ИССЛЕДОВАНИЕ ВЛИЯНИЯ РАЗЛИЧНЫХ РЕЖИМОВ АДАПТИВНОГО ПОВЕДЕНИЯ AI-ЧАТБОТА DUMB-AI НА КАЧЕСТВО ОТВЕТОВ ПРИ РЕШЕНИИ ЗАДАЧ РАЗНЫХ ТИПОВ

Автор: Лайша Арсений

Класс: 9 «А»

Колодищанская средняя школа № 2 — Минск, Беларусь

2026

1. ВВЕДЕНИЕ

Искусственный интеллект является одним из наиболее активно развивающихся направлений современной информационной технологии. Особое место среди систем искусственного интеллекта занимают языковые модели и AI-чатботы, способные отвечать на вопросы пользователя, объяснять информацию и решать задачи различных типов.

При этом качество ответа чатбота зависит не только от используемой модели, но и от способа организации её поведения. Изменение параметров адаптивного поведения может влиять на способность системы учитывать условие задачи, выполнять рассуждения и формировать правильный итоговый ответ.

В рамках данного исследования рассматривается разработанный AI-чатбот DUMB-AI, в котором предусмотрены различные режимы адаптивного поведения. Для экспериментальной проверки были использованы три режима: DUMB, NORMAL и GENIUS, соответствующие различной степени сложности поведения системы.

Цель исследования: определить влияние различных режимов адаптивного поведения AI-чатбота DUMB-AI на качество ответов при решении задач разных типов.

Для достижения цели были поставлены следующие задачи: изучить подходы к оцениванию качества ответов языковых моделей; определить критерии оценки ответов DUMB-AI; провести эксперимент с использованием трёх режимов; сравнить количество правильных и неправильных ответов; классифицировать ошибки; определить различия между режимами и оценить практическую значимость результатов.

Объект исследования: AI-чатбот DUMB-AI.

Предмет исследования: зависимость качества ответов DUMB-AI от выбранного режима адаптивного поведения.

Гипотеза исследования: изменение режима адаптивного поведения AI-чатбота влияет на качество его ответов, при этом повышение уровня адаптивного поведения может приводить к увеличению доли правильных решений.

Методы исследования: анализ литературы, эксперимент, сравнение результатов, количественный анализ, классификация ошибок и интерпретация данных.

Практическая значимость заключается в возможности использовать результаты эксперимента для совершенствования DUMB-AI, в частности для разработки механизма выбора режима ответа в зависимости от типа и сложности задачи.

2. ОБЗОР ЛИТЕРАТУРЫ

Современные языковые модели используют методы обработки естественного языка для анализа пользовательских запросов и формирования ответов. Одним из наиболее значимых этапов развития таких систем стало появление архитектуры Transformer, основанной на механизме внимания. Она стала основой для большого количества современных языковых моделей.

Высокая способность языковой модели генерировать текст сама по себе не гарантирует правильность результата. Поэтому важной задачей является разработка методов оценки качества AI-систем.

Одним из примеров комплексного подхода является HELM — система оценки языковых моделей, предполагающая использование нескольких показателей и различных сценариев тестирования. Такой подход показывает, что качество модели целесообразно оценивать с учётом характера выполняемых задач.

Вопросам оценки и управления рисками AI-систем также уделяется внимание в документах NIST. AI Risk Management Framework рассматривает подходы к измерению, оценке и документированию характеристик AI-систем. Для генеративного искусственного интеллекта разработан отдельный профиль, учитывающий особенности подобных систем.

Для DUMB-AI в настоящем исследовании используется более простой экспериментальный подход. Качество работы системы оценивается прежде всего по правильности конечного ответа, а дополнительно анализируются типы ошибок и качество рассуждения.

3. МЕТОДИКА ИССЛЕДОВАНИЯ

Для проведения эксперимента были сформированы задания различных типов, требующие от AI-чатбота выполнения вычислений, анализа условий, логических рассуждений и формирования итогового вывода.

Каждое задание последовательно решалось в трёх режимах DUMB-AI: **DUMB** — режим с минимальным уровнем адаптивного поведения; **NORMAL** — стандартный режим; **GENIUS** — режим с повышенным уровнем адаптивного поведения.

Изначально экспериментальный набор включал 20 заданий. Все задания были протестированы во всех трёх режимах. Для количественной статистической оценки использовались 17 заданий. Задание №6 исключено из-за неоднозначности условия. Задания №19 и №20 не использовались для статистического сравнения в соответствии с планом эксперимента.

Для каждого режима определялось количество правильных и неправильных конечных ответов. Точность рассчитывалась по формуле:

Точность = количество правильных ответов / количество оцениваемых заданий × 100%.

Дополнительно проводился анализ характера ошибок. Были выделены категории: неправильное понимание условия, логическая ошибка, вычислительная ошибка, неправильный вывод, противоречие в рассуждении, избыточное усложнение и неполное выполнение задания.

Отдельно учитывалось различие между правильностью конечного ответа и качеством рассуждения. Если система получала правильный результат, но в объяснении допускала существенную логическую ошибку, этот случай отмечался отдельно.

Ограничения исследования: небольшая выборка и ограниченное количество типов заданий. Поэтому результаты характеризуют проведённый эксперимент и не могут автоматически распространяться на все AI-системы.

3.1. Организация эксперимента

Эксперимент проводился на одном и том же наборе входных данных: формулировка каждого задания сохранялась неизменной при переходе между режимами. Это позволяет сравнивать режимы между собой и уменьшает влияние различий в условиях тестирования.

Задания были подобраны так, чтобы проверить не только арифметические вычисления, но и работу с последовательностями, логическими ограничениями, процентами, геометрией и задачами, в которых из условия нельзя сделать однозначный вывод. Такой набор позволяет проверить разные стороны поведения чатбота.

Для каждого ответа фиксировались: конечный результат, соответствие правильному ответу, тип ошибки и краткий комментарий. В случае правильного результата дополнительно проверялось рассуждение, поскольку совпадение итогового числа с эталоном не всегда означает корректность объяснения.

После завершения всех прогонов результаты были перенесены в электронную таблицу. На её основе рассчитаны показатели точности и проведена классификация ошибок. Таким образом, анализ включал как количественную, так и качественную часть.

4. РЕЗУЛЬТАТЫ

В итоговую статистику вошли 17 заданий для каждого режима. Полученные результаты представлены в таблице 1.

Таблица 1 — Результаты оценки точности

Режим Верные Неверные Всего Точность

DUMB 11 6 17 64,7%

NORMAL 14 3 17 82,4%

GENIUS 16 1 17 94,1%

Рисунок 1 — Сравнение точности ответов

Полученные данные показывают последовательное увеличение точности при переходе от DUMB к NORMAL и GENIUS. Точность DUMB составила 64,7%, NORMAL — 82,4%, GENIUS — 94,1%.

Таким образом, переход от DUMB к NORMAL дал увеличение точности на **17,7 процентного пункта**. Переход от DUMB к GENIUS дал увеличение на **29,4 процентного пункта**, а GENIUS превысил NORMAL на **11,7 процентного пункта**.

4.1. Сводка по отдельным заданиям

Ниже приведена сводка по 17 заданиям, вошедшим в статистику. Знак «✓» означает правильный конечный ответ, знак «✗» — неправильный.

№ Тип задания DUMB NORMAL GENIUS

1 Уравнение ✓ ✓ ✓

2 Проценты ✗ ✓ ✓

3 Геометрия ✓ ✓ ✓

4 Последовательность ✗ ✓ ✓

5 Движение ✓ ✓ ✓

7 Последовательность ✓ ✓ ✓

8 Логический вывод ✗ ✓ ✓

9 Проценты ✓ ✓ ✓

10 Комбинаторика / геометрия ✗ ✗ ✓

11 Временные интервалы ✓ ✓ ✓

12 Логическое распределение ✓ ✗ ✓

13 Последовательное изменение % ✗ ✓ ✓

14 Логический вывод ✗ ✓ ✓

15 Последовательность ✓* ✗ ✓

16 Рукопожатия ✓ ✓ ✗

17 Последовательное изменение цены ✓ ✓ ✓

18 Недостаточность данных ✓ ✓ ✓

* В задании №15 конечный ответ DUMB совпал с эталоном, однако в объяснении была обнаружена логическая ошибка. Поэтому случай учитывается как правильный по основной метрике, но отмечается при качественном анализе.

5. АНАЛИЗ ОШИБОК И ОБСУЖДЕНИЕ

В режиме DUMB было получено 6 неправильных ответов: в заданиях №2, №4, №8, №10, №13 и №14. Среди выявленных проблем встречались неправильное понимание условия, логические и вычислительные ошибки, а также неправильные выводы.

В режиме NORMAL было выявлено 3 неправильных ответа — в заданиях №10, №12 и №15. Количество ошибок уменьшилось по сравнению с DUMB.

В режиме GENIUS статистически был зафиксирован один неправильный конечный ответ — в задании №16. При этом анализ показал, что увеличение сложности рассуждения не всегда означает его безошибочность.

Важным результатом является различие между правильностью конечного ответа и качеством рассуждения. Например, в отдельных заданиях система могла прийти к правильному результату, несмотря на ошибочный или противоречивый промежуточный ход рассуждения. Поэтому одной точности недостаточно для полной характеристики качества AI-ответа.

В проведённом эксперименте повышение уровня адаптивного поведения сопровождалось увеличением доли правильных ответов. Это согласуется с выдвинутой гипотезой, однако из-за небольшого объёма выборки полученный результат следует рассматривать как результат конкретного эксперимента, а не как универсальное свойство языковых моделей.

5.1. Сводка выявленных проблем

Наиболее заметное отличие режимов проявилось в количестве ошибок. DUMB допустил ошибки в шести статистически оцениваемых заданиях, NORMAL — в трёх, GENIUS — в одном. При этом ошибки распределялись по разным причинам: от неверной интерпретации условия до вычислительных и логических ошибок.

Особенно важны случаи, когда более сложный режим формирует чрезмерно сложное рассуждение. В задании №16 GENIUS сначала получил правильное количество рукопожатий, но затем ошибочно умножил результат на 5, из-за чего конечный ответ стал неверным. Это показывает, что увеличение сложности рассуждения не должно само по себе считаться гарантией качества.

6. ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ

Результаты эксперимента могут быть использованы при дальнейшем развитии DUMB-AI. Одним из возможных направлений является создание механизма автоматического выбора режима в зависимости от типа и предполагаемой сложности запроса.

Например, для простых повседневных запросов может использоваться менее сложный режим, а для задач, требующих вычислений и последовательного

рассуждения, — более высокий режим. Это потенциально позволит одновременно учитывать качество ответа и вычислительную нагрузку системы.

В дальнейшем целесообразно увеличить количество заданий, расширить набор типов задач и использовать дополнительные показатели оценки: качество объяснения, логическую последовательность, устойчивость результата и время формирования ответа.

6.1. Связь результатов с разработкой DUMB-AI

Эксперимент имеет непосредственное значение для проектирования DUMB-AI. Если режимы отличаются по качеству решения задач, выбор режима становится не только элементом интерфейса, но и частью алгоритма работы системы.

На следующем этапе можно реализовать адаптивный маршрутизатор запросов. Он может предварительно определять тип задачи: простой информационный запрос, вычислительная задача, логическая задача или задача, требующая последовательного рассуждения. После этого система выбирает подходящий режим.

Полученные данные позволяют рассматривать GENIUS как режим с наибольшей точностью на данной выборке, а DUMB — как более простой режим, который может быть достаточен для части элементарных запросов. При этом окончательное правило выбора режима необходимо проверять на значительно большей выборке.

7. ЗАКЛЮЧЕНИЕ

В ходе исследования было изучено влияние различных режимов адаптивного поведения AI-чатбота DUMB-AI на качество решения задач. Был проведён эксперимент с 20 заданиями, из которых 17 использовались для итогового статистического сравнения.

Результаты показали рост точности от 64,7% в режиме DUMB до 82,4% в режиме NORMAL и 94,1% в режиме GENIUS. Максимальное увеличение относительно DUMB составило 29,4 процентного пункта.

Таким образом, в рамках проведённого эксперимента выдвинутая гипотеза получила подтверждение: изменение режима адаптивного поведения сопровождалось изменением качества ответов, а более высокий уровень режима характеризовался большей долей правильных решений.

Цель исследования достигнута, а поставленные задачи выполнены. Полученные результаты могут быть использованы для дальнейшего совершенствования DUMB-AI и разработки адаптивного механизма выбора режима ответа.

8. ИСТОЧНИКИ

Vaswani A. et al. *Attention Is All You Need*. 2017.

Liang P. et al. *Holistic Evaluation of Language Models (HELM)*. 2022.

Tabassi E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST, 2023.

Autio C. et al. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST, 2024.

NIST AI Resource Center. *Measure* — материалы по оценке и измерению характеристик AI-систем.

ПРИЛОЖЕНИЕ А. СТАТИСТИЧЕСКАЯ ТАБЛИЦА

Показатель DUMB NORMAL GENIUS Примечание

Правильные ответы 11 14 16 17 оцениваемых заданий

Неправильные ответы 6 3 1 17 оцениваемых заданий

Оцениваемых заданий 17 17 17 №6, №19–20 исключены

Точность 64,7% 82,4% 94,1%

Изменение относительно DUMB — +17,7 п.п. +29,4 п.п.

ПРИЛОЖЕНИЕ Б. ОСНОВНЫЕ ТИПЫ ОШИБОК

Неправильное понимание условия — система неверно интерпретирует исходные данные или требования задания.

Логическая ошибка — нарушение логической последовательности рассуждения.

Вычислительная ошибка — ошибка при выполнении арифметических или иных вычислений.

Неправильный вывод — итоговое утверждение не соответствует условию или полученным данным.

Противоречие в рассуждении — отдельные части объяснения противоречат друг другу.

Избыточное усложнение — система вводит лишние действия или преобразования, не требуемые условием.

Неполное выполнение — система выполняет только часть поставленного задания.