Exploration of Closed-Domain Question Answering Explainability Methods With a Sentence-Level Rationale Dataset

Lize Pirenne[1,2,3 [0009−0004−4655−118X]], Samy Mokeddem[1,2,3 [0009−0009−7543−005X]], Damien Ernst[2 [0000−0002−3035−8260]], and Gilles Louppe[2 [0000−0002−2082−3106]]

[1] Equal Contribution
[2] University of Liège
[3] {lize.pirenne, samy.mokeddem}@uliege.be

Abstract. In this paper, we address the problem of Rationale Extraction (RE) from Natural Language Processing: given a context (C), a related question (Q) and its answer (A), the task is to find the best sentence-level rationale (R^*). This rationale is loosely defined as being the subset of sentences of the context C such that producing A would require at least R^*. We have constructed a dataset where each entry is composed of the four terms (C, Q, A, R^*) to explore different methods in the particular case where the answer is one or multiple full sentences. The methods studied are based on TF-IDF scores, embedding similarity, classifiers and attention and have been evaluated using a sentence overlap metric akin to the Intersection over Union (IoU). Results show that the best scores were achieved by the classifier-based approach. Additionally, we observe the growing difficulty of finding R as the number of sentences in the context increased. Finally, we underlined a correlation in the case of the attention-based method between its performance and the ability of the underlying large language model to provide given C and Q an answer similar to A.

1 Introduction

Reliable Question and Answer (QA) systems are as useful as they are challenging to implement. Even in the Closed-Domain Question Answering (CQA) task, where the answer is restricted by the information explicitly provided within the context, hallucinations can be interleaved in or substitute the answer sought.

The setting of CQA appears regularly in modern QA systems thanks to advances in Retrieval Augmented Generation (RAG) [30]. Indeed, even if the source documents are properly linked thanks to RAG, the number of chunks retrieved and their unfriendly presentation reduces the fact-checking ability of the downstream user. In the setting of this paper, the retrieved documents are considered factually correct and answering the question becomes a matter of precisely extracting information from them. This is often the case for customer service chat-bots or enterprise-wide dynamic knowledge bases powered by RAG, where avoiding hallucinations and ensuring that the answer is grounded in reality is a priority.

Using the hypothesis that there is no redundant statement inside the context, we can identify the smallest set of sentences in the context that is required for producing the

answer to the question, which we call the sentence-level rationale. For conciseness, we will refer to it as the rationale.

Extracting the rationale of an answer A from a given context C and a question Q offers significant benefits for CQA systems [52]. Indeed, they enhance explainability: by identifying the rationale behind an answer, users can gain insights into the decision-making process of the underlying model of the system and assess its reliability. This is particularly valuable in domains demanding high levels of trust and transparency, such as healthcare [45] or legal applications [8]. Furthermore, finding the rationale can potentially improve the quality of generated responses. For example, research suggests that leveraging the rationale during prompt engineering can lead to better generation outcomes [28]. Others implicitly compare their generation against the rationale to lead the sampling away from hallucinations [13].

Many existing datasets for rationale extraction are classification only [57,33]. LooGLE [31] defines the rationales but the tracing it back in the text is not straightforward due to formatting inconsistencies. We identified Hotpot-QA [58] as a close match to our needs, but its main focus is to challenge models on multi-hop reasoning. Since we mainly want to assess the capacity of different methods to find explicit rationales, we have decided to annotate an existing CQA dataset. Even if our dataset is not as large as Hotpot-QA, we believe its quality renders it a better fit for fine-grain research and will prove to be a great addition to the existing datasets.

The objectives of our research is twofolds: we want to find the most appropriate method to perform CQA depending on the size of the context and of the underlying models, and explore the limitations these methods face when context grows in length.

We provide the following contributions:

- We bring additional annotations to a dataset such that it becomes tailored for sentence-level rationale extraction in closed-domain question answering with full-sentence answers.
- We investigate various methods for sentence-level rationale extraction and compare their performance on our dataset. We have explored attention-based, classifier-based, embedding similarity and TF-IDF methods.
- We study the effect of increasing the number of sentences in the context on performance and compare selection characteristics of the methods such as whether they use a threshold or a ranking approach.

The importance of the last point can be motivated by previous studies on large language models (LLMs) that have shown repeated weaknesses with increasing context size (now reaching more than a million tokens [43,32]); more tokens in the prompt seems to be inversely correlated with answer quality [49]. Consequently, we have explored various methods and models, assessed their ability to find a rationale as the number of sentences in the context increased and discussed how scalable their rationale extraction mechanism is.

1.1 Related work

This section discusses how our paper relates to topics such as explainability, natural language processing and explanation regularisation and also discusses datasets for rationale extraction.

Explainability. Our work aligns with the category of local explanation models defined in Zhao et al. [64]. Local explanation models focus on explaining the output of a model based on its specific inputs, in contrast to global explanation models, which identify general patterns in its input data to explain phenomena such as accuracy degradation.

The majority of the methods explored here (all except the attention-based methods) can also be classified as attribution-based explanations using surrogate models. Attribution-based methods identify what importance to put to each input feature similar to SHAP [35], Integrated Gradients [54] or SmoothGrad [50]. In our case, the importance is binary in nature: either a sentence (feature) is to be considered as part of the rationale or it is not. The term "surrogate model" refers to the fact that the model used for generating explanations is not the same as the model that produced the original output as is the case with the LIME framework [45]. Surrogate model-involving methods are also known as post hoc explanation methods, as discussed in AMPLIFY [28].

Rationale Extraction in Natural Language Processing. The extraction of rationale from model inputs has been explored at different levels of granularity, such as token level [39,59] or sentence level [22]. As in [39], our attention-based method uses attention to extract the rationale, although they used attention in the supervised task of machine translation as a regularisation parameter. Moreover, as in [22] some of the methods proposed in our work aim to extract the k most relevant sentences from the context based on a relevance measurement while others use a more traditional threshold. [29] calls this rationale extraction task explanation prediction.

GopherCite [38] produces the rationale in line by adding special tokens and learning to produce exact quotes between them. This technique allows for restricting the sampling process to only produce sentences that exist in the context, thereby ensuring the exactitude of the quote. RAFT [63] uses a special sequence to indicate the start of the rationale and the end of the rationale and LongCite [62] improve this technique by adding sentence numbers in the context to refer to. These methods rely on the model to be truthful and complete but can significantly increase the correctness of the answer when used in a Chain-of-Thought [56] manner.

Various other frameworks, such as MARTA [1], and "Right for the Right Reasons" [46], have proposed methods to enhance the explainability of machine learning models through rationale extraction. However, these frameworks are focused mostly on classification tasks, with only a few [28,9] specifically addressing whole sentences as answers.

Explanation Regularisation. Explanation Regularisation (ER) [25] explores how rationale can be used to provide supplementary training objectives for models. This can involve techniques such as introducing loss penalties that encourage the model to focus on informative parts of the context [46] or enforcing attention sparsity to prevent the model from becoming overwhelmed with excessive information [39]. Frameworks like UNIREX [9] demonstrate how these methods that leverage rationales can be integrated into a larger system for improved CQA performance. Similarly, in our reinforcement learning attention-based method, we have regularised the reward (METEOR [3]) by adding our explanatory metric.

Datasets for rationale extraction. There exists a number of datasets specialised in providing rationales. Excluding datasets that are limited to classification (like MultiRC [26], FEVER [55] or Rationales-Movies [60]), or those that encompass more than sentences [7], and to the best of our knowledge, we have found four relevant datasets. There is QED [29] that has the strong assumption of there being only one sentence for the rationale which is rarely the case in our own examples. There are also QuoRef [16] and QuAC [11] which are more focused on solving co-references. Finally, Hotpot-QA [58] is good for our task but is quite challenging for smaller models due to the objective of using multi-hop reasoning.

2 Problem statement

Given the triplet question-context-answer (Q, C, A), we are interested in finding a method that uses this triplet to produce a good approximation R of the best rationale $R^* \subset C$ to explain A. More formally, let $\mathcal{M}$ the set of all methods taking (Q, C, A) as input and outputting a subset of sentences R in the context C ($R \subset C$). The objective is to find the method $M \in \mathcal{M}$ that provides a good approximation $R = M(Q, C, A)$ of R^*.

To identify a high-performing method $M \in \mathcal{M}$, we have at our disposal a training set $TS = \{(Q_i, C_i, A_i, R_i^*)\}_{i=1}^N$ where each sample is composed of the (i) question, (ii) context, (iii) answer and (iv) rationale.

Moreover, given R^*, the quality of the approximated rationale R will be assessed using the Intersection-over-Union (IoU) score defined by

$$IoU(R,\ R^*) = \frac{|R \cap R^*|}{|R \cup R^*|}, \tag{1}$$

where the operator $|\ |$ in Equ. 1 gives the number of character in all sentences in the set it operates on, $\cup$ outputs the set of sentences that appear in at least one operand, and $\cap$ computes the set of sentences appearing in both operands. We note that this IoU is equivalent to $\frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} - 1}$ if both precision and recall are also defined on a per character basis. This metric is also known as the Jaccard coefficient/index/similarity. This choice was motivated by its use in [19] but we differ in that we work on characters rather than tokens and our R and R^* always correspond to complete sentences. We prefer working on sentences because we believe they are more interpretable for end-users and are easier to annotate.

The IoU score will be used in the training (reward regularisation), validation and evaluation sets to improve and assess the performance of a method M.

3 Methods

In this section, we will explain the four different methods, named Embedding Similarity, TF-IDF, LLM classifier and LLM attention, that will be later used in the experiments.

3.1 Embedding similarity

The first method tested is a sentence-embedding method based on LLMs pre-trained for Semantic Textual Similarity, which aims to determine the degree of similarity between two pieces of text.

To generate the embedding of a sentence using an LLM, the most commonly employed approaches are to either average the final hidden vectors (before the classification layer of a classical causal Transformer LLM) of the tokens in the sentence, or simply pool the final hidden vector of the special first token (the [CLS] token). We have chosen to use the latter.

We have defined two methods using two different cut-off functions: the first chooses the top k_{emb} sentences with the highest scores and the other picks all sentences above a certain threshold t_{emb}. They can more succinctly be presented as in Equations 2 and 3, where Emb() is a function that takes a sentence as input and returns a vector $v \in \mathbb{R}^{d_{emb}}$, with d_{emb} being the size of the embedding and cos_sim designating the cosine similarity function.

$$\begin{aligned} &\textbf{Embedder}_{\text{Top-k}}(A,\, C) \\ &= \underset{s_j \text{ in } C}{\text{Top-k}}(\text{cos\_sim}(\text{Emb}(A), \text{Emb}(s_j))) \end{aligned} \tag{2}$$

$$\begin{aligned} &\textbf{Embedder}_{\text{Threshold}}(A,\, C) \\ &= \left\{ s \in C \;\middle|\; \text{cos\_sim}\Big(\text{Emb}(A), \text{Emb}(s)\Big) > t_{emb} \right\} \end{aligned} \tag{3}$$

3.2 TF-IDF

The second class of methods tested uses Term Frequency - Inverse Document Frequency (TF-IDF) [47] rather than an LLM to produce embeddings but otherwise operates the same as the previous method. Each column of the TF-IDF matrix corresponds to a term (word) in the vocabulary and each row is a document. The value represents the TF-IDF score of the corresponding term in the document. This score, derived from the Term Frequency (TF) and Inverse Document Frequency (IDF) values, highlights terms that are prevalent within a document but rare across the corpus, thereby underlining their significance within that document.

The construction of the TF-IDF score is described in Appendix D.

As before, we have tried both a threshold and a ranking approach, described in Equations 4 and 5, where TF-IDF() is a function that takes a sentence as input and returns a vector $v \in \mathbb{R}^{d_{voc}}$, with d_{voc} being the size of the vocabulary.

$$\begin{aligned} &\textbf{NG}_{\text{Top-k}}(A,\, C) \\ &= \underset{s \text{ in } C}{\text{Top-k}}(\text{cos\_sim}(\text{TF-IDF}(A), \text{TF-IDF}(s))) \end{aligned} \tag{4}$$

$$\begin{aligned}&\mathbf{NG}_{\text{Threshold}}(A,\, C)\\&= \left\{ s \in C \;\middle|\; \text{cos\_sim}\Big(\text{TF-IDF}(A), \text{TF-IDF}(s)\Big) > t_{tf-idf} \right\}\end{aligned} \tag{5}$$

3.3 LLM classifier

The third class of method that will be used in our experiments is inspired by [53] and [6]. It involves fine-tuning a pre-trained LLM for binary text classification. The objective is to determine whether a sentence in the context is part of the rationale or not.

The input of the classifier consists in the concatenation of the sentence to be classified (s) surrounded by its neighbouring sentences ($N(s)$), the answer text (A), and the question text (Q). This method can be formalised as:

$$\begin{aligned}&\mathbf{LLM}_{\text{Classifier}}(Q,\, C,\, A)\\&= \left\{ s \in C \;\middle|\; \text{classify}((N(s),\, A,\, Q))^p > 0.5 \right\}\end{aligned} \tag{6}$$

where p denotes the positive label of the soft-maxed output of classify$((s,\, A,\, Q))$, representing the LLM classifier. We only use the first left and right neighbouring sentences of s to avoid the model being overwhelmed by the context but this assumption was not tested.

We provide an illustration of the classification procedure in Figure 1.

Context:
This small mansion has medieval origins and is surrounded by a large landscaped park.<!>The present building was constructed in 1634 by Evan Edwards, a member of a well established Flintshire family which traced its descent from the Welsh king Hywel Dda.<!> He most likely incorporated an older medieval house into the north wing of the current building.
Answer:
Rhual was constructed in 1634 by Evan Edwards
Question:
When was Rhual constructed?

(a) Positive example

Context:
Rhual is a Grade I listed building in Flintshire. <!>This small mansion has medieval origins and is surrounded by a large landscaped park.<!> The present building was constructed in 1634 by Evan Edwards, a member of a well established Flintshire family which traced its descent from the Welsh king Hywel Dda.
Answer:
Rhual was constructed in 1634 by Evan Edwards
Question:
When was Rhual constructed?

(b) Negative example

Fig. 1: Examples of the formatted input fed to the LLM classifier. The sentence to classify is highlighted within the context window.

3.4 LLM attention

This last method is based upon the attention mechanism present in most LLMs. The attention relates two parts of the input together with a numerical value, akin to a correlation matrix. Incidentally, it is of the form $N \times N$, where N is the number of parts in the input (nicely explained in Transformer Explainer [10]). These parts are called tokens.

The attention mechanism is replicated multiple times in a single layer, all with different weights (multi-head attention). This means that for a given model with L layers, H heads and for each token produced, there are $L \times H$ attention results to consider, each attending to different parts of the input and enabling it to understand different linguistic features [14].

Our goal with this method is to produce a view of this matrix where we only consider how the context is related to the answer. Therefore we produce an aggregation over the tokens of the answer. An example of these aggregated (mean) values is shown in Figure 2 as the grey opacity.

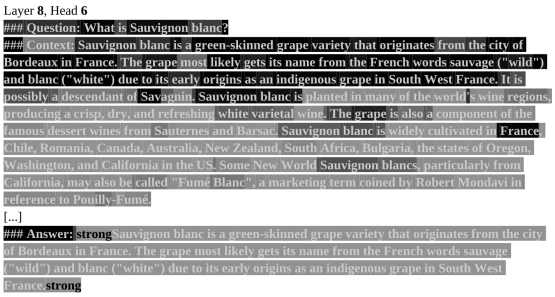


Fig. 2: Average attention weights over a generation by Google/gemma-2b, colourised (darker is higher).

In essence, we transform the matrix presented in Figure 3 into the compression of the light-grey components along the ordinates. Then we map the tokens and strings together to be able to average over sentences.

We start from the internal values of attention $a(i,j)$ per token of the answer $i \in T(A)$ and of the context $j \in T(C)$, where $T()$ is the tokenizer. We average these values over the tokens of the answer to have only one per token of the context by following the equation: $\mathtt{A}(j) = \frac{1}{|T(A)|} \sum_{i \in T(A)} a(i,j)$. We get the following criterion:

$$\begin{aligned} &\mathbf{LLM}_{Attention}(A, Q, C) \\ &= \underset{s \text{ in } C}{\text{Top-K}} \left(\left\{ s \;\middle|\; \left(\frac{1}{|T(s)|} \sum_{j \in T(s)} \mathtt{A}(j) \right) > t_{att} \right\} \right) \end{aligned} \tag{7}$$

.

We note that the LLM can itself generate an answer A' that replaces A using only Q and C. Thus producing the equation $\mathbf{LLM}_{Attention}(Q, C) = \mathbf{LLM}_{Attention}(A', Q, C)$.

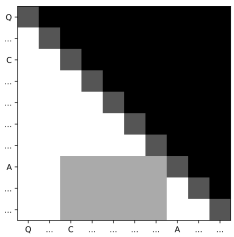


Fig. 3: Representation of the attention matrix for each token of the question Q, context C and answer A. Black is the causal mask, dark grey is the predicted token and light grey represents the tokens the attention will be averaged on (along the axis of ordinates then by parts on the abscissas).

4 Experiments

In this section we will describe the dataset that was used to evaluate our methods and how they were concretely implemented.

4.1 Dataset

The following paragraphs will elaborate on the construction of the reference dataset from which the training, validation and evaluation sets were extracted.

Data source and filtering. We chose the closed-QA part of the `databricks-dolly-15k` [15] as our base CQA dataset. We filtered the triplets (Q,C,A) in the dataset by excluding those where the answer A did not respond to the question Q strictly using the context C or when the answer was wrong (see Appendix A). When little change was required to avoid discards (e.g., deleting a sentence, adding a word,...), we tried applying those instead. This filtered dataset contains 1595 triplets.

Construction. From the filtered CQA dataset, each triplet has undergone human annotation to form our Rationale Databricks Dolly CQA (RDD) dataset. The annotation process involves linking each (Q, A) pair to the relevant sentences within the context C. These form the rationale and will be denoted as R^*. We have labelled in the context only complete sentences rather than segments of sentences. In cases where multiple questions existed within the same example, each sub-question has been labelled separately. The annotation process was done in this manner so that a more complex problem

statement could be created: in this new problem, the goal is to produce multiple sub-rationales corresponding to multiple sub-questions and answers; there is often a combination of questions in a single Q, such as inquiries for *what*, *who* and *when*, and the current statement ignores all these subdivisions. For the rest of this paper, we will consider that a data point i is represented as a quadruplet (Q_i, C_i, A_i, R_i) where the input x is the triplet (Q_i, C_i, A_i) and the targeted output y is the union of all sub-rationales R_i^n: $R_i = \bigcup_{n=1}^{N} R_i^n$. The tool we used to annotate the dataset is `Doccano` [41].

On average, the length of the context is 7.68 sentences while the rationale has 2.02 sentences. There is on average 1.15 question per Q (often in the form of "Who, where and what ...?").

Dataset utilisation. The RDD dataset has been shuffled using the same random seed for all experiments to ensure consistency, and then divided into three sets: the training, validation and evaluation set. They represent respectively 80%, 10% and 10% of the original dataset. The training set has been used to train the methods, the validation set to fine-tune the parameters, and lastly the evaluation to compare their performance.

We decided to split the data points into four categories based on the number of sentences in the context as shown in Figure 4. To do so, we created four intervals $[1; 3]$, $[4; 6]$, $[7; 10]$ and $[11; \inf]$ of different sizes to keep the number of samples in each one comparable. In particular, the last category covers all triplets above ten sentences that have less than 2048 tokens for Q and C; its largest member has 75 sentences in the context. The token restriction removes two triplets that had 88 and 127 sentences.

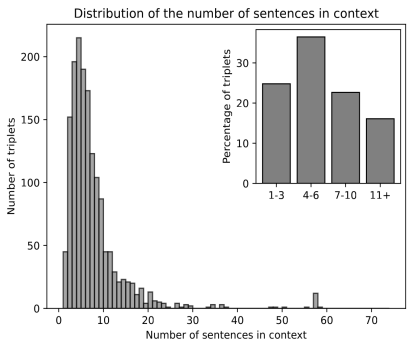


Fig. 4: Truncated (max. 2048 tokens for Q and C) triplet distribution by number of sentences in the context.

4.2 Training

In this paragraph, we will discuss the different choices that have been made to run each experiments.

All methods explored were able to run on our two Nvidia 2080tis and will be further explained in this section. To achieve this, for the methods necessitating training, we have used Low Rank Adaptation (LoRA) [24] together with quantisation (called QLoRA [18]). LoRA is a technique to reduce the number of trainable parameters and quantisation reduces the representation space of the parameters to fit on a smaller number of bytes.

Embedding. For the experiments based on the embeddings, we utilised two pre-trained LLMs: `Sentence-Bert` [44], one of the pioneering models for text similarity embedding based on LLM, and `SFR-Embedding-Mistral` [37], the current state-of-the-art (SOTA) of open-source models for textual similarity tasks according to [40].

To choose the appropriate hyper-parameters, we have swept over $k \in [1 \dots 5]$ and $t \in [0.1 \dots 0.9]$ (90 steps).

TF-IDF. Since this method does not centre on a model, we did not have to make any specific choice other than the hyper-parameters, for which we have made the same sweep as for the embedding method.

LLM Classifier. The different pre-trained LLMs used to train classifiers are: `DistilBERT` [48], `RoBERTa-Base` [34], and `Gemma-2B` [21].

We have fine-tuned the different pre-trained models over 20 epochs using the standard cross-entropy loss. The selection of the best models and checkpoint is based on accuracy since the IoU metric does not apply to the input of the classifier; our metric is only used during evaluation.

The hyper-parameter selection is primarily based on empirical results, the final parameters used for all fine-tuned models are the following: Learning rate=5e-5, LoRa rank=4, alpha LoRa=4, LoRa dropout=0.1.

LLM Attention. For the class of methods using attention, we only used `Gemma-2B` [21].

There are three variations of the attention-based method. The first (suffixed *Base* in tables and graphs) is the base pre-trained model. The second (*FT*) is continually pre-trained (via Huggingface:Trainer) on the base CQA dataset by performing a standard causal language modelling training with Q, C and A always present concurrently. The third (*RL*) is an RL-tuned (via HuggingFace:TRL:PPOTrainer) version of the second, where the reward is the average of the IoU score and the METEOR metric with flat penalty for not including the *EOS* token.

Due to hardware constraints, we have limited the sizes of the examples to $|(Q, C)| < 2048$ and $|A| < 542$ such that the total number of tokens respected $|(Q, C, A)| < 2600$. For the RL training, these values are respectively 450, 50 and 500.

The pre-training of the initial model is continued on the base dataset, ignoring the rationale. The prompts have been formatted by adding "### Question: ", "### Context: ", "### Answer: " and "### End" separated by double line breaks in order to provide a clear description of the task. The training was continued for one epoch.

The attention computation of Equation 7 requires the *head*, *layer*, *threshold* and k to be set. To do so, we have swept over all eight heads, 18 layers for $k \in [0 \dots 4]$ (0 indicates no restriction) and $t \in [0.006, \dots, 0.001, 0.0005, 0.0003, \dots, 0.0]$ to compute the average score on the validation set and took the best combination of parameters. The range for t was motivated by an analysis of reoccurring values.

4.3 Results

In this section we will show and comment on the results given by our simulations. These have been obtained by running the best model we obtained on the evaluation set once.

Best method. As can be seen in Table 1 summarising the performance of all methods over the entire test set, the best method on the dataset is the Gemma classifier. However, when more sentences are present in the context, the attention method seems to scale better.

Influence of model size. The aforementioned results suggest that larger models performs better at our small scale, or that models able to generate such answers are also more likely to find the correct rationale. This is also the case for embedding-based methods, as SFR-Embedding-Mistral consistently gets a higher IoU score than Sentence-Bert.

Influence of hyper-parameter. The sweep of hyper-parameters has shown that methods using a ranking approach perform best with small k values (i.e., 1 or 2), except the attention method once it has had the opportunity to train. This is likely due to the skew of the dataset for smaller numbers of sentences, as shown in Figure 4. For methods using Top-k in particular, we can observe that they are capped around a 0.7 IoU score in the first group (1 to 3), likely because $k = 1$ restricts them from retrieving additional sentences needed for a higher score. In contrast, threshold methods do not have this limitation and can theoretically achieve an IoU score of 1 (i.e., the maximum score) although only the classifier methods reach higher scores. The small values of k being preferred may also be the reason why TF-IDF achieves better scores than the embedding methods.

Influence of training on attention. The results of Table 1 demonstrate that the attention method performs better after training the LLM model when generating the answer. The impact of the training is less noticeable when the golden answer is provided, meaning modern LLMs can have good results as post-hoc checkers. Still in the generation setting, RL training does not significantly improve the results on the evaluation or test set despite showing a 5% absolute increase at training time. This can indicate that the SFT training was sufficient and going beyond would only over-fit.

Influence of the number of sentences in the context. As shown in Figure 5, for all methods, as the number of sentences in the context increases, the performance of the models decrease. This supports the results of [2] who reported (converted from Precision and Recall) over three datasets of increasing size IoUs of 0.89, 0.66 and 0.59 (see Appendix B). For comparison, the projected, context size weighted, scores of our RL attention would be 0.80, 0.71 and 0.70. In terms of time complexity, while the TF-IDF is significantly faster with its linear dependency on the number of sentences and lack of underlying model, the attention, embedding and classifier use the sentences in parallel to diminish that gap. Since the generative attention method produces the answer as well, it is the slowest of all methods.

Model	Size	IoU
Sentence-Bert-large (k_{emb}=1)	109M	$0.61_{\pm 0.05}$
Sentence-Bert-large (t_{emb}=0.68)	109M	$0.54_{\pm 0.06}$
SFR-Embedding-Mistral (k_{emb}=1)	7.11B	$0.65_{\pm 0.05}$
SFR-Embedding-Mistral (t_{emb}=0.72)	7.11B	$0.59_{\pm 0.06}$
TF-IDF (k_{tf-idf}=1)	/	$0.64_{\pm 0.05}$
TF-IDF (t_{tf-idf}=0.25)	/	$0.66_{\pm 0.05}$
Classifier DistilBERT	67M	$0.64_{\pm 0.06}$
Classifier RoBERTa	125M	$0.75_{\pm 0.05}$
Classifier Gemma	2.51B	$\mathbf{0.79}_{\pm \mathbf{0.04}}$
Gemma Base (L=5, H=4, k=0, t=0.002)	2.51B	$0.74_{\pm 0.05}$
Gemma FT (L=8, H=6, k=0, t=0.002)	2.51B	$0.75_{\pm 0.05}$
Gemma RL (L=5, H=4, k=0, t=0.002)	2.51B	$0.76_{\pm 0.05}$
+ Generation of Answer (Gen)		
Gemma Base (L=14, H=7, k_{att}=2, t_{att}=0.000)	2.51B	$0.53_{\pm 0.05}$
Gemma FT (L=8, H=6, k_{att}=0, t_{att}=0.002)	2.51B	$0.60_{\pm 0.06}$
Gemma RL (L=8, H=6, k_{att}=0, t_{att}=0.002)	2.51B	$0.61_{\pm 0.06}$

Table 1: Summary table of experiment results, including the number of parameters for each method (size) and the average IoU score obtained on the evaluation set, presented with a 95% confidence interval assuming a student-t distribution.

4.4 Limitations

In this section we will discuss some limitations we encountered and/or are aware of.

Concerning the dataset, it has only been annotated by us and thus may not have been reviewed impartially and/or in sufficient depth. Additionally, we could have extended the dataset by using existing datasets for rationale extraction in classification and procedurally generated appropriate outputs. However, this approach would decrease the variety of answers and the impact on performance of such a decrease has, to our knowledge, not yet been studied. We could also have used HotpotQA [58] to broaden the dataset.

Despite our best attempt at exploring a widely applicable array of methods, they still come with intrinsic restrictions. For example, the attention-based method cannot

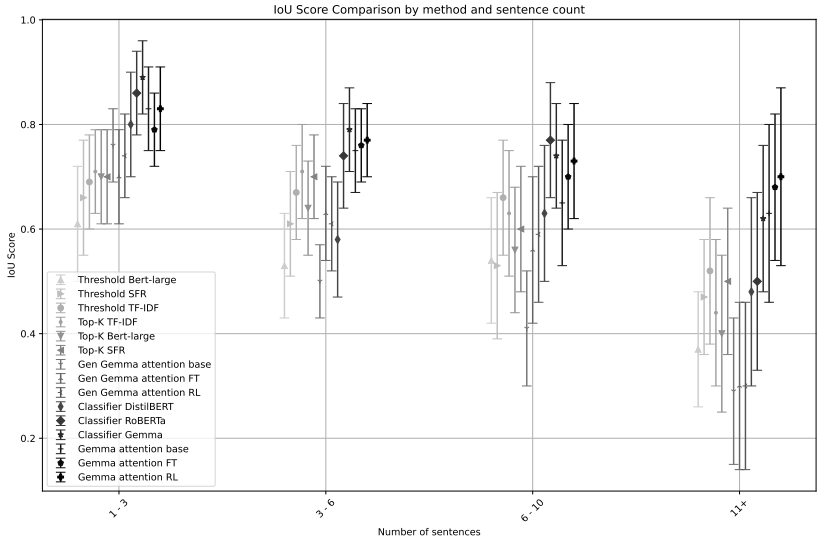


Fig. 5: IoU scores with 95% confidence error bars (student-t).

be extended easily to sub-quadratic [27,20,36,12,51], or attention-less [23,61,17,42,5] LLMs. The exploration of gradient-based methods would have been more broadly applicable but would require many additional backward passes inducing a higher compute requirement. This particular choice has been at the center of debates as discussed in [4].

The methods studied also do not cover all existing ones and should not be regarded as an exhaustive comparison but rather an educated guess of potentially well-performing methods.

The models used also were limited by our hardware, being two Nvidia 2080ti. Some runs were carried out on a cluster to hasten the experiments but the parameters have been kept the same such that any of these can be exactly replicated on the original hardware requirements.

Finally, the tuning of hyper-parameters required a lot of compute, so we have restricted the range of values to those we estimated which would be most pertinent. The linear/grid search approach could also be revised to other optimisation techniques.

5 Conclusion

In this paper, we have constructed the CQA Rationale Databricks Dolly dataset for the express purpose of improving sentence-level rationale extraction in closed-domain question answering where the answer appears as full sentences. This dataset will be provided under the same license (CC BY-SA 3.0) as the original for the community along with all the code used for the experiments. This dataset allowed us to study a range of methods and we hope will foster interest in the subject. Evaluations were perform via our IoU metric.

Concerning our results, we have underlined the difficulty in reaching satisfying scores as the scale of the context grows. Nonetheless, we have found that classifier models could achieve an IoU of **79%** and the attention with reinforcement learning followed closely with **76%**, which is on a par with previous work for smaller sizes of contexts (a few sentences) but is projected to scale slightly better as the number of sentences grows. Still, achieving satisfying results on lengthy documents remains an unresolved challenge.

This research calls for several future works. First, in the produced dataset there have been numerous instances of questions that contained sub-questions while taking care of differentiating the sub-answers and corresponding sub-rationales. This is of particular interest because no other dataset reviewed seemed to differentiate this case.

Second, it may be interesting to see other approaches compete, such as gradient-based methods, to see how the scores could be improved. This includes the incorporation of the optimal answer at evaluation time for the attention method to see if it would become competitive with the others. Conversely, as a third point, it might be interesting to address only the issue of finding both A and R.

Finally, conjugating both the quality of the approximated rationale and of the generated answer is an interesting challenge leading to a unique and capable model. To do this, reward regularisation, context sampling or compression might be the most straightforward of approaches as has been shown empirically in related works.

Acknowledgments.

Lize Pirenne gratefully acknowledges the financial support of the Walloon Region for Grant No. 2010235 – ARIAC by DW4AI.

Disclosure of Interests.

The authors declare no conflict of interest.

A RDD Dataset

Our enhanced dataset is composed of the textual fields "question", "context", "citation" and "answer" corresponding to (Q_i, C_i, R_i, A_i). These are expanded upon by the fields "sub_question", "sub_citation" and "sub_answer" that list each j triplet (Q_i^j, R_i^j, A_i^j). Each element has its corresponding bounds in "sub_question_index", "sub_citation_index", "sub_answer_index" to avoid searching. The additional fields "id", "num_sub_question" (#("sub_question_index")) and "citation_index" ($\bigcup$"sub_citation_index") are provided for convenience.

A.1 Guidelines

The guidelines for the annotation of the dataset are as follows:

- Highlight a question in the question part, its answer and all the sentences in the context that are necessary to answer the question.
- If the rationale for the answer is not in the context, skip.
- If the rationale only contains pronouns, add the closest sentence defining the pronoun.

An example of this would be the sentences “Marta is a politician. She is a member of the Green Party.” with the question “What party is Marta a member of?” and the answer “Green Party”. The rationale would be “Marta is a politician. She is a member of the Green Party.”. This was done to avoid ambiguity if only the rationale was presented to a user or in the case multiple context were presented at a time.

– Repeat the process for all questions in the question part.

A.2 Filtered examples

Some answers in the dataset did not strictly depend on the context and where filtered out:

Q: What is Cryptocurrency?

C: A cryptocurrency, crypto-currency, or crypto is a digital currency designed to work as a medium of exchange through a computer network that is not reliant on any central authority, such as a government or bank, to uphold or maintain it. It is a decentralised system for verifying that the parties to a transaction have the money they claim to have, eliminating the need for traditional intermediaries, such as banks, when funds are being transferred between two entities.

A: Cryptocurrency is the digital currency built on the blockchain or other network which allows people to exchange and trade for goods. There are over 1000 different kinds of cryptocurrency around the world and each of them is designed differently. Currently, the popular cryptocurrencies are Bitcoin, Ethereum, and Dogecoin.

Others were simply false (here the retirement age should be 65 or 66 depending on the exact date of birth):

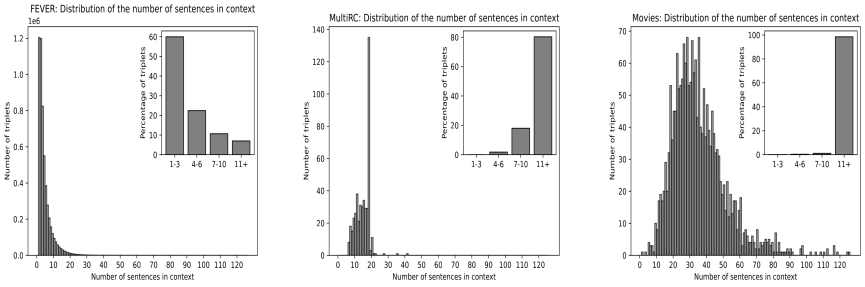
Q: Given this short paragraph about the life of George K. McGunnegle, how old was he when he retired and what month of the year did that happen?

C: As a member of the 3rd Infantry Regiment and 17th Infantry Regiment, McGunnegle served in the Philippines during the Philippine–American War. After receiving promotion to colonel in 1906, he commanded the 1st Infantry Regiment during service in Washington, the Philippines, and Hawaii. In 1914, he was assigned to command the recruit depot at Fort McDowell, California. He served until recaching the mandatory retirement age of 64 in June 1918. Because World War I was ongoing, McGunnegle agreed to defer retirement and remained on duty at Fort McDowell. The war ended in November 1918, and he remained in command of Fort McDowell to muster out troops as the army demobilized. He retired in July 1919.

A: George K. McGunnegle was 64 when he retired in July of 1919.

B Dataset comparison

We provide a few figures to illustrate the distribution of the number of sentences in the context for three other datasets: FEVER, MultiRC and Movies. They are respectively shown in Figure 6a, Figure 6b and Figure 6c.



(a) FEVER (b) MultiRC (c) Movies

Fig. 6: Triplet distribution by number of sentences in the context.

C IoU, Precision and Recall

We will quickly show that IoU is equivalent to $\frac{1}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}-1}$

The precision is defined as $\frac{TP}{TP+FP}$ and the recall as $\frac{TP}{TP+FN}$ where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. The IoU is defined as $\frac{TP}{TP+FP+FN}$ because predicted $= TP + FP$ and true $= TP + FN$, thus the union of the two is $TP + FP + FN$ and the intersection is TP.

We can write $\frac{1}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}-1} = \frac{1}{\frac{TP+FP}{TP}+\frac{TP+FN}{TP}-1} = \frac{1}{\frac{TP+FP+TP+FN-TP}{TP}} = \frac{TP}{TP+FP+FN}$.

D Algorithm for TF-IDF

Here follows the algorithm used for the TF-IDF method.

Algorithm 1 TF-IDF Fit and Transform

Inputs:
D : The corpus composed of n documents D_i

Outputs:
TF-IDF : The embedding matrix of the corpus

1: $V = \{\text{split\_terms}(D)\}$
2: $\text{TF}(t, d) = \frac{1}{|S|} \sum_{w \in S} [w = V_t], S = \text{split\_terms}(D_d)$
3: $\text{IDF}(t) = \log\left(\frac{1}{|D|} \sum_{d=0}^{|D|} [\text{TF}(t, d) > 0]\right) + 1$
4: $\text{TF-IDF}(t, d) = \text{TF}(t, d) * \text{IDF}(t)$

The brackets in the algorithm refers to the Iverson brackets, they produce "1" if the inside is true, and "0" otherwise. The function split_terms extracts each term composing its argument.

References

1. Arous, I., Dolamic, L., Yang, J., Bhardwaj, A., Cuccu, G.: Marta: Leveraging human rationales for explainable text classification. Proceedings of the AAAI Conference on Artificial Intelligence **35**(7), 5868–5876 (5 2021). https://doi.org/10.1609/aaai.v35i7.16734, https://ojs.aaai.org/index.php/AAAI/article/view/16734
2. Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I.: Diagnostics-guided explanation generation. Proceedings of the AAAI Conference on Artificial Intelligence **36**(10), 10445–10453 (Jun 2022). https://doi.org/10.1609/aaai.v36i10.21287, https://ojs.aaai.org/index.php/AAAI/article/view/21287
3. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C.R. (eds.) Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005. pp. 65–72. Association for Computational Linguistics (2005)
4. Bastings, J., Filippova, K.: The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (2020). https://doi.org/10.18653/V1/2020.BLACKBOXNLP-1.14
5. Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O.: xlstm: Extended long short-term memory (2024), https://arxiv.org/abs/2405.04517
6. Chae, Y., Davidson, T.: Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation (2023)
7. Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I.: Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. North American Chapter of the Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.NAACL-MAIN.22
8. Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I.: Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 226–241. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.22, https://aclanthology.org/2021.naacl-main.22
9. Chan, A., Sanjabi, M., Mathias, L., Tan, L., Nie, S.: UNIREX: A unified learning framework for language model rationale extraction. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2867–2889. PMLR (7 2022)
10. Cho, A., Kim, G.C., Karpekov, A., Helbling, A., Wang, Z.J.: Transformer explainer: Interactive learning of text-generative models (2024), https://arxiv.org/abs/2408.04619
11. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.t.: QuAC: Question answering in context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2174–2184. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov

2018). https://doi.org/10.18653/v1/D18-1241, https://www.aclweb.org/anthology/D18-1241
12. Choromanski, K.M., Likhosherstov, V., Dohan, D., Song, X., Gane, A.: Rethinking Attention with Performers. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
13. Chuang, Y.S., Qiu, L., Hsieh, C.Y., Krishna, R., Kim, Y.: Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps (2024), https://arxiv.org/abs/2407.07071
14. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What Does BERT Look At? An Analysis of BERT's Attention (6 2019)
15. Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J.: Free dolly: Introducing the world's first truly open instruction-tuned LLM. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm (2023)
16. Dasigi, P., Liu, N.F., Marasović, A., Smith, N.A., Gardner, M.: Quoref: A reading comprehension dataset with questions requiring coreferential reasoning (2019), https://arxiv.org/abs/1908.05803
17. De, S., Smith, S.L., Fernando, A., Botev, A., Cristian-Muraru, G.: Griffin: Mixing gated linear recurrences with local attention for efficient language models (2 2024). https://doi.org/10.48550/arXiv.2402.19427
18. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 10088–10115. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf
19. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C.: Eraser: A benchmark to evaluate rationalized nlp models (2020), https://arxiv.org/abs/1911.03429
20. Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S.: LongNet: Scaling Transformers to 1, 000, 000, 000 Tokens (2023). https://doi.org/10.48550/arXiv.2307.02486
21. Gemma, T., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S.: Gemma: Open Models Based on Gemini Research and Technology (3 2024)
22. Glockner, M., Habernal, I., Gurevych, I.: Why do you think that? exploring faithful sentence-level rationales without supervision. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1080–1095. Association for Computational Linguistics, Online (11 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.97, https://aclanthology.org/2020.findings-emnlp.97
23. Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces (12 2023). https://doi.org/10.48550/arXiv.2312.00752
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y.: LoRA: Low-Rank Adaptation of Large Language Models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
25. Joshi, B., Chan, A., Liu, Z., Nie, S., Sanjabi, M.: ER-test: Evaluating explanation regularization methods for language models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. pp. 3315–3336. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.242
26. Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., Roth, D.: Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In: Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL) (2018)

27. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: The Efficient Transformer. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020)
28. Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S.: Post hoc explanations of language models can improve language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/ce65173b994cf7c925c71b482ee14a8d-Abstract-Conference.html
29. Lamm, M., Palomaki, J., Alberti, C., Andor, D., Choi, E.: QED: A Framework and Dataset for Explanations in Question Answering. Transactions of the Association for Computational Linguistics **9**, 790–806 (08 2021). https://doi.org/10.1162/tacl_a_00398, https://doi.org/10.1162/tacl_a_00398
30. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020)
31. Li, J., Wang, M., Zheng, Z., Zhang, M.: Loogle: Can long-context language models understand long contexts? arXiv preprint arXiv:2311.04939 (2023)
32. Liu, H., Zaharia, M., Abbeel, P.: Ring attention with blockwise transformers for near-infinite context. CoRR **abs/2310.01889** (2023). https://doi.org/10.48550/ARXIV.2310.01889, https://doi.org/10.48550/arXiv.2310.01889
33. Liu, Y., Cao, J., Liu, C., Ding, K., Jin, L.: Datasets for large language models: A comprehensive survey (2024), https://arxiv.org/abs/2402.18041
34. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M.: Roberta: A robustly optimized bert pretraining approach (2019)
35. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
36. Ma, X., Zhou, C., Kong, X., He, J., Gui, L.: Mega: Moving Average Equipped Gated Attention. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023)
37. Meng, R., Liu, Y., Joty, S.R., Xiong, C., Zhou, Y.: Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog (2024), https://blog.salesforceairesearch.com/sfr-embedded-mistral/, accessed: 2024-05-28
38. Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F.: Teaching language models to support answers with verified quotes (2022), https://arxiv.org/abs/2203.11147
39. Moradi, P., Kambhatla, N., Sarkar, A.: Measuring and improving faithfulness of attention in neural machine translation. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2791–2802. Association for Computational Linguistics, Online (4 2021). https://doi.org/10.18653/v1/2021.eacl-main.243
40. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark (2023), https://arxiv.org/abs/2210.07316
41. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: doccano: Text annotation tool for human. https://github.com/doccano/doccano (2018), accessed: 2024-03-02

42. Peng, B., Alcaide, E., Anthony, Q.G., Albalak, A., Arcadinho, S.: RWKV: Reinventing RNNs for the Transformer Era. Conference on Empirical Methods in Natural Language Processing (2023). https://doi.org/10.48550/arXiv.2305.13048
43. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (3 2024)
44. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. Conference on Empirical Methods in Natural Language Processing (2019). https://doi.org/10.18653/v1/D19-1410
45. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
46. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 2662–2670. ijcai.org (2017). https://doi.org/10.24963/IJCAI.2017/371
47. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5), 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0
48. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020)
49. Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D.: Large Language Models Can Be Easily Distracted by Irrelevant Context. In: Proceedings of the 40th International Conference on Machine Learning. pp. 31210–31227. PMLR (7 2023)
50. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise (2017), https://arxiv.org/abs/1706.03825
51. Song, K., Wang, X., Cho, S., Pan, X., Yu, D.: Zebra: Extending Context Window with Layerwise Grouped Local-Global Attention (12 2023). https://doi.org/10.48550/arXiv.2312.08618
52. Sun, J., Swayamdipta, S., May, J., Ma, X.: Investigating the benefits of free-form rationales. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. pp. 5867–5882. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.432, https://doi.org/10.18653/v1/2022.findings-emnlp.432
53. Sun, X., Li, X., Li, J., Wu, F., Guo, S.: Text classification via large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023. pp. 8990–9005. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.603, https://doi.org/10.18653/v1/2023.findings-emnlp.603
54. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/sundararajan17a.html
55. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: NAACL-HLT (2018)

56. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
57. Wiegreffe, S., Marasović, A.: Teach me to explain: A review of datasets for explainable natural language processing (2021), https://arxiv.org/abs/2102.12060
58. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
59. Yu, M., Zhang, Y., Chang, S., Jaakkola, T.: Understanding interlocking dynamics of cooperative rationalization. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 12822–12835. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/6a711a119a8a7a9f877b5f379bfe9ea2-Paper.pdf
60. Zaidan, O.F., Eisner, J., Piatko, C.: Machine learning with annotator rationales to reduce annotation cost. In: Proceedings of the NIPS*2008 Workshop on Cost Sensitive Learning (December 2008)
61. Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H.: An Attention Free Transformer (2021)
62. Zhang, J., Bai, Y., Lv, X., Gu, W., Liu, D.: Longcite: Enabling llms to generate fine-grained citations in long-context qa. arXiv preprint arXiv:2409.02897 (2024)
63. Zhang, T., Patil, S.G., Jain, N., Shen, S., Zaharia, M.: Raft: Adapting language model to domain specific rag. arXiv preprint arXiv: 2403.10131 (2024)
64. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H.: Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology **15**(2) (2 2024). https://doi.org/10.1145/3639372