Take Home Research Focused RU

В рамках задания на эту неделю, пожалуйста, изучите и внедрите следующую перспективную технологию в области больших языковых моделей (LLM):

GitHub Репозиторий: https://github.com/ericlbuehler/mistral.rs

Ваши конкретные задачи следующие:

Реализация: Настройте и внедрите демонстрацию для «интерактивного режима с потоковой передачей» с использованием модели Phi 3 128k mini с квантованием через ISQ в Q4K.

Разработка интерфейса: Разработайте простой пользовательский интерфейс (UI) для демо, который позволит пользователям вводить текст и получать ответы в формате чата, эффективно создавая базового чат-бота.

После завершения реализации, пожалуйста, напишите подробный анализ и дайте детализированные ответы на следующие вопросы:

Запуск интерфейса: Предоставьте чёткое пошаговое руководство по настройке и запуску интерфейса. Это руководство должно включать любые проблемы, с которыми вы столкнулись в процессе реализации, и способы их решения. Убедитесь, что инструкции доступны другим членам команды, которые могут быть не знакомы с исходной документацией по технологии.

Преимущества: Каковы преимущества использования данного интерфейса по сравнению с другими в данной области?

Недостатки: Какие ограничения или сложности вы заметили при работе с этим интерфейсом, как на этапе настройки, так и с точки зрения функциональности?

Влияние: Каковы более широкие последствия этой технологии? Рассмотрите её потенциальное влияние на сферу ИИ, технические достижения или будущие применения.

Бизнес-кейсы: Определите и опишите потенциальные бизнес-применения, где эта технология может быть использована для создания ценности.

Будущие разработки: Предложите возможные направления дальнейшего развития или исследования на основе этой технологии. Как её можно улучшить или расширить в будущих итерациях?

Пожалуйста, убедитесь, что ваш анализ тщательно проработан и хорошо задокументирован, так как он будет передан исследовательской команде для дальнейшего рассмотрения и обсуждения.