Разработка мультиагентной ансамблевой системы рассуждения малых языковых моделей по принципу суда присяжных

Муллагалиев Дениз Радикович

Студент

Казанский (Приволжский) Федеральный Университет, Институт ИИРСИ, Казань, Россия

E-mail: mullagalievdeniz@yandex.ru

Научный руководитель *— Смольникова Камилла Рустемовна*

Малые языковые модели открывают большие возможности для анализа данных при ограниченных вычислительных ресурсах, однако из-за небольшой размерности им крайне свойственны галлюцинации. В одном из тестов ошибочные утверждения содержались в 48.8% ответов для Llama-3.1-8B, 76.9% — Gemma-3-4b, 63.7% — Ministral-8B [1]. Эффективный способ борьбы с галлюцинациями LLM — рассуждение (reasoning), когда модель задает сама себе проверочные вопросы. Есть методы, где в рассуждении участвуют несколько агентов с разными ролями; такой подход бывает эффективнее стандартного рассуждения [2]. Применение этой концепции в задачах верификации ответов путём рассуждения малых языковых моделей мы рассмотрим в этой работе.

Архитектура мультиагентной системы построена по принципу "суда присяжных". Выделены следующие агенты: **Истец** (Первичная генерация ответа), **Критик** (Поиск галлюцинаций), **Библиотекарь** (Поиск информации), **N агентов присяжных** (Независимая оценка), **Судья** (Итоговая оценка итерации рассуждения и принятие решения) и **Финализатор** (Подготовка финального ответа).

Рассуждение ведётся итеративно, ответы агентов накапливаются в общем контексте. Сначала система получает вопрос: истец формирует первичный ответ, критик выявляет слабые места ответа, библиотекарь ищет недостающую информацию в Интернете и делает краткую выжимку. Затем N присяжных независимо оценивают ответ, а судья на основе контекста итерации решает, завершать ли рассуждение или продолжать. В конце итерации финализатор формирует итоговый ответ либо рекомендации, как улучшить ответ на следующей итерации для истца. Передача контекста между итерациями обеспечивается механизмом итеративной памяти — при продолжении контекст суммаризируется.

Число агентов-присяжных N выбирается из условия, что по бино-

миальному распределению вероятность корректного решения (большинство из N) превышает заданный порог точности k_{target}:

$$\sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{i} p^i (1-p)^{N-i} \; > \; k_{\text{target}}$$

где p — вероятность корректного ответа одного агента, N — число присяжных, k_{target} — требуемый порог доверия.

Техническая реализация выполнена на LangGraph: используется граф состояний StateGraph, узлы — агенты. Роли и поведение агентов задаются промпт-инжинирингом. Веб-поиск у агента-библиотекаря реализован через LangSearch Web Search API. Для доступа к SLM используется OpenRouter, также есть поддержка локальных моделей через transformers. Реализация: `https://github.com/qxdeniz/MultiagentJuryReasoningSLMSystem`

Для проверки подхода было проведено тестирование, где SLM генерировала ответы на вопросы в разных доменах сама и с мультиагентной системой рассуждения. Для теста использовалась малая языковая модель gemma-3-4b-it с температурой 0,6. В каждом домене содержалось 3 вопроса. Качество ответов оценивалось с помощью LLM GPT-5 по принципу LLM as Judge с использованием методики Claim-UQ. Ответы оценивались по следующим метрикам: Understanding, Justification Quality, Factual Precision, Refutation Strength, Hallucination Rate. Результаты в таблице приведены средние значения метрик по 3 вопросам в формате «без рассуждения / с рассуждением».

Домен	**Unders.**	**Justif.**	**Fact.**	**Refut.**	**Hall.**
ИТ	0.65/0.8	0.5/0.75	0.75/0.85	0.55/0.80	0.1/0.03
Математика	0.7/0.9	0.55/0.86	0.80/0.95	0.6/0.9	0/0.05
Логика	0.35/0.85	0.22/0.80	0.2/0.75	0.3/0.82	0.45/0.01
История	0.82/0.88	0.76/0.81	0.78/0.85	0.55/0.70	0.14/0.02
Экономика	0.60/0.85	0.50/0.80	0.55/0.82	0.45/0.78	0.15/0.05

Литература

1. Dentan J., Canesse A., Buscaldi D., Shabou A., Vanier S. MUCH: A Multilingual Claim Hallucination Benchmark // arXiv.zorg, 2025. arXiv:2511.17081v1. `https://arxiv.org/html/2511.17081v1`
2. Hegazy M. Diversity of Thought Elicits Stronger Reasoning Capabilities in Multi-Agent Debate Frameworks // arXiv.org, 2024. arXiv:2410.12853v2. `https://arxiv.org/abs/2410.12853`