УДК 111.11

МЕТОД ИЗМЕРЕНИЯ СТАБИЛЬНОСТИ МОДЕЛИ К ИЗМЕНЕНИЯМ ВХОДНЫХ ПРОМПТОВ

Абдурахимов М.А (Университет ИТМО)
Научный руководитель – к.т.н., старший научный сотрудник исследовательского центра в сфере искусственного интеллекта
Ходорченко М.А. **(**Университет ИТМО**)**

Введение. В последние годы промптинг, как методология направленного взаимодействия с моделями машинного обучения, приобретает все большую популярность. Этот метод предоставляет возможность контролировать вывод моделей путем предоставления им определенных текстовых вводных данных или промптов. В данной работе наша основная цель заключается в улучшении оценки языковых моделей, исходя из предположения о том, что хорошая модель должна быть устойчива к вариативности промпта. В нашей работы мы предлагаем коэффициент стабильности, который позволяет более эффективно оценивать и интерпретировать результаты работы языковых моделей в зависимости от их устойчивости к изменениям входного промпта.

Основная часть. В контексте оценки языковых моделей и их реакции на быстрые изменения важно установить надежные оценочные показатели, которые точно отражают производительность и стабильность моделей. Традиционные метрики оценки в задачах обработки языка, такие как оценка BLEU и оценки классификации, дают ценную информацию о производительности модели, но могут не полностью отражать нюансы, связанные с вариациями промптов [1]. Одним из ключевых вопросов, который необходимо решить, является учет стабильности и надежности моделей перед лицом изменения входных данных или промптов. В настоящее время не существует стандартного подхода к учету стабильности модели в ответ на изменение предложения, что ограничивает нашу способность осмысленно оценивать и сравнивать модели.

Коэффициент стабильности можно рассчитать путем измерения сходства или согласованности результатов модели при представлении разных промптов для одной и той же задачи. В частности, это может включать вычисление косинусного сходства или других показателей сходства между выходными данными, генерируемыми моделью в ответ на различные промпты. Более высокий коэффициент стабильности указывает на то, что модель дает более последовательные ответы независимо от изменения промпты, что отражает ее стабильность при обработке различных входных условий.

Мы представили предложенную методологию для проверки гипотезы о коэффициенте устойчивости языковых моделей к изменениям в промптах и оценили ее потенциальные последствия для практического использования и понимания проблемы. Для этого мы провели серию экспериментов на задачах из бенчмарка MERA [2].

Для оценки производительности моделей мы использовали модели Vikhr и TinyLlama с температурой 0, чтобы исключить случайные ответы и сосредоточиться на их результаты генерации на различные промпты. В процессе анализа мы применили методы интерпретации, включая SHAP (SHapley Additive exPlanations), чтобы визуализировать, на какие части входных промптов модели обращали внимание при формировании ответов. Этот подход позволил нам получить информацию о том, как внимание и понимание модели могли изменяться в зависимости от вариаций промптов.

Выводы. В результате нашего исследования мы предложили новую метрику оценки - коэффициент стабильности, которая учитывает надежность и стабильность моделей в условиях, изменяющихся промптов. Эта метрика позволяет более полноценно оценивать производительность языковых моделей и принимать обоснованные решения о их использовании.

Список использованных источников:

1. "The Art of Prompting: How to Win Conversations and Influence LLMs." Alkymi Blog, URL: https://www.alkymi.io/resources/blog/the-art-of-prompting-how-to-win-conversations-and-influence-llms. Дата доступа: 01.02.2024
2. Официальный вебсайт MERA Benchmark. URL: https://mera.a-ai.ru/ru. Дата доступа: 01.02.2024
3. Beurer-Kellner, Luca, Marc Fischer, and Martin Vechev. "Prompting Is Programming: A Query Language for Large Language Models." arXiv:2212.06094 (2023)

Абдурахимов М.А. (автор)

Ходорченко М.А. (научный руководитель)