УДК 004.85

ОЦЕНКА КАЧЕСТВА КОНТЕКСТА В АГЕНТНЫХ LLM-СИСТЕМАХ: СОПОСТАВЛЕНИЕ ИНЖЕНЕРНЫХ ПРИНЦИПОВ И СОВРЕМЕННЫХ БЕНЧМАРКОВ

Рыбинская З. В.[1], Иржанова Ю. И.[1]
Научный руководитель – к. т. н., доцент Хлопотов М. В.[1]
[1]Университет ИТМО
zltrbn@gmail.com

Введение

В последние годы наблюдается смещение фокуса в развитии больших языковых моделей от увеличения параметров и масштабирования обучающих данных к исследованию их способности к работе в сложных, динамически изменяющихся контекстах. Появление специализированных бенчмарков, ориентированных на оценку контекстного обучения и рассуждения внутри заданного информационного окружения, отражает данный сдвиг. В частности, бенчмарк CL-bench [1] направлен на оценку способности моделей к работе в динамическом контексте. В агентных LLM-системах контекст выступает не просто входными данными, а рабочей средой принятия решений. При этом успешность работы модели определяется не только её внутренними параметрами, но и качеством предоставляемого контекста.

В предыдущем этапе исследования были сформулированы инженерные принципы управления качеством контекста, включающие релевантность, актуальность, непротиворечивость, полноту и трассируемость. Настоящая работа направлена на сопоставление данных принципов с современными подходами к оценке способности моделей к контекстному обучению.

Основная часть

Современные бенчмарки, ориентированные на оценку контекстного обучения (context learning), проверяют способность модели использовать новую информацию, обновлять интерпретацию данных и выполнять рассуждения в условиях длинного и потенциально противоречивого контекста. В частности, CL-bench предлагает задачи, в которых корректность ответа зависит не от запомненных знаний, а от способности адаптивно интерпретировать предоставленный контекст.

Анализ структуры подобных бенчмарков показывает, что успешность модели определяется несколькими факторами: корректной обработкой релевантной информации, игнорированием шума, согласованием противоречивых данных и устойчивостью к изменениям условий задачи [2]. Эти требования непосредственно соотносятся с ранее сформулированными инженерными принципами качества контекста.

Так, принцип релевантности обеспечивает снижение перегрузки контекстного окна; принцип актуальности – соответствие информации текущему состоянию среды; принцип непротиворечивости – согласованность источников; принцип полноты – достаточность данных для корректного вывода; принцип трассируемости – возможность анализа и воспроизводимости решений.

Сопоставление инженерных принципов и требований современных бенчмарков позволяет выявить разрыв между архитектурной организацией контекста в прикладных системах и метриками, применяемыми для оценки способностей моделей. В частности, существующие бенчмарки в большей степени оценивают поведение модели, чем качество самого контекстного окружения.

Выводы

Проведённый анализ показывает, что качество контекста является критическим фактором способности агентных LLM-систем к контекстному обучению и динамическому

рассуждению. Современные бенчмарки фиксируют смещение парадигмы от статического запоминания к адаптивному использованию информации, однако при этом недостаточно учитывают инженерные аспекты формирования и управления контекстом.

Результаты работы демонстрируют необходимость интеграции принципов управления качеством контекста в процедуры оценки LLM-систем. Дальнейшие исследования могут быть направлены на разработку формальных метрик качества контекста и их включение в систему бенчмаркинга агентных моделей.

Литература

1. CL-bench: бенчмарк для оценки контекстного обучения / Доу Ш., Чжан М., Инь Ч., Хуан Ч. – arXiv, 2026 [Электронный ресурс] – Режим доступа: https://doi.org/10.48550/arXiv.2602.03587, свободный. Яз. англ. (дата обращения 15.02.2026).

2. Lost in the Middle: как языковые модели используют длинный контекст / Лю Н. Ф., Лин К., Хьюитт Дж. – arXiv, 2003 [Электронный ресурс]. – Режим доступа: https://doi.org/10.48550/arXiv.2307.03172, свободный. Яз. англ. (дата обращения 15.02.2026).