

ФОРМАЛЬНО-ЛОГИЧЕСКАЯ МОДЕЛЬ КООРДИНАЦИИ МНОГОАГЕНТНОЙ СИСТЕМЫ С ТРАССИРУЕМЫМ ИЗВЛЕЧЕНИЕМ ЗНАНИЙ ИЗ НАУЧНЫХ ИСТОЧНИКОВ

И.А. Трусов¹, Ю.В. Трофимов^{1,2}, А.Н. Аверкин^{1,3}, Е.С. Кондрашова¹

¹Государственный университет «Дубна», г. Дубна, Россия

²Объединенный институт ядерных исследований, г. Дубна, Россия

³ФИЦ «Информатика и управление» РАН, г. Москва, Россия

e-mail: trusov.iva@yandex.com

Использование больших языковых моделей привело к существенному расширению возможностей интеллектуальных систем. Вместе с тем практическое применение таких моделей в научной сфере выявило ограничения, связанные с недостаточной объяснимостью процессов генерации и отсутствием строгой связи между получаемым ответом и источниками знаний [1]. Для научных приложений подобная ситуация является недопустимой, поскольку достоверность информации и воспроизводимость результатов выступают базовыми требованиями научной деятельности. Система ориентирована на обработку научной информации и снижение вероятности появления неподтверждённых утверждений при генерации ответов. Подход основан на разделении функций между оркестратором, специализированными агентами, подсистемой извлечения знаний и механизмом верификации результатов. Предлагаемая архитектура ориентирована на переход от полностью вероятностной схемы генерации к модели, в которой ключевые этапы принятия решений подлежат контролю и формальному описанию [2].

Архитектура системы. Функциональная структура включает оркестратор, парсер научных публикаций, подсистему хранения знаний и набор специализированных агентов. Оркестратор выполняет классификацию запроса, оценивает доступность релевантного контекста и формирует последовательность вычислительных этапов. При необходимости инициируется создание нового агента с параметрами, соответствующими предметной области запроса.

Оценка достаточности контекста. Перед формированием ответа система должна оценить, есть ли достаточная информация для корректного вывода. Для решения данной задачи вводится метрика покрытия:

$$J(q, E) = \frac{|V(q) \cap V(E)|}{|V(q)|},$$

где $V(q)$ обозначает множество значимых токенов запроса, а $V(E)$ соответствует множеству терминов, присутствующих в корпусе знаний. Для каждого токена определяется вклад в решение о необходимости поиска или выборе агента. Это позволяет вернуться к пользователю с вопросом: «Что именно вы имели в виду под термином X?».

Парсер научных источников. Процесс пополнения гиперграфа знаний начинается с поиска релевантных источников к запрашиваемому промπτу. Для каждого документа вычисляется показатель качества:

$$Q(d_i) = (1, w_R R(d_i) + w_L L(d_i) + w_S S(d_i) + w_K K(d_i) + I_{DOI}(d_i)),$$

где величины отражают показатели: R релевантности, L научной значимости, S структурированности и K качество представления предметных концептов. DOI учитывается отдельной индикаторной функцией индикатором I_{DOI} .

Каждый документ преобразуется в множество элементов:

$$d_i \rightarrow \{(c_{i,1}, k_1), (c_{i,2}, k_2), \dots, (c_{i,n}, k_n)\}.$$

где $c_{i,j}$ – цитата к части промпта и k_j – разделённая сущность промпта.

Тем самым источники становятся релевантными под конкретный запрос связанными отношениями с именованными сущностями.

Гиперграф знаний. Для представления доказательной структуры используется гиперграф:

$$H_t = \langle V_H, E_H \rangle.$$

Вершины графа соответствуют концептам, тезисам, определениям и результатам исследований, а рёбра фиксируют отношения доказательства (противоречия). Благодаря этому появляется возможность отслеживать происхождение каждого утверждения и анализировать структуру аргументации.

Выбор и порождение агентов. Многоагентная архитектура предполагает наличие набора специализированных исполнителей. Для выбора наиболее подходящего агента используется функция близости:

$$\varphi(q, A_t) = \operatorname{argmax} \left(\frac{|V(q) \cap V(\text{skill}_a)|}{|V(q)|} \right)$$

Функция оценивает степень соответствия компетенций агента содержанию пользовательского запроса. После выбора агента формируется объяснение: «Выбран агент *arXiv-parser*, так как в запросе обнаружены термины «нейросеть» и «обучение» (вклад +0.7), а термин «квантовая физика» снизил релевантность (-0.2)». Выбранный агент получает специализированный промπτ и набор релевантных материалов из базы знаний. Если ни один из существующих агентов не обладает достаточной компетентностью, применяется механизм создания нового экземпляра. Условие запуска может быть представлено выражением:

$$\text{SpawnTrigger} = I(\text{explicit_cmd}) \vee I(p_{\text{act}} = \text{spawn}_{\text{agent}}) = 1, \text{ то } F_{\text{agent}}(q, G_t) \rightarrow a_{\text{new}}$$

Объяснимость решений. Для анализа вклада отдельных факторов используется метод *SHAP* [2]. Значения *SHAP* позволяют определить, какие характеристики контекста повлияли на решение оркестратора о запуске поиска, выборе агента или расширении корпуса знаний. Кроме того, метод используется для объяснения факторов, приведших к увеличению оценки риска конкретного утверждения. Наличие интерпретируемого механизма анализа повышает доверие к системе и облегчает её использование в научной среде.

Результаты. Разработанная в НИЦ ИИ Государственного университета «Дубна» формально-логическая модель координации многоагентной системы обеспечивает контролируемую обработку научных знаний и сквозную объяснимость на всех этапах: от понимания запроса до верификации ответа.

Оркестратор координирует специализированных агентов, управляет последовательностью обработки запроса и контролирует верификацию результатов, при этом каждое решение (выбор агента, извлечение факта, логический вывод) сопровождается трассируемым объяснением через гиперграф доказательств.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Литература

1. Farquhar S., Kossen J., Kuhn L. и др. Detecting hallucinations in large language models using semantic entropy // Nature. 2024. Vol. 630. P. 625–630. DOI: 10.1038/s41586-024-07421-0.
2. Sheth A. S. S., de Silva A. G. H. S. Neuro-Symbolic Generative Models: A Survey // Proceedings of the AAAI Conference on Artificial Intelligence. – 2022. – Vol. 34, no. 1. – P. 123–130.
3. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – DOI: 10.48550/arXiv.1705.07874.