ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Т. 8, № 1(25)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.8 DOI: 10.54708/SIIT-2026-no1-p3 EDN: QHZOLM

Оптимизация математической основы IRT с помощью LLM-моделей

Д. С. Алексеева, Е. В. Пальчевский, В. В. Антонов, В. А. Суворова

Уфимский университет науки и технологий
МИРЭА — Российский технологический университет

Оптимизация математической основы IRT (Item Response Theory) основанная на теории реакции на предметы, с помощью LLM (Large Language Models) — это довольно новая и многообещающая область исследований, которая сочетает в себе статистику, психологию, образование и машинное обучение. LLM могут быть использованы для анализа данных прошлых тестов и выявления тенденций и закономерностей. Модели могут помочь в оценке наиболее эффективных типов вопросов для оценки определенных навыков или знаний. Они могут оценить, как изменения в формулировках вопросов или формате тестирования могут повлиять на ответы испытуемых. Модели могут предсказать, как изменения в содержании теста могут повлиять на уровень знаний и успеваемость студентов. Также модели могут использоваться для создания адаптивных тестов, которые подстраиваются под уровень знаний и способностям испытуемых. Оптимизация математической основы IRT с применением LLM может привести к более точным и эффективным тестам. Это может быть полезно в различных областях, где требуется оценка уровня знаний или навыков, таких как образование, медицина, психология и другие

IRT, LLM, Large Language Models, большие языковые модели, оптимизация, нейронные сети, тестирование, адаптивные методики.

Введение

Item Response Theory (IRT), или теория отклика на задание, представляет собой статистический подход, широко применяемый в психометрике для анализа результатов тестов и конструирования эффективных оценочных материалов [Lor68, Ham91, Lin97]. Он основан на моделировании вероятности правильного ответа на тестовый вопрос как функции двух главных факторов: латентной способности испытуемого и параметров самого задания (например, сложности). Такое моделирование позволяет более точно измерять уровень знаний и навыков на всем диапазоне способностей испытуемых, обеспечивая надежную оценку как для сильных, так и для слабых студентов. Одним из ключевых приложений IRT является адаптивное тестирование, при котором сложность последующих вопросов динамически подстраивается под уровень подготовки учащегося на основе его предыдущих ответов [Lin97, Che25, Hua25]. Это дает возможность значительно повысить эффективность и справедливость оценивания по сравнению с традиционными подходами, одновременно выявляя и устраняя возможные перекосы теста (например, избыточную сложность отдельных заданий).

Строгая математическая основа IRT задает формальные связи между характеристиками испытуемого и задания, позволяя параметризовать их на общей шкале. В простейшем случае вероятность правильного ответа может быть выражена с помощью логистической функции [Ras60]:

$$P(\text{правильного ответа} \mid \theta, \beta) = \frac{1}{1+e^{-(\theta-\beta)}},$$

Алексеева Д. С., Пальчевский Е. В., Антонов В. В., Суворова В. А. Оптимизация математической основы IRT с помощью LLM-моделей // СИИТ. 2026. Т. 8, № 1(25). С. 3-18. DOI: 10.54708/SIIT-2026-no1-p3. EDN: QHZOLM.

Alekseeva D. S., Palchevsky E. V., Antonov V. V., Suvorova V. A. “Optimization of the mathematical basis of IRT using LLM models” // SIIT. 2026. Vol. 8, no. 1(25), pp. 3-18. DOI: 10.54708/SIIT-2026-no1-p3. EDN: QHZOLM (In Russian).

где θ – уровень способности (латентной переменной) учащегося, а β – сложность данного тестового задания. Подобная модель (известная как одномерная логистическая модель [Ray25]) позволяет одновременно оценивать уровень подготовленности ученика и трудность вопросов теста. Это, в свою очередь, дает возможность сопоставлять способности разных учащихся и сложности разных заданий напрямую, что особо ценно для адаптивного тестирования и точной калибровки тестов. Таким образом, IRT выступает фундаментом современных компьютеризированных систем тестирования знаний, обеспечивая математически обоснованное и объективное измерение результатов.

Несмотря на очевидные преимущества IRT, актуальной задачей является дальнейшее совершенствование ее математической основы с учетом растущих объемов данных и новых возможностей анализа. В последние годы появление методов искусственного интеллекта, в особенности больших языковых моделей (LLM), открыло новые перспективы для улучшения обработки тестовой информации и оптимизации оценочных моделей. Использование LLM в науке и технике уже стало новым этапом развития методов анализа данных и поддержки принятия решений. Особенно заметен их потенциал в образовательной сфере, где требуется разработка инновационных подходов к тестированию и повышение точности оценки знаний учащихся. Интеграция возможностей LLM с методологией IRT рассматривается сейчас как перспективное направление, способное усилить классические психометрические модели за счет мощных инструментов анализа данных.

Большая языковая модель (Large Language Model, LLM) – это класс моделей искусственного интеллекта, предназначенных для анализа и генерации текстовой информации. Такие модели обучаются на огромных массивах текстов, благодаря чему они способны распознавать контекст, выявлять смысловые закономерности и порождать осмысленные фрагменты естественного языка. За счет использования нейросетевых архитектур с чрезвычайно большим числом параметров и обучающих примеров, LLM относят к категории «больших» моделей. Их ключевым преимуществом является способность эффективно обрабатывать огромные объемы текстовых данных, извлекая из них существенные сведения и представляя результаты в удобной для пользователя форме. LLM уже нашли применение во множестве областей, где требуется работа с текстом: от автоматизированного перевода и контент-анализа до генерации осмысленных ответов на произвольные запросы. В контексте образования LLM зарекомендовали себя, например, как инструмент для автоматического создания черновых вариантов тестовых заданий на основе учебных материалов, существенно сокращающий трудоемкость подготовки экзаменов. Подобные модели способны формировать уникальные наборы вопросов, адаптированные под содержание курса и уровень обучающихся, тем самым снижая повторяемость тестов и способствуя индивидуализации обучения.

Применение LLM в области тестирования открывает качественно новые возможности для оптимизации IRT-моделей. Прежде всего, большие языковые модели могут быть использованы для глубокого анализа накопленных эмпирических данных тестирования. Обрабатывая результаты прошлых экзаменов и опросов, LLM способны выявлять скрытые тенденции и статистические закономерности в ответах испытуемых. На основании таких данных модель может помогать определить, какие типы вопросов наиболее эффективно оценивают те или иные навыки, а также анализировать, как изменения в формулировках заданий или формате теста влияют на ответы участников. Более того, LLM позволяют прогнозировать эффект модификации тестового контента, например, оценить, как включение новых тем или изменение порядка вопросов скажется на успеваемости и уровне знаний студентов. Еще одним перспективным направлением является генерация адаптивных тестов с помощью LLM: на практике это означает, что нейросеть динамически подбирает последующие вопросы на основе предыдущих ответов экзаменуемого, автоматически подстраивая сложность и тематику под его уровень знаний и способности. Таким образом, интеграция LLM дает возможность не только ускорить анализ тестовых данных, но и сделать сами тесты более «умными», гибкими и персонализированными в реальном времени.

Сочетание традиционной методологии IRT с новейшими LLM-технологиями формирует междисциплинарное направление исследований, объединяющее статистику, психометрику, педагогические измерения и машинное обучение. Ожидается, что оптимизация математической основы IRT с использованием LLM позволит повысить точность и эффективность тестовых измерений, что найдет применение в различных сферах: от образования до медицины и психологии, т. е. везде, где требуется объективная оценка уровня знаний и навыков. Таким образом, цель данного исследования заключается в применении больших языковых моделей для совершенствования и оптимизации метода IRT, развивая новые подходы к анализу результатов тестирования и адаптивному формированию тестов.

Предметная область и архитектура LLM

Калибровка тестовых заданий в парадигме Item Response Theory (IRT) традиционно опирается на анализ табличных откликов «правильно / неправильно», однако современная образовательная практика генерирует куда более богатый контент: текстовые формулировки вопросов и дистракторов, развернутые ответы студентов, экспертные комментарии, журналы попыток и т.д. — все это остается практически невостребованным в классических IRT-процедурах, хотя именно здесь скрыты ключевые индикаторы сложности β, дискриминационной способности α, потенциальной предвзятости пункта и латентной способности θ испытуемого. Чтобы извлечь эти признаки, нужны инструменты, способные автоматически «читать» неструктурированный текст, выделять релевантные семантические паттерны и количественно описывать их для дальнейшего включения в IRT-модели.

Большие языковые модели (LLM) заполняют этот пробел: они извлекают семантические признаки из текстов заданий, прогнозируя параметры β (сложность) и α (дискриминация) еще до пилотирования, группируют родственные формулировки, устраняя дубликаты, и автоматически генерируют альтернативные версии вопросов для A/B-тестирования. Интегрируя такие признаки в логистические функции IRT, платформа повышает информативность теста и уточняет онлайн-оценку способности θ испытуемого в адаптивных алгоритмах, тем самым ускоряя весь цикл «создание → калибровка → эксплуатация» банка заданий. Примерами LLM-моделей на рынке служат ChatGPT от OpenAI [Ope24], YandexGPT от «Яндекса» [Yan25], GigaChat от «Сбера» [Sbe25], Bard от Google [Goo23], DeepSeek от китайской компании «DeepSeek» [Sha24].

Bard предоставляет возможность работать с синтаксисом программного кода, решать математические задачи и генерировать тексты. Тем не менее, ответы, предоставляемые ботом, требуют дополнительной проверки, так как получаемая информация может быть некорректной.

В основе **Bard** лежит модель PaLM 2, которая обучена более чем на 100 разговорных языках и 20 языках программирования, а также способна воспринимать двусмысленность и фразеологические выражения. Модель проходила обучение на множестве научных публикаций, веб-страницах и учебниках по математике и физике. Одним из преимуществ является возможность развертывания модели в различных приложениях, так как она доступна в разных размерах (Gecko, Otter, Bison и Unicorn). Например, Gecko является облегченной версией, которая может работать на мобильных устройствах и обеспечивает быструю работу для интерактивных приложений, включая автономный режим.

GigaChat от «Сбер» — это сервис, разработанный компанией «Сбер», который позволяет работать с текстом, писать программный код и создавать изображения. Модель демонстрирует высокое качество работы с русским языком и «запоминает» детали ранее поставленных задач, формируя ответы на основе предыдущих запросов. Однако она не реагирует на корректировки неверных ответов, что может привести к получению компиляции реальных и вымышленных фактов при запросе актуальной информации. Архитектура GigaChat основана на нейросетевом ансамбле модели NeONKA (NEural Omnimodal Network with Knowledge-Awareness), которая

включает различные нейросетевые модели и методы supervised fine-tuning и reinforcement learning with human feedback.

YandexGPT от «Яндекс» — это нейросеть семейства GPT, предназначенная для работы с текстовой информацией, кодом и изображениями, а также для использования внешней информации при формировании ответов [Yan25]. В соответствии с определением самой нейросети, она заявляет (рис. 1): «Я — модель искусственного интеллекта, созданная для решения текстовых задач и помощи людям в поиске информации. Меня зовут Ассистент, и я здесь, чтобы ответить на ваши вопросы, предложить идеи, помочь с исследованиями или просто поддержать разговор».

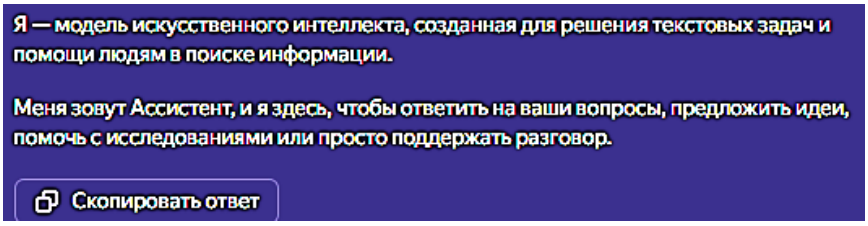

Рис. 1 Ответ от YandexGPT

Модель YandexGPT от компании «Яндекс» функционирует в двух режимах: асинхронном и синхронном. Асинхронный режим предназначен для решения сложных задач и оптимально подходит для бизнес-приложений, тогда как синхронный режим обеспечивает ответы в реальном времени, что делает его более подходящим для задач, связанных с виртуальными ассистентами.

Архитектура YandexGPT представлена на рис. 2 и содержит:

- входной слой: текстовые данные, которые поступают в модель;
- асинхронный режим: блок, который обрабатывает сложные задачи;
- синхронный режим: блок, который отвечает на запросы в реальном времени;
- выходной слой: ответы, генерируемые моделью.

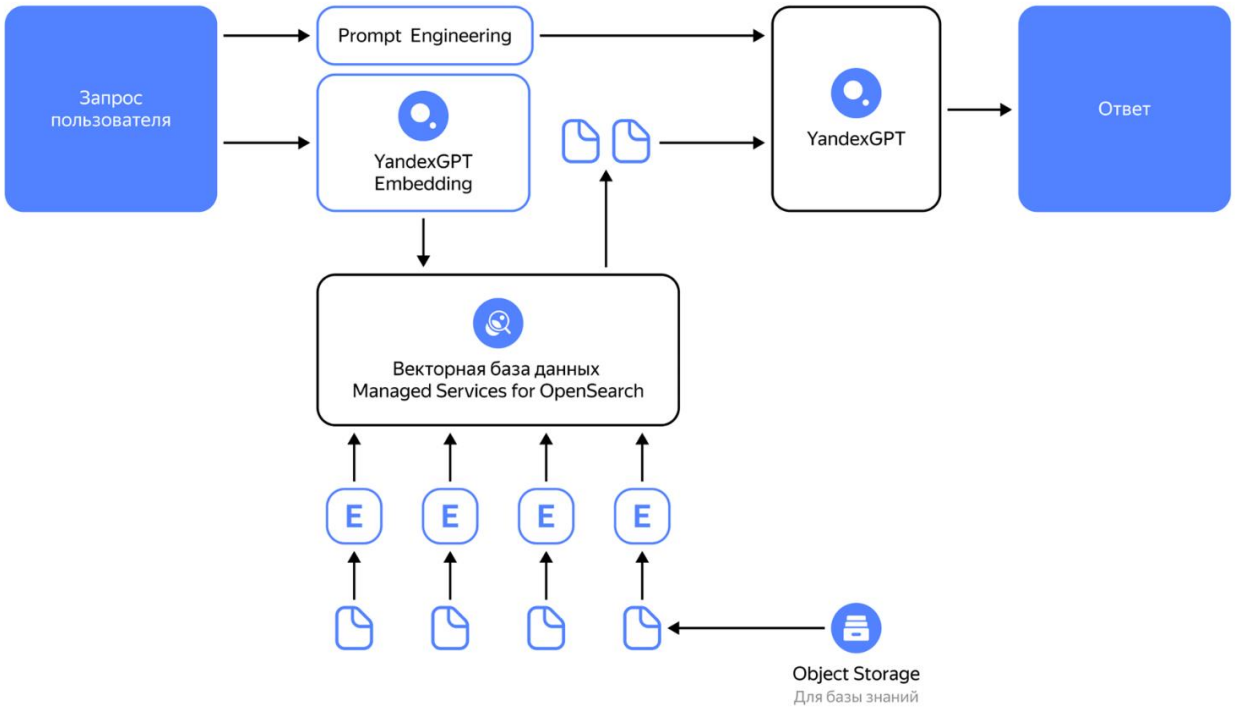

Рис. 2 Архитектура YandexGPT [Yan25]

DeepSeek, разработанный одноименной китайской организацией, в настоящее время является одним из самых популярных сервисов на рынке наряду с ChatGPT от OpenAI. Нейросеть DeepSeek построена на архитектуре «mixture of experts» (MoE), которая предполагает раздельную обработку данных с использованием знаний экспертов только в соответствующих областях. Эта архитектура включает два ключевых компонента [Sha24]:

1. Sparse MoE Layers. Эти слои заменяют плотные слои в архитектуре трансформера, что позволяет значительно уменьшить количество параметров и повысить эффективность обработки данных.

2. Gate Network. Этот компонент определяет, какие токены обрабатываются теми или иными экспертами, что позволяет оптимизировать использование ресурсов и улучшить качество обработки информации.

Формально, работа Gate Network может быть описана следующим образом:

$$y_i = \sum_{j=1}^{n} g_{ij} x_j,$$

где y_i — выходной вектор, g_{ij} — весовой коэффициент, определяемый сетью, x_j— входные токены.

DeepSeek отличается от ChatGPT тем, что предоставляет бесплатный доступ не только к генерации текста и кода, но и к анализу документов и интернет-поиску. Однако ChatGPT демонстрирует более точное понимание фраз и контекста, а также отличается высоким качеством генерации текста. Архитектура DeepSeek представлена на рис. 3 и содержит [Sha24]:

- входной слой: неструктурированные данные;
- sparse MoE Layers: блок, который заменяет плотные слои и обрабатывает данные;
- gate Network: блок, который определяет, какие токены обрабатываются экспертами;
- экспертные модели: несколько параллельных моделей, каждая из которых обрабатывает определенные токены;
- выходной слой: генерируемый текст или код.

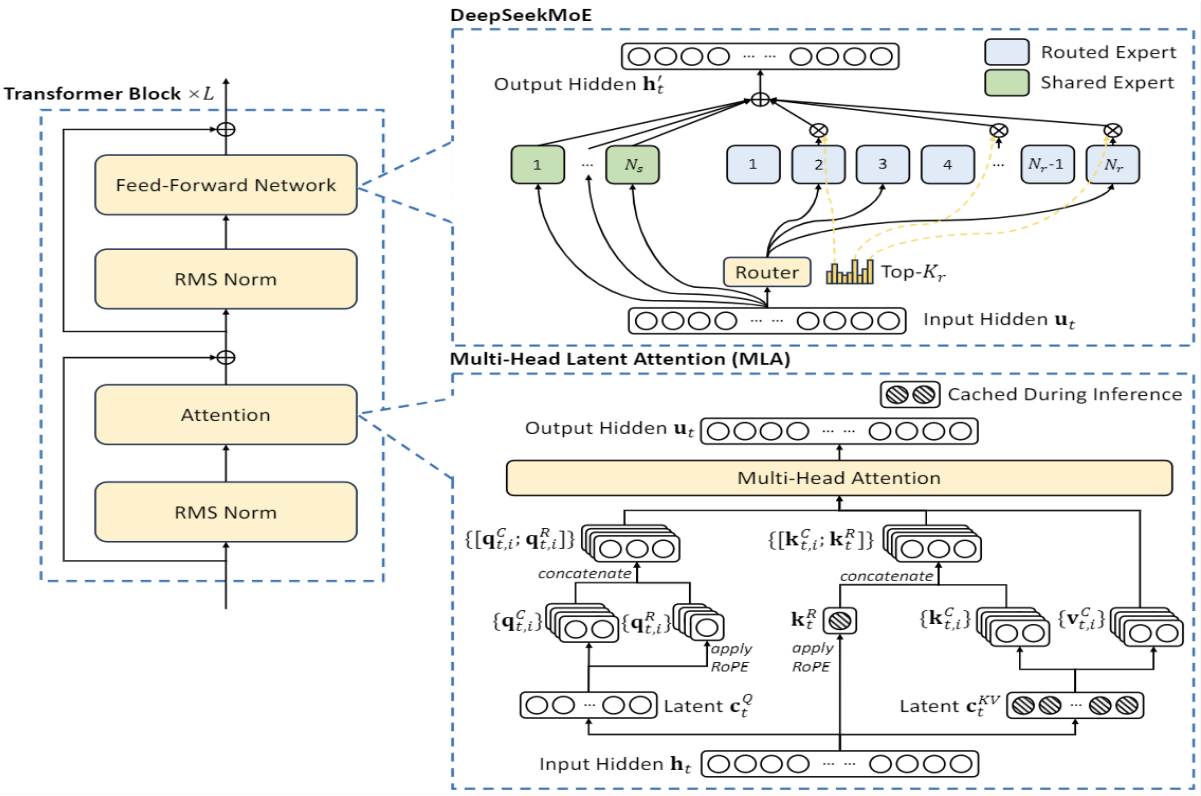

Рис. 3 Архитектура DeepSeek [Sha24]

ChatGPT от OpenAI — это сервис с генеративным искусственным интеллектом, который работает с текстовой информацией, программным кодом, математическими задачами и изображениями (ChatGPT 4) [Ope24]. Архитектура GPT представляет собой разновидность трансформаторной модели, которая в значительной степени зависит от механизма внимания. Трансформеры произвели революцию в обработке естественного языка (NLP) благодаря своей способности обрабатывать долгосрочные зависимости в тексте и их эффективности при обучении на больших наборах данных.

В основе ChatGPT лежат следующие компоненты:

1. Механизм многоглавого самовнимания (Multi-Head Self-Attention Mechanism). Данный механизм инициирует несколько параллельных потоков самовнимания с различными весовыми коэффициентами, что предоставляет модели возможность интегрировать контекстную информацию из каждого блока обучения. Математически это можно выразить как

$$\text{Attention}(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

где Q, K, и V — матрицы запросов, ключей и значений соответственно, а d_k— размерность ключей.

2. Нейронная сеть с прямой связью. Этот тип нейронной сети (НС) не содержит циклов в узлах и может иметь один или несколько скрытых слоев. Нейронная сеть с прямой связью наиболее эффективна для фильтрации шумовых данных, что является значительным преимуществом при работе с большими текстовыми массивами.

Математическая основа GPT

Одним из наиболее популярных решений на рынке является ChatGPT.

GPT – Generative Pre-trained Transformer – тип нейронных языковых моделей, которые обучаются на больших объемах текстовых данных и способны генерировать текст, схожий с тем, что пишет человек. Архитектура трансформера представлена на рис. 4 [Vas17].

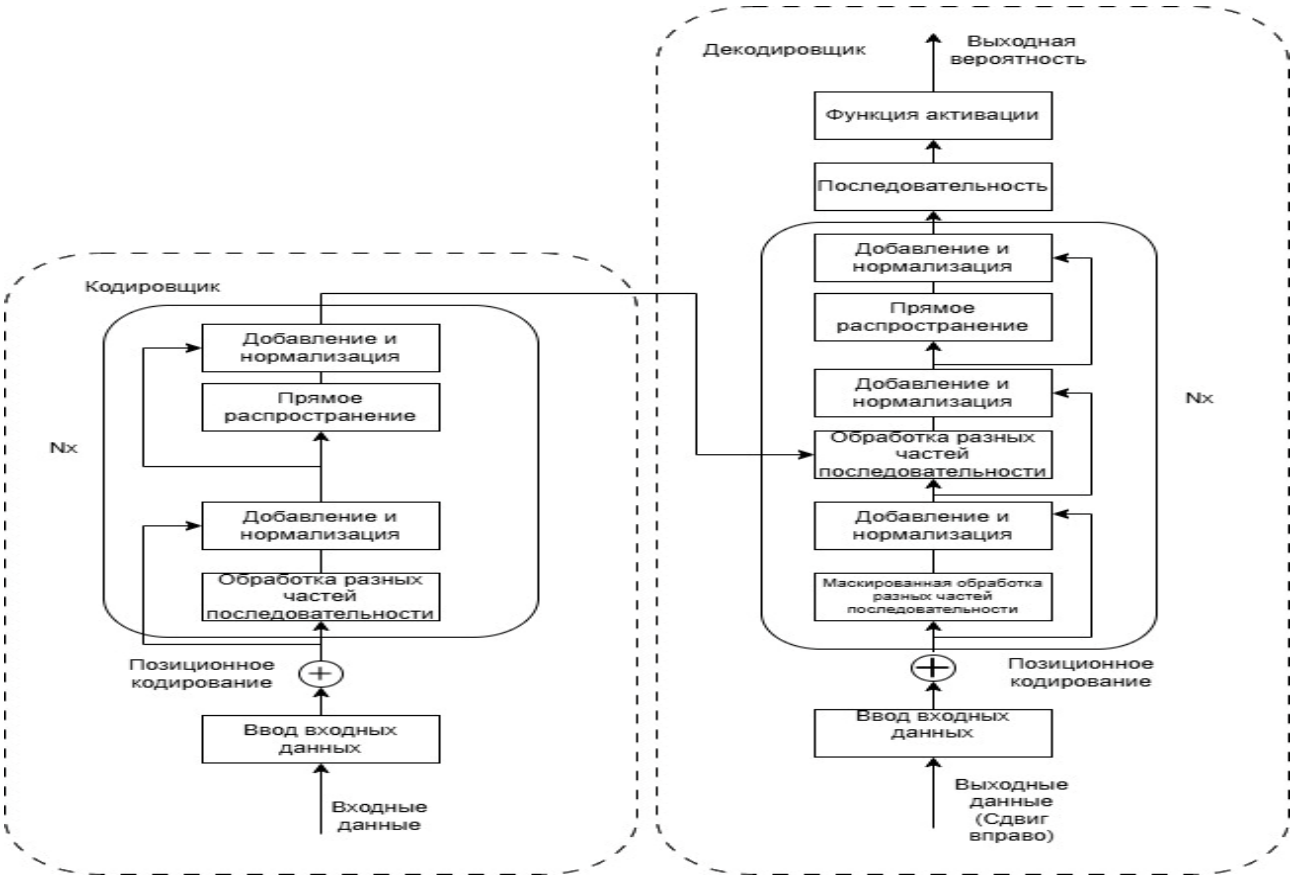

Рис. 4 Архитектура трансформера

Трансформер состоит из кодировщика и декодировщика. На вход трансформер получает текст, в дальнейшем кодировщик преобразуется его в векторную последовательность. Кодировщик учитывает позицию слов в предложении, для того чтобы модель могла «запомнить» на каком мете было конкретное слово. При условии перестановки модель будет «помнить» где было конкретное слово. Декодировщик получает на вход часть векторной последовательности после чего происходит обратное преобразование.

Кодировщик и декодировщик представлены на рис. 5 и 6 соответственно [Vas17].

Структура кодировщика и декодировщика отличается наличием механизма маскировки.

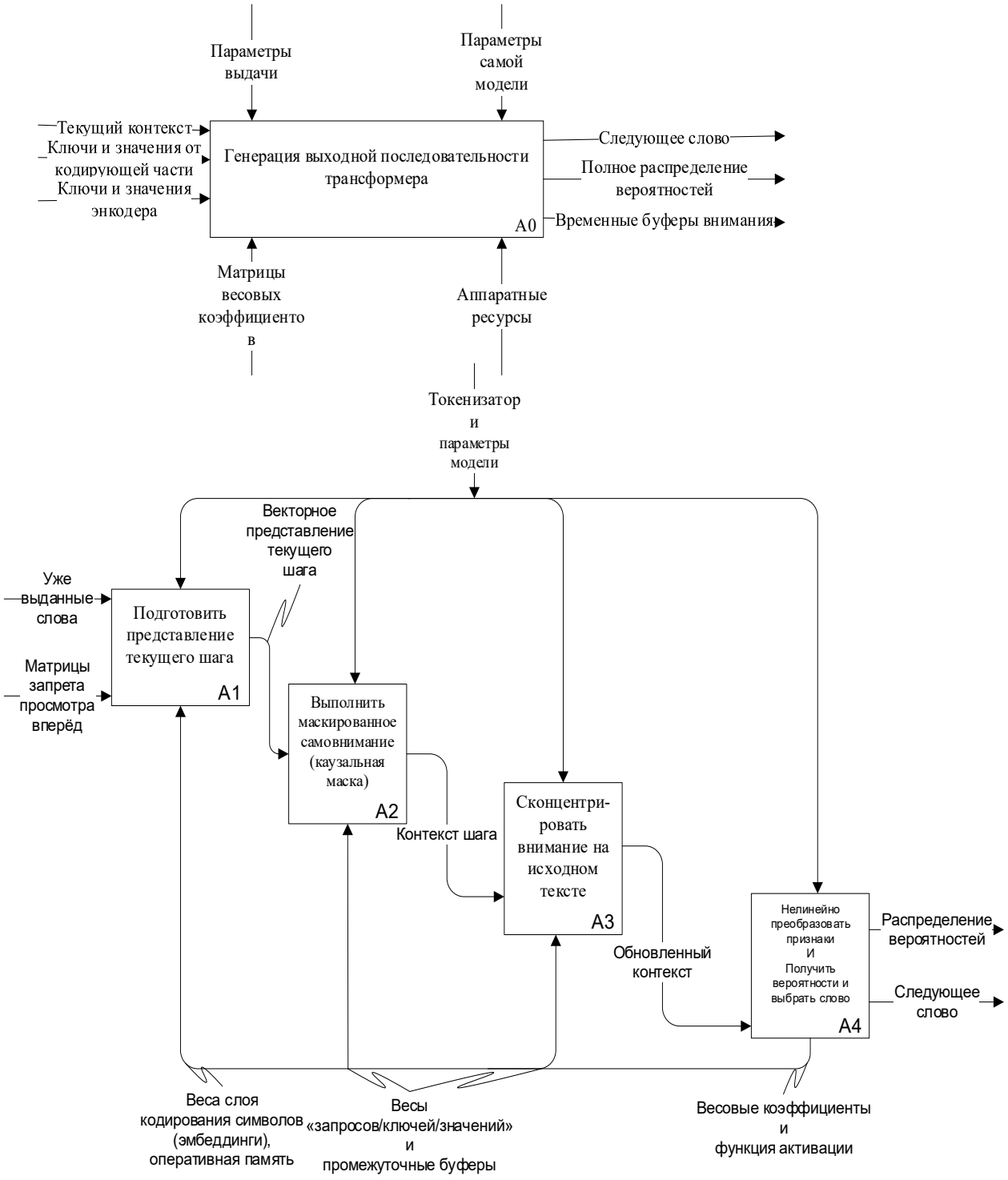

Рис. 5. IDEF0-диаграмма декодера трансформера

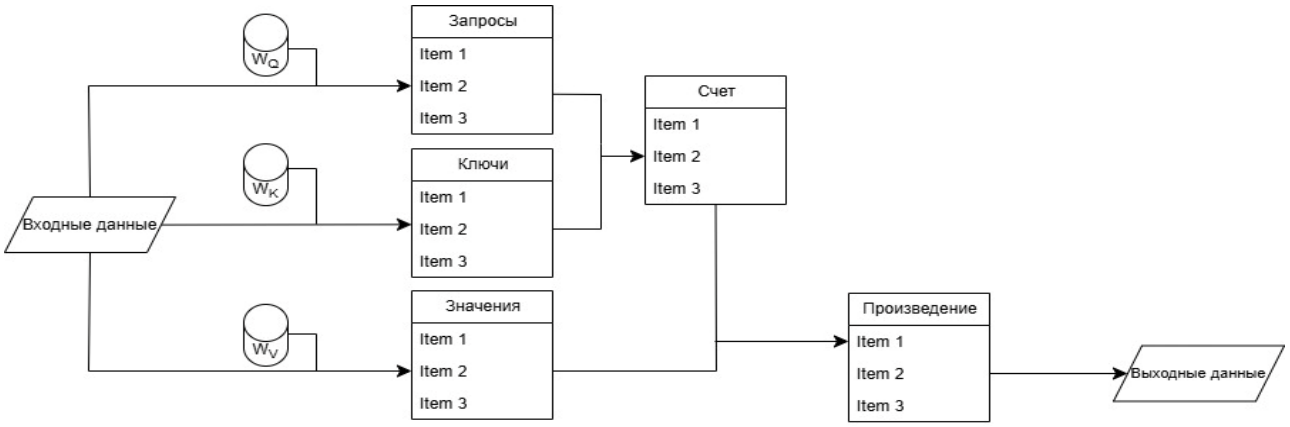

Рис. 6 Обучаемые матрицы

Процесс декодирования начинается с передачи токена <start>. Далее генерируется слово, проверяется позиция, предполагается следующее слово. Процесс продолжается пока не будет сгенерирована полная фраза.

Создатели ChatGPT используют комбинированный подход: предобучение без учителя и дообучение с частичным привлечением учителя.

В процессе вычисления механизма внимания используются три обучаемые матрицы: W_Q, W_K и W_V. Данные матрицы используются для получения трех сущностей: Query (Запрос), Key (Ключ) и Value (Значение). Эти матрицы служат для получения трех компонентов: Query (Запрос), Key (Ключ) и Value (Значение). Первые два компонента определяют попарные взаимосвязи между элементами последовательности, в то время как последний компонент предоставляет контекст для анализа рассматриваемого элемента. (см. рис. 6).

Каждый элемент входной последовательности умножается на эти матрицы, в результате чего формируются векторы-строки: запросы, ключи и значения. Сходство запроса и ключа определяется с использованием скалярного произведения.

Затем итоговый вектор преобразуется в вероятностное распределение по всем потенциальным следующим токенам. Для достижения этого вектор умножается на матрицу весов, в результате чего формируются логиты (ненормализованные логарифмические вероятности) для каждого из возможных токенов. Рассмотренный алгоритм представлен на рис. 7.

Наконец, происходит предсказание следующего слова. Каждый логит делится на параметр, известный как «температура». Температура является гиперпараметром, который оказывает влияние на уровень случайности при выборе следующего слова из распределения.

Алгоритм обратного распространения ошибки (Backpropagation) является ключевым методом, используемым для обучения нейронных сетей, включая модели на основе архитектуры GPT (Generative Pre-trained Transformer). Этот метод позволяет корректировать параметры модели с целью минимизации ошибки на обучающих данных, что можно формализовать следующим образом:

$$\text{Loss} = \frac{1}{n}\sum_{i=1}^{n} L\,(y_i, y_i^{'}),$$

где L – функция потерь, y_i – фактические выходные значения, $y_i^{'}$ – предсказанные выходные значения, а n — количество обучающих примеров.

Процесс обучения с использованием алгоритма обратного распространения ошибки можно разделить на несколько этапов:

1. **Прямое распространение.** На этом этапе входные данные x проходят через все слои нейронной сети. На каждом слое осуществляется вычисление выходных значений нейронов, что можно формализовать следующим образом: $a^l = f(W^l a^{l-1} + b^l)$, где f – функция активации, l – номер слоя, a^l – активации на l-м слое, W^l – веса на l-м слое, b^l – смещения на l-м слое.

2. **Вычисление ошибки.** После этапа прямого распространения осуществляется оценка ошибки модели на выходных данных. Эта ошибка представляет собой разницу между фактическими выходными значениями и ожидаемыми результатами: $\boldsymbol{E = y - \hat{y}}$.

3. **Обратное распространение (Backward Propagation)**. Ошибка E распространяется обратно через слои нейронной сети. Для каждого слоя вычисляются градиенты функции потерь L относительно параметров этого слоя. Градиенты можно вычислить с использованием правила цепочки: $\frac{\partial L}{\partial W^{(l)}} = \frac{\partial L}{\partial a^{(l)}} \cdot \frac{\partial a^{(l)}}{\partial W^{(l)}}$.

4. **Обновление параметров (Parameter Update)**. Параметры модели обновляются с использованием алгоритма градиентного спуска. Обновление весов можно выразить следующим образом: $\boldsymbol{W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}}}$, где η – скорость обучения.

5. **Повторение (Iteration)**. Шаги 1–4 повторяются до тех пор, пока ошибка не станет достаточно малой, что можно формализовать как $\textbf{Stop if } \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{L}\,(\boldsymbol{y_i}, \boldsymbol{y_i'}) < \boldsymbol{\varepsilon}$, где ε – заданный порог ошибки.

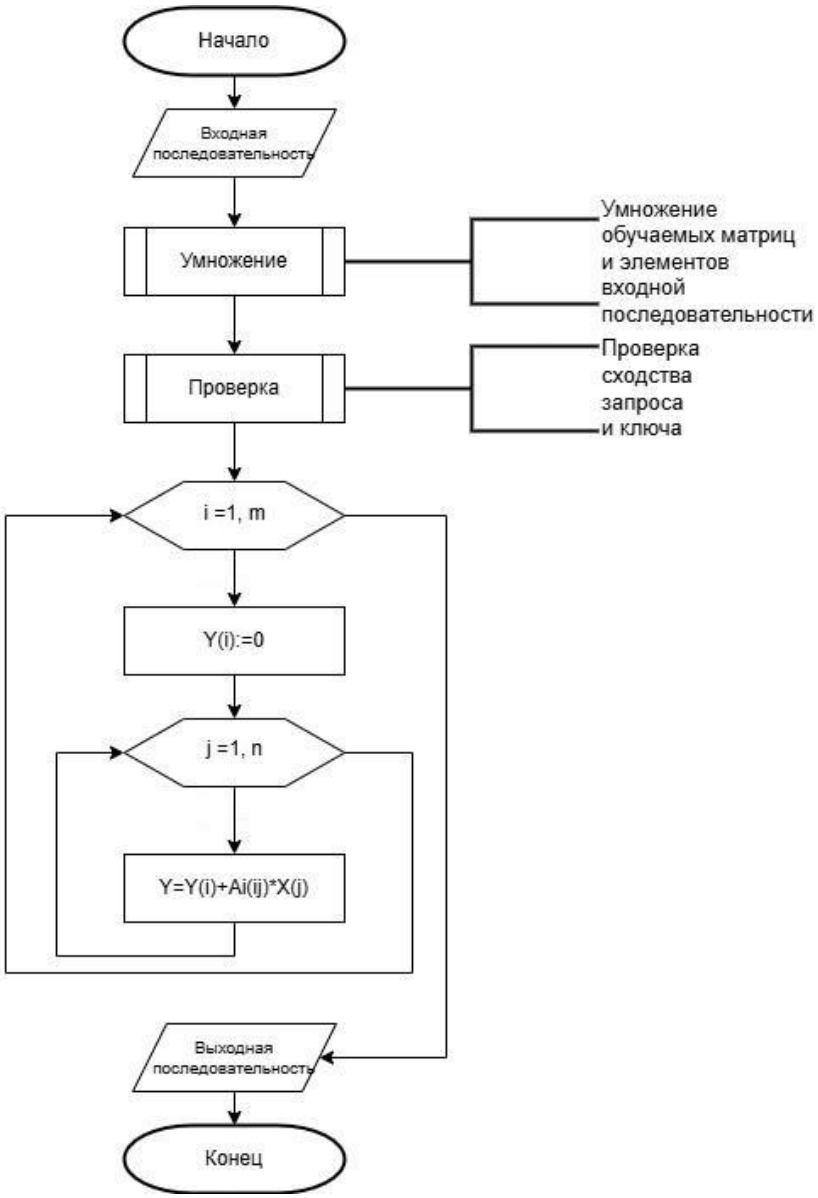

Рис. 7. Алгоритм Backpropagation [Rum86]

Таким образом, алгоритм обратного распространения ошибки является основным механизмом, позволяющим моделям, таким как GPT, эффективно обучаться на больших объемах данных, минимизируя ошибку предсказания и улучшая качество генерации текстов.

Таким образом, алгоритм back-propagation является ключевым механизмом, позволяющим крупным языковым моделям, таким как GPT, эффективно обучаться на больших корпусах данных, сокращая предсказательную ошибку и повышая качество генерируемого текста.

Математическая модель процесса тестирования

Прежде чем использовать LLM необходимо учитывать, что и сам процесс тестирования имеет математические модели. Одной из таких моделей является IRT.

IRT – Item Response Theory (Теории тестовых заданий) – набор методов, позволяющий оценить вероятность правильного ответа испытуемых на задания различной трудности. Она используется для того, чтобы избавиться от плохих вопросов в опроснике, оценки взаимосвязи латентных конструктов между собой и с наблюдаемыми переменными, оптимизации предъявления заданий респондентам.

IRT основывается на теории латентно-структурного анализа (ЛСА), созданной П. Лазарсфельдом и его последователями. В классической теории теста уровень свойства считается некоторым постоянным значением. В IRT латентный параметр трактуется как непрерывная переменная.

Преимущества модели IRT [Lin97]:

- *Более точная оценка.* IRT позволяет более точно оценить уровень знаний и способностей испытуемых, так как учитывает индивидуальные различия и сложность заданий. Это помогает создать более справедливые и объективные тесты.
- *Анализ качества заданий.* IRT предоставляет информацию о качестве каждого задания, включая его сложность, дискриминационную способность и другие характеристики. Это позволяет улучшить качество тестов и удалить неэффективные задания.
- *Адаптивное тестирование.* IRT может использоваться для создания адаптивных тестов, которые подстраиваются под уровень знаний испытуемого. Это сокращает время тестирования и повышает его эффективность.
- *Стандартизация оценок.* IRT обеспечивает стандартизацию оценок, что позволяет сравнивать результаты разных тестов и групп испытуемых. Это особенно полезно в образовании и профессиональной оценке.
- *Прогнозирование результатов.* IRT может прогнозировать результаты будущих тестов на основе предыдущих данных. Это помогает планировать обучение и развитие, а также оценивать эффективность образовательных программ.

IRT позволяет решить такие ключевые задачи, как:

- найти параметры заданий;
- найти параметры респондентов.

Известна также и трехпараметрическая модель (3PL), в которой третий параметр учитывает способность (вероятность) испытуемого угадать ответ на задание (параметр угадывания) с учетом характеристик самого задания [Nov24]:

$$3\mathrm{PL}_i(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - \delta_i)}}{1 + e^{a_i(\theta_j - \delta_i)}}, \quad (1)$$

где $3\mathrm{PL}_i(\theta_j)$ – вероятность правильного ответа j-го испытуемого на i-е задание; α_i – дифференцирующая способность задания (параметр дискриминации i-го задания), показывает, насколько хорошо задание различает испытуемых с разными уровнями способностей. Чем выше значение a_i, тем сильнее задание дифференцирует между теми, кто хорошо знает материал, и теми, кто знает его хуже; θ_j – параметр, описывающий латентную характеристику j-го испытуемого; δ_i – характеристика сложности i-го пункта теста (задания), указывает на уровень

способности, необходимый для успешного выполнения задания. Более высокие значения δ_i соответствуют более сложным заданиям; c_i – параметр угадывания i-го задания, оценивает вероятность того, что испытуемый угадает правильный ответ на задание, даже если он не обладает необходимыми знаниями.

Усредненная трехпараметрическая модель дает средний показатель способности множества испытуемых (усредняем значение вероятности правильного ответа для каждого испытуемого, учитывая параметры конкретного тестового элемента):

$$3\mathrm{PL}_{\mathrm{sr}} = \frac{\sum_{j=1}^{n}(c_i+(1-c_i)\frac{e^{a_i(\theta_j-\delta_i)}}{1+e^{a_i(\theta_j-\delta_i)}})}{n} \quad (2)$$

где n – количество испытуемых; $e^{a_i(\theta_j-\delta_i)}$ – функция, моделирующая влияние способностей испытуемого и сложности тестового элемента на вероятность правильного ответа (т. е. чем выше способность испытуемого, тем выше вероятность правильного ответа). Формула

$$c_i+(1-c_i)\frac{e^{a_i(\theta_j-\delta_i)}}{1+e^{a_i(\theta_j-\delta_i)}}$$

моделирует вероятность правильного ответа i-го тестового элемента для j-го испытуемого.

Оптимизированная математическая модель тестирования с применением LLM

Для оптимизации математический модели тестирования в том числе необходимо рассмотреть характеристику надежности вариантов тестирования. Эту характеристику также называют надежностью параллельных форм.

Надежность тестов, в частности надежность параллельных форм, может быть оценена с помощью различных статистических методов. Одним из таких методов является использование коэффициента надежности, который измеряет степень согласованности между различными формами теста.

Одной из наиболее распространенных моделей для оценки надежности является коэффициент альфа Кронбаха (α), который можно использовать для оценки надежности параллельных форм. Однако, если говорить конкретно о надежности параллельных форм, то можно использовать следующую формулу, получившую название «формула надежности параллельных форм» [Cro51]:

Если X_1 и X_2 — это результаты двух параллельных форм теста, то надежность rxx' может быть выражена следующим образом:

$$r_{xx'} = \frac{\mathrm{Cov}\,(X_1,X_2)}{\sqrt{\mathrm{Var}\,(X_1)\,\mathrm{Var}\,(X_2)}},$$

где $\mathrm{Cov}\ (X_1, X_2)$ – ковариация между результатами двух форм теста; $\mathrm{Var}\ (X_1)\,\mathrm{Var}\ (X_2)$ — дисперсии результатов по каждой из форм теста.

Для практического применения важно учитывать, что надежность теста может варьироваться в зависимости от уровня способностей тестируемых, поэтому рекомендуется проводить анализ надежности на разных подгруппах. На практике применяется надежность по однородности. Для его определения используют статистический метод для оценки надежности тестов Кьюдера–Ричардсона, который используется для тестов с бинарными (правильный/неправильный) ответами. Он позволяет оценить степень согласованности между элементами теста и является одним из наиболее распространенных методов для определения надежности по однородности. Суть метода заключается в определении надежности, основанном на однократном предъявлении теста. Он использует данные о выполнении испытуемыми каждого задания. Формула Кьюдера–Ричардсона имеет следующий вид [Kud37]:

$$r_{KR} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{s_v^2}\right), \quad (3)$$

где p_j — доля правильных ответов на j-е задание; q_j — доля неправильных ответов на j-е задание, которая вычисляется как $q_j = 1 - p_j$; S_v^2 — дисперсия по распределению наблюдаемых баллов; n — число заданий теста.

Приведем некоторые пояснения к компонентам формулы (3):

- p_j и q_j отражают, сколько испытуемых ответили правильно и неправильно на конкретное задание. Эти значения находятся в пределах от 0 до 1;
- сумма произведений p_j и q_j показывает, сколько испытуемых располагались между правильными и неправильными ответами, что является показателем сложности заданий.

Дисперсия баллов S_v^2 вычисляется как мера разброса результатов тестирования. Она показывает, насколько сильно результаты теста варьируются от среднего значения. Если дисперсия мала, это может указывать на то, что большинство испытуемых получили схожие баллы, что может быть признаком низкой надежности теста.

Коэффициент надежности также принимает значения от 0 до 1. Значение, близкое к 1, указывает на высокую надежность теста, в то время как значение, близкое к 0, свидетельствует о низкой надежности. Если значение меньше 0,7, это может указывать на необходимость пересмотра или улучшения теста.

Используя все рассмотренные выше формулы, можно представить оптимизированную математическую модель тестирования с применением LLM в следующем сокращенном виде:

$$M = \frac{kx + 3\mathrm{PL}_{sr}}{r_{KR}}, \quad (4)$$

где M – оптимизированная модель тестирования (это итоговая оценка или метрика, которая характеризует эффективность тестирования); k – параметр LLM – модели (метода top-k; выбирается из топ-k наиболее вероятных слов, независимо от их совокупной вероятности), может отражать вес или значимость определенных факторов в модели. Этот коэффициент может быть настроен в зависимости от целей теста или специфики контекста; x – векторизованная последовательность входного значения задания (другими словами – это переменная, представляющая конкретный аспект тестирования. Это может быть, например, количество заданий, уровень сложности теста или другие количественные показатели); $3\mathrm{PL}_{sr}$ – среднее значение трехпараметрической модели 3PL; r_{KR} – характеристика надежности параллельных форм.

Данная модель позволяет адаптировать набор тестов под конкретное множество студентов и в том числе определить вероятность правильного ответа студента на конкретное задание на основе его уровня подготовки и сложности задания:

- После ответа на очередное задание вероятность его правильного решения оценивается формулой (2) трехпараметрической модели 3PL, а апостериорная способность студента θj уточняется байесовским шагом.
- Из банка заданий LLM предварительно вычисляет параметры сложности β_i и дискриминации α_i, что ускоряет калибровку и позволяет включать свежесгенерированные вопросы.
- На каждом шаге система выбирает пункт с максимальной информационной функцией $I_i(\theta_j)$ – это минимизирует стандартную ошибку оценки для данного студента.
- Такая процедура в среднем сокращает длину теста при том же доверительном интервале способности, тем самым повышая точность и снижая нагрузку на испытуемого.

Обсуждение

Применение больших языковых моделей (LLM) в образовании охватывает перевод, генерацию пояснительных материалов и другие задачи. Однако именно в тестировании они уже демонстрируют наибольший практический эффект по ряду причин [Gui25]:

- *Генерация тестовых сценариев.* LLM способны создавать тестовые сценарии на основе требований к программному обеспечению или описания функциональности, что позволяет ускорить разработку тестов и повысить их полноту.
- *Автоматическое создание вопросов.* LLM могут генерировать вопросы и задачи для проверки знаний, варьируя сложность и типы ответов.
- *Анализ результатов.* LLM выявляют закономерности и тенденции, предлагая улучшения в процессе тестирования.
- *Улучшение тестовой документации.* LLM помогают создавать и улучшать тестовую документацию, делая ее более понятной и информативной.
- *Поддержка в обучении.* LLM могут служить виртуальными наставниками, объясняя сложные концепции и помогая студентам с подготовкой к экзаменам.
- *Интеграция с системами управления тестированием.* LLM помогают автоматизировать задачи, такие как планирование тестов, управление дефектами и генерация отчетов.

LLM могут помочь в выявлении и анализе когнитивных процессов, лежащих в основе ответов на тесты, что может привести к более глубокому пониманию того, как разные группы учащихся взаимодействуют с тестовыми заданиями.

LLM может быть использован для создания симуляционных моделей, которые позволяют предсказывать последствия изменений в тестовых заданиях. Это может быть полезно для совершенствования образовательных программ и стандартов оценки знаний.

Например, модель может:

- анализировать данные прошлых тестов и результаты, чтобы выявить тенденции и закономерности;
- оценивать, как изменения в формулировках вопросов или формате тестирования могут повлиять на ответы испытуемых;
- предсказывать, как изменения в содержании теста могут повлиять на уровень знаний и успеваемость студентов;
- тестировать новые подходы к оценке знаний, такие как адаптивные тесты;
- прогнозировать, какие изменения в тестовых заданиях могут улучшить их валидность и надежность.

Вместе с тем у применения языковых моделей существует и ряд ограничений, в том числе этических и ресурсных.

Существует понятие AI-галлюцинации, которое требует обязательной перепроверки выдаваемых данных, так как при недостаточном обучении модель может генерировать правдоподобную, но ложную информацию. Также с точки зрения информационной безопасности нельзя не учитывать риск утечки обучающего датасета, что может привести к последующему некорректному анализу.

С точки зрения ресурсных ограничений возникает вопрос объема вычислительных мощностей. Начиная от высокого требования к пропускной способности сети, мощных графических процессоров, больших объемов оперативной памяти, объемного датасета. Заканчивая энергетическими затратами.

Необходимо учитывать, что LLM не могут полностью заменить математические расчеты и специализированные знания. Они могут быть использованы в качестве инструмента для анализа данных и прогнозирования результатов, но окончательное решение о содержании и структуре теста должно приниматься на основе профессиональной экспертизы.

Благодарности и поддержка

Работа поддержана Министерством науки и высшего образования Российской Федерации в рамках базовой части государственного задания для высших учебных заведений # FRRR-2026-0006.

Заключение

Представленное исследование подтверждает, что объединение методов Item Response Theory с большими языковыми моделями значительно повышает точность и гибкость компьютеризированного тестирования. Предложенный алгоритм динамически обновляет оценку способностей испытуемого и подбирает задания в соответствии с его текущим уровнем подготовки, позволяя сократить общее количество вопросов без потери надёжности. Использование языковой модели для предварительной калибровки новых заданий упрощает расширение банка тестовых материалов и сохраняет их психометрические характеристики.

Дальнейшее развитие работы связано с учётом времени отклика, применением более сложных многопараметрических моделей и интеграцией алгоритма в образовательные платформы. Планируется расширить экспериментальную базу за счёт школьных и корпоративных курсов, а также исследовать устойчивость метода к культурным и языковым различиям. Полученные результаты демонстрируют перспективность синергии IRT и LLM для создания адаптивных, научно обоснованных систем оценки знаний в цифровой среде.

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

[Che25] Chen Y., Li X., Liu J., Ying Z. Item response theory — a statistical framework for educational and psychological measurement. LSE Research Online Documents on Economics, No 120810, London School of Economics and Political Science, 2025. DOI: 10.1214/23-STS896.

[Cro51] Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika. 1951;16(3):297-334. DOI: 10.1007/BF02310555. EDN: EGXXRL.

[Goo23] Google. Bard and new AI features in Search — official Google AI blog update [Электронный ресурс]. 2023. URL: https://blog.google/technology/ai/bard-google-ai-search-updates (дата обращения: 12.10.2025).

[Gui25] Guizani, S., Mazhar, T., Shahzad, T. et al. A systematic literature review to implement large language model in higher education: issues and solutions. Discov Educ 4, 35 (2025). DOI: 10.1007/s44217-025-00424-7. EDN: YEAJWD.

[Ham91] Hambleton R. K., Swaminathan H., Rogers H. J. Fundamentals of Item Response Theory. Newbury Park, CA: Sage, 1991.

[Hua25] Huang S., Luo J., Jeon M. A response time-based mixture item response theory model for dynamic item-response strategies. Behavior Research Methods. 2025; 57(1):54. DOI: 10.3758/s13428-024-02555-5. EDN: HPEEXF.

[Kud37] Kuder GF, Richardson MW. The Theory of the Estimation of Test Reliability. Psychometrika. 1937;2(3):151-160. DOI: 10.1007/BF02288391. EDN: KOOOMO.

[Lin97] Linden W. J. van der, Hambleton R. K. (Eds.). Handbook of Modern Item Response Theory. New York: Springer, 1997. DOI: 10.1007/978-1-4757-2691-6.

[Lor68] Lord F. M., Novick M. R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

[Nov24] Noventa S, Ye S, Kelava A, Spoto A. On the Identifiability of 3- and 4-Parameter Item Response Theory Models From the Perspective of Knowledge Space Theory. Psychometrika. 2024;89(2):486-516. DOI: 10.1007/s11336-024-09950-z. EDN: XHSHYW.

[Ope24] OpenAI. GPT-4 technical report [Электронный ресурс]. 2024. 98 с. URL: https://arxiv.org/abs/2303.08774 (дата обращения: 12.10.2025).

[Ras60] Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche.

[Ray25] Raykov, T., & Zhang, B. (2025). The One-Parameter Logistic Model Can Be True With Zero Probability for a Unidimensional Measuring Instrument: How One Could Go Wrong Removing Items Not Satisfying the Model. Educational and Psychological Measurement, 0(0). DOI: 10.1177/00131644251345120.

[Rum86] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986). DOI: 10.1038/323533a0.

[Sbe25] Sber Developers. GigaChat API: Large Language Models — What They Are and How They Work [Электронный ресурс]. 2025. URL: https://developers.sber.ru/docs/ru/gigachat_api (дата обращения: 12.10.2025).

[Sha24] Shao Z., Wang P., Zhu Q. et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models [Электронный ресурс]. arXiv preprint arXiv:2402.03300. 2024. 37 с. URL: https://arxiv.org/abs/2402.03300 (дата обращения: 12.10.2025).

[Vas17] Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems — 2017. Vol. 30. P. 5998–6008. URL: https://arxiv.org/abs/1706.03762 (дата обращения: 12.10.2025).

[Yan25] Yandex Cloud. YandexGPT 5 — Generative AI for Business [Электронный ресурс]. 2025. URL: https://yandex.cloud/en/services/yandexgpt (дата обращения: 12.10.2025).

[Кор24] Коровин Е. А., Чиглинцева С. А., Сазонова Е. Ю., Сметанина О. Н. Медицинская рекомендательная система на основе автоматического извлечения знаний из текстов // СИИТ. 2024. Т. 6, № 4(19). С. 111-121. DOI: 10.54708/2658-5014-SIIT-2024-no4-p111. EDN: OTVTXR.

[Куч24] Кучкарова Н. В. Оценка актуальных угроз и уязвимостей объектов критической информационной инфраструктуры с использованием технологий интеллектуального анализа текстов // СИИТ. 2024. Т. 6, № 2(17). С. 50-65. DOI: 10.54708/2658-5014-SIIT-2024-no2-p50. EDN: NLDWBE.

[Мор25] Морозов М. И. Предсказание наступления страхового случая с помощью трансформерных нейросетей // СИИТ. 2025. Т. 7, № 2(21). С. 96-102. DOI: 10.54708/2658-5014-SIIT-2025-no2-p100. EDN: FEFPBE.

[Рез25] Резников Г. А., Синицын Р. Д., Шулик А. М. Современные архитектуры нейронных сетей для тегирования и аннотирования изображений: достижения, вызовы и перспективы // СИИТ. 2025. Т. 7, № 2(21). С. 78-85. DOI: 10.54708/2658-5014-SIIT-2025-no2-p82. EDN: TJFUGV.

[Шал23] Шалфеева Е. А. Методология производства жизнеспособных систем доверительного искусственного интеллекта // СИИТ. 2023. Т. 5, № 4(13). С. 28-49. DOI: 10.54708/2658-5014-SIIT-2023-no3-p114. EDN: CJTKQH.

[Шир25] Ширинов Р. А., Гардашова Л. А. г., Богданова Д. Р. Краткий анализ методов Deep Learning для распознавания эмоционального состояния человека для принятия решений // СИИТ. 2025. Т. 7, № 2(21). С. 68-77. DOI: 10.54708/2658-5014-SIIT-2025-no2-p68. EDN: DAWGEI.

ОБ АВТОРАХ | ABOUT THE AUTHORS

АЛЕКСЕЕВА Дарья Сергеевна
Уфимский университет науки и технологий, Россия.
ads.stat@mail.ru ORCID: 0009-0002-8955-9849.
Аспирантка, кафедра автоматизированных систем управления.

ПАЛЬЧЕВСКИЙ Евгений Владимирович
МИРЭА — Российский технологический университет, Россия.
teelxp@inbox.ru ORCID: 0000-0001-9033-5741.
Доцент каф. индустриального программирования. Магистр (Уфимск. гос. авиац. техн. ун-т, 2019). Канд техн. наук (Уфимск. ун-т науки и технологий, 2024). Иссл. в обл. интеллектуальных вычислений и систем.

АНТОНОВ Вячеслав Викторович
Уфимский университет науки и технологий, Россия.
antonov.v@bashkortostan.ru ORCID: 0000-0002-5402-9525
Зав. каф. автоматизированных систем управления, профессор. Инженер (Башкирск. гос. ун-т, 1979). Д-р техн. наук (Уфимск. гос. авиац. техн. ун-т, 2015). Иссл. в обл. интеллектуальных систем.

СУВОРОВА Вероника Александровна
Уфимский университет науки и технологий, Россия.
Milana_da@mail.ru
Доцент каф. автоматизированных систем управления, доцент. Уфимск. гос. авиац. техн. ун-т, 2004. Канд техн. наук (Уфимск. гос. авиац. техн. ун-т, 2010). Иссл. в обл. интеллектуальных систем.

ALEKSEEVA Daria Sergeevna
Ufa University of Science and Technology, Russia.
ads.stat@mail.ru ORCID: 0009-0002-8955-9849.
Postgraduate student of the Department of Automated Control Systems.

PALCHEVSKY Evgeny Vladimirovich
MIREA — Russian Technological University (RTU MIREA), Russia.
teelxp@inbox.ru ORCID: 0000-0001-9033-5741.
Associate Professor, Department of Industrial Programming.
M.Sc. (Ufa State Aviation Technical University, 2019).
Ph.D. (Tech.) – Ufa University of Science and Technology, 2024.
Research interests: intelligent computing and systems.

ANTONOV Vyacheslav Viktorovich
Ufa University of Science and Technology, Russia.
antonov.v@bashkortostan.ru ORCID: 0000-0002-5402-9525
Head of the Department of Automated Control Systems, Prof. Eng. (Bashkir State Univ., 1979). Doctor of Engineering Sciences (Ufa State Aviation Technical Univ., 2015). Research in the field of intelligent systems.

SUVOROVA Veronika Aleksandrovna
Ufa University of Science and Technology, Russia.
Milana_da@mail.ru .
Associate Professor of the Department of Automated Control Systems, (Ufa State Aviation Technical University, 2004), Ph.D. (Tech.) (Ufa State Aviation Technical University, 2010). Research in the field of intelligent systems.

МЕТАДАННЫЕ | METADATA

Заглавие: Оптимизация математической основы IRT с помощью LLM-моделей.

Авторы: Алексеева Д. С., Пальчевский Е. В., Антонов В. В., Суворова В. А.

Аннотация: Оптимизация математической основы IRT (Item Response Theory) основанная на теории реакции на предметы, с помощью LLM (Large Language Models) — это довольно новая и многообещающая область исследований, которая сочетает в себе статистику, психологию, образование и машинное обучение. LLM могут быть использованы для анализа данных прошлых тестов и выявления тенденций и закономерностей. Модели могут помочь в оценке наиболее эффективных типов вопросов для оценки определенных навыков или знаний. Они могут оценить, как изменения в формулировках вопросов или формате тестирования могут повлиять на ответы испытуемых. Модели могут предсказать, как изменения в содержании теста могут повлиять на уровень

Title: Optimization of the mathematical basis of IRT using LLM models.

Authors: Alekseeva D. S., Palchevsky E. V., Antonov V. V., Suvorova V. A.

Abstract: Optimizing the mathematical basis of IRT (Item Response Theory) based on the theory of reaction to objects using LLM (Large Language Models) is a fairly new and promising field of research that combines statistics, psychology, education and machine learning. LLMs can be used to analyze data from past tests and identify trends and patterns. Models can help in evaluating the most effective types of questions to assess certain skills or knowledge. They can assess how changes in the wording of the questions or the format of the test may affect the responses of the subjects. The models can predict how changes in the content of the test may affect students' level of knowledge and academic performance. Models can also be used to create adaptive

знаний и успеваемость студентов. Также модели могут использоваться для создания адаптивных тестов, которые подстраиваются под уровень знаний и способностям испытуемых. Оптимизация математической основы IRT с применением LLM может привести к более точным и эффективным тестам. Это может быть полезно в различных областях, где требуется оценка уровня знаний или навыков, таких как образование, медицина, психология и другие.

Ключевые слова: IRT, LLM, Large Language Models, большие языковые модели, оптимизация, нейронные сети, тестирование, адаптивные методики.

Язык: Русский.

Статья поступила в редакцию 24 сентября 2025 г.

tests that adapt to the level of knowledge and abilities of the subjects. Optimizing the mathematical basis of IRT using LLM can lead to more accurate and effective tests. This can be useful in various fields where an assessment of the level of knowledge or skills is required, such as education, medicine, psychology, and others.

Key words: IRT, LLM, Large Language Models, optimization, neural networks, testing, adaptive methods.

Language: Russian.

The article was received by the editors on 24 September 2025.