|  |  |
| --- | --- |
| Наименование Проекта с ИРН номером  | AP23489782 «AutoCorrection: обучение с подкреплением от искусственного интеллекта для разработки само-корректирующихся языковых моделей». |
| Актуальность/ Абстракт | Применение методики обучения с подкреплением на основе обратной связи от человека (Reinforcement Learning with Human Feedback, RLHF) (Ziegler et.al., 2019) оказало значительное влияние на развитие сферы языковых моделей. В то время как подход RLHF в большей степени опирается на человеческую обратную связь для директивного усовершенствования моделей, настоящее исследование предлагает инновационный подход, включающий в себя модели, настраиваемые на основе инструкций. Эти модели, вдохновленные методом Self-Instruct (Wang et.al., 2022), предназначены для улучшения эффективности и расширения области применения как RLHF, так и Reinforcement Learning from AI Feedback (RLAIF) (Lee et.al., 2023). Данный проект фокусируется на изучении RLAIF в контексте моделей, основанных на инструкциях, с целью расширения их обобщающих способностей и минимизации необходимости обширной человеческой обратной связи. |
| Цель (согласно заявке) | Целью проекта является исследование и разработка учебно-ориентированной языковой модели со встроенными механизмами самокоррекции, которая при этом объединяет ключевые аспекты RLAIF для минимизации зависимости от обратной связи с человеком, что приводит к автономности и адаптивности языковых моделей. |
| Ожидаемые результаты | В результате реализации данного проекта предполагается следующее:1. Разработка Усовершенствованных Языковых Моделей: Создание более продвинутых моделей, способных эффективно следовать инструкциям и самостоятельно корректировать свои ответы на основе встроенных механизмов обучения с подкреплением от искусственного интеллекта.2. Улучшение Автономности и Адаптивности: Разработанные модели будут демонстрировать повышенную автономию и способность к адаптации, что позволит им обрабатывать широкий спектр задач без значительной человеческой вмешательства.3. Минимизация Зависимости от Человеческой Аннотации: Сокращение потребности в обширной человеческой обратной связи и аннотациях, что упрощает процесс обучения и повышает эффективность моделей.4. Расширение Области Применения Языковых Моделей: Открытие новых возможностей для использования языковых моделей в различных областях, включая образование, обслуживание клиентов, здравоохранение и другие.5. Усиление Научно-Технического Потенциала: Внесение значительного вклада в научно-технический прогресс Казахстана, повышая конкурентоспособность научных организаций.6. За весь период реализации проекта будет опубликовано 2(две) статья и (или) обзор в рецензируемом научным издании, индексируемом в Science Citation Index Expanded базы Web of Science и (или) имеющий процентиль по CiteScore в базе Scopus не менее 65 (шестьдесят пять); и будет опубликовано не менее 1 (одной) статьи или обзора в рецензируемом зарубежном или отечественном издании, рекомендованном КОКСНВО.7. По технологической готовности данный проект находится на уровне № 2. По результатам выполнения задач проекта уровень готовности технологии будет соответствовать уровню № 4. |
| Список полных ФИО членов исследовательской группы с их идентификаторами (Scopus Author ID, Researcher ID, ORCID, если имеются) и ссылками на соответствующие профили | Научный руководитель проекта -Пак Александр Александрович, Кандидат технических наук, Ассоц. проф.Индекс Хирша 5; ResearcherID Web of Science: W-4002-2018; ORCID https://orcid.org/0000-0002-8685-9355; Scopus Author ID – 35362583100Жасандықызы Марал, PhD ComputerScience Индекс Хирша 4;ResearcherID Web of Science: ORCID<https://orcid.org/0000-0003-2374-7468> Scopus Author ID57105686400Зияден Атабай, магистр технических наук, PhD студентИндекс Хирша 2; ResearcherID Web of Science: ORCID [https://orcid.org/0000-0002-4878-8971 Scopus Author ID 57211658770](https://orcid.org/0000-0002-4878-8971%20Scopus%20Author%20ID%2057211658770)Сапарғали Есболат Еділұлы, Магистр Технических наукWeb of Science Web of Science Researcher ID: IWU-9683-2023Orcid: https://orcid.org/ 0000-0002-9072-3389Scopus Author ID –57320278100Жиенбеков Жалғас Жұмағалиұлы, Магистр технических наук, лектор ORCID: <https://orcid.org/0009-0009-0291-9938>Мираш Айтуған Мирболатұлы, бакалавр ORCID: <https://orcid.org/0009-0008-5320-082X>Меркебаев Азат Галымжанулы |