

# БЕНЧМАРК ДЛЯ ОЦЕНКИ ДОВЕРЕННОСТИ LLM

Каталог ИИ-решений



Современные бенчмарки для оценки доверенности имеют принципиальное ограничение: все задания в них разработаны для английского или китайского языков. Русскоязычные бенчмарки способны лишь частично оценить доверенность, но их главная цель — оценка общих навыков LLM

На текущем этапе разработки бенчмарка доверенности LLM для русскоязычных задач мы:

- ✓ адаптировали 12 русскоязычных и англоязычных наборов данных для 14 задач проверки критериев доверенности
- ✓ провели тестирование доверенности 14 открытых LLM на основе разработанных задач

Исправление недостатков, выявленных бенчмарком, повысит как общее качество информационных систем на основе LLM, так и доверие к ним

## 01 Справедливость

- Выявление стереотипов
- Распознавание стереотипов
- Согласие со стереотипом

## 03 Безопасность

- Устойчивость атакам
- Защита от ненадлежащего исполнения
- Чрезмерная безопасность

## 05 Достоверность

- Проверка усвоенных знаний
- Проверка использования внешних источников

## 02 Этичность

- Определение этических концепций
- Выявление нарушения этических норм

## 04 Приватность

- Осведомленность о конфиденциальности
- Защита от утечек данных

## 06 Надежность

- Выявление OOD
- Устойчивость естественному шуму

# СПАСИБО ЗА ВНИМАНИЕ!



ФЦПРИИ.РФ



t.me/fcprii