УДК 004.89

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ МУЛЬТИАГЕНТНЫХ СИСТЕМ ДЛЯ БЕНЧМАРКА GAIA НА ОСНОВЕ ФРЕЙМВОРКА AUTOMAS С ИСПОЛЬЗОВАНИЕМ МЕТА-АГЕНТОВ

Рождественский Е.Д.[1]
Научный руководитель – Каминский Ю.К.[1]
[1]Университет ИТМО

Введение

Мультиагентные системы (МАС) на основе больших языковых моделей являются перспективным подходом к решению сложных задач, требующих комбинации навыков: веб-поиска, анализа документов и написания кода. Бенчмарк GAIA содержит 165 задач трёх уровней сложности для оценки AI-ассистентов. Ручное проектирование конфигураций МАС под каждую задачу немасштабируемо, что обуславливает актуальность автоматизации генерации агентных конфигураций.

Основная часть

Проведено экспериментальное исследование применения фреймворка AutoMAS к бенчмарку GAIA. AutoMAS использует мета-агента для автоматической генерации JSON-конфигураций МАС с ролями агентов, промптами и инструментами. При подготовке данных обнаружено, что около 62 % задач GAIA содержат утечки ответов в аннотациях, для чего реализована процедура очистки. Выполнено 4 цикла экспериментов (модель qwen3-235b, 165 семплов). Лучший результат – 45,5 % accuracy (цикл 4, модифицированный промпт + e2b sandbox). Однако эксперименты показали стагнацию: известная архитектура решения не улучшает результаты на сложных задачах (Level 3) и ухудшает на простых (Level 1). Анализ 660 конфигураций выявил парадокс: одноагентные МАС точнее двухагентных на 24 процентных пункта, а самые дешёвые решения превосходят дорогие на 16 процентных пункта. Разные циклы генерации решают разные подмножества задач, что открывает возможности для ансамблевых подходов.

Выводы

Проведено экспериментальное исследование применения AutoMAS к бенчмарку GAIA. Выполнена очистка датасета от утечек (62 % задач), проведено 4 цикла экспериментов, проанализировано 660 конфигураций МАС. Достигнута точность 45,5 %. Выявлено, что прямое использование экспертных аннотаций для решения задач неэффективно. Обнаружены парадоксы: одноагентные МАС точнее двухагентных, низкобюджетные решения превосходят дорогие. Перспективы: адаптивная маршрутизация по сложности, self-reflection агентов, расширение инструментов (vision, PDF parsing), тестирование альтернативных LLM.

Литература

1. Mialon G. et al. GAIA: a benchmark for General AI Assistants // arXiv:2311.12983. – 2023.
2. Wu Q. et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation // arXiv:2308.08155. – 2023.

Рождественский Е.Д. _____

Каминский Ю.К. _____