УДК 004.82

ИССЛЕДОВАНИЕ ТОПОЛОГИИ ГРАФОВ ИНСТРУМЕНТОВ КАК МЕТРИКИ СЛОЖНОСТИ БЕНЧМАРКОВ ДЛЯ МУЛЬТИАГЕНТНЫХ LLM-СИСТЕМ

Файрушин Б. Р.[1], Амерханова Н. А.[1], Исхаков К. Ф.[1]
Научный руководитель – канд. техн. наук, доцент Калюжная А. В.[1]
[1] Университет ИТМО
bulat.fairushin@niuitmo.ru

Введение

Для оценки качества мультиагентных LLM-систем уже существуют множество бенчмарков – графов инструментов, специально созданных для проверки качества модели, но пока не определена единая метрика, сопоставляющая бенчмарки друг с другом, не оценивается сложность графов для задач ToolCalling. Это делает сравнение результатов тестирования и сопоставление гипотез на разных данных ненадежным. Нельзя определить специфичен ли полученный результат только для используемых для эксперимента бенчмарков, или он будет наблюдаться системно. В качестве такой характеристики сложности предлагается топологическая структура графа инструментов, а именно содержание им определенных подструктур (chain, star, tree, cycle, hourglass и др.). Влияние топологических паттернов на взаимодействие между агентами изучалось в статье MultiAgentBenchc[1]. Авторы использовали паттерны для описания архитектуры коммуникации между инструментами в процессе планирования, но не рассматривали их определяющие факторы сложности всего графа.

Основная часть

Была сформулирована гипотеза о том, что различные структурные паттерны создают разную степень сложности как для идентификации нужных инструментов, так и для определения корректного порядка их вызова. Исследование проведено в два этапа. С целью формирования экспериментальной базы был проведён топологический анализ существующего бенчмарка: были извлечены и записаны представители выделенных типов подграфов [2]. Для каждой найденной подструктуры с помощью LLM генерировался вопрос пользователя, который естественным образом требует вызова именно тех инструментов, которые входят в подструктуру.

Эксперименты проводились в двух группах, различающихся использованной моделью, подходами к генерации и последующей валидации. В первой группе LLM-модель явно получала на вход порядок ожидаемого вызова инструментов в соответствии с исходной структурой графа. Для валидации множество инструментов каждой подструктуры дополнялось дистракторами – инструментами из графа, не связанными с задачей. Модель получала вопрос и расширенный список инструментов, из которого должна была выбрать подходящие именно для текущего вопроса и определить последовательность их вызова. Полученный ответ сравнивался с эталоном по нескольким метрикам: Recall, Strict Recall (доля полных совпадений внутри типа паттерна), Accuracy (доля инструментов в правильной позиции).

Во второй группе модель, генерирующая инструкции, получала только множество инструментов и самостоятельно планировала зависимости между ними. Полученные вопросы были отфильтрованы по семантическому сходству между описанием инструкции и соответствующих ей инструментам. Далее для каждого отобранного вопроса проверялась топология полученных связей (Precision, Recall по инструментам и зависимостям).

Экспериментальные результаты показали, что тип структуры подграфа может быть предиктором сложности вызова входящих в него инструментов, при этом влияние количества дистракторов выражено в разы слабее.

Выводы

Следующей задачей после формулировки метрики для оценки сложности бенчмарков является разработка генератора синтетических графов инструментов с управляемой структурой с соблюдением семантической и логической связности, что позволит создать базу воспроизводимых бенчмарков для анализа эффективности работы LLM с графами различной структуры.

Литература

1. Zhu K, Du H, Hong Z, Yang X, Guo S, Wang DZ, Wang Z, Qian C, Tang R, Ji H, You J. Multiagentbench: Evaluating the collaboration and competition of llm agents. // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025 , P. 8580-8622.
2. Lumer E, Basavaraju PH, Mason M, Burke JA, Subbiah VK. Graph rag-tool fusion. arXiv preprint arXiv:2502.07223. 2025 Feb 11.