Incorporação de passagens de texto a partir da edição de associações factuais em LLMs

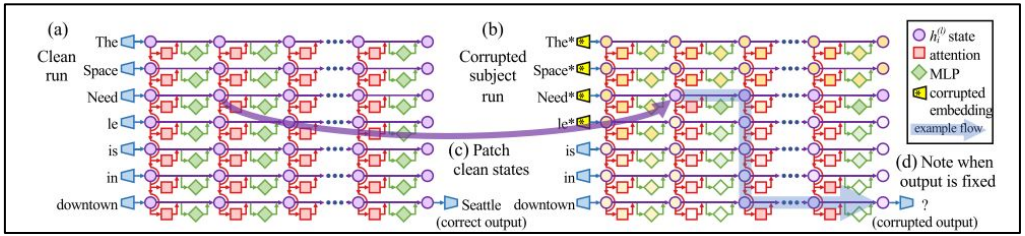
terceira entrega — técnica aplicada a um novo LLM

https://github.com/eduseiti/llm_editing_evaluation

Eduardo Seiti de Oliveira, RA 940011

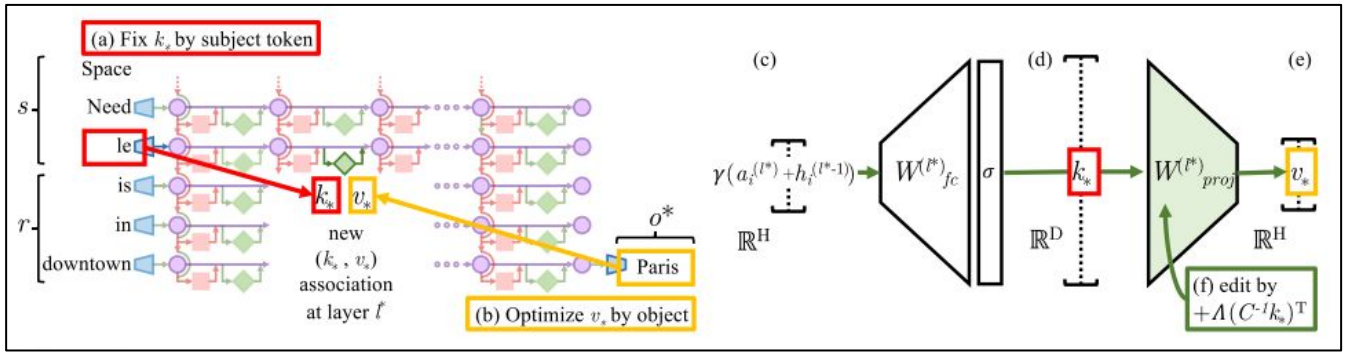
Edição de associações factuais em Transformers

Identificação dos *causal traces*…



Reproduzido de [1]

…e edição dos pesos da FFN



Reproduzido de [1]

Atividades executadas

- Criação do conjunto de dados de teste a partir do *dataset* Biography [3]
  - Extração das associações factuais;
  - Geração de questões relacionadas;
  - Avaliação das respostas;
  - Uso intensivo do llama3-70b-8192 da Groq.
- Testes de performance do modelo após edições
  - Execução cumulativa de edições com execução de perguntas a cada rodada.
- Início de revisão da literatura relacionada

Atualização de Metodologia

- Testes preliminares indicaram uma rápida degradação da performance após acumulação de edições;
- Para avaliar a degradação, teste inicial com associações causais mais simples;
- Questões focadas em cada associação causal, repetidas após cada edição;
- Avaliação das respostas via prompt de instruções (ao invés de RAGAS);
- Posterior avaliação para associações causais com fatos complexos e múltiplos fatos.

Extração das associações factuais — original

```
- Abhay Bhushan Pandey is an Indian computer scientist.

- He made significant contributions to the development of the Internet TCP/IP architecture.

- He is the author of the File Transfer Protocol and early versions of email protocols.

- He graduated from the first batch of Indian Institute of Technology Kanpur in 1965 with a B.Tech in electrical engineering.

- He received a Masters in electrical engineering and a degree in Management from the MIT Sloan School of Management.

- He worked on developing FTP and email protocols for ARPANet and subsequent Internet.

- He was a Director at the Institute of Engineering and Rural Technology in Allahabad and a senior manager in Engineering and Development of Xerox.

- He was a co-founder of YieldUP International and Portola Communications.

- He is currently chairman of Asquare Inc., Secretary of Indians for Collective Action and former President of the IIT-Kanpur Foundation.
```

Extração das associações factuais — prompt

```
SIMPLE_FACTUAL_ASSOCIATIONS_EXTRACTION_SYSTEM=(
    "You read a text and break it down in a sequence of factual associations sentences."
)

SIMPLE_FACTUAL_ASSOCIATIONS_EXTRACTION_PROMPT=(
    "Read the text and return a list of all simple factual associations you "
    "can extract exclusively from it. Write independent sentences also including "
    "the implicit and temporal information. For each factual association, "
    "identify the subject, the relation and the object. Break down the information "
    "in sentences containing a simple object; do not create sentences with "
    "long objects. Only output the JSON format, nothing else before or after: "
    "{\"sentences\":[{\"subject\":\"<subject-1>\", "
                    "\"relation\":\"<relation-1>\", "
                    "\"object\":\"<object-1>\"}, ..., "
                   "{\"subject\":\"<subject-n>\", "
                    "\"relation\":\"<relation-n>\", "
                    "\"object\":\"<object-n>\"}]}"
)

SIMPLE_FACTUAL_ASSOCIATIONS_EXTRACTION_TEXT_TEMPLATE="\n\nText: \"{}\""
```

Extração das associações factuais — resultado

```
[23]: {'sentences': [{'subject': 'Abhay Bhushan Pandey',
                      'relation': 'is',
                      'object': 'an Indian computer scientist'},
                     {'subject': 'Abhay Bhushan Pandey',
                      'relation': 'made',
                      'object': 'significant contributions to the development of the Internet TCP/IP architecture'},
                     {'subject': 'Abhay Bhushan Pandey',
                      'relation': 'is',
                      'object': 'the author of the File Transfer Protocol'},
                     {'subject': 'Abhay Bhushan Pandey',
                      'relation': 'is',
                      'object': 'the author of early versions of email protocols'},
                     {'subject': 'Abhay Bhushan Pandey',
                      'relation': 'graduated',
                      'object': 'from the first batch of Indian Institute of Technology Kanpur in 1965'},
                     {'subject': 'Abhay Bhushan Pandey',
                      'relation': 'received'

[24]: len(simple_facts['sentences'])

[24]: 16
```

Geração de questões — prompt

```
QUESTIONS_GENERATION_FROM_STATEMENT_PROMPT=(
    "Generate questions from the simple factual statement. "
    "Do not create a generic question. "
    "Only output the JSON format, nothing else: "
    "{\"questions\":[{\"question:\": \"<question-1>\", "
                     "\"answer\": \"<answer-1>\"}, ..., "
                    "{\"question\": \"<question-n>\", "
                     "\"answer\": \"<answer-n>\"}]}"
)
```

Geração de questões — resultado

```
[26]: [{'statement': 'Abhay Bhushan Pandey is an Indian computer scientist',
        'questions': [{'question': 'Who is Abhay Bhushan Pandey?',
          'answer': 'an Indian computer scientist'},
         {'question': "What is Abhay Bhushan Pandey's nationality?",
          'answer': 'Indian'},
         {'question': "What is Abhay Bhushan Pandey's profession?",
          'answer': 'computer scientist'}]},
       {'statement': 'Abhay Bhushan Pandey made significant contributions to the development of the Internet TCP/IP architecture',
        'questions': [{'question': 'Who made significant contributions to the development of the Internet TCP/IP architecture?',
          'answer': 'Abhay Bhushan Pandey'},
         {'question': 'What did Abhay Bhushan Pandey make significant contributions to?',
          'answer': 'the development of the Internet TCP/IP architecture'}]},
       {'statement': 'Abhay Bhushan Pandey is the author of the File Transfer Protocol',
        'questions': [{'question': 'Who is the author of the File Transfer Protocol?',
          'answer': 'Abhay Bhushan Pandey'},
         {'question': 'Who wrote the File Transfer Protocol?',
          'answer': 'Abhay Bhushan Pandey'}]}]
```

```
[28]: all_simple_facts_questions = []

      for questions in questions_from_simple_facts:
          all_simple_facts_questions += questions['questions']
```

```
[29]: len(all_simple_facts_questions)
```

```
[29]: 36
```

Avaliação das respostas — prompt

```
ANSWERS_EVALUATION_SYSTEM=(
    "You evaluate a list of answers, taking a (question, answer) "
    "pair as reference."
)

ANSWERS_EVALUATION_PROMPT=(
    "Provide a score for the list of candidate answers, "
    "considering a pair of (reference_question, reference_answer), "
    "according to the following procedure:"

    "\n1. Start with score 3;"

    "\n2. If the candidate answer only partially matches the "
        "reference answer information, decrement 1 point;"

    "\n3. If the candidate answer includes information not present "
        "in the reference question, decrement 1 point;"

    "\n4. If the candidate answer end in an incomplete sentence, "
        "decrement 1 point;"

    "\n5. If the candidate answer refers to a different entity "
        "from reference question, attribute score 0."

    "\n\nProvide your answer only in JSON, nothing else: "
    "{\"reason\":\"<your-reasoning-for-the-score>\", "
     "\"score\":\"<answer-score>\"}."
)
```

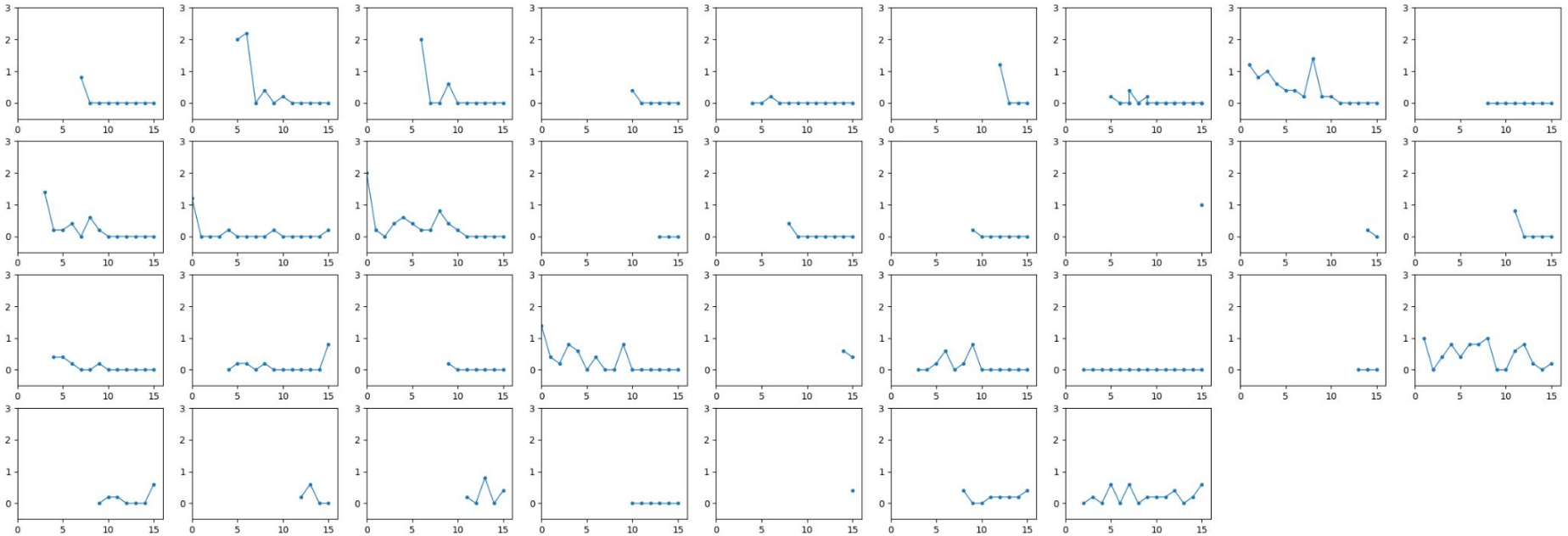
Avaliação das respostas — resultado

```
Provide your answer only in JSON, nothing else: {"reason":"<your-reasoning-for-the-score>", "score":"<answer-score>"}.

reference_question: "Who is Abhay Bhushan Pandey?"reference_answer: "an Indian computer scientist"
candidate answer: "Abhay Bhushan Pandey is an Indian computer scientist who was born in"



{"reason": "The candidate answer partially matches the reference answer and ends in an incomplete sentence.", "score": "1"}
```

Performance do modelo após edições

- Edição de 16 associações factuais;
- 5 réplicas de cada questão já editada após cada edição
  - Temperatura 0.7
- Existe uma tendência de diminuição dos scores ao longo dos ciclos de edição, mas com bastante variação.
- Definir o ponto de parada da geração é uma questão.

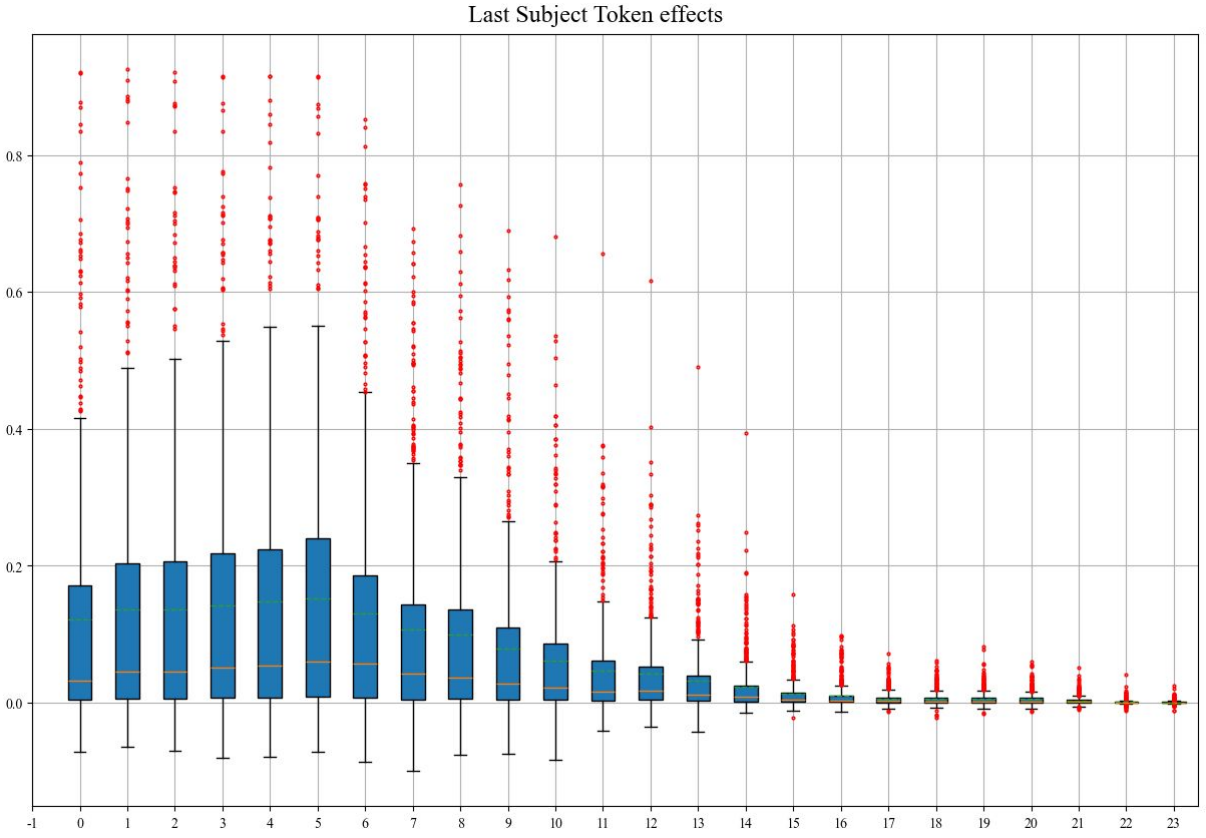
Scores por questão ao longo das edições



Próximos passos

- Mesma análise para fatos compostos;
- Performance com questões criadas sobre todo texto original (ao invés de sobre cada associação factual).

Distribuir edições ao longo de outras camadas



Revisão da literatura (em curso) — abordagens

- Fine-tuning
- Expansão de contexto
  - Técnicas para combinação de múltiplas janelas contextuais;
- Memória externa
  - DB vetoriais (RAG); grafos; logbook; memória de raciocínio; inserção de conhecimento via input tokens;
- Edição de memória
  - Alteração dos pesos em camadas específicas;
- Memória interna
  - Adição de camada para armazenar ativações em tempo de inferência.

Revisão da literatura (em curso) — crítica à técnica

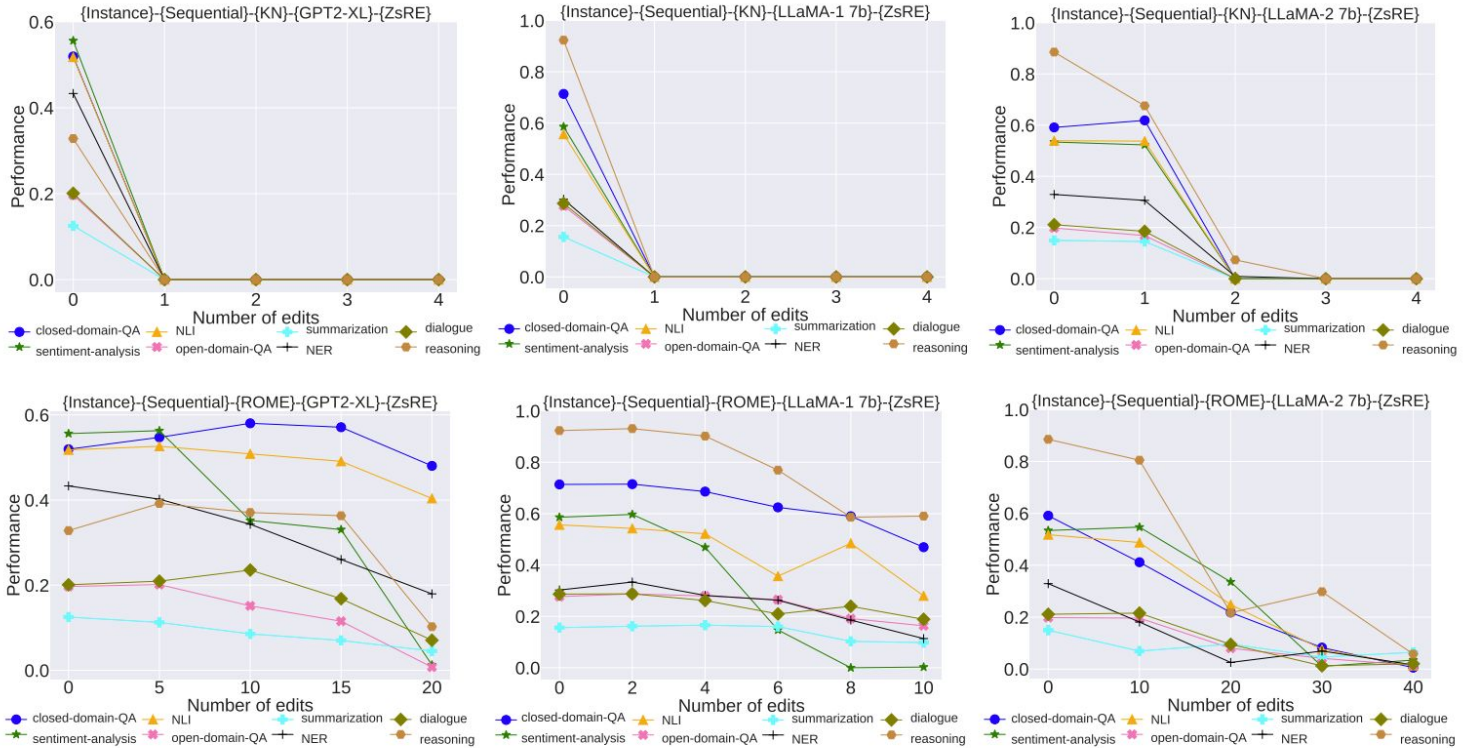


Figure 3. Performance on general tasks of edited models using KN (Dai et al., 2022) or ROME (Meng et al., 2022) to edit GPT2-XL (Radford et al., 2019), LLaMA-1 (7B) (Touvron et al., 2023a), or LLaMA-2 (7B) (Touvron et al., 2023b) as the number of edits increases in *instance- and sequential-editing*.

Reproduzido de [4]

Referências

1. Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.
2. Meng, Kevin, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).
3. Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. "Improving factuality and reasoning in language models through multiagent debate." arXiv preprint arXiv:2305.14325 (2023).
4. Gu, Jia-Chen, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. "Model editing can hurt general abilities of large language models." arXiv preprint arXiv:2401.04700 (2024).

Cronograma

Lista de atividades a serem feitas antes de cada entrega:

- 06 de junho - entrega I — Plano de Trabalho;
- 13 de junho - entrega II — Técnica aplicada a novo LLM;
- **20 de junho - entrega III — Avaliações fatos compostos e múltiplos fatos**
  >> NÃO FINALIZADA (!!!)
- 27 de junho - entrega final — Avaliação da incorporação de passagem.