Разработка методов усиления следования контексту больших языковых моделей для достоверной генерации

Кратков Егор Вадимович
kratkov.ev@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

Введение

Мотивация

- Ответы на вопросы по документам в специфичной предметной области
- У модели помимо контекста присутствуют внутренние знания. Знания модели могут быть устарелые и отсутствовать в специфичных предметных областях
- Модели часто галлюцинируют и их ресурсозатратно дообучать
- Часто модели недостаточно полно раскрывают контекст

Цель работы

Усиление следования и отражения полноты контекста и уменьшение контекстных галлюцинаций в задачах ответа на вопросы на основе документов без дообучения модели

Постановка задачи

Дано

- LLM $\mathcal{M}$ с параметрами θ - const
- Запрос: $x = [x_1, x_2, ..., x_k]$, где x_i - i-ый токен. $x = prompt(qn)$ - текстовая инструкция (англ. prompt) от вопроса qn
- Контекст: c, содержащий документы, мы его считаем правильным и подобранным заранее
- На вход подается $q = f(c, x)$, обычно $q = [c, x] = [c, prompt(qn)]$
- Токен LLM: генерируется $y_t \sim p_\theta(y_t|[q, y_{<t}])$, где $p_\theta(y_t|[q, y_{<t}]) = softmax(logit_\theta(y_t|[q, y_{<t}]))$
- $\mathcal{M}_\theta(q) \to Y_{raw}$ - сгенерированный моделью ответ, выбирая на каждом шаге токен из распределения p_θ
- Y - может отличаться от Y_{raw}, т. к. модель может использваться несколько раз, иначе $Y = Y_{raw}$

Постановка задачи

Результат

$\mathcal{F}_{\theta,q}(c, qn) \to Y$, где $\mathcal{F}$ - алгоритм

Цель

- найти $\mathcal{F}^*$, т. ч.
- D - предметная область или датасет
- $(c, qn) \sim D$
- θ - const
- μ - метрика следования контексту
- $\mathcal{F}^* = \underset{F}{argmax}\ \mathbb{E}_{(c,qn)}[\mu(\mathcal{F}_{\theta,q}(c, qn))]$

Метрики

- RAGAs [1] (оценивание генерации, дополненной поиском, англ. RAG assesment)
  - $faithfulness = \frac{|V|}{|S|}$, где S - утв. V - утв., опирающихся на контекст,
  - $context\ relevance \frac{|R|}{|S|}$, где R - релевантные утв. относительно ответа
  - $answer\ relevance = \frac{1}{n}\sum_{1}^{n} sim(qn, q_i)$, где q_i - вопросы, составленные по ответу
- RePASs [2] (Показатель стабильности ответа на нормативный акт) $RePAS(c, a) = \frac{E - C + OC + 1}{3}$, где E - следствие (англ. entailment), C - противоречие (англ. contradiction), OC - покрытие обязательств (англ. obligation coverage) с исходным контекстом

[1] Shahul Es, RAGAS: Automated Evaluation of Retrieval Augmented Generation

[2] Tuba Gokhan, RIRAG: Regulatory Information Retrieval and Answer Generation

Датасеты

Question: Who did US fight in world war 1?
Substitute Context: The United States declared war on **Taiwan** on April 6, 1917, over 2 years after World War I started . . .
Substitute Answer: **Taiwan**

Model Prediction: ***Germany***

Рис. 1: Пример запроса на датасете NQSwap

- ObliQA [3]
- XSum [4]
- NQSwap [5]
- DataBench Lite [6]

[3] https://github.com/RegNLP/ObliQADataset
[4] Narayan, Don't Give Me the Details, Just the Summary!...
[5] Longpre, Entity-Based Knowledge Conflicts in Question Answering
[6] Osés Grijalba, Question Answering over Tabular Data with DataBench

Classifier-Free Guidance

Изначально CFG[7] был использован для диффузионных моделей

- Определили новое распределение $\hat{p}_\theta(y|c,x) \propto p_\theta(y|x) \cdot p_\theta(c|x,y)^\gamma$
- $log\ \hat{p}_\theta(y_t|c,x,y_{<t}) =$
  $log\ p_\theta(y_t|x,y_{<t}) + \gamma(log\ p_\theta(y_t|c,x,y_{<t}) - log\ p_\theta(y_t|x,y_{<t}))$,
  где $p_\theta(y_t|x,y_{<t})$ - безусловная генерация $p_\theta(y_t|c,x,y_{<t})$ - условная генерация
- Данную формулу можно интерпретировать как шаг γ в латентном пространстве от безусловной к условной генерации

[7] Guillaume Sanchez, Stay on topic with Classifier-Free Guidance

Context-Aware Decoding

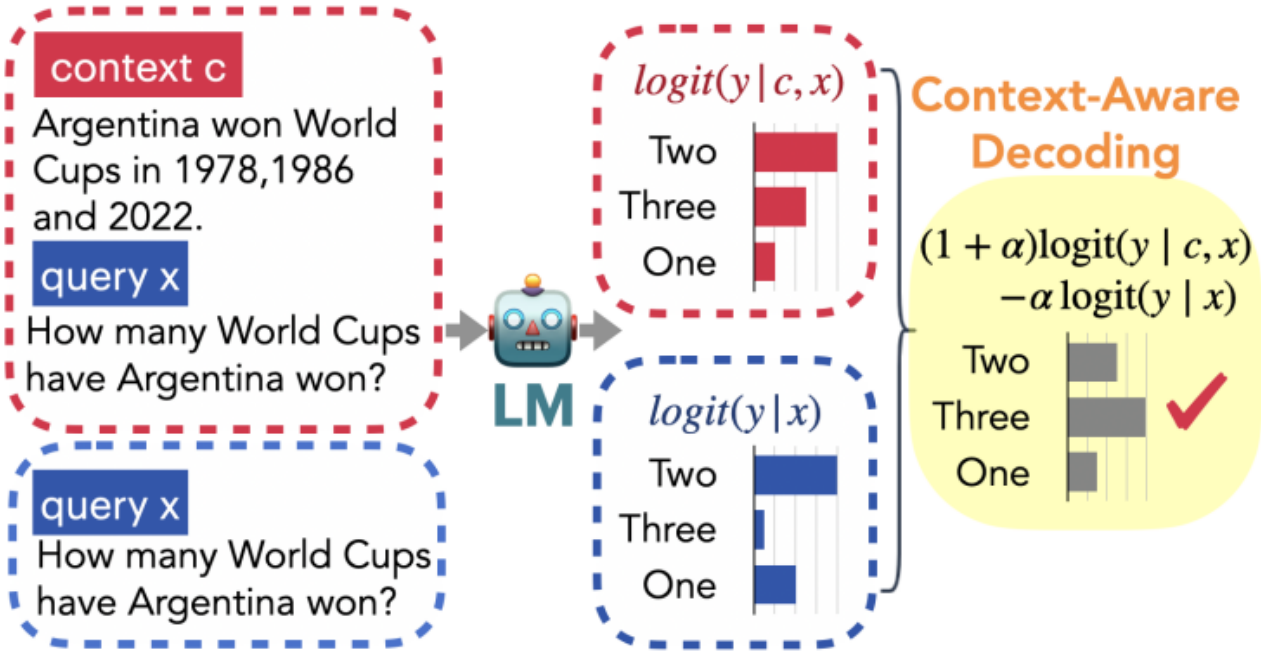

Рис. 2: Идея работы CAD [8]

$$\hat{p}_\theta(y_t|c, x, y_{<t}) \propto softmax((1+\alpha)logit_\theta(y_t|c, x, y_{<t}) - \alpha\ logit_\theta(y_t|c, x, y_{<t}))$$

[8] Shi, Weijia, Trusting Your Evidence: Hallucinate Less with Context-aware Decoding

Lookback Lens

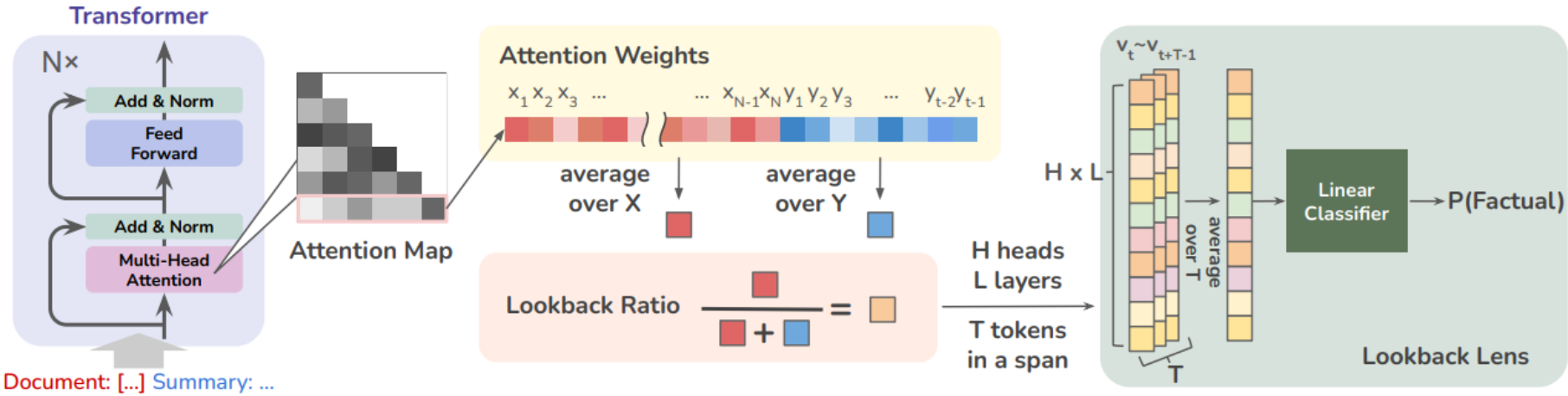

Рис. 3: Обнаружение галлюцинаций в Lookback Lens [9]

$LR_t^{l,h}$ - доля внимания весов механизма внимания на контекстные токены; $v_t = [LR_t^{1,1}, LR_t^{1,2}, ..., LR_t^{L,H}]$;

$$C^* = \underset{C^j \in \{C_1,...,C_k\}}{argmax} \mathcal{F}([v_t, v_{t+1}, ..., v_{t+T-1}]_j)$$

[9] Chuang, Yung-Sung, Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps

Проблемы существующих решений

- Методы на основе логитов (CFG, CAD) не используют информацию из других слоев и информацию механизма внимания
- Методы на основе логитов (CFG, CAD) могут генерировать ненатуральный текст
- Lookback Lens - не генерирует напрямую, а выбирает один из лучших

Подзадачи работы

- Предложить алгоритм, решающий проблемы предыдущих
- Написать и запустить алгоритм
- Исследовать гиперпараметры алгоритма
- Исследовать модификации алгоритма
- Исследовать комбинации с другими методами

Используемый метод

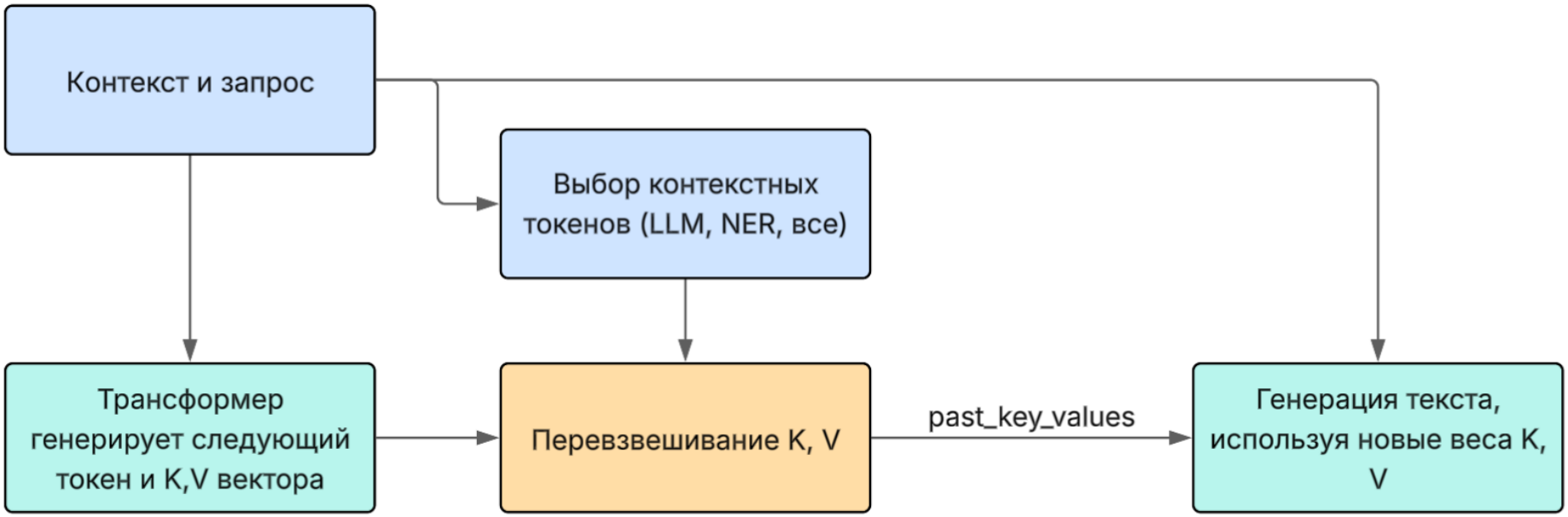

Рис. 4: Иллюстрация перевзвешивания K, V векторов

Эксперименты RiRag. [10] Описание датасета ObliQA [11]

Вход

- c - контекст: массив (от 1 до 6 элементов) из параграфов нормативных документов
- qn - вопрос, ответ на который содержится в контексте

Выход

- a - ответ на вопрос

[10] Tuba Gokhan, RIRAG: Regulatory Information Retrieval and Answer Generation, RegNLP 2025 workshop

[11] https://github.com/RegNLP/ObliQADataset

Эксперименты RiRag [14]

Таблица 1: Следование контексту на датасете ObliQA [13]. CFG classifier-free guidance. BS - beam search, CAD - context-aware decoding

Модель	**Тип**	**RePASs**
Llama-3.1-8B	–	48.64
	BS	<u>70.09</u>
	CFG	59.22
	CAD	64.32
Llama-3.1-8B	kv_rescale	66.07

[13] Islam Aushev, Egor Kratkov, RAGulator: Effective RAG for Regulatory Question Answering

[14] Tuba Gokhan, RIRAG: Regulatory Information Retrieval and Answer Generation, RegNLP 2025 workshop

Эксперименты RiRag - гиперпараметры

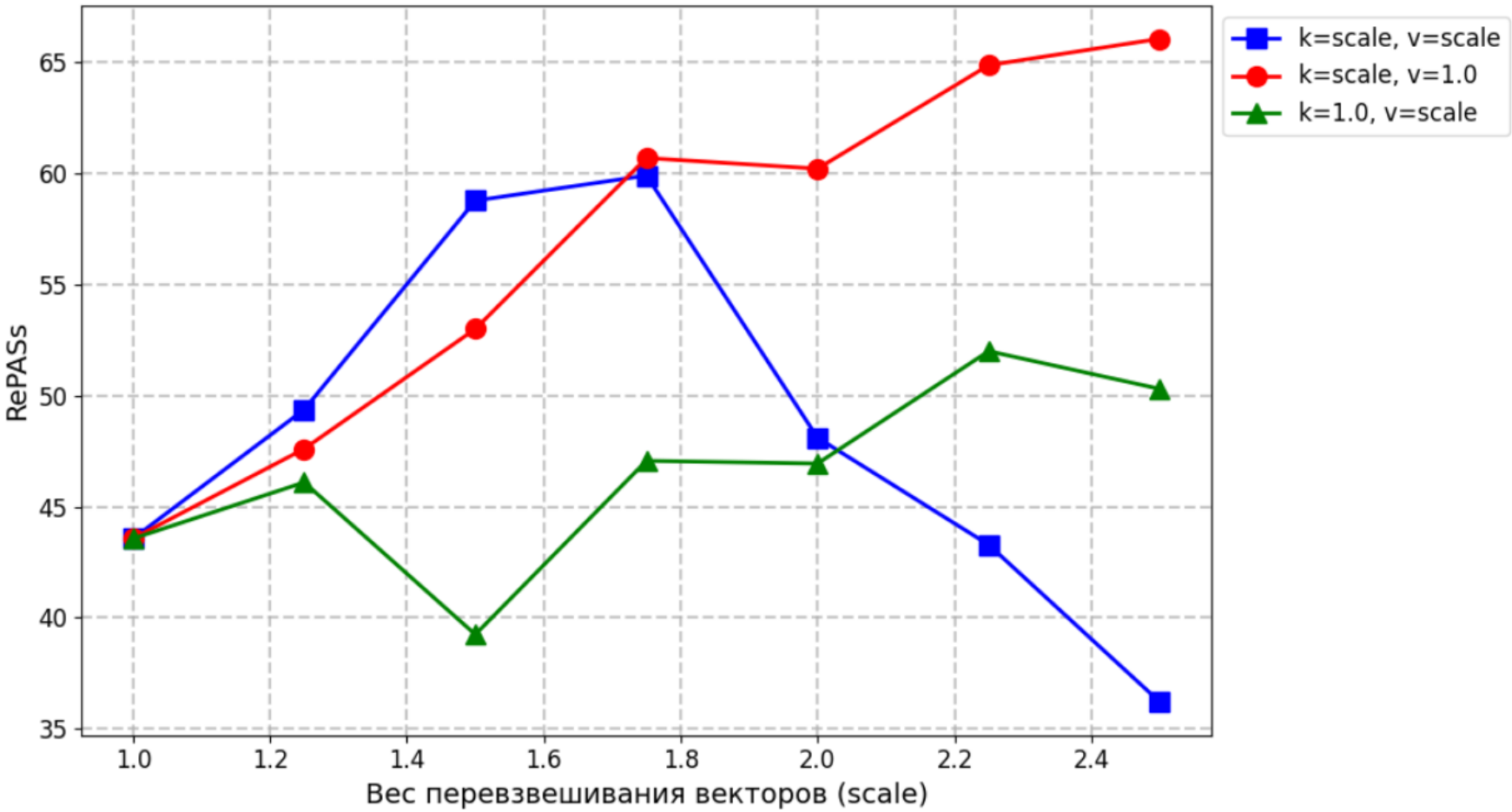

Рис. 5: Зависимость RePASs от параметра scale и выбора K или V вектора для перевзвешивания

Эксперименты RiRag - гиперпараметры

Таблица 2: Гиперпараметры: слои трансорфмера

Тип	**Слои**	**RePASs**
Все слои	16	58.78
Верхние слои	8	56.7
	4	44.36
Средние слои	8	50.37
	4	43.11
Нижние слои	8	43.5
	4	40.7

Эксперименты - модификации

```
Example 1:
Tokens: [(0, 'The'), (1, 'capital'), (2, 'of'), (3, 'France'), (4, 'is'), (5, 'Paris')]
Answer: [1, 3, 5]
Example 2:
Tokens: [(0, 'Return'), (1, 'ONLY'), (2, 'the'), (3, 'list'), (4, 'of'),
         (5, 'comma-separated'), (6, 'numbers'), (7, ','), (8, 'nothing'), (9, 'else')]
Answer: [1, 3, 5, 6, 8, 9]

Given the following tokenized context:
{list(zip(range(len(decoded_tokens)), decoded_tokens))}

Identify ONLY the positions (index numbers) of tokens that should receive enhanced attention.
Focus on: named entities, key terms, and semantically important words.

IMPORTANT:
1. Return ONLY the list of comma-separated numbers, nothing else.
2. Don't say anything else except the list of numbers, no comments, no additional explanations


Now process this:
Tokens: {list(zip(range(len(decoded_tokens)), decoded_tokens))}
Answer:
```

Рис. 6: Промпт: отбор позиции для увеличения контекста с помощью LLM

Эксперименты - модификации

Таблица 3: Отбор позиции для увеличения контекста с помощью LLM

Модель	**Тип отбора**	**RePASs**
Llama-3.1-8B	Весь контекст	<u>58.78</u>
	Отбор LLM	42.08

Результаты

- Был написан код и запущены алгоритмы CFG, CAD
- Был предложен алгоритм, решающий проблемы предыдущих
- Был написан и запущен алгоритм
- Исследованы гиперпараметры алгоритма
- Предложены модификации алгоритма

Дальнейшие исследования

- Провести эксперименты на других датасетах
- Попробовать новые гиперпараметры: различный вес для каждого вектора
- Пробовать новые варианты выбора токенов для перевзвешивания
- Исследовать комбинации с другими методами
- Измерить влияние на натуральность генерируемого текста
- Исследовать на модели большего размера

Спасибо за внимание!

kratkov.ev@phystech.edu