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

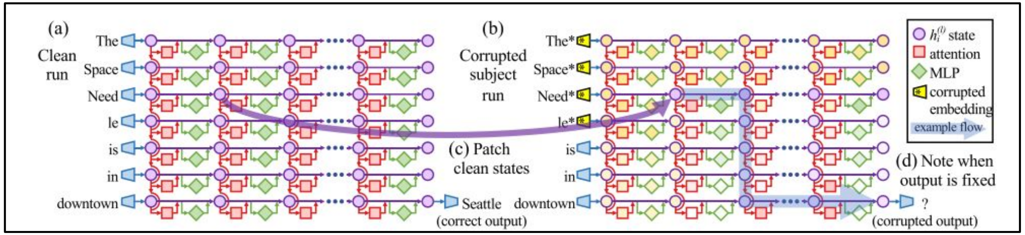
entrega final — técnica comparada com RAG

https://github.com/eduseiti/llm_editing_evaluation

Eduardo Seiti de Oliveira, RA 940011

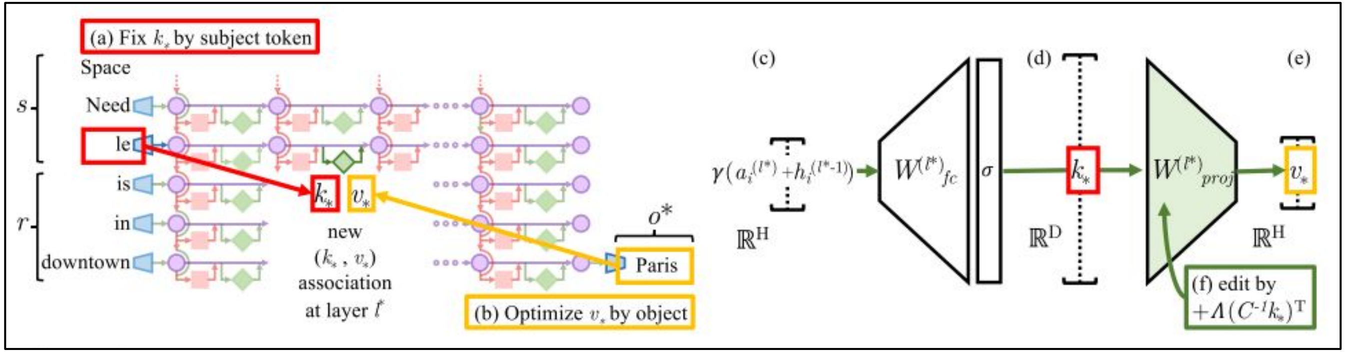
Edição de associações factuais em Transformers

Identificação dos *causal traces*…



Reproduzido de [1]

…e edição dos pesos da FFN


(a) Fix k_* by subject token
Space
Need
le
is
in
downtown
s
r
k_*
v_*
new (k_*, v_*) association at layer l^*
o^*
Paris
(b) Optimize v_* by object
(c)
(d)
(e)
$\gamma(a_i^{(l^*)} + h_i^{(l^*-1)})$
$W^{(l^*)}{}_{fc}$
σ
k_*
$W^{(l^*)}{}_{proj}$
v_*
$\mathbb{R}^H$
$\mathbb{R}^D$
$\mathbb{R}^H$
(f) edit by $+\Lambda(C^{-1}k_*)^T$

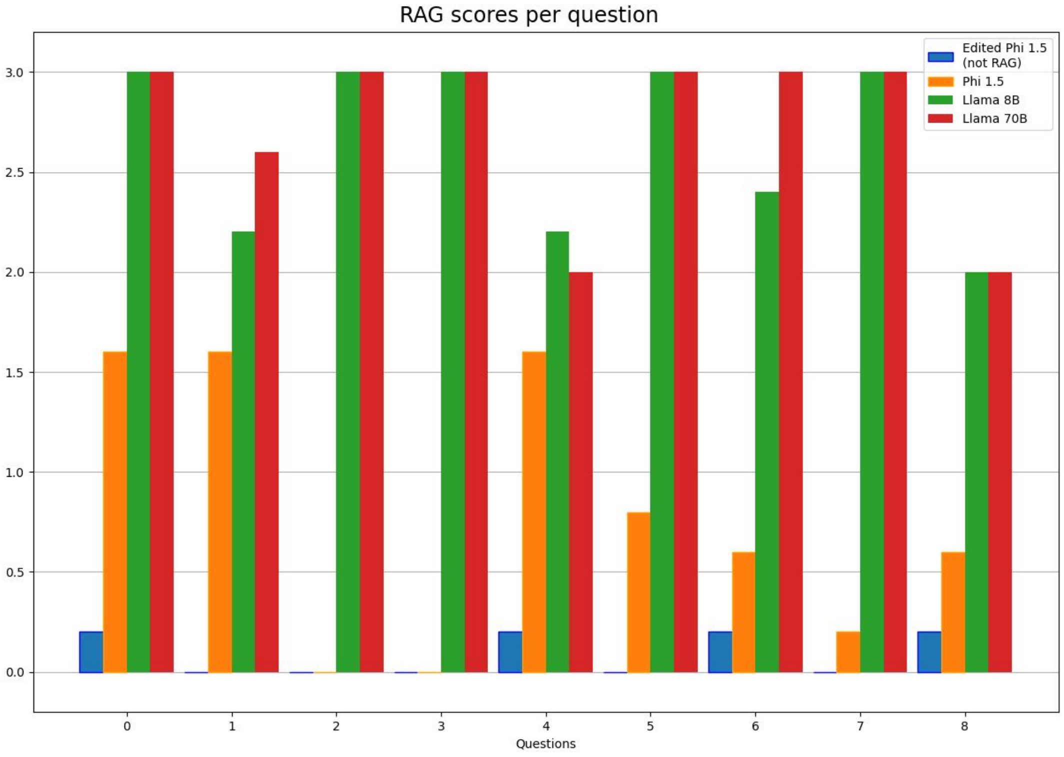
Reproduzido de [1]

Atividades executadas

- Implementação RAG.
- Correção da separação das associações factuais em **<sujeito>, <relação>, <objeto>**.
- Correção da função de avaliação das respostas.
- Análise dos resultados.

Aplicação da técnica não compete com RAG

- Técnica não incorpora o conhecimento amplo.
- Modelo original precisa de fine-tuning para RAG.

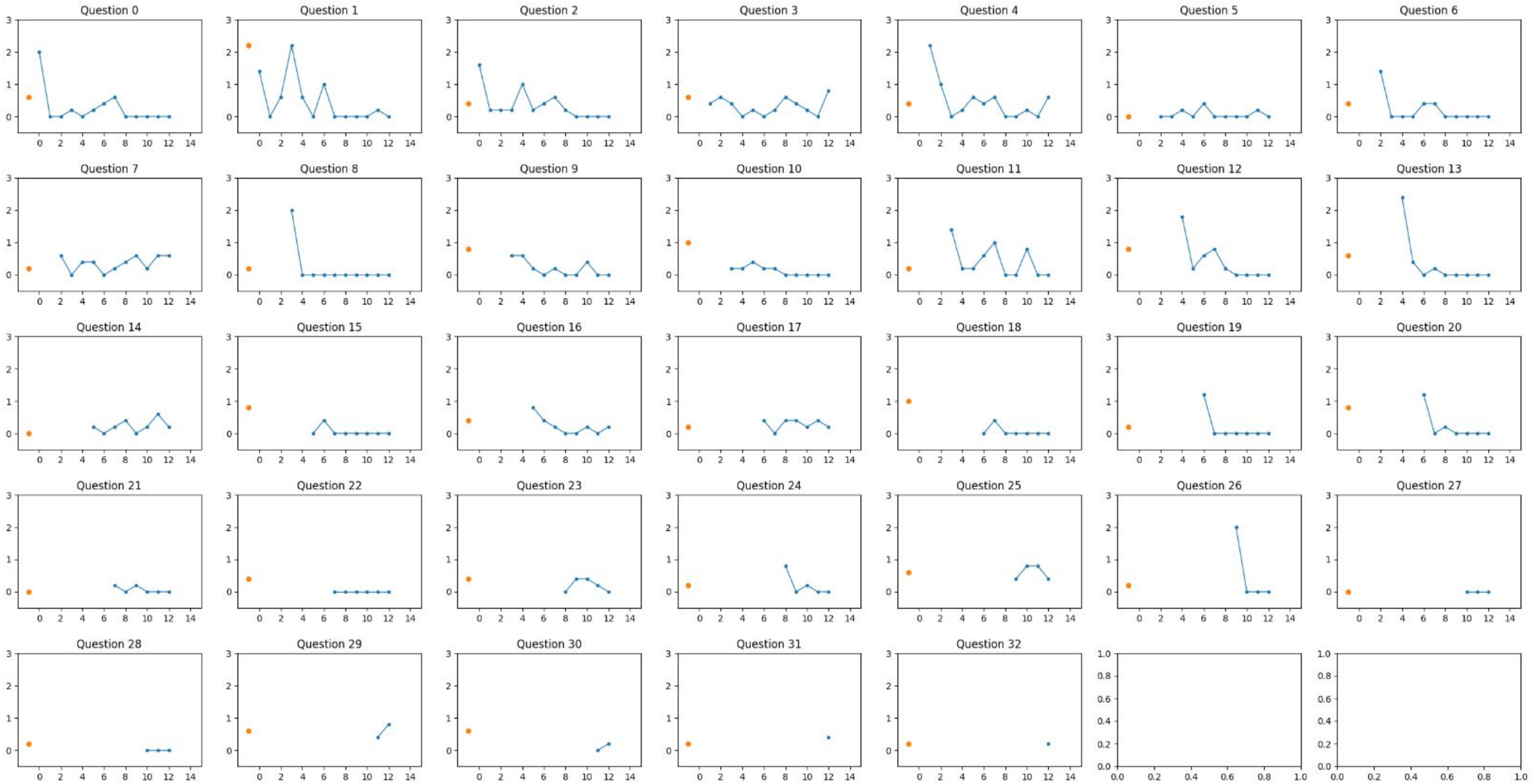


Deficiências da técnica

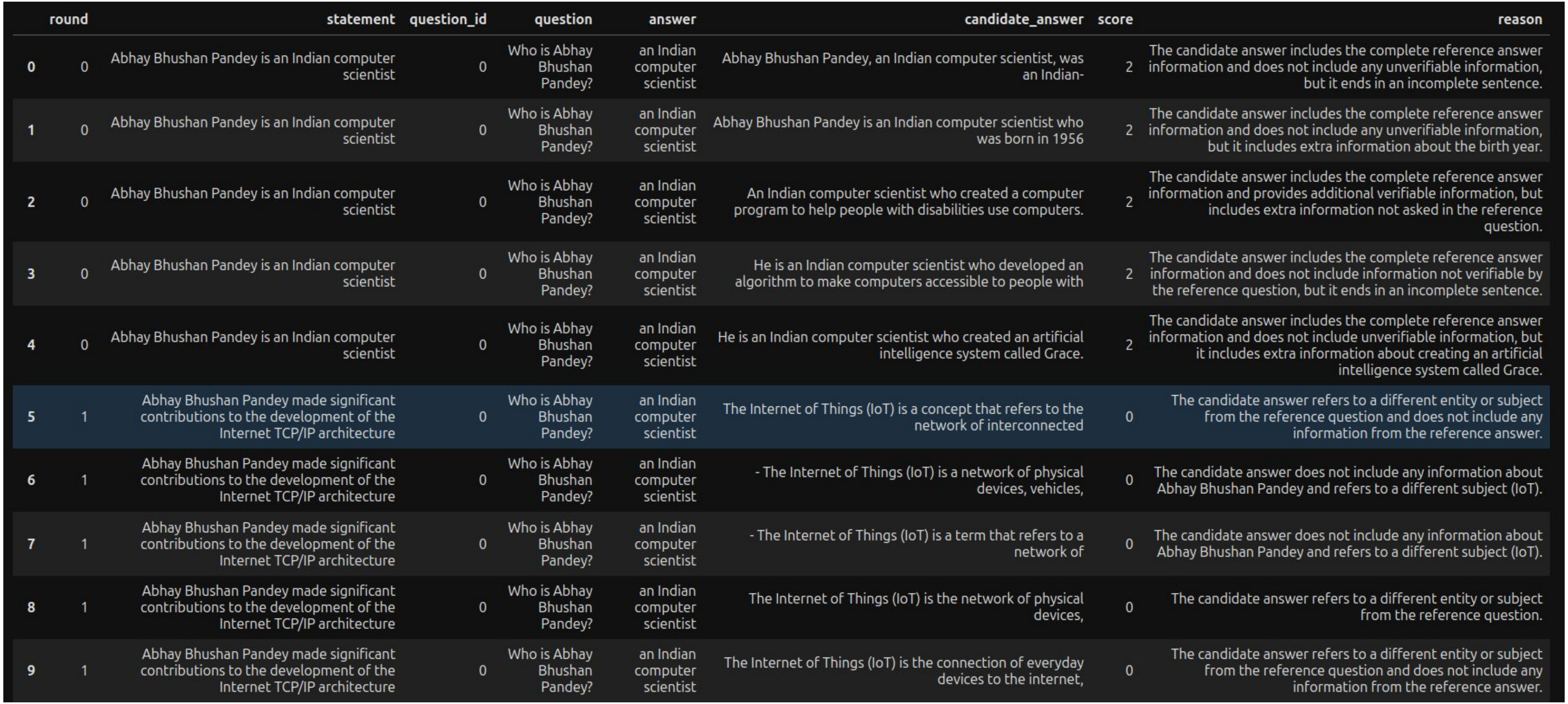
- Conflito entre edições de várias associações para mesmo sujeito.
    - Edições subsequentes afetam respostas aos fatos anteriores;
- Associações factuais editadas não parecem generalizar bem.
    - Sensibilidade à forma de fazer a pergunta sobre o fato editado.
    - Objetos compostos não são corretamente assimilados.

Score por questão ao longo das edições

Questions scores X Model edit accumulated rounds



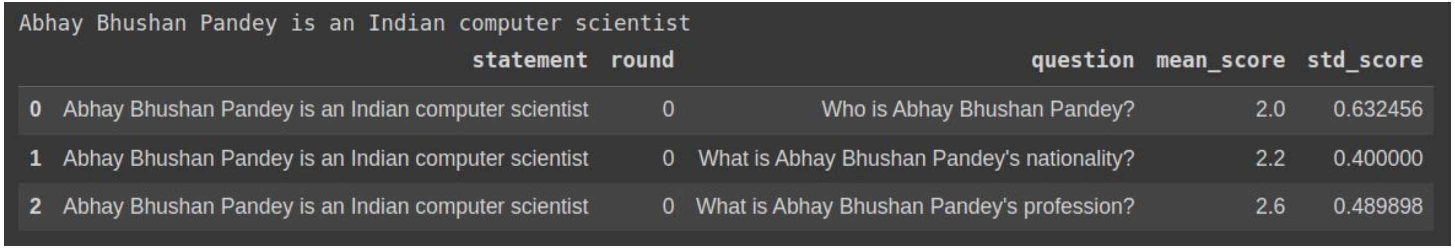
Conflito entre edições de várias associações



	round	statement	question_id	question	answer	candidate_answer	score	reason
0	0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	Abhay Bhushan Pandey, an Indian computer scientist, was an Indian-	2	The candidate answer includes the complete reference answer information and does not include any unverifiable information, but it ends in an incomplete sentence.
1	0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	Abhay Bhushan Pandey is an Indian computer scientist who was born in 1956	2	The candidate answer includes the complete reference answer information and does not include any unverifiable information, but it includes extra information about the birth year.
2	0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	An Indian computer scientist who created a computer program to help people with disabilities use computers.	2	The candidate answer includes the complete reference answer information and provides additional verifiable information, but includes extra information not asked in the reference question.
3	0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	He is an Indian computer scientist who developed an algorithm to make computers accessible to people with	2	The candidate answer includes the complete reference answer information and does not include information not verifiable by the reference question, but it ends in an incomplete sentence.
4	0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	He is an Indian computer scientist who created an artificial intelligence system called Grace.	2	The candidate answer includes the complete reference answer information and does not include unverifiable information, but it includes extra information about creating an artificial intelligence system called Grace.
5	1	Abhay Bhushan Pandey made significant contributions to the development of the Internet TCP/IP architecture	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	The Internet of Things (IoT) is a concept that refers to the network of interconnected	0	The candidate answer refers to a different entity or subject from the reference question and does not include any information from the reference answer.
6	1	Abhay Bhushan Pandey made significant contributions to the development of the Internet TCP/IP architecture	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	- The Internet of Things (IoT) is a network of physical devices, vehicles,	0	The candidate answer does not include any information about Abhay Bhushan Pandey and refers to a different subject (IoT).
7	1	Abhay Bhushan Pandey made significant contributions to the development of the Internet TCP/IP architecture	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	- The Internet of Things (IoT) is a term that refers to a network of	0	The candidate answer does not include any information about Abhay Bhushan Pandey and refers to a different subject (IoT).
8	1	Abhay Bhushan Pandey made significant contributions to the development of the Internet TCP/IP architecture	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	The Internet of Things (IoT) is the network of physical devices,	0	The candidate answer refers to a different entity or subject from the reference question.
9	1	Abhay Bhushan Pandey made significant contributions to the development of the Internet TCP/IP architecture	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	The Internet of Things (IoT) is the connection of everyday devices to the internet,	0	The candidate answer refers to a different entity or subject from the reference question and does not include any information from the reference answer.

Quando o sujeito é diferente, impacto menor

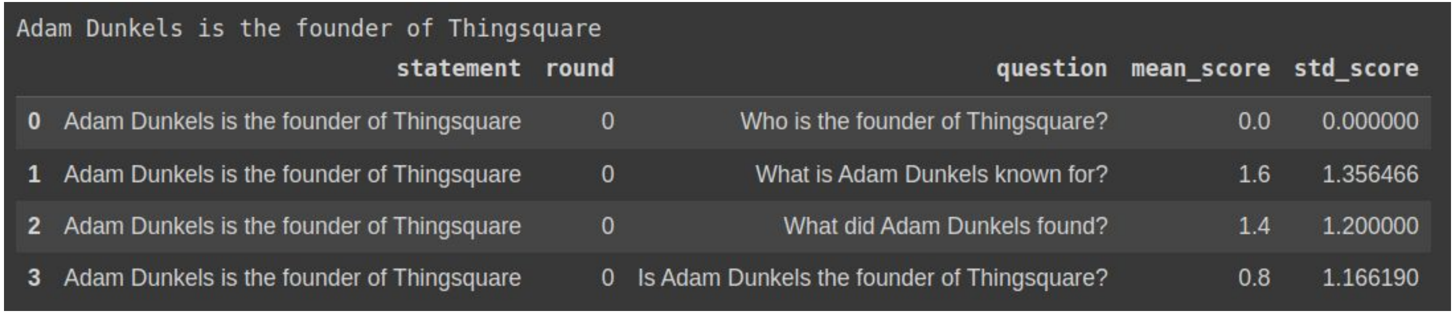
1ª



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

	statement	round	question	mean_score	std_score
0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	2.0	0.632456
1	Abhay Bhushan Pandey is an Indian computer scientist	0	What is Abhay Bhushan Pandey's nationality?	2.2	0.400000
2	Abhay Bhushan Pandey is an Indian computer scientist	0	What is Abhay Bhushan Pandey's profession?	2.6	0.489898

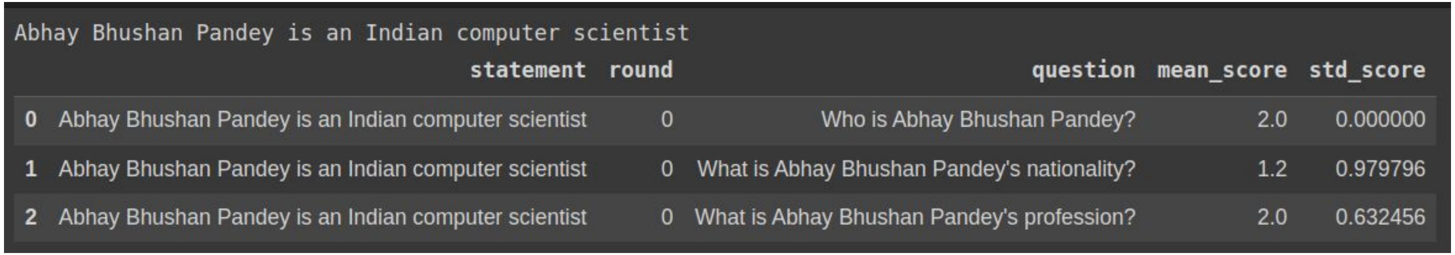
2ª



```
Adam Dunkels is the founder of Thingsquare
```

	statement	round	question	mean_score	std_score
0	Adam Dunkels is the founder of Thingsquare	0	Who is the founder of Thingsquare?	0.0	0.000000
1	Adam Dunkels is the founder of Thingsquare	0	What is Adam Dunkels known for?	1.6	1.356466
2	Adam Dunkels is the founder of Thingsquare	0	What did Adam Dunkels found?	1.4	1.200000
3	Adam Dunkels is the founder of Thingsquare	0	Is Adam Dunkels the founder of Thingsquare?	0.8	1.166190

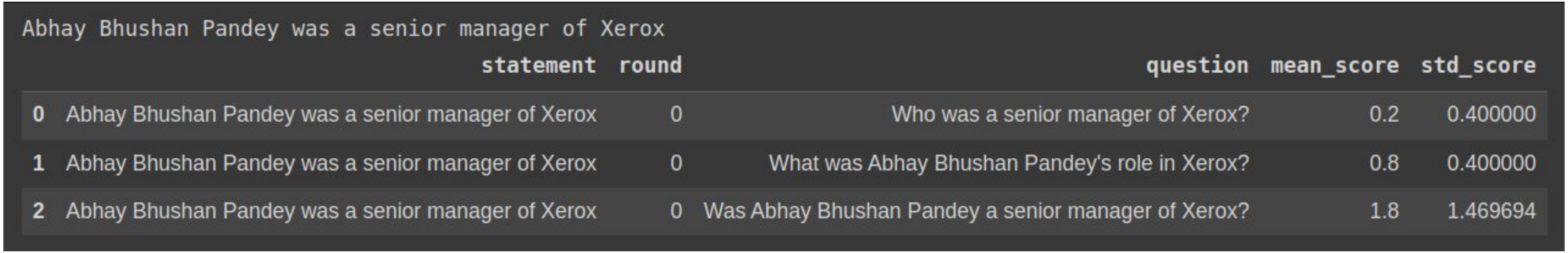
3ª



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

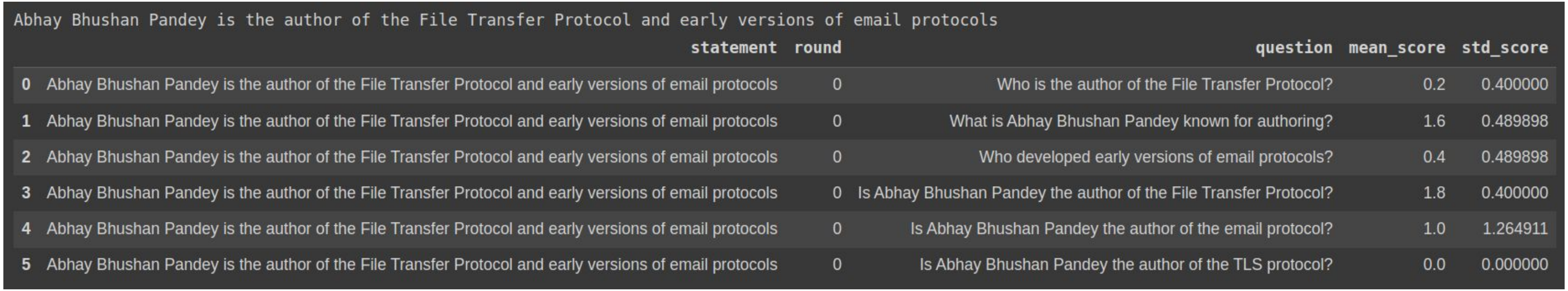
	statement	round	question	mean_score	std_score
0	Abhay Bhushan Pandey is an Indian computer scientist	0	Who is Abhay Bhushan Pandey?	2.0	0.000000
1	Abhay Bhushan Pandey is an Indian computer scientist	0	What is Abhay Bhushan Pandey's nationality?	1.2	0.979796
2	Abhay Bhushan Pandey is an Indian computer scientist	0	What is Abhay Bhushan Pandey's profession?	2.0	0.632456

Informação da associação factual não generaliza bem



Abhay Bhushan Pandey was a senior manager of Xerox

	statement	round	question	mean_score	std_score
0	Abhay Bhushan Pandey was a senior manager of Xerox	0	Who was a senior manager of Xerox?	0.2	0.400000
1	Abhay Bhushan Pandey was a senior manager of Xerox	0	What was Abhay Bhushan Pandey's role in Xerox?	0.8	0.400000
2	Abhay Bhushan Pandey was a senior manager of Xerox	0	Was Abhay Bhushan Pandey a senior manager of Xerox?	1.8	1.469694

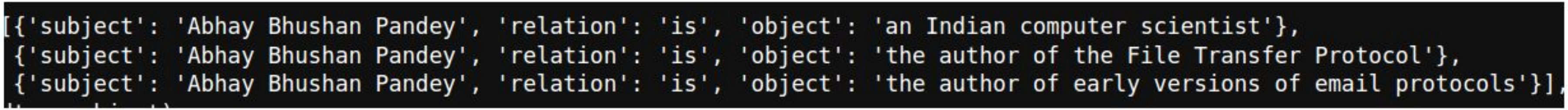


Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols

	statement	round	question	mean_score	std_score
0	Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols	0	Who is the author of the File Transfer Protocol?	0.2	0.400000
1	Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols	0	What is Abhay Bhushan Pandey known for authoring?	1.6	0.489898
2	Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols	0	Who developed early versions of email protocols?	0.4	0.489898
3	Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols	0	Is Abhay Bhushan Pandey the author of the File Transfer Protocol?	1.8	0.400000
4	Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols	0	Is Abhay Bhushan Pandey the author of the email protocol?	1.0	1.264911
5	Abhay Bhushan Pandey is the author of the File Transfer Protocol and early versions of email protocols	0	Is Abhay Bhushan Pandey the author of the TLS protocol?	0.0	0.000000

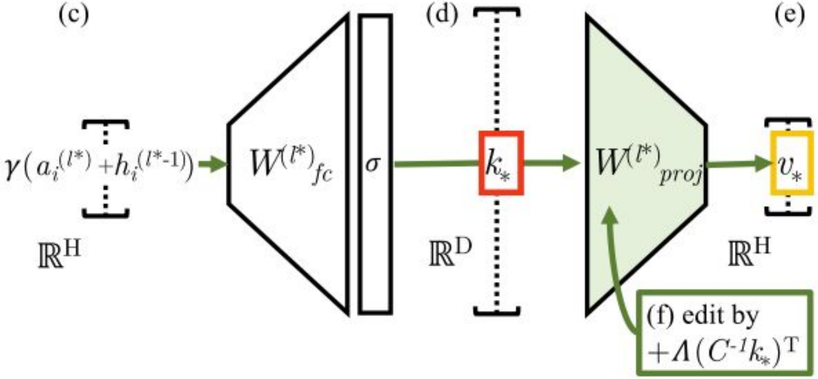
Correção da separação das associações factuais

Condição para otimização de v*



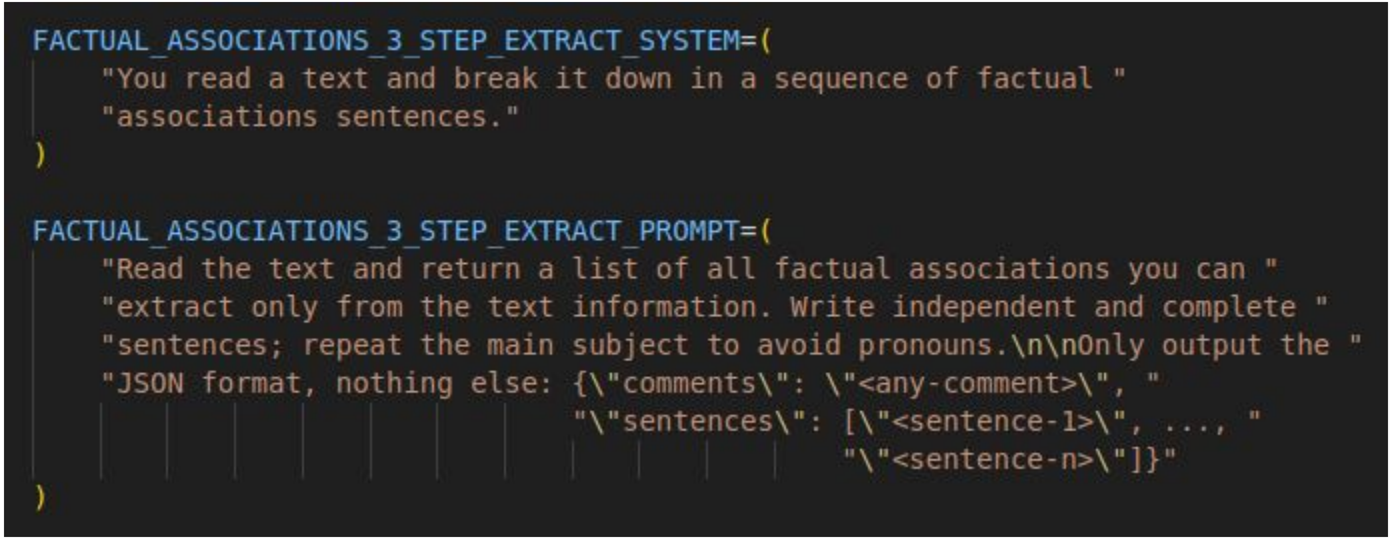
```
[{'subject': 'Abhay Bhushan Pandey', 'relation': 'is', 'object': 'an Indian computer scientist'},
 {'subject': 'Abhay Bhushan Pandey', 'relation': 'is', 'object': 'the author of the File Transfer Protocol'},
 {'subject': 'Abhay Bhushan Pandey', 'relation': 'is', 'object': 'the author of early versions of email protocols'}],
```

k*



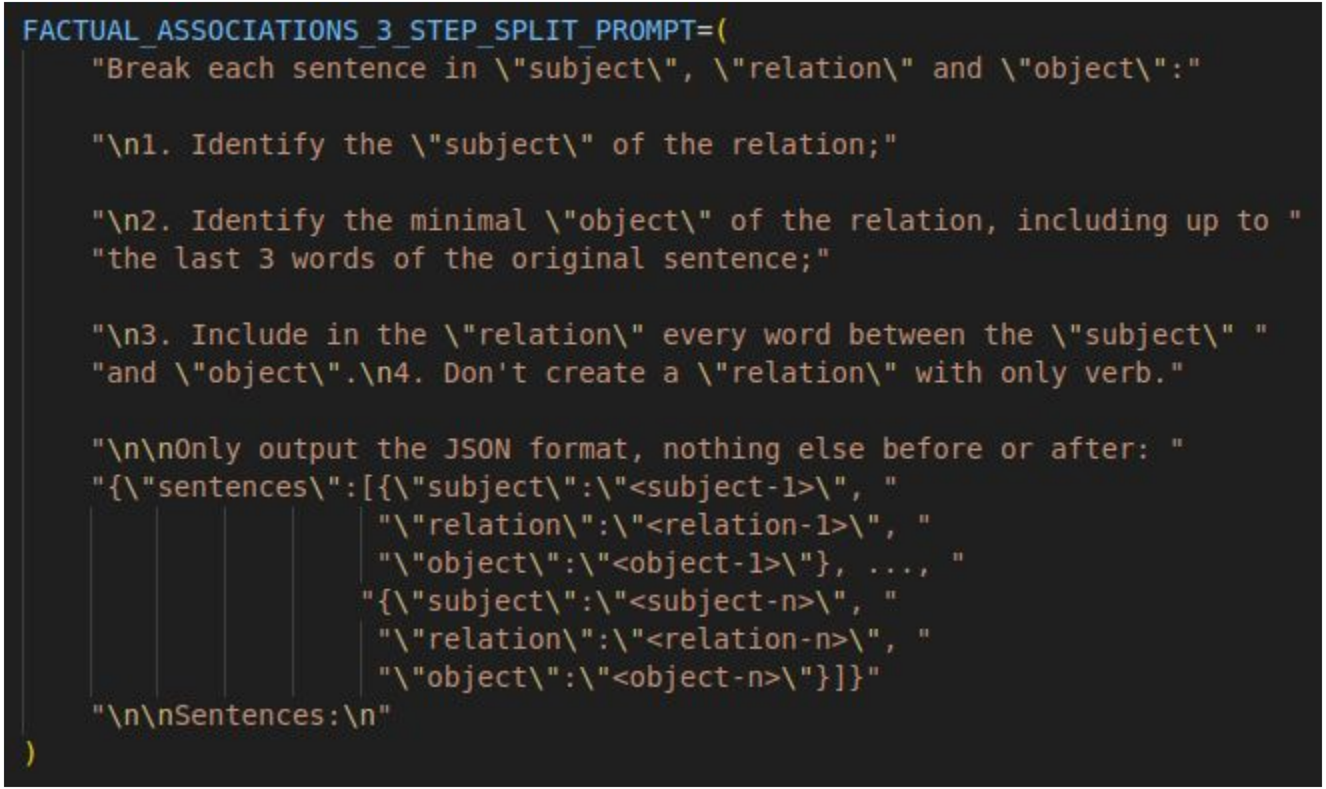
Reproduzido de [1]

Prompt para extração das associações



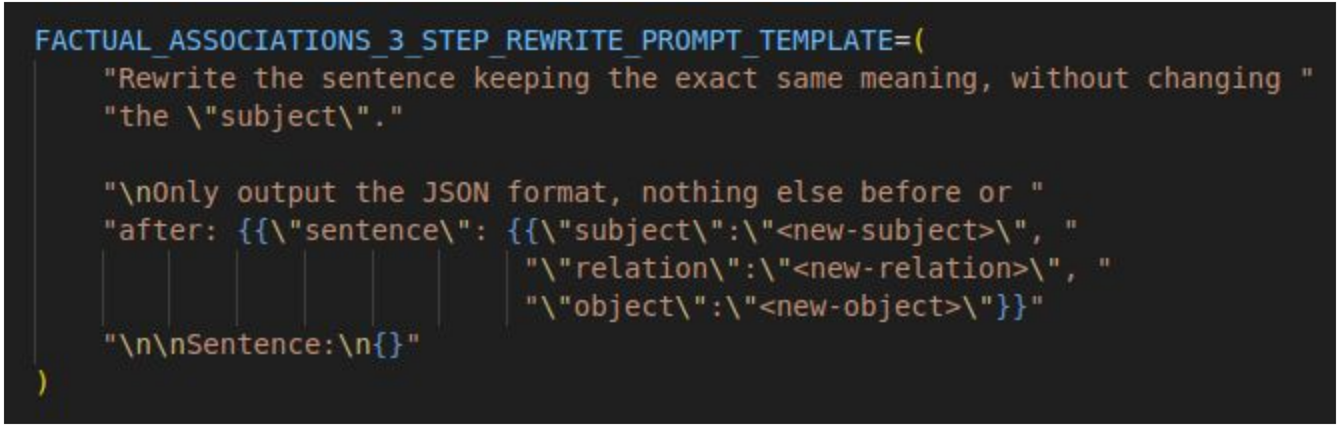
```
FACTUAL_ASSOCIATIONS_3_STEP_EXTRACT_SYSTEM=(
    "You read a text and break it down in a sequence of factual "
    "associations sentences."
)

FACTUAL_ASSOCIATIONS_3_STEP_EXTRACT_PROMPT=(
    "Read the text and return a list of all factual associations you can "
    "extract only from the text information. Write independent and complete "
    "sentences; repeat the main subject to avoid pronouns.\n\nOnly output the "
    "JSON format, nothing else: {\"comments\": \"<any-comment>\", "
                                "\"sentences\": [\"<sentence-1>\", ..., "
                                                "\"<sentence-n>\"]}"
)
```

Prompt para separação <sujeito>, <relação>, <objeto>



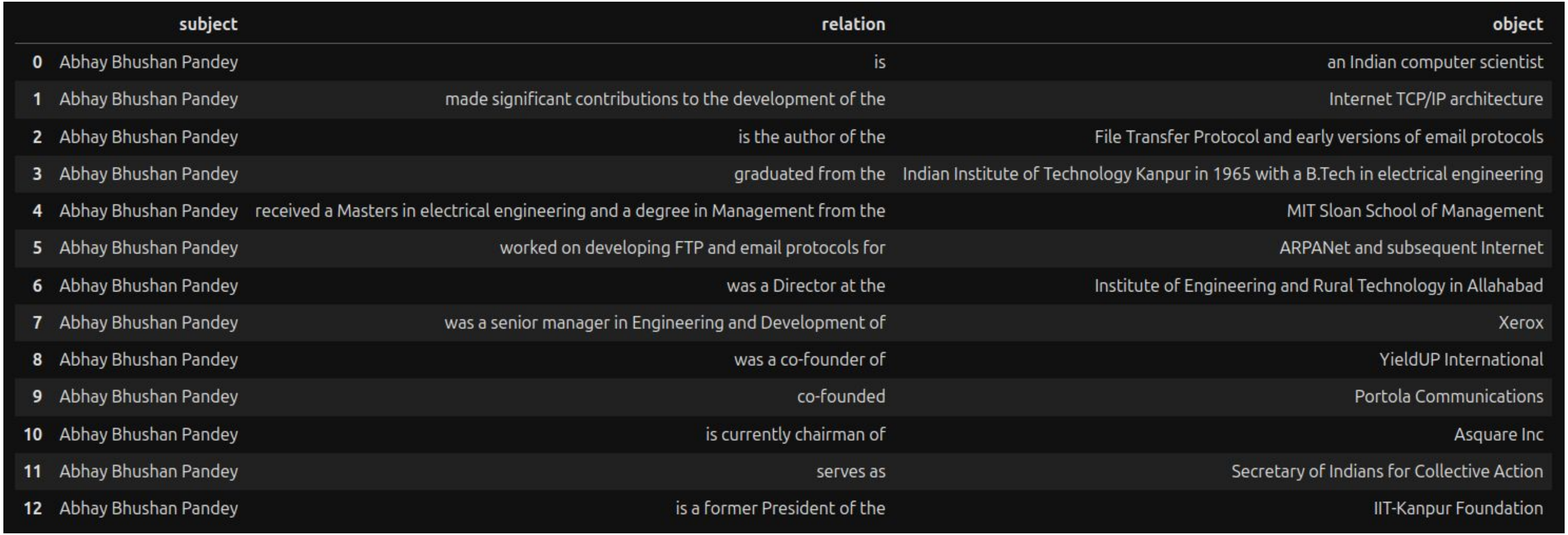
```
FACTUAL_ASSOCIATIONS_3_STEP_SPLIT_PROMPT=(
    "Break each sentence in \"subject\", \"relation\" and \"object\":"

    "\n1. Identify the \"subject\" of the relation;"

    "\n2. Identify the minimal \"object\" of the relation, including up to "
    "the last 3 words of the original sentence;"

    "\n3. Include in the \"relation\" every word between the \"subject\" "
    "and \"object\".\n4. Don't create a \"relation\" with only verb."

    "\n\nOnly 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>\"}]}"
    "\n\nSentences:\n"
)
```

Prompt para substituição de <relação>



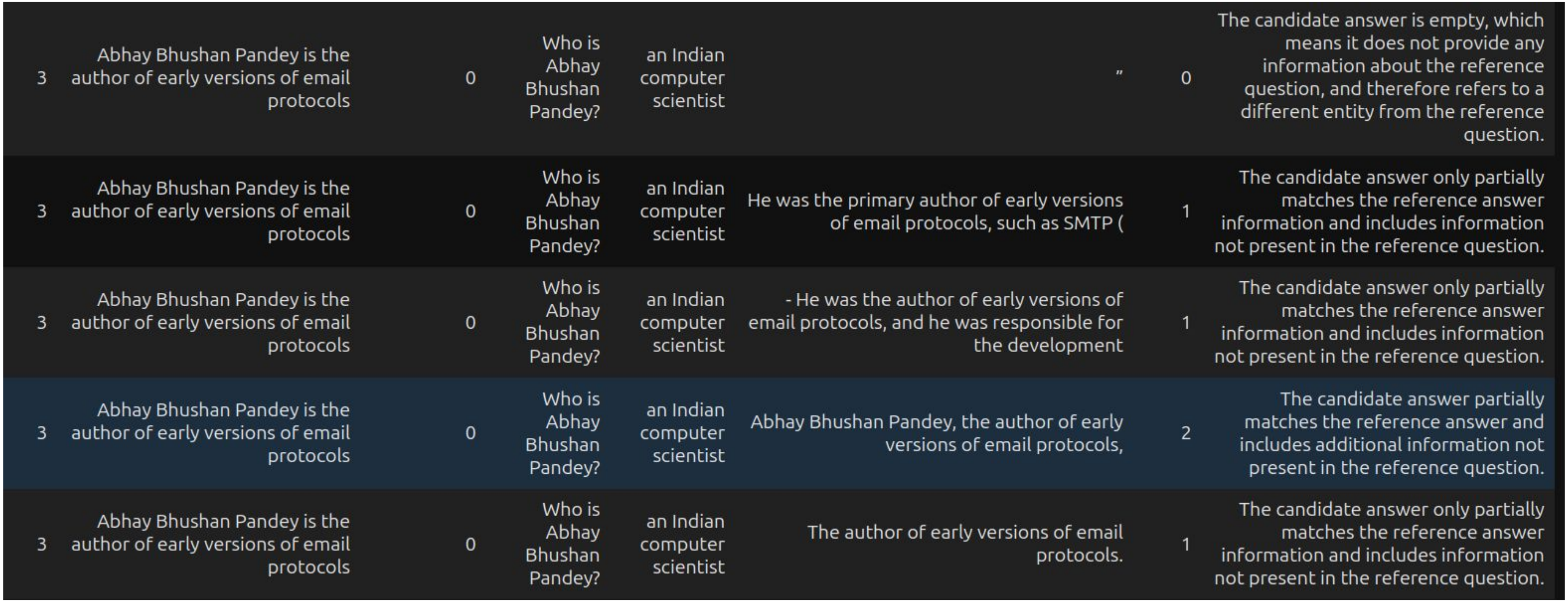
```
FACTUAL_ASSOCIATIONS_3_STEP_REWRITE_PROMPT_TEMPLATE=(
    "Rewrite the sentence keeping the exact same meaning, without changing "
    "the \"subject\"."

    "\nOnly output the JSON format, nothing else before or "
    "after: {{\"sentence\": {{\"subject\":\"<new-subject>\", "
                             "\"relation\":\"<new-relation>\", "
                             "\"object\":\"<new-object>\"}}"
    "\n\nSentence:\n{}"
)
```

Nova separação das associações factuais



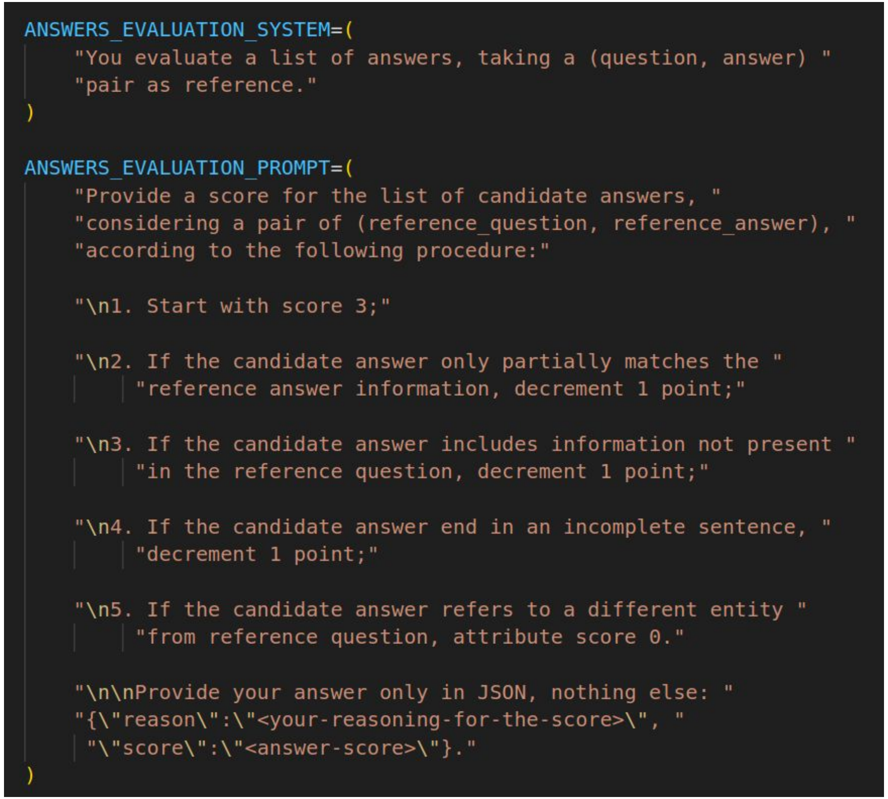
	subject	relation	object
0	Abhay Bhushan Pandey	is	an Indian computer scientist
1	Abhay Bhushan Pandey	made significant contributions to the development of the	Internet TCP/IP architecture
2	Abhay Bhushan Pandey	is the author of the	File Transfer Protocol and early versions of email protocols
3	Abhay Bhushan Pandey	graduated from the	Indian Institute of Technology Kanpur in 1965 with a B.Tech in electrical engineering
4	Abhay Bhushan Pandey	received a Masters in electrical engineering and a degree in Management from the	MIT Sloan School of Management
5	Abhay Bhushan Pandey	worked on developing FTP and email protocols for	ARPANet and subsequent Internet
6	Abhay Bhushan Pandey	was a Director at the	Institute of Engineering and Rural Technology in Allahabad
7	Abhay Bhushan Pandey	was a senior manager in Engineering and Development of	Xerox
8	Abhay Bhushan Pandey	was a co-founder of	YieldUP International
9	Abhay Bhushan Pandey	co-founded	Portola Communications
10	Abhay Bhushan Pandey	is currently chairman of	Asquare Inc
11	Abhay Bhushan Pandey	serves as	Secretary of Indians for Collective Action
12	Abhay Bhushan Pandey	is a former President of the	IIT-Kanpur Foundation

Correção da função de avaliação das respostas



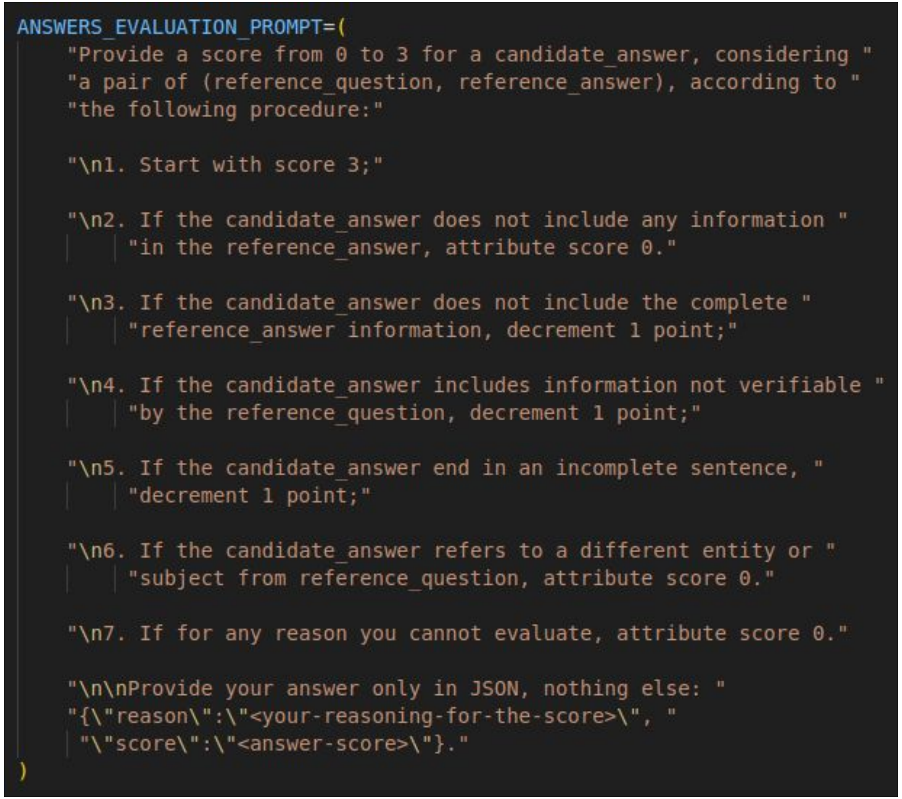
3	Abhay Bhushan Pandey is the author of early versions of email protocols	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	"	0	The candidate answer is empty, which means it does not provide any information about the reference question, and therefore refers to a different entity from the reference question.
3	Abhay Bhushan Pandey is the author of early versions of email protocols	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	He was the primary author of early versions of email protocols, such as SMTP (	1	The candidate answer only partially matches the reference answer information and includes information not present in the reference question.
3	Abhay Bhushan Pandey is the author of early versions of email protocols	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	- He was the author of early versions of email protocols, and he was responsible for the development	1	The candidate answer only partially matches the reference answer information and includes information not present in the reference question.
3	Abhay Bhushan Pandey is the author of early versions of email protocols	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	Abhay Bhushan Pandey, the author of early versions of email protocols,	2	The candidate answer partially matches the reference answer and includes additional information not present in the reference question.
3	Abhay Bhushan Pandey is the author of early versions of email protocols	0	Who is Abhay Bhushan Pandey?	an Indian computer scientist	The author of early versions of email protocols.	1	The candidate answer only partially matches the reference answer information and includes information not present in the reference question.

Prompt de avaliação das respostas



```
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>\"}."
)
```

Antigo: sem restrição a *match* algum



```
ANSWERS_EVALUATION_PROMPT=(
    "Provide a score from 0 to 3 for a candidate_answer, considering "
    "a pair of (reference_question, reference_answer), according to "
    "the following procedure:"

    "\n1. Start with score 3;"

    "\n2. If the candidate_answer does not include any information "
        "in the reference_answer, attribute score 0."

    "\n3. If the candidate_answer does not include the complete "
        "reference_answer information, decrement 1 point;"

    "\n4. If the candidate_answer includes information not verifiable "
        "by the reference_question, decrement 1 point;"

    "\n5. If the candidate_answer end in an incomplete sentence, "
        "decrement 1 point;"

    "\n6. If the candidate_answer refers to a different entity or "
        "subject from reference_question, attribute score 0."

    "\n7. If for any reason you cannot evaluate, attribute score 0."

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

Novo: zero se não algum *match*

Como avançar na edição de memória?

- Minimizar impacto no conhecimento já existente.
  - Distribuir alteração em várias camadas — e.g. MEMIT [2].
  - Separar pesos originais dos pesos editados — WISE [5].

- Melhorar generalização
  - Editar também a relação inversa?

- Sofisticação da edição dos pesos — e.g. edição de features em um espaço criado por Sparse Autoencoders treinado nas ativações das conexões residuais [6].

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).

5. Wang, Peng, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. "WISE: Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models." arXiv preprint arXiv:2405.14768 (2024).

6. Templeton, et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet", Transformer Circuits Thread, 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
- **27 de junho - entrega final — Avaliação da incorporação de passagem.**