УДК 004.83

Модель памяти для LLM-агента на основе темпорального графа знаний

Меньщиков М. А.[1]
Научный консультант – д.ф.-м.н., профессор Бурнаев Е. В.[1,2]
[1]Сколковский институт науки и технологий
[2]Институт AIRI
m.menschikov@skoltech.ru

Работа выполнена при поддержке гранта для научно-исследовательских центров в области ИИ, предоставленного Министерством экономического развития Российской Федерации в соответствии с соглашением 000000C313925P4F0002 и соглашением со Сколтехом № 139-10-2025-033.

Введение

В последнее время для расширения знаний LLM все чаще применяется подход поисковой дополненной генерации (RAG), при котором процесс генерации текста сочетается с извлечением информации из внешней базы знаний. Однако традиционный RAG-подход полагается исключительно на семантическое сходство и слабоструктурированные независимые фрагменты информации, что затрудняет его использование при обработке составных запросов, для которых требуется учитывать глобальный контекст и связи между фактами.

Для преодоления этих ограничений в качестве внешнего хранилища или “памяти” для LLM перспективно использование графов знаний [1,2]. Графовая модель памяти позволяет LLM-агенту не только находить отдельные факты, но и учитывать их контекст, а также взаимосвязи между ними, что повышает обоснованность и точность его выводов.

Основная часть

В рамках одного из проведённых исследований [3] был предложен и реализован вариант модели памяти для LLM-агента на основе Text-Attributed-графа. Строящийся в автоматическом режиме граф состоит из вершин, семантика которых инспирирована исследованиями из области нейробиологии на тему способа хранения информации в человеческом мозге: (1) episodic - исходные фрагменты текста для сохранения в модель памяти; (2) thesis - независимые полные краткие единицы знаний в виде текста на естественном языке; (3) object - именованные сущности, объекты/субъекты реального мира. Вершины thesis- и object-типов извлекаются из episodic-вершин и далее связываются с соответствующими эпизодическими воспоминаниями за счёт ненаправленных гипер-рёбер. Из полученных thesis-вершин также извлекаются object-вершины, которые далее связываются с соответствующими тезисными воспоминаниями за счёт ненаправленных гипер-рёбер. Так же при извлечении object-вершин из эпизодических воспоминаний извлекаются предикаты в виде текста на естественном языке, которые связывают направленным ребром соответствующую пару вершин и образуют simple-триплет. Можно выделить следующие недостатки текущего решения:

1. Данная модель памяти позволяет только неявно учитывать темпоральные характеристики хранимой информации: можно сохранить желаемую отметку времени в атрибуты соответствующих триплетов; при переводе триплета в строковое представление содержащиеся в нём атрибуты также будут отражены и LLM сможет обуславливаться на эту дополнительную информацию при генерации ответа на вопрос, однако из-за “Lost in the Middle”-дилеммы часть важной информации может быть упущена, что, в свою очередь, негативно повлияет на качестве полученного результата.
2. Онтология генерируемого графа знаний достаточно простая. Предложенного количества характеристик, по котором можно осуществлять поиск и фильтрацию информации недостаточно, что, в свою очередь, негативно влияет на эффективности и производительности QA-алгоритмов поиска.

В рамках текущего исследования для устранения/смягчения влияния вышеуказанных недостатков и повышения эффективности/производительности существующего решения предлагается выполнить ряд модификаций модели памяти:

1. Тезисным воспоминаниям (thesis-вершинам) присваивается две категории меток: episode и temporal. Episode-метка характеризует формулировку тезиса: (1) fact - утверждения, которые являются объективными и могут быть независимо проверены или опровергнуты с помощью соответствующих доказательств; (2) opinion - утверждения, содержащие личное мнение, чувства, ценности или суждения конкретного объекта реального мира, которые не поддаются проверке без контекста; (3) prediction - неопределенные утверждения о будущем, касающиеся возможных событий, гипотетических результатов, непроверенные заявления. Temporal-метка характеризует временные пределы актуальности содержащейся в тезисе единицы знания: (1) static - утверждения, которые действительны со дня, когда они произошло и никогда не теряющие свою актуальность; (2) dynamic - утверждения, которые действительны в течение определенного периода времени и обычно теряющие свою актуальность в случае возникновения какого-либо статического факта, отмечающего конец события или начало нового, противоречащего ему; (3) atemporal - утверждения, которые всегда будут истинными вне зависимости от времени.
2. Для каждой thesis–вершины определяется/фиксируется интервал времени, в рамках которого содержащаяся в ней единица знания считается актуальной/валидной. Заполняются следующие поля: t_created, t_valid, t_invalid, t_expired, invalidated_by. Выполняется явное извлечение отметки времени из тезисного воспоминания и её приведение к абсолютному значению.
3. Возможные значения текстовых полей у предикатов в simple-триплетах фиксированы заданным списком и имеют (предикаты) нейтральную форму ко времени (time-neutal). В object-вершине, помимо сущности, содержится её тип (type) и краткое описание (description). В предикате simple-триплета, помимо name-поля присутствует текстовое поле с определением данного отношения.

Для оценки эффективности (за счёт качества решения QA-задач) предложенной модели памяти на основе темпорального графа знаний выбраны следующие датасеты: TimeQA, DiaASQ, HotpotQA и MuSiQue. В качестве основной метрики для оценки близости между generated- и golden-ответами используется LLM-as-a-Judge. В качестве LLM-моделей для построения графа и осуществления поиска в рамках QA-алгоритма выбраны следующие варианты: Gemma2 9B и Qwen2.5 7B.

Выводы

Совокупность предлагаемых модификаций направлена на создание интеллектуального агента с внешней памятью, пригодного для внедрения в различные отрасли экономики, включая банковскую сферу, который будет обладать следующим функционалом: (1) поддержка длительных диалогов с пользователем; (2) предоставлять точные и безопасные ответы на сложные вопросы; (3) учитывать актуальность имеющейся информации; (4) обеспечивать персонализацию рекомендаций и советов.

Литература

1. Rasmussen P. et al. Zep: a temporal knowledge graph architecture for agent memory //arXiv preprint arXiv:2501.13956. – 2025.
2. Anokhin P. et al. Arigraph: Learning knowledge graph world models with episodic memory for llm agents //arXiv preprint arXiv:2407.04363. – 2024.
3. Menschikov M. et al. PersonalAI: A Systematic Comparison of Knowledge Graph Storage and Retrieval Approaches for Personalized LLM agents //arXiv preprint arXiv:2506.17001. – 2025.