—
Английский язык	Нет данных	—	—	—
Биология	Нет данных	—	—	—
Всемирная История	Нет данных	—	—	—
География	Нет данных	—	—	—
Геометрия	Нет данных	—	—	—
Глобальные компетенции	Нет данных	—	—	—
Графика и проектирование	Нет данных	—	—	—
Естествознание	Естественно-научная	32	0 / 0,00%	0 / 1
Информатика	Нет данных	—	—	—
История Казахстана	Нет данных	—	—	—
Казахская литература	Нет данных	—	—	—
Казахский язык	Читательская	16	1 / 6,25%	1 / 0
Казахский язык и литература	Нет данных	—	—	—
Краеведение	Нет данных	—	—	—
Математика	Математическая	32	0 / 0,00%	0 / 0
Музыка	Нет данных	—	—	—
Начальная военная и технологическая подготовка	Нет данных	—	—	—
Начальная военная подготовка	Нет данных	—	—	—
Обучение грамоте	Нет данных	—	—	—
Основы права	Нет данных	—	—	—
Основы предпринимательства и бизнеса	Нет данных	—	—	—
Познание мира	Нет данных	—	—	—
Русская литература	Нет данных	—	—	—
Русский язык	Читательская	16	0 / 0,00%	0 / 2
Русский язык и литература	Нет данных	—	—	—
Самопознание	Нет данных	—	—	—
Светскость и основы религиоведения	Нет данных	—	—	—
Уйгурская литература	Нет данных	—	—	—
Уйгурский язык	Нет данных	—	—	—
Физика	Нет данных	—	—	—
Химия	Нет данных	—	—	—

В каждой строке соответствие сложности: **не измерялось**. Нет данных — предмет доступен в интерфейсе, но отдельного подтверждённого замера по нему пока нет.

PISA: диаграмма по всем предметам

Синяя отметка и процент — проверенный предмет. «Нет данных» означает отсутствие замера, а не нулевой процент дефектов.



Естествознание — название контекста проверочных заданий для классов: 6, 8, 10, 11. Результаты не приписаны физике, химии или биологии и не подтверждают доступность этого школьного предмета во всех этих классах.

GPT-6.1 Sol: проверка через API

2,08%

Подтверждённых дефектов: 2 / выданных заданий: 96. План: 96 заданий. gpt-6.1-sol · reasoning: medium · режим: standard.

Предметов: 8, по 12 заданий; типов: 6. Русский: 48, казахский: 48.

Предмет	Выдано	S+M	Доля	P	Спорно
Биология	12	0	0,00%	2	1
Химия	12	0	0,00%	2	0
Алгебра	12	0	0,00%	0	0
Физика	12	1	8,33%	2	0
Русский язык	12	1	8,33%	0	0
Казахская литература	12	0	0,00%	2	0
Всемирная история	12	0	0,00%	0	0
История Казахстана	12	0	0,00%	2	1
Итого	96	2	2,08%	10	2



Два подтверждённых дефекта — несколько допустимых порядков действий при единственном ключе. Неверных ключей по содержанию не выявлено.

Отдельная генерация через API вне платформы aiTalim. Использованы исходные инструкции доступного API-контроля; число заданий в одном ответе уменьшено с пяти до двух.

Это отдельный диагностический набор с независимой проверкой ответов. Его нельзя объединять с таблицей тестов aiTalim или считать той же выборкой, что API-ориентир 110/600. Результат не подтверждает универсального преимущества одной модели.

Почему выбран reasoning medium

Основная генерация тестов aiTalim; вспомогательные этапы имеют отдельные настройки.

Основание рабочего решения. Reasoning medium сохранён по практическому наблюдению руководителя исследования и предварительной архивной сверке. В прежних проверках учебных заданий более высокие уровни reasoning чаще давали ошибки, а medium работал лучше. Это основание рабочего решения; величина преимущества и причина различия не установлены.

Что означает параметр. Параметр reasoning.effort задаёт объём рассуждения модели перед ответом. В официальной документации OpenAI низкие уровни связаны с меньшей задержкой и расходом токенов; для выбора между medium и high рекомендована оценка на своих задачах, а xhigh — при измеренном выигрыше. Из этого не следует гарантия меньшего числа дефектов учебных заданий при повышении уровня.

Источник: [OpenAI Docs — Reasoning models, Reasoning effort](#). Проверено 12.10.2026.

Что подтверждает архив. В архиве есть сравнение low, medium, high и xhigh одной модели на одинаковых предметах, темах и типах, с двумя повторами. Предварительная сводка по широкой диагностической рубрике показала более благоприятный результат medium, чем high и xhigh; это не доля финально подтверждённых S+M и не причинное доказательство. Сравнение относится к прежней версии генерации и содержит непроверенные ключи. Судье были скрыты модель и уровень reasoning, но доступны авторские ключи. Сбросы части цепочек и дальнейшие изменения генерации ограничивают сопоставимость.

Актуальная доля дефектов aiTalim измерена при medium вместе с действующими инструкциями, контекстом и проверками. Она не измеряет отдельный эффект этого параметра. Утверждать, что medium универсально точнее high или xhigh, по имеющимся данным нельзя.

Гипотеза, а не установленная причина. Одна из рабочих гипотез: при большем объёме рассуждения модель в отдельных заданиях может добавлять неоговорённые предпосылки или переосмысливать прямое условие. Внутренние рассуждения не исследовались, механизм не установлен. Возможны и другие объяснения: состав небольшого набора, ошибки проверки, разные истории цепочек и отбор после повторных попыток.

Следующий эксперимент. Отдельное исследование сопоставит поддерживаемые моделью low, medium, high и xhigh на одних источниках и заданиях. Зафиксируем модель, режим, инструкции, проверки, лимиты и повторы; распределим порядок запусков случайно и сбалансируем предметы и типы. Проверяющий сначала решит задание без авторского ключа и без сведений об уровне reasoning, затем сверит ответ. Учтём дефекты среди выданных заданий, невыданные слоты, спорные случаи, расход токенов и время. Опубликуем парные различия с интервалами неопределённости. Новые контрольные источники будут отделены от материала настройки.

Протокол подготовлен; новый эксперимент reasoning не проводился. Требуется отдельный бюджет.

Отчёт будет расширяться по мере накопления проверенных данных: сравнения уровней reasoning, новых предметных замеров и оценки сложности. Планируемый объём около 30 страниц — ориентир структуры, а не основание добавлять неподтверждённые выводы.

AIREQ Bench: единый стандарт проверки

AI Reliability & Educational Quality

Надёжность ИИ и качество учебных заданий

Для чего. Сравнить новые версии генерации и модели на одинаковых входных материалах, по заранее зафиксированным критериям.

Что проверяем. Факты и вычисления; решаемость условия; полнота и однозначность ключей; допустимые альтернативные ответы; объяснения; методика; критерии, дескрипторы и баллы; качество языка; соответствие предмету, теме, классу и учебной цели. Для PISA — также вид грамотности, доступность контекста и соответствие заявленной когнитивной сложности.

Как сравниваем. Тесты и PISA составляют отдельные направления одного бенчмарка. Для каждой системы фиксируются одинаковые запросы, источники, классы, языки, форматы и уровни сложности. Проверяющий сначала решает задание без авторского ключа, затем сверяет решение с ключом и критериями. Подтверждённые существенные и спорные случаи проходят повторный разбор.

Как считаем. Публикуются два основных результата: доля подтверждённых содержательных и методических дефектов среди выданных заданий и доля полностью проверенных пригодных заданий среди всех запланированных. Отказы, остановки, редакторские недостатки, спорные случаи и непроверенные критерии показываются отдельно.

Версии и расширение. Набор будет расширяться по темам, классам, языкам и типам заданий. Раздельно версионироваться протокол, набор и критерии. Результаты разных версий сравниваются на неизменном общем ядре; увеличение выборки само по себе не доказывает улучшение генерации.

Границы оценки. Оценка соответствия заявленной сложности отличается от калибровки по ответам учеников. Последняя требует отдельного исследования. Национальная репрезентативность и соответствие шкалам OECD не заявляются.

Статус AIREQ 1.0. Первая версия протокола подготовлена; парный замер по замороженному набору ещё не проведён.

Методика, сложность и дальнейшая работа

Основная генерация	Модель	Reasoning	Режим
Тесты aiTalim	gpt-6-luna	medium	standard
API-ориентир	gpt-6-luna	medium	standard
Новый API-контроль Sol	gpt-6.1-sol	medium	standard

Настройки основной генерации подтверждены сохранёнными запросами и ответами API. Вспомогательные проверки могут иметь отдельные настройки; reasoning medium не приписан всем этапам обработки.

Что проверялось. Решаемость по условию и доступному ученику тексту, факты, ключи, достаточность данных и критерии оценивания. Проверочные ИИ-агенты сначала фиксировали собственное решение, затем сравнивали его с авторским ключом. В ранее проведённых проверках aiTalim варианты допустимых ответов и критерии эссе могли быть доступны; это не полностью скрытая проверка всех полей. Для нового API-контроля Sol все авторские ключи и объяснения были скрыты при самостоятельном решении всех 96 заданий; собственные ответы зафиксированы до открытия ключей. В последних замерах казахского языка, казахской литературы и истории Казахстана все ключи, образцы и источники учителя раскрывались только после фиксации самостоятельных решений. Подтверждённые существенные дефекты проходили вторую проверку. В последнем замере биологии и географии выдано заданий: 264; слотов: 336; не выдано: 72. Все выданные задания решены со скрытыми ключами; решения зафиксированы до их раскрытия. Подтверждённые дефекты прошли вторую проверку. Подтверждено S+M: 2; спорных случаев: 3.

Грамотности PISA. Сейчас активны: читательская, математическая, естественно-научная. Неактивные направления не включены в измеренный результат. Предметы не образуют новые виды грамотности.

Сложность PISA. В проверенном наборе запрашивались общие уровни: 3; индивидуальные уровни: 2, 3. Отдельной оценки соответствия когнитивной сложности этим уровням нет; уровни 1–6 полностью не покрыты. Соответствие сложности и эмпирическая калибровка по ответам учеников пока не измерялись.

Границы результатов. В проверках использованы конкретные темы и источники, часть сюжетов повторялась по классам. Контексты последних замеров казахского языка, казахской литературы и истории Казахстана уже использовались при доработке; совершенно новые источники требуют отдельной проверки. Биология и география: фиксированные тематические контексты с условными учебными моделями. Диагностический набор не является случайной выборкой; сложность и соответствие программе не подтверждены. Это внутренний диагностический контроль с участием ИИ, без независимой экспертной комиссии преподавателей. Национальная репрезентативность и калибровка OECD не заявляются. Проверенные классы PISA: 6, 8, 10, 11; начальные классы и PIRLS требуют отдельного протокола.

Что улучшаем дальше

Продолжаем улучшать саму генерацию: качество учебного контекста и инструкций, корректность условий, ключей и критериев, а также полноту текста, доступного ученику. Приоритет — казахская и русская литература, казахский, русский и английский языки, история Казахстана, биология, география, PISA и основы права. Цель — менее 2% подтверждённых дефектов по каждому предмету.

Для PISA расширяем проверку на остальные предметы, темы и классы; отдельно оцениваем соответствие сложности и допускаем альтернативные верные решения. Новые результаты будут заменять соответствующие строки актуального среза. Изменения проходят повторную проверку на заданиях, которые не использовались для настройки.

Evaluation служит проверке изменений и выявлению повторных дефектов; само по себе оно не обучает модель. Дообучение весов модели в описанной версии не выполнялось. COP и СОЧ остаются на паузе.

Подготовлено для публичного ознакомления как внутренний документ проекта aiTalim.

Технический директор aiTalim,
куратор исследования



Пернебаев Б.Ш.