УДК 004.8

ДИНАМИКА НЕОПРЕДЕЛЁННОСТИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В МНОГОШАГОВЫХ ПРОЦЕССАХ АНАЛИЗА ДАННЫХ

Гротов К. Ю.[1]
Научный руководитель – канд. техн. наук, доцент Малых В. А.[1]
[1]Университет ИТМО
konstantin.grotov@gmail.com

Введение

Широкое распространение агентных систем на основе больших языковых моделей[1] (БЯМ) в задачах анализа данных порождает новые требования к надёжности таких систем[2]. В отличие от однократной генерации, агентные рабочие процессы представляют собой многошаговые траектории, в которых модель последовательно формирует рассуждения, вызывает внешние инструменты (Python, SQL, Bash) и интерпретирует полученные результаты. Ошибка на ранних этапах траектории детерминированно распространяется на последующие шаги, что может привести к некорректным аналитическим выводам даже при внешне связном и грамматически правильном ответе. При этом существующие методы оценки неопределённости, основанные на вероятностях отдельных токенов, сэмплировании или вербализованной уверенности, разработаны преимущественно для одношаговых задач и не учитывают зависимости между шагами и распространение ошибок в многошаговых сценариях[3]. Актуальность разработки методов количественной оценки неопределённости, адаптированных к агентным аналитическим рабочим процессам, обусловлена высокими рисками принятия ошибочных решений на основе результатов таких систем в бизнес-среде.

Основная часть

В данной работе исследуется динамика неопределённости БЯМ в рамках бенчмарка DA-Code[4], набора задач по анализу данных, требующих многошагового взаимодействия агента с инструментами и источниками данных. В качестве метрики неопределённости используется пошаговая энтропия распределения следующего токена, вычисляемая на основе логарифмических вероятностей, предоставляемых моделью в процессе генерации. Для каждой траектории выполняется раздельный анализ энтропии на сегментах рассуждений (reasoning) и вызовов инструментов (tool call), что позволяет выявить характерные шаблоны динамики неопределённости в зависимости от типа генерируемого контента. Эксперименты проведены на четырёх моделях различных семейств и масштабов: Qwen3-8B, Qwen3-30B-A3B, Qwen2.5-7B и Ministral-8B, а также с двумя форматами вызова инструментов. Анализ динамики средней энтропии по шагам траектории демонстрирует устойчивый шаблон: сегменты вызовов инструментов характеризуются значительно более низкой энтропией по сравнению с сегментами рассуждений, что свидетельствует о систематической самоуверенности моделей при генерации действий. Данный эффект наблюдается независимо от размера и семейства модели, а также от формата вызова инструментов, что указывает на его модельно-независимую природу. Дополнительно исследуется распределение доли токенов рассуждений (reasoning ratio) в зависимости от позиции шага в траектории: в начальных шагах (0–5) модели демонстрируют более высокую долю рассуждений, тогда как на поздних шагах (12+) наблюдается её снижение. Сравнение энтропии по исходам траектории (корректное завершение, самоуверенная ошибка, корректный провал) показывает, что траектории, завершившиеся ошибкой при высокой уверенности модели, не отличаются по уровню энтропии от успешных, что

подтверждает недостаточность токен-уровневых сигналов для детекции ошибок в многошаговых сценариях.

Выводы

Проведённое исследование демонстрирует, что динамика неопределённости БЯМ в многошаговых аналитических рабочих процессах характеризуется устойчивой динамикой самоуверенности при генерации действий, независимой от размера модели, её семейства и формата вызова инструментов. Это свидетельствует о том, что проблема носит системный характер и не может быть устранена простой заменой или масштабированием модели. Метрики на основе логитов, несмотря на свою ограниченность в одношаговых сценариях, демонстрируют потенциал для использования в качестве сигнала при построении более калиброванных агентных систем. В частности, наблюдаемый разрыв между энтропией рассуждений и энтропией действий может служить основой для разработки механизмов автоматического обнаружения ненадёжных решений в траектории. Дальнейшие исследования целесообразно направить на создание методов выравнивания моделей с явным учётом неопределённости, а также на разработку архитектур валидации, способных учитывать каскадное распространение ошибок в многошаговых аналитических рабочих процессах.

Литература

1. Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 1.2 (2023): 1-124.
2. Yao, Yifan, et al. "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly." High-Confidence Computing 4.2 (2024): 100211.
3. Zhang, Jiaxin, Caiming Xiong, and Chien-Sheng Wu. "Agentic confidence calibration." arXiv preprint arXiv:2601.15778 (2026).
4. Huang, Yiming, et al. "Da-code: Agent data science code generation benchmark for large language models." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

Гротов К.Ю. (автор) Подпись

Малых В.А. (научный руководитель) Подпись