

УДК 004.8

Сравнительный анализ архитектур многоагентных дебатов крупных языковых моделей в задаче поддержки принятия решений

Ш. М. Магомедов¹

¹ Дагестанский государственный технический университет, Махачкала, Россия,
shakhmanay.magomedov@mail.ru

Аннотация. В статье представлен сравнительный анализ пяти архитектур многоагентных дебатов крупных языковых моделей в задаче поддержки принятия решений при оценке качества образовательных тестов. Цель исследования состояла в выявлении зависимости между топологией взаимодействия агентов и качеством решения, вычислительными затратами, временем ответа и устойчивостью результата, а также в определении условий, при которых применение многоагентных дебатов в системах поддержки принятия решений (СППР) является практически оправданным. На едином прикладном полигоне сопоставлялись базовая, линейная, циклическая, перекрестная архитектуры и архитектура с исключением модели. Корпус включал 60 тестов по четырем дисциплинам основного общего образования с разделением на корректные тесты и тесты с внесенными ошибками; ансамбль моделей был фиксирован и включал Qwen 2.5 72B Instruct, Grok 4.1 Fast и Gemma 4 26B-A4B IT. Для оценки использовались macro-F1, макро-средняя абсолютная ошибка, число токенов, стоимость на один тест и средняя длительность ответа. Установлено, что наибольшее наблюдаемое значение macro-F1 в рамках проведенного эксперимента показала архитектура с исключением модели, тогда как базовая архитектура оказалась лучшей по макро-средней абсолютной ошибке и сохранила сильные позиции по операционной эффективности. Дополнительно выявлено, что часть архитектур значительно чувствительна к типу теста, а увеличение числа раундов циклических дебатов после определенного порога не обеспечивает устойчивого прироста качества. Полученные результаты позволяют рассматривать архитектуру дебатов как управляемый параметр СППР, а не как нейтральную техническую деталь.

Ключевые слова: крупные языковые модели, многоагентные дебаты, поддержка принятия решений, образовательные тесты, языковая модель-судья, macro-F1, архитектура взаимодействия

1. Введение

Многоагентные дебаты крупных языковых моделей рассматриваются сегодня как один из путей повышения надежности рассуждений и оценок по сравнению с одиночным выводом модели [1, 5]. Работы по многоагентным дебатам показали, что коллективное обсуждение может улучшать фактическую корректность и качество рассуждений, однако более поздние исследования продемонстрировали, что результат чувствителен к самой схеме коммуникации между агентами и не определяется только их количеством [1, 2]. Дополнительную сложность вносит использование LLM в роли судьи: такие схемы масштабируемы, но потенциально уязвимы к систематическим смещениям, что особенно важно для задач поддержки решений [3, 4].

В образовательной диагностике это ограничение приобретает прикладной характер. Качество тестового задания определяется не только фактом наличия ошибок, но

и соответствием дисциплине, уровню обучающихся, корректностью эталонных ответов, ясностью формулировок и полнотой тематического покрытия [7–10]. Поэтому оценка таких объектов требует сопоставления архитектур дебатов на материале, где возможна эталонная разметка и где цена ложноположительных и ложноотрицательных решений различна. Образовательный контекст также уже используется в современных исследованиях LLM как самостоятельный класс многоагентных сценариев, что подтверждает методологическую уместность выбранного полигона [6].

Цель статьи – выполнить сравнительный анализ архитектур многоагентных дебатов крупных языковых моделей в задаче поддержки принятия решений при оценке качества образовательных тестов и определить, какие схемы взаимодействия дают более удачный баланс между качеством, затратами и устойчивостью.

В проведенном исследовании на едином

корпусе сопоставлялись пять архитектур: базовая, линейная, циклическая, перекрестная и архитектура с исключением модели. Корпус был организован как восемь файлов формата JSON по четырем дисциплинам основного общего образования Российской Федерации – история России, математика, информатика и русский язык – с разделением на корректные тесты и тесты с внесенными ошибками. Каждый тест содержал десять вопросов с четырьмя вариантами ответа, что дало 60 тестов и 600 заданий в сумме. Итоговая числовая оценка качества формировалась по пяти критериям C1–C5: соответствие дисциплине, соответствие уровню обучающихся, корректность эталонных ответов, ясность формулировок и полнота тематического покрытия; сами критерии и их веса задавались заранее на нормативной и методической основе [9, 10]. Фиксированный ансамбль включал Qwen 2.5 72B Instruct, Grok 4.1 Fast и Gemma 4 26B-A4B IT, а основными метриками выступали macro-F1 для классификации проблемных вопросов, макро-средняя абсолютная ошибка для числовой оценки качества теста, число токенов, стоимость на один тест и длительность ответа.

Тестовый корпус был сформирован автором как контролируемая экспериментальная выборка на основе содержания программ основного общего образования Российской Федерации для 7–9-х классов. Стандартные экзаменационные банки ОГЭ, ЕГЭ или международные наборы не использовались напрямую, поскольку для данной задачи требовалась парная структура «корректная версия / версия с внесенными ошибками», заранее известные проблемные вопросы и эталонные оценки по критериям C1–C5; такие поля в открытых экзаменационных наборах обычно отсутствуют. Дисциплины были

выбраны так, чтобы совместить формализованные задания (математика, информатика) и задания с более выраженным контекстным анализом (история России, русский язык). В качестве типовых объектов корпуса использовались, например, задания на вычисление или выбор корректного алгоритмического утверждения, а также вопросы на понимание исторического факта или языковой формулировки; в версии с внесенными ошибками отдельно закладывались неверные эталонные ответы, несоответствие уровню класса, неоднозначная формулировка и выход за пределы заявленной дисциплины. Доступ к моделям осуществлялся через API-агрегатор OpenRouter; все вызовы выполнялись с фиксированными параметрами генерации и сохранялись в структурированных JSON-логах.

Такой дизайн позволяет сравнивать не только точность, но и операционные свойства разных топологий дебатов на одном прикладном полигоне. Тем самым исследование направлено не на подтверждение общей полезности многоагентных дебатов, а на определение условий их рационального применения в СППР.

2. Результаты исследования

Первый этап анализа связан с проверкой основного предположения исследования: действительно ли усложнение архитектуры дебатов приводит к более качественному решению. Для этого архитектуры сопоставлялись не только по значению macro-F1, но и по макро-средней абсолютной ошибке, числу токенов, стоимости и времени ответа. В таблице 1 приведены сводные результаты сравнения архитектур многоагентных дебатов; для сопоставимости стоимостные показатели нормированы к одному тесту.

Таблица 1. Сводные показатели архитектур многоагентных дебатов

Архитектура	Macro-F1	Макро-средняя абсолютная ошибка	Токенов / тест	Стоимость / тест, USD	Длительность, с
Архитектура с исключением модели	0,772	0,099	24 155	0,0071	188,5
Базовая архитектура	0,701	0,052	15 208	0,0043	92,5
Циклическая архитектура	0,686	0,088	20 252	0,0054	140,7
Перекрестная архитектура	0,616	0,095	21 126	0,0058	165,0
Линейная архитектура	0,566	0,102	9 643	0,0027	94,0

Как видно из таблицы 1, наибольшее наблюдаемое значение макро-F1 получила архитектура с исключением модели – 0,772. Базовая архитектура показала 0,701, циклическая – 0,686, перекрестная – 0,616, линейная – 0,566. При этом по макро-средней абсолютной ошибке лучшей оказалась базовая архитектура – 0,052, тогда как у архитектуры с исключением модели эта метрика составила 0,099. Линейная архитектура потребовала наименьшего количества ресурсов – 9643 токена и 0,0027 USD на тест, – но уступила остальным по качеству решения. Следовательно, уже на уровне сводного сравнения видно, что архитектура дебатов влияет не только на точность, но и на тип компромисса между качеством и затратами. При этом различия

следует интерпретировать аккуратно: по данным попарных сравнений с критерием Вилкоксона и поправкой Холма статистически устойчивым в этой серии осталось только отличие между линейной архитектурой и архитектурой с исключением модели. Поэтому окончательное ранжирование остальных схем в рамках текущего корпуса носит прежде всего описательный характер.

Для оценки прикладной рациональности архитектур далее рассмотрим соотношение качества и стоимости. На рисунке 1 показано положение архитектур в пространстве «макро-F1 – стоимость на один тест».

Как видно из рисунка 1, на Парето-фронт попадают линейная, базовая архитектура и архитектура с исключением модели.

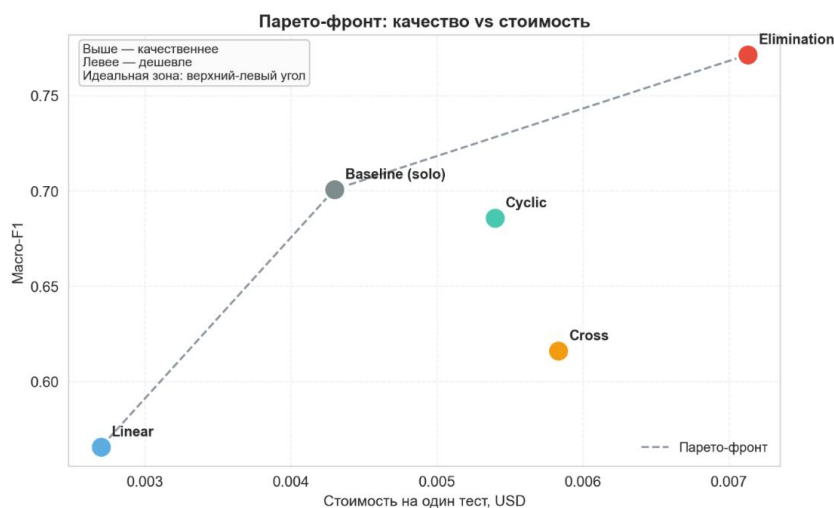


Рис. 1. Парето-сравнение архитектур в пространстве «макро-F1 – стоимость на тест»

Это означает, что при разных бюджетных режимах рациональными оказываются разные решения: линейная архитектура – при минимальных расходах, базовая – как умеренный и устойчивый компромисс, архитектура с исключением модели – при приоритете качества. Циклическая и перекрестная архитектуры в текущем эксперименте остаются вне Парето-фронта, поскольку требуют больше ресурсов без соответствующего выигрыша по макро-F1.

Практический вывод из этого состоит в том, что рост сложности взаимодействия между агентами не тождествен росту полезности архитектуры для СППР. Иными словами, более «плотная» коммуникация между моделями еще не означает более выгодное решение для реальной системы.

В таблице 2 приведены значения макро-F1 по двум срезам корпуса – корректные и с внесенными ошибками.

Таблица 2. Значения макро-F1 по типам тестов

Архитектура	Корректные	С внесенными ошибками	Δ , п.п.
Базовая архитектура	0,689	0,713	+2,4
Линейная архитектура	0,433	0,698	+26,5
Циклическая архитектура	0,567	0,805	+23,9
Перекрестная архитектура	0,433	0,799	+36,6
Архитектура с исключением модели	0,767	0,776	+1,0

Этот разрез важен потому, что одна и та же средняя метрика может скрывать различную чувствительность архитектур к ложным тревогам и пропуску реальных дефектов.

Как видно из таблицы 2, наиболее ровное поведение демонстрирует архитектура с исключением модели: 0,767 на корректных тестах и 0,776 на тестах с внесенными ошибками. Базовая архитектура также остается сравнительно устойчивой: 0,689 и 0,713 соответственно. Линейная, циклическая и перекрестная архитектуры ведут себя иначе: они заметно лучше работают на тестах с внесенными ошибками, чем на корректных тестах. Наиболее выраженная асимметрия наблюдается у перекрестной архитектуры – 0,433 на корректных против 0,799 на тестах с внесенными ошибками, то есть разрыв составляет 36,6 процентного пункта. Это позволяет заключить, что часть дебатных схем

подходит скорее для режима активного поиска дефектов, чем для режима осторожной проверки уже подготовленного банка заданий, где цена ложноположительного решения выше. Для практической СППР это критично: одна и та же архитектура может быть полезной в режиме «найти проблему любой ценой» и одновременно неудачной в режиме «не тревожить эксперта без достаточных оснований».

Отдельно был проверен вопрос о том, повышает ли увеличение числа раундов качество дебатов. На рисунке 2 изображена динамика macro-F1 и стоимости для циклической архитектуры при изменении числа раундов от 1 до 5.

Как видно из рис. 2, увеличение числа раундов не приводит к монотонному росту качества.

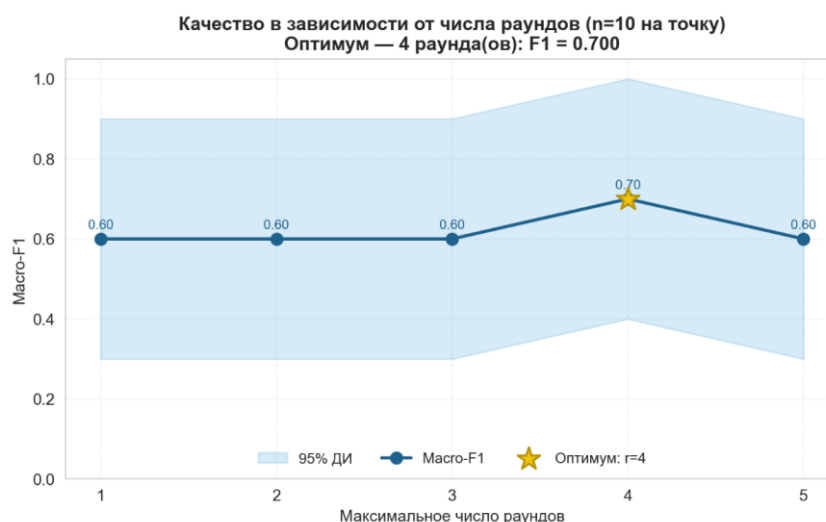


Рис. 2. Влияние числа раундов в циклической архитектуре на macro-F1 и стоимость на тест

В проведенном эксперименте значения macro-F1 для циклической архитектуры составляли 0,600 при 1–3 раундах, возрастали до 0,700 на 4 раундах и вновь снижались до 0,600 на 5 раундах; одновременно продолжали расти стоимость и время ответа. Стоимость увеличивалась от 0,0485 до 0,0847 USD на тест, а средняя длительность ответа – от 158,5 до 301,8 секунды. Следовательно, число раундов следует рассматривать как ограничиваемый параметр, а не как величину, которую нужно автоматически увеличивать. Для СППР это означает, что дополнительные итерации оправданы только там, где потенциальный прирост качества действительно перекрывает рост затрат и времени ответа.

3. Заключение

В рамках проведенного исследования установлено, что архитектура взаимодействия между агентами является существенным фактором качества решения в многоагентных дебатах крупных языковых моделей. Наибольшее наблюдаемое значение macro-F1 показала архитектура с исключением модели, однако базовая архитектура сохранила преимущество по макро-средней абсолютной ошибке и оказалась сильным компромиссным вариантом при учете стоимости и времени ответа. Следовательно, многоагентные дебаты нельзя рассматривать как гарантированный способ улучшения ответа. Их эффективность зависит от топологии

взаимодействия, типа проверяемого материала, допустимых затрат и цены ошибки. Тем самым достигнута цель исследования: определены не только различия между архитектурами, но и условия, при которых их применение в СППР является рациональным.

Практическая применимость результатов состоит в возможности выбора архитектуры под конкретный сценарий СППР. Если приоритетом является максимизация качества и допустимы дополнительные затраты, наиболее оправданной выглядит архитектура с исключением модели. Если критичны стоимость и скорость ответа, сильным рабочим решением остается базовая архитектура. В режиме активного поиска дефектов допустимы более чувствительные схемы, поскольку в этом случае важнее не пропустить реальную ошибку. В режиме осторожной проверки готового банка заданий, напротив, следует избегать архитектур с выраженной асимметрией между корректными тестами и тестами с внесенными ошибками, так как они могут создавать лишние ложноположительные решения. Следовательно, дебаты целесообразно применять в спорных и ответственных случаях: при высокой цене

пропуска ошибки, необходимости дополнительной проверки сомнительных заданий и наличии бюджета на более длительную обработку. От сложных дебатных схем следует отказываться при массовой первичной проверке, жестких ограничениях по времени ответа, низкой цене ошибки и отсутствии значимого выигрыша по качеству.

Ограничения исследования связаны с одной предметной областью, фиксированным ансамблем из трех моделей и ограниченной статистической мощностью части сравнений. Кроме того, в тех схемах, где суждения одних моделей воздействуют на итог других, нельзя игнорировать известные риски смещений, возникающих при использовании языковой модели в роли судьи [3, 4]. Поэтому полученные выводы не следует автоматически переносить на иные области, языки или составы ансамблей. Перспективными направлениями дальнейшей работы являются расширение корпуса, внешняя экспертная валидация, отдельные эксперименты с парными и более крупными ансамблями, а также проверка архитектур на других задачах поддержки принятия решений.

Comparative analysis of multi-agent debate architectures of large language models in decision support

Sh. M. Magomedov

Abstract. The paper presents a comparative analysis of five multi-agent debate architectures of large language models in a decision-support task related to evaluating the quality of educational tests. The study aims to identify how communication topology affects decision quality, computational cost, latency, and robustness. A unified experimental setting was used to compare a baseline, linear, cyclic, cross, and model-elimination architecture. The corpus comprised 60 tests from four school disciplines split into correct and errors subsets, while the model ensemble included Qwen 2.5 72B Instruct, Grok 4.1 Fast, and Gemma 4 26B-A4B IT. The main evaluation metrics were macro-F1 for detecting problematic questions, macro-MAE for the numerical assessment of test quality, and token usage, cost per test, and response time for operational comparison. The results show that the model-elimination architecture achieved the highest observed macro-F1 within the experiment, while the baseline architecture remained superior in macro-MAE and cost-efficiency. The study also demonstrates that some debate architectures are highly sensitive to the type of test and that increasing the number of rounds does not guarantee sustained quality improvement. The findings support treating debate architecture as a controllable design parameter of decision-support systems rather than as a neutral implementation detail.

Keywords: large language models, multi-agent debate, decision support, educational tests, LLM-as-a-Judge, macro-F1, interaction architecture

Литература

1. Du Y., Li S., Torralba A., Tenenbaum J. B., Mordatch I. Improving factuality and reasoning in language models through multiagent debate [Электронный ресурс] // arXiv.org. – 2023. – arXiv: 2305.14325. – URL: <https://arxiv.org/abs/2305.14325> (дата обращения: 20.02.2026). – Текст: электронный.
2. Li Y., Du Y., Zhang J. [et al.]. Improving multi-agent debate with sparse communication topology // Findings of the Association for Computational Linguistics: EMNLP 2024. – 2024. – P. 7281–7294. – URL: <https://arxiv.org/abs/2406.11776> (дата обращения: 30.06.2026).
3. Zheng L., Chiang W.-L., Sheng Y. [et al.]. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena [Электронный ресурс] // arXiv.org. – 2023. – arXiv: 2306.05685. – URL: <https://arxiv.org/abs/2306.05685> (дата обращения: 22.02.2026). – Текст: электронный.
4. Zhu L., Wang X., Wang X. JudgeLM: fine-tuned large language models are scalable judges [Электронный ресурс] // arXiv.org. – 2023. – arXiv: 2310.17631. – URL: <https://arxiv.org/abs/2310.17631> (дата обращения: 25.02.2026). – Текст: электронный.
5. Guo T., Chen X., Wang Y. [et al.]. Large language model based multi-agents: a survey of progress and challenges // Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. – 2024. – URL: <https://arxiv.org/pdf/2402.01680> (дата обращения: 30.06.2026).
6. Shi Y., Liang R., Xu Y. EducationQ: evaluating LLMs' teaching capabilities through multi-agent dialogue framework // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. – 2025. – P. 32799–32828. – URL: <https://arxiv.org/html/2504.14928> (дата обращения: 30.06.2026).
7. Аванесов В. С. Форма тестовых заданий: учебное пособие для учителей школ, лицеев, преподавателей вузов и колледжей. – 2-е изд., перераб. и расшир. – Москва: Центр тестирования, 2005. – 156 с.
8. Чельшкова М. Б. Теория и практика конструирования педагогических тестов. – Москва: Логос, 2002. – 432 с.
9. Приказ Министерства просвещения Российской Федерации от 31.05.2021 № 287 «Об утверждении федерального государственного образовательного стандарта основного общего образования» [Электронный ресурс]. URL: <https://publication.pravo.gov.ru/Document/View/0001202107050027> (дата обращения: 13.03.2026). – Текст: электронный.
10. Федеральный институт педагогических измерений. Аналитические и методические материалы по разработке и оцениванию контрольных измерительных материалов [Электронный ресурс]. – URL: <https://fipi.ru/ege/analiticheskie-i-metodicheskie-materialy> (дата обращения: 13.03.2026). – Текст: электронный.

Поступила в редакцию / Received: 15.06.2026.

Поступила после рецензирования / Revised: 05.06.2026.

Принята к печати / Accepted: 06.07.2026.