Aria

Andrew Ustinov

October 2024

Аннотация

В данной работе представлен подробный анализ архитектуры и функциональных возможностей мультимодальной языковой модели Aria, разработанной компанией Rhythmes-AI. Aria объединяет обработку текста, изображений и видео в единой архитектуре на основе Mixture of Experts (MoE), обеспечивая высокую производительность и эффективность. Рассмотрены ключевые компоненты модели, включая языковую модель, визуальный энкодер и мультимодальный проектор, а также особенности обработки мультимодальных данных. Проведен сравнительный анализ Aria с конкурентными моделями DeepSeek-V2 и Pixtral 12B, продемонстрированы преимущества и ограничения Aria. Кроме того, обсуждены бизнес-кейсы применения модели в различных отраслях, таких как автоматизация обслуживания клиентов, анализ контента и образовательные технологии. В заключение рассматриваются перспективы дальнейших разработок и влияние Aria на развитие мультимодальных систем искусственного интеллекта.

1 Архитектура и особенности Aria

Aria представляет собой LLM, объединяющую обработку текста, изображений и видео в единой архитектуре. Модель демонстрирует высокую производительность в широком спектре мультимодальных, языковых и кодовых задач.

1.1 Ключевые компоненты

- **Языковая модель**: Основана на архитектуре Mixture of Experts (MoE) с 3.9B активируемых параметров на токен для текстовых данных и 3.5B для визуальных. Общее количество параметров достигает 24.9B.
- **Визуальный энкодер**: Легковесный энкодер на основе Vision Transformer (ViT) с 438M параметров для обработки изображений и видеокадров.
- **Мультимодальный проектор**: Модуль для объединения визуальных и текстовых признаков, основанный на механизме внимания.
- **Контекстное окно**: Поддерживает до 64K токенов, что позволяет обрабатывать длинные последовательности мультимодальных данных.

1.2 Обработка мультимодальных данных

Процесс обработки визуальных и текстовых данных в Aria происходит следующим образом:

1. **Предобработка изображений**: Входные изображения или видеокадры масштабируются до трех возможных разрешений:
   - Среднее разрешение: длинная сторона 490 пикселей
   - Высокое разрешение: длинная сторона 980 пикселей
   - Сверхвысокое разрешение: изображение динамически разбивается на несколько изображений высокого разрешения
2. **Визуальное кодирование**: Подготовленные изображения обрабатываются ViT, который преобразует их в последовательность визуальных токенов.
3. **Проекция**: Мультимодальный проектор объединяет визуальные токены с текстовыми эмбеддингами, используя механизм кросс-внимания и полносвязный слой.
4. **Языковая обработка**: Объединенное представление обрабатывается основной языковой моделью MoE, которая генерирует выходные данные.

1.3 Особенности архитектуры MoE

Модель Aria использует усовершенствованную архитектуру MoE:

- **Мелкозернистые эксперты**: 66 экспертов в каждом слое MoE, из которых 2 являются общими для всех входных данных.
- **Динамическая маршрутизация**: Для каждого токена активируются 6 специализированных экспертов, выбранных маршрутизатором.
- **Эффективное использование параметров**: Несмотря на общее количество 24.9B параметров, для обработки каждого токена используется только 3.9B или 3.5B параметров.
- **Балансировка нагрузки**: Применяются специальные техники для предотвращения коллапса маршрутизации и обеспечения равномерного использования экспертов.

Такая архитектура позволяет Aria эффективно обрабатывать разнородные мультимодальные данные, адаптируясь к специфике каждого входного токена.

2 Преимущества Aria

Мультимодальная модель Aria обладает рядом существенных преимуществ, которые выделяют ее среди конкурентов и открывают новые возможности в области искусственного интеллекта.

2.1 Архитектурные преимущества

- **Эффективная архитектура MoE**: Использование мелкозернистой архитектуры Mixture of Experts позволяет модели адаптироваться к различным типам входных данных, активируя только релевантных экспертов.
- **Оптимизированный визуальный энкодер**: Легковесный ViT-энкодер (438M параметров) обрабатывает изображения и видео без значительного увеличения общего размера модели.
- **Гибкая обработка изображений**: Поддержка различных разрешений и динамическое разбиение изображений со сверхвысоким разрешеним позволяет работать с широким спектром визуальных данных.
- **Большое контекстное окно**: Способность обрабатывать до 64K токенов позволяет Aria работать с длинными последовательностями мультимодальных данных, что критично для задач с видео и длинными документами.

2.2 Производительность и эффективность

- **Высокая эффективность параметров**: Несмотря на общее количество 24.9B параметров, модель активирует только 3.9B или 3.5B параметров на токен, что сильно снижает вычислительные затраты.
- **Конкурентоспособность с более крупными моделями**: Aria демонстрирует производительность на уровне с моделями с большим количеством параметров, таких как GPT-4 и Gemini-1.5, на различных мультимодальных задачах.

2.3 Мультимодальные возможности

- **Нативная мультимодальность**: Aria изначально спроектирована для работы с текстом, изображениями и видео, что обеспечивает глубокую интеграцию различных модальностей.
- **Превосходная производительность в видеозадачах**: Модель демонстрирует высокие результаты в задачах понимания длинных видео. У большинства проприетарных моделей нет возможности обрабатывать видео.
- **Гибкость в обработке визуальных данных**: Способность работать с изображениями различных размеров и форматов, а также с видеопоследовательностями различной длины.

2.4 Гибкость и масштабируемость

- **Адаптивность к различным задачам**: Благодаря архитектуре MoE, Aria может эффективно адаптироваться к широкому спектру задач без необходимости полного переобучения.
- **Возможности тонкой настройки**: Поддержка методов адаптации, например LoRA, позволяет быстро настраивать модель под специфические задачи.
- **Масштабируемость**: Архитектура Aria позволяет относительно легко масштабировать модель, увеличивая количество экспертов или размер базовой модели, без пропорционального роста вычислительных требований при инференсе.

3 Сравнение с конкурентами

3.1 Aria vs DeepSeek-V2

3.1.1 Архитектура и размер модели

Метрика	Aria	DeepSeek-V2
Общее количество параметров	25.3B	236B
Активируемые параметры	3.9B (виз.), 3.5B (текст.)	21B
Количество слоев	28	60
Размер скрытого состояния	2560	5120
Количество голов внимания	16	128
Размер головы внимания	128	128

Таблица 1: Сравнение архитектурных особенностей Aria и DeepSeek-V2

DeepSeek-V2 значительно превосходит Aria по общему количеству параметров, но Aria демонстрирует более эффективное использование параметров, активируя меньшее их количество на токен. Это может указывать на лучшую оптимизацию архитектуры Aria.

3.1.2 Контекстное окно и обучающие данные

Метрика	Aria	DeepSeek-V2
Длина контекста	64K токенов	128K токенов
Объем предобучающих данных	6.4T язык., 400B мультимод.	8.1T токенов

Таблица 2: Сравнение контекстного окна и объема обучающих данных

DeepSeek-V2 имеет преимущество в длине контекста и общем объеме обучающих данных, что потенциально может обеспечить лучшее понимание длинных последовательностей и более широкий охват знаний.

3.1.3 Производительность на бенчмарках

Бенчмарк	Aria	DeepSeek-V2
MMLU (5-shot)	73.3	78.5
TriviaQA (5-shot)	86.7	79.9
NaturalQuestions (5-shot)	53.4	38.7
ARC-Challenge (25-shot)	92.3	92.4
GSM8K (8-shot)	92.2	79.2
MATH (4-shot)	53.9	43.6
HumanEval (0-shot)	81.1	48.8
C-Eval (5-shot)	78.0	81.7
CMMLU (5-shot)	81.6	84.0

Таблица 3: Сравнение результатов на различных бенчмарках

Результаты варьируются в зависимости от задачи. DeepSeek-V2 показывает лучшие результаты на MMLU, C-Eval и CMMLU, но Aria превосходит в TriviaQA, NaturalQuestions, GSM8K, MATH и HumanEval. Сильная сторона Aria - задачи, требующие глубокого понимания и рассуждения.

3.2 Aria vs Pixtral 12B

3.2.1 Архитектура и размер модели

Метрика	Aria	Pixtral 12B
Общее количество параметров	24.9B	12B
Активируемые параметры	3.9B (виз.), 3.5B (текст.)	Не указано
Количество слоев	28	Декодер: 40, Энкодер: 24
Размер скрытого состояния	2560	Декодер: 5120, Энкодер: 1024
Количество голов внимания	66 экспертов	Декодер: 32, Энкодер: 16
Размер головы внимания	1664 (FFN эксперта)	Декодер: 128, Энкодер: 64

Таблица 4: Сравнение архитектурных особенностей Aria и Pixtral 12B

Aria имеет больше параметров, но использует архитектуру Mixture-of-Experts (MoE), активируя только часть параметров для каждого входа. Pixtral 12B использует стандартную трансформерную архитектуру с более глубоким декодером.

3.2.2 Контекстное окно и обучающие данные

Метрика	Aria	Pixtral 12B
Длина контекста	64K токенов	128K токенов
Объем предобучающих данных	6.4T язык., 400B мультимод.	400B мультимод.

Таблица 5: Сравнение контекстного окна и объема обучающих данных

Pixtral 12B предлагает вдвое большую длину контекста, что может быть преимуществом при обработке длинных документов или сложных мультимодальных входных данных.

3.2.3 Производительность на бенчмарках

Бенчмарк	Aria	Pixtral 12B
MMLU	73.3	69.2
TriviaQA	86.7	Не указано
GSM8K	92.2	48.1
MATH	53.9	48.1
HumanEval	81.1	72.0

Таблица 6: Сравнение результатов на различных бенчмарках

Aria демонстрирует превосходство над Pixtral 12B на всех доступных для сравнения бенчмарках, особенно в задачах, требующих рассуждений и решения проблем (GSM8K, MATH, HumanEval).

3.2.4 Мультимодальные возможности

Обе модели позиционируются как мультимодальные, способные обрабатывать текст и изображения. Однако Aria также заявляет о способности работать с видео и документами, что расширяет ее потенциальные области применения.

3.3 Общий вывод по трем моделям

Сравнивая Aria, DeepSeek-V2 и Pixtral 12B, мы можем сделать следующие выводы:

1. **Размер и эффективность**: DeepSeek-V2 является самой крупной моделью (236B параметров), за ней следует Aria (24.9B), и Pixtral 12B (12B). Но Aria, благодаря архитектуре MoE, демонстрирует наиболее эффективное использование параметров, активируя лишь небольшую часть для каждого входа.

2. **Контекстное окно**: DeepSeek-V2 и Pixtral 12B предлагают большее контекстное окно (128K токенов) по сравнению с Aria (64K токенов), что может быть преимуществом при работе с длинными последовательностями.

3. **Производительность**: Aria показывает превосходные результаты на многих бенчмарках, особенно в задачах, требующих рассуждений и решения проблем. DeepSeek-V2 лидирует в некоторых языковых задачах, особенно связанных с китайским языком. Pixtral 12B, несмотря на меньший размер, демонстрирует конкурентоспособные результаты.

4. **Архитектурные инновации**: Aria выделяется использованием архитектуры MoE, что потенциально обеспечивает более эффективное использование вычислительных ресурсов. DeepSeek-V2 и Pixtral 12B используют более традиционные трансформерные архитектуры, но с различными оптимизациями.

Каждая из этих моделей имеет свои сильные стороны:

- **Aria** выделяется эффективностью использования параметров, превосходными результатами в задачах рассуждения и расширенными мультимодальными возможностями.
- **DeepSeek-V2** предлагает наибольшую мощность благодаря своему размеру и показывает отличные результаты в языковых задачах, особенно на китайском языке.
- **Pixtral 12B** демонстрирует, что можно достичь конкурентоспособной производительности с меньшим количеством параметров, предлагая хороший баланс между эффективностью и возможностями.

3.4 Сравнение с другими моделями

Помимо DeepSeek-V2 и Pixtral 12B, модель Aria была сравнена с рядом других современных языковых и мультимодальных моделей. Рассмотрим результаты этих сравнений на основе данных из таблиц 1, 3 и 4.

3.4.1 Производительность на мультимодальных и языковых бенчмарках

Бенчмарк	**Aria**	**Llama3.2-11B**	**GPT-4V**	**GPT-4o**	**Gemini-1.5 Flash**	**Gemini-1.5 Pro**
MMMU (val)	54.9	50.7	56.4	59.4	56.1	62.2
MathVista (testmini)	66.1	51.5	-	54.7	58.4	63.9
DocVQA (test)	92.6	88.4	88.4	-	89.9	93.1
MMBench-1.1	80.3	-	79.8	76.0	-	73.9
LongVideoBench (test)	65.3	45.7	60.7	58.8	62.4	64.4
VideoMME (w subs)	72.1	50.2	63.3	68.9	75.0	81.3
MMLU (5-shot)	73.3	69.4	86.4	-	78.9	85.9
HumanEval	73.2	72.6	67.0	87.2	74.3	84.1

Таблица 7: Сравнение результатов на различных мультимодальных и языковых бенчмарках

Анализ результатов показывает:

- Aria превосходит Llama3.2-11B почти на всех бенчмарках, демонстрируя значительное улучшение в мультимодальных задачах.
- В сравнении с GPT-4V и GPT-4o, Aria показывает конкурентоспособные результаты, превосходя их в некоторых задачах (например, DocVQA, MMBench-1.1).
- Gemini-1.5 Pro демонстрирует лучшие результаты на многих бенчмарках, но Aria конкурентоспособна и даже превосходит его в некоторых задачах (например, MMBench-1.1).
- Aria особенно сильна в задачах, связанных с обработкой документов (DocVQA) и пониманием видео (LongVideoBench).

3.4.2 Производительность на задачах с длинным контекстом

Модель	**Параметры**	**LongVideoBench**	**VideoMME**	**MMLongBench-Doc**
Aria	3.9B (25.3B)	65.3	72.1	28.3
Qwen2-VL-7B	7B	56.8	69.0	21.3
InternVL2-40B	40B	60.6	62.4	18.2
Gemini-1.5-Pro	-	64.4	81.3	28.2
GPT-4o	-	66.7	77.2	42.9

Таблица 8: Сравнение результатов на задачах с длинным контекстом

Анализ результатов:

- Aria демонстрирует отличные результаты в задачах с длинным контекстом, особенно учитывая ее меньший размер по сравнению с некоторыми конкурентами.
- На LongVideoBench Aria превосходит большинство моделей, уступая лишь немного GPT-4o.
- В задаче VideoMME Aria показывает конкурентоспособные результаты, хотя и уступает Gemini-1.5-Pro и GPT-4o.
- На MMLongBench-Doc Aria демонстрирует результаты на уровне Gemini-1.5-Pro, но уступает GPT-4o.

3.4.3 Оценка способности следовать инструкциям

Модель	MIA-Bench (Multimodal)	MT-Bench (Language)
Aria	8.76	8.53
Phi-3 Vision	7.60	6.27
Qwen2-VL-7B	8.07	6.41
Pixtral-12B	8.43	7.68
GPT-4o	8.86	-

Таблица 9: Оценка способности следовать инструкциям

Анализ результатов:

- Aria демонстрирует превосходные результаты в способности следовать инструкциям как в мультимодальных (MIA-Bench), так и в чисто языковых (MT-Bench) задачах.
- На MIA-Bench Aria уступает только GPT-4o, превосходя все остальные модели.
- На MT-Bench Aria показывает лучший результат среди представленных моделей, что указывает на ее сильные языковые способности.

3.4.4 Общий вывод

На основе представленных данных можно сделать следующие выводы:

1. **Мультимодальные возможности**: Aria демонстрирует исключительные результаты в мультимодальных задачах, часто превосходя более крупные модели и приближаясь к производительности ведущих проприетарных моделей.
2. **Эффективность архитектуры**: Несмотря на меньшее количество активируемых параметров (3.9B из 25.3B), Aria показывает конкурентоспособные или превосходящие результаты по сравнению с моделями, имеющими значительно больше параметров.
3. **Обработка длинного контекста**: Модель хорошо справляется с задачами, требующими понимания длинного контекста, особенно в видео и документах, что подтверждает эффективность ее 64K контекстного окна.
4. **Языковые способности**: Высокие результаты на MT-Bench и других языковых задачах указывают на сильные языковые возможности Aria, несмотря на ее мультимодальную природу.
5. **Следование инструкциям**: Aria демонстрирует отличные способности в следовании сложным инструкциям как в мультимодальных, так и в чисто текстовых задачах.
6. **Конкурентоспособность**: Хотя некоторые проприетарные модели (например, GPT-4o и Gemini-1.5 Pro) показывают лучшие результаты в определенных задачах, Aria остается высококонкурентной, особенно учитывая ее открытость и эффективность.

4 Недостатки и ограничения Aria

4.1 Архитектурные ограничения

- **Сложность архитектуры MoE**: Использование архитектуры Mixture of Experts может усложнить процесс настройки и оптимизации модели, особенно для исследователей и разработчиков, не имеющих опыта работы с MoE.

- **Ограниченное контекстное окно**: 64K токенов меньше, чем у некоторых конкурентов (например, 128K у DeepSeek-V2 и Pixtral 12B), что может ограничивать производительность на сверхдлинных последовательностях.
- **Эффективность использования параметров**: Aria активирует меньше параметров (3.9B для визуальных и 3.5B для текстовых токенов), это больше, чем у DeepSeek-V2 (21B активированных параметров) в относительных величинах.

4.2 Вычислительные и технические ограничения

- **Вычислительные требования**: Несмотря на эффективность MoE, общий размер модели в 24.9B параметров может требовать больших вычислительных ресурсов для обучения и инференса.
- **Потенциальная несбалансированность экспертов**: В архитектурах MoE существует риск недоиспользования некоторых экспертов, что может привести к неэффективному использованию модели.
- **Сложности с распараллеливанием**: Архитектура MoE может создавать трудности при распараллеливании вычислений, особенно в распределенных системах.
- **Отсутствие информации о KV-кеше**: В отличие от DeepSeek-V2, который значительно уменьшил размер KV-кеша, для Aria нет подробной информации об оптимизации KV-кеша, что затрудняет сравнение эффективности инференса.

5 Потенциальные недостатки модели Aria

На основе предоставленных данных можно выделить следующие возможные недостатки модели Aria:

- **Большее количество активируемых параметров по сравнению с другими моделями**: Несмотря на использование архитектуры Mixture-of-Experts (MoE) для повышения эффективности, Aria активирует больше параметров, чем некоторые другие MoE модели, такие как DeepSeek-V2. DeepSeek-V2 активирует 21B параметров на токен, тогда как у Aria активируются 3.9B параметров для визуальных токенов и 3.5B для текстовых токенов. Это может негативно сказываться на эффективности обучения и инференса модели, так как модели с меньшим количеством активируемых параметров обычно требуют меньше вычислительных ресурсов, что приводит к более быстрой тренировке, снижению затрат, уменьшению объема памяти и ускорению инференса.
- **Ограниченная информация о специфике обработки изображений**: Модель Aria обрабатывает изображения различных разрешений с использованием разных режимов разрешения и динамического разбиения, отсутствуют детальные сведения о конкретных техниках обработки изображений. В отличие от Pixtral, который использует ROPE-2D для обработки изображений в их исходном разрешении и с сохранением соотношения сторон, у Aria нет подробных данных о подобных методах. Это затрудняет оценку гибкости и эффективности обработки изображений разного размера и пропорций по сравнению с моделями, использующими такие техники, как ROPE-2D.
- **Отсутствие прозрачности в маршрутизации экспертов**: Несмотря на упоминание специализации экспертов в Aria в зависимости от модальности, отсутствует подробное описание механизма маршрутизации экспертов в MoE декодере. Неясно, использует ли модель статическую или динамическую маршрутизацию и применяются ли техники балансировки нагрузки для предотвращения перегрузки отдельных экспертов. Отсутствие этих данных затрудняет полное понимание внутренней работы архитектуры MoE в Aria. .

6 Влияние Rythmes-AI/Aria на сферу ИИ

6.1 Развитие мультимодальных систем ИИ

Aria является первой открытой мультимодальной моделью с архитектурой Mixture of Experts (MoE). Это стимулирует интерес к области интеграции различных типов данных — текста, изображений, видео и кода — в единую модель. Это позволяет создавать более универсальные системы ИИ, способные решать комплексные задачи, требующие одновременной обработки нескольких модальностей.

6.2 Оптимизация вычислительных ресурсов и энергоэффективность

Архитектура MoE существенно снижает вычислительные затраты и потребление энергии по сравнению с традиционными моделями, где активируются все параметры независимо от задачи. Такая оптимизация не только снижает затраты на инференс и обучение модели, но и способствует созданию более экологически устойчивых ИИ-систем.

6.3 Расширение доступности передовых ИИ-технологий

Aria выпущена под лицензией Apache 2.0 позволяет организациям и отдельным исследователям без затрат и ограничений интегрировать её в свои проекты.

6.4 Установление новых стандартов для открытых моделей

Высокие показатели производительности Aria устанавливают новые стандарты для открытых мультимодальных моделей. Это поощряет разработчиков и исследователей создавать более эффективные и мощные решения, стимулируя здоровую конкуренцию в сообществе ИИ. Установление таких стандартов способствует быстрому развитию области, ускоряя темпы внедрения новых технологий и улучшений в существующие модели.

6.5 Влияние на сообщество

Разработчики предоставили код для полного дообучения модели Aria и обочения LoRa, а также полный доступ к архитектуре модели. Это открывает возможность даже для тех, кто не слишком глубоко разбирается в принципах машинного обучения, попробовать свои силы в улучшении качества моделей на узкоспециализированных доменах.

7 Бизнес-кейсы

7.1 Автоматизация обслуживания клиентов

С применением мультимодальных возможностей Rhymes-AI/Ari компании могут разрабатывать умных чат-ботов и виртуальных ассистентов, которые работают не только с текстом, но и с изображениями и видео. Это даёт возможность предоставлять более оперативную и качественную поддержку клиентам, например, решая проблемы с продуктами через загрузку фотографий дефектов или видеоинструкций.

7.2 Анализ и управление контентом

Медиа-компании и платформы для обмена контентом могут использовать Aria для автоматической классификации, категоризации и анализа мультимедийных данных. Способность модели обрабатывать текст, изображения и видео позволяет создавать эффективные системы для управления большими объёмами информации, улучшения поиска и рекомендаций контента.

7.3 Образовательные технологии

В сфере образования Aria может стать основой для создания интерактивных учебных материалов, объединяющих текст, изображения и видео. Модель помогает генерировать адаптивные обучающие материалы, предоставляет мультимодальные объяснения и оценивает выполнение заданий, что способствует лучшей персонализации и качеству обучения.

7.4 Безопасность и видеонаблюдение

Компании, работающие в области безопасности, могут внедрять Aria в системы видеонаблюдения для автоматического выявления и анализа подозрительных действий. Модель обрабатывает видео в реальном времени, распознаёт объекты и их поведение, а также формирует текстовые отчёты, что делает мониторинг и реагирование на инциденты более эффективными.

8 Будущие разработки

Модель Aria от Rhymes AI уже демонстрирует значительные достижения в области мультимодальных ИИ-систем. Однако для дальнейшего улучшения и расширения её возможностей существуют несколько направлений, которые могут быть реализованы в будущих итерациях.

8.1 Улучшение архитектуры Mixture-of-Experts (MoE)

- **Динамическая балансировка нагрузки**: Разработка более продвинутых алгоритмов маршрутизации позволит равномерно распределять задачи между экспертами. Это поможет избежать перегрузки отдельных экспертов и повысит общую производительность модели.

- **Адаптивное количество экспертов**: Введение механизма, который позволяет динамически изменять количество активируемых экспертов в зависимости от сложности задачи, повысит гибкость и эффективность модели. Для простых задач может активироваться меньше экспертов, что снизит вычислительные затраты, а для сложных задач — больше экспертов, обеспечивая более точные и качественные результаты.

8.2 Расширение масштабируемости и оптимизация вычислительных ресурсов

- **Квантование и прунинг**: Квантование Aria позволит уменьшить её размер и ускорить процесс инференса без значительной потери качества. Прунинг поможет снизить вычислительные затраты и улучшить эффективность модели.

8.3 Расширение функциональных возможностей и интеграция с другими технологиями

- **Дообучение на tool calling**: Внедрение механизма tool calling позволит модели Aria взаимодействовать с внешними инструментами и сервисами. Такой подход повысит гибкость модели и расширит её область применения.

8.4 Расширение международной поддержки и локализации

- **Многоязычная поддержка**: Расширение языковой базы модели позволит поддерживать большее количество языков, что сделает Aria доступной для пользователей из разных регионов.

9 Заключение

- **Архитектура**: Aria от Rhythmes-AI использует архитектуру Mixture of Experts (MoE) для интеграции и обработки текста, изображений и видео. Это позволяет модели активировать только часть параметров для каждого входного токена, что снижает нагрузку на систему.
- **Сравнение с конкурентами**: По сравнению с моделями DeepSeek-V2 и Pixtral 12B, Aria активирует меньше параметров (3.9B для визуальных и 3.5B для текстовых данных) и при этом сохраняет производительность на уровне или выше конкурентов.
- **Ограничения**: Несмотря на достоинства, модель имеет ряд ограничений:
  - Высокие вычислительные требования — большой объём параметров требует значительных ресурсов.
  - Сложность архитектуры MoE — её настройка и оптимизация могут вызывать трудности.