Peering into the Mind of Language Models: An Approach for Attribution in Contextual Question Answering

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami
Balaji Vasan Srinivasan
Adobe Research, India
{phukan, shsomasu, apoorvs, koustavag, balsrini}@adobe.com

Abstract

With the enhancement in the field of generative artificial intelligence (AI), contextual question answering has become extremely relevant. Attributing model generations to the input source document is essential to ensure trustworthiness and reliability. We observe that when large language models (LLMs) are used for contextual question answering, the output answer often consists of text copied verbatim from the input prompt which is linked together with "glue text" generated by the LLM. Motivated by this, we propose that LLMs have an inherent awareness from where the text was copied, likely captured in the hidden states of the LLM. We introduce a novel method for attribution in contextual question answering, leveraging the hidden state representations of LLMs. Our approach bypasses the need for extensive model retraining and retrieval model overhead, offering granular attributions and preserving the quality of generated answers. Our experimental results demonstrate that our method performs on par or better than GPT-4 at identifying verbatim copied segments in LLM generations and in attributing these segments to their source. Importantly, our method shows robust performance across various LLM architectures, highlighting its broad applicability. Additionally, we present VERIFIABILITY-GRANULAR[1], an attribution dataset which has token level annotations for LLM generations in the contextual question answering setup.

1 Introduction

The surge in the capabilities of Large Language Models (LLMs) has revolutionized natural language understanding. Their ability to comprehend and generate human-like text has resulted in their widespread adoption across various industries. A prominent and pivotal application of these models is question answering, particularly in contextual settings. The ability of LLMs to parse, interpret, and respond to queries within a given context has facilitated efficient information retrieval and comprehension.

Despite the remarkable strides made in contextual question answering, challenges persist within LLMs. While LLMs excel in generating informative responses, they often fall short in providing explicit references or attributions to the specific sections or sources within the context from which their answers derive (Liu et al., 2023b). The absence of this attribution impedes the ability to verify the authenticity and accuracy of the generated information. Attribution not only enables verification (Rashkin et al., 2023) but also increases user trust and confidence in the responses generated by these models, thus fostering their broader acceptance and utilization across diverse domains (Bohnet et al., 2022).

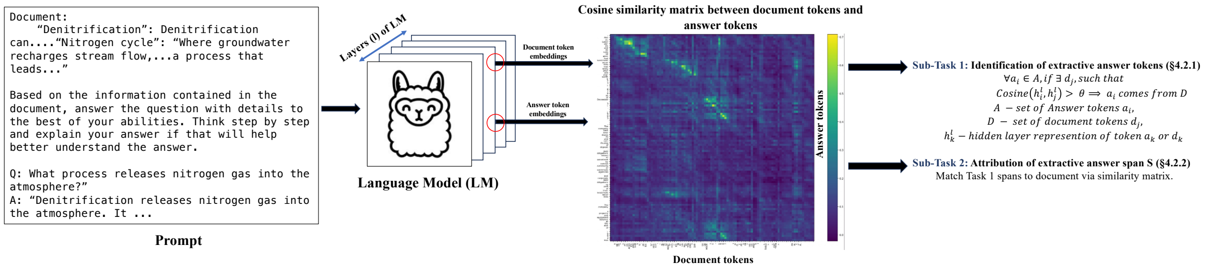
Existing methods for attribution typically align within three primary categories (Li et al., 2023): Systems fine-tuned or trained explicitly for attribution tasks (Weller et al., 2023; Gao et al., 2023b), retrieve-then-read (Chen et al., 2017; Lee et al., 2019) and approaches relying on post-generation attribution (Gao et al., 2023a; Huo et al., 2023). However, each of these approaches encounters substantial challenges in achieving accurate and granular attribution, thereby limiting their effectiveness.

The initial category, which consists of systems tailored or explicitly trained for attribution tasks, poses a significant challenge (Huang and Chang, 2023). This challenge primarily arises because the training process demands extensive resources, including time, computational power, and a large corpus of annotated data.

Furthermore, the continuous development of more sophisticated models requires retraining, perpetuating an unending cycle of adaptation. Additionally, the need for regression testing to ensure

[1] Dataset is available at `https://github.com/Anirudh-Phukan/verifiability-granular`.

Findings of the Association for Computational Linguistics ACL 2024, pages 11481–11495
August 11-16, 2024 ©2024 Association for Computational Linguistics



```
Document:
    “Denitrification”: Denitrification
can....“Nitrogen cycle”: “Where groundwater
recharges stream flow,...a process that
leads...”

Based on the information contained in the
document, answer the question with details to
the best of your abilities. Think step by step
and explain your answer if that will help
better understand the answer.

Q: What process releases nitrogen gas into the
atmosphere?”
A: “Denitrification releases nitrogen gas into
the atmosphere. It ...
```

Sub-Task 1: Identification of extractive answer tokens (§4.2.1)

$\forall a_i \in A, if\ \exists\ d_j, such\ that$

$Cosine(h_i^l, h_j^l) > \theta \Longrightarrow a_i\ comes\ from\ D$

$A\ -\ set\ of\ Answer\ tokens\ a_i,$

$D\ -\ set\ of\ document\ tokens\ d_j,$

$h_k^l - hidden\ layer\ represention\ of\ token\ a_k\ or\ d_k$

Sub-Task 2: Attribution of extractive answer span S (§4.2.2)

Match Task 1 spans to document via similarity matrix.

Figure 1: Overall picture of our proposed methodology; our method utilizes hidden layer representations from both the document and answer to determine the attribution of answer tokens. Initially, we identify extractive answer tokens (§ 4.2.1) through a cosine similarity matrix between document and answer tokens. Subsequently, we map these tokens to document token sequences by identifying anchor tokens and generating candidates, later ranked based on their cosine similarity to achieve attribution. (§ 4.2.2)

answer quality amplifies the burden, creating a considerable bottleneck in practical application and scalability.

On the other hand, the second category—post-generation attribution—offers an alternative but not without its own set of drawbacks. This approach relies on leveraging retrieval models to trace the sources of the generated answers. However, this method suffers from the overhead associated with these retrieval models, which often demand substantial computational resources and time. Furthermore, the granularity of attribution is constrained by the chunk size at which retrieval is performed, limiting the precision of attribution.

Lastly, the third category—retrieve then read methods—also grapples with its unique set of complications. Similar to the post-generation attribution approach, this method faces the problem of retrieval method overhead and chunking issues. The concept here is to utilize retrieved evidence as a basis for generating answers and consequently, attributions. However, as outlined in (Gao et al., 2023a), retrieval does not equate to attribution due to the potential integration of external knowledge.

Our proposed method aims to surmount the limitations posed by existing attribution approaches by adopting a distinct strategy that eliminates the need for additional training while offering granular attributions. By delving into the inner workings of the LLM during its generation process, we sidestep the resource-intensive training requirements.

The crux of our approach lies in accessing the hidden state representations of tokens produced by the LLM when generating responses. Leveraging the contextual cues provided by the input—context, question, and generated answer—we extract these hidden state representations via a forward pass through the model. These hidden representations which happen to be contextual embeddings of tokens are then matched to perform attribution. An overview of this process is shown in Figure 1.

Operating at a token-level granularity mirrors the natural generation process of LLMs. Unlike conventional post generation attribution methods that grapple with decisions on chunking at paragraph or sentence levels, our approach bypasses this dilemma. By steering clear of chunk-based attributions, we are liberated from arbitrary segmentation of the context, which often leads to imprecise referencing and diluted contextual connections.

The existing public datasets curated for the task of attribution (attribution of LLM generations) have annotations only at the response level or at the sentence level (for every sentence within the response). (Kamalloo et al., 2023; Liu et al., 2023b; Malaviya et al., 2023). To assess our method's effectiveness, token-level annotations are essential. Consequently, we process the data collected by Liu et al. (2023b) to generate token-level annotations. We release this dataset to facilitate further research on token-level attribution.

In essence, our approach offers a unique and efficient solution to the problem of attribution in contextual question answering. Additionally, we also introduce a dataset that facilitates token-level attribution. With this in mind, we are excited to share our primary contributions in this paper.

- We introduce a pioneering method for attribution in contextual question answering. The strengths of our approach include its token-level granularity ensuring precise attributions, lack of additional retrieval model overhead and its training-free nature which ensures consistent answer quality.

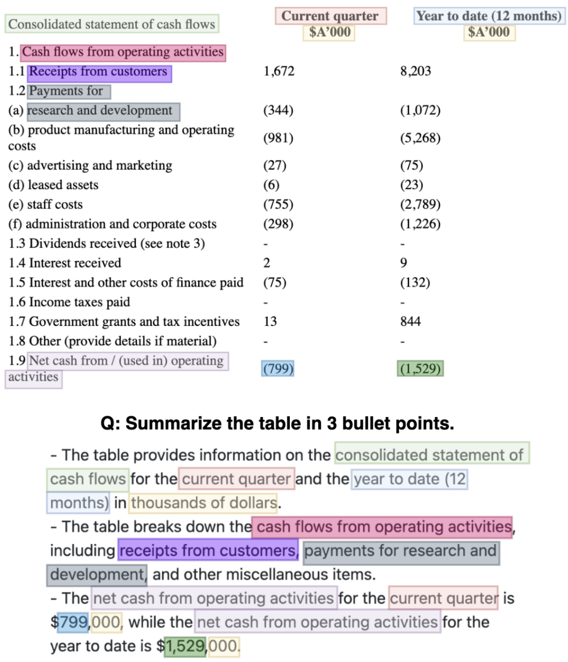
- Our experimental findings demonstrate the efficacy of our method across various model families. This indicates that leveraging hidden layer representations for attribution can be broadly applied across different LLM architectures, highlighting the wide-ranging applicability of our approach.

- Additionally, we release the VERIFIABILITY-GRANULAR dataset, which contains token-level attributions for LLM generations in the contextual question answering setup.

2 Related Work

Researchers have employed different approaches for the attribution of generated or identified text spans. A shared task was organized by Thorne et al. (2018) to encourage researchers to build systems capable of performing fact verification and attribution to the source texts. Evans et al. (2021) emphasized the role of source attribution in fostering truthful and responsible AI. The growing concern of fake news detection in AI-based news generation has also been considered as an attribution task (Pomerleau and Rao, 2017; Ferreira and Vlachos, 2016). In addition to automatic attribution, studies on manual attribution have been performed by domain specific individuals (Borel, 2023; Li et al., 2015). While we discuss about fact verification as an attribution task, it is important to note that user interactions have also been found to require attribution (Dziri et al., 2022). Petroni et al. (2022) has highlighted the fact that wikipedia articles needs validation and that the citations need to be attributed. On the other hand Sarti et al. (2023b) introduced a python library based on GPT-2 capable of identifying feature attributions generated from the Insequence model; capable of performing in multilingual settings (Sarti et al., 2023a). Recently, researchers performed attribution task on multimodal systems as well (Ancona et al., 2017; Holzinger et al., 2021; Zhao et al., 2023).

However, what is missing in prior art is the utilisation of the contextual guidance inherently present during generation for the attribution task. In our work, we tackle this problem by utilising the contextual information encoded during generation by the LLM and attribute spans to semantically relevant parts of the source document.



Consolidated statement of cash flows	Current quarter $A'000	Year to date (12 months) $A'000
1. Cash flows from operating activities		
1.1 Receipts from customers	1,672	8,203
1.2 Payments for		
(a) research and development	(344)	(1,072)
(b) product manufacturing and operating costs	(981)	(5,268)
(c) advertising and marketing	(27)	(75)
(d) leased assets	(6)	(23)
(e) staff costs	(755)	(2,789)
(f) administration and corporate costs	(298)	(1,226)
1.3 Dividends received (see note 3)	-	-
1.4 Interest received	2	9
1.5 Interest and other costs of finance paid	(75)	(132)
1.6 Income taxes paid	-	-
1.7 Government grants and tax incentives	13	844
1.8 Other (provide details if material)	-	-
1.9 Net cash from / (used in) operating activities	(799)	(1,529)

Q: Summarize the table in 3 bullet points.

- The table provides information on the consolidated statement of cash flows for the current quarter and the year to date (12 months) in thousands of dollars.
- The table breaks down the cash flows from operating activities, including receipts from customers, payments for research and development, and other miscellaneous items.
- The net cash from operating activities for the current quarter is $799,000, while the net cash from operating activities for the year to date is $1,529,000.

Figure 2: Semi Extractive answers by LLMs

3 Problem Statement

We observe a phenomenon in contextual question answering using LLMs which results in a comprehensive answer typically characterized by factual spans, replicated verbatim from various segments of the provided context, interwoven with "glue text". An example of this phenomenon is shown in Figure 2. Yang et al. (2023) also note this pattern and use references to losslessly speed up LLM inference.

Building upon the aforementioned observations, we dissect the challenge of attribution in the contextual generation setting into two distinct yet interconnected sub-problems. The first sub-problem involves the identification of tokens within the output that have been directly copied from the provided context.

The second sub-problem delves deeper into the attribution of these identified tokens. This step involves mapping these tokens back to their original positions within the document. In essence, this implies tracing back the tokens to the exact sections within the document from which they were copied verbatim.

By doing this, we aim to establish a clear flow of information from the source document to the generated output, thereby enabling a more nuanced understanding of the context generation process.

The task at hand involves two distinct sub-tasks within the context of attributing the output of a large language model (LLM). We are given a document D represented as a sequence of tokens $(d_1, d_2, \ldots, d_n)$, a question Q, and an answer A represented as a sequence of tokens $(a_1, a_2, \ldots, a_m)$. Our objective comprises two interconnected subtasks:

Subtask 1: Token Attribution Identification. This subtask involves identifying a subset of tokens in A, denoted as $A_{\text{attr}} \subseteq A$, that require attribution to tokens in D. Mathematically, this can be expressed as selecting tokens $a_i \in A$ that are directly influenced or copied from D. In our current study, we restrict the scope to tokens that are verbatim copied from D.

Subtask 2: Corresponding Token Mapping. For each token in A_{attr}, this subtask aims to find a mapping function $f : A_{\text{attr}} \to D$ such that for every token $a_i \in A_{\text{attr}}$, there exists a corresponding token or sequence of tokens in D to which a_i can be attributed.

Formally, the mapping function can be expressed as:

$$\begin{aligned}
&f : A_{\text{attr}} \to D, \quad \text{s.t. } \forall a_i \in A_{\text{attr}} \\
&f(a_i) = (d_j, \ldots, d_k) \iff \\
&a_i \text{ is attributed to } (d_j, \ldots, d_k) \text{ in } D
\end{aligned}$$

In this formulation, the function f establishes a connection between the tokens in the answer requiring attribution and their corresponding tokens or token sequences in the document D.

4 Proposed Work

4.1 Motivation

Humans easily identify which parts of a document they use to answer questions. This skill to trace information back to its source is key for effective communication, especially when discussing complex ideas. Similarly, we propose that Large Language Models (LLMs) have an inherent awareness of the document parts they use while generating answers. This awareness is likely captured within the hidden states of the LLM, which encode token information during the generation process.

Our work is based on the idea that if LLMs can generate responses combining copied segments and self-generated "glue text", they must inherently differentiate between copied content and self-generated content. This differentiation is what we aim to uncover. By accessing and analyzing the hidden states of an LLM, we seek to reveal how it uses and attributes source information in its responses.

This approach allows us to trace how the LLM processes and uses the provided context at a detailed, token-level granularity. We believe that fully understanding how LLMs generate responses requires examining the process at this fundamental level - where each token's role in the response is clarified and attributed to its original source. Not only does this give us an understanding of a LLM's inner workings, but also accomplishes the useful task of attributing outputs, which helps humans easily verify/navigate the LLM outputs.

4.2 Methodology

Given a language model, denoted as M, we construct a prompt P by concatenating the document D, represented as a sequence of tokens $(d_1, d_2, \ldots, d_n)$, the question Q, and the answer A, represented as a sequence of tokens $(a_1, a_2, \ldots, a_m)$. This concatenation is formally expressed as $P = D + Q + A$, where '+' denotes the concatenation operation. The prompt P is then passed through the model M in a forward pass to obtain the hidden layer representations for each token in P. These representations capture the contextual information encoded by the model for each token. An illustrative example of this process is provided below.

```
[INST]
Document:
{document}
Based on the information contained in
    the document, answer the question
    with details to the best of your
    abilities. Think step by step and
    explain your answer if that will
    help better understand the answer.
Q: {question} A:
[/INST]
{answer}
```

Let's denote the hidden layer representation of each token t_i for a specific layer l as h_i^l.

4.2.1 Identifying extractive output tokens

For the first sub-task of identifying tokens in answer A that are directly copied from document D, we perform the following operation. For any specific layer l and for all tokens $a_i \in A$, we say that a token a_i comes from D if there exists a token $d_j \in D$ such that the cosine similarity between h_i^l and h_j^l is greater than a threshold θ.

This can be formally represented as:

$$\forall a_i \in A, \quad \text{if } \exists d_j \in D \text{ s.t.}$$
$$\text{Cosine}(h_i^l, h_j^l) > \theta \implies a_i \text{ comes from } D$$

4.2.2 Attributing Extractive Spans

Given a span S in A to be attributed, consisting of tokens $a_1, \ldots, a_n$, we compute the average hidden layer representation h_S for each token $a_i \in S$ as:

$$h_S = \frac{1}{n} \sum_{i=1}^{n} h_i^l$$

Next, we use h_S to identify anchor tokens in D. For each token $d_j \in D$, we compute the cosine similarity between h_S and h_j^l and select tokens with the highest similarities as anchor tokens, denoted as D_T.

For each anchor token $d_a \in D_T$, we explore windows of tokens around d_a, up to a maximum length L. We calculate the average hidden layer representation h_W for each window W and identify the window with the highest similarity to h_S. The highest-ranked window is considered the final attribution for the span S.

In cases where D is segmented into evidence spans $e \in E$, the score for each e is the similarity between h_S and the best window within e.

5 Experimental Setup

In this section, we outline the details of our experimental setup; the dataset, the evaluation metrics; and the baselines.

5.1 Datasets

QuoteSum (Schuster et al., 2023)**:**[2] This dataset consists of questions, relevant passages, and human-written semi-extractive answers. In the process of dataset construction, the human annotators were tasked to answer multiple source questions by combining information from various sources. They were instructed to explicitly extract factual spans and weave them together into a coherent, well-grounded passage. The dataset comprises of 4,009 semi-extractive answers to 1,376 unique questions in total and is split into train, validation and test sets with ratios 60%, 7%, 33%, respectively.

Verifiability-Granular: We also curate a dataset with token-level attributions called VERIFIABILITY-GRANULAR (VERI-GRAN) by processing the dataset introduced by (Liu et al., 2023b)[3]. The original dataset contains commercial generative search engines generated responses to input queries along with the retrieved content used for the generation. Annotations map sentences in the response to portions of the source text.

We split the source text using *nltk.sent_tokenize*[4] and filter the samples that map a response statement to only one sentence of the source text. Characters are matched using *diff_match_patch*[5] and tokens are annotated based on if all characters in the corresponding tokens match. Some additional post-processing (removing spans containing only punctuation and stop words) is done to obtain annotations in the same format as the QuoteSum dataset. The dataset contains 170, 197 annotated statements and 272, 320 annotated spans in the dev and test set respectively.

5.2 Metrics

The evaluation metrics for the two sub-tasks are described separately below.

5.2.1 Metrics for Sub-task 1

For the first sub-task, which involves the identification of spans that have been directly copied from the context, we use token-level precision, recall, and F1 score as our metrics. The ground truth for this task is the set of tokens marked as spans explicitly extracted from the context.

5.2.2 Metrics for Sub-task 2

For the second sub-task, which entails attributing the identified spans to their original positions within the document, we use accuracy as our metric. Accuracy is computed as the fraction of instances where the system correctly predicts the paragraph from which the span was extracted. This allows us to measure how effectively the system can trace the origin of the spans within the document.

5.3 Baselines

In the context of our study, we identify and utilize a number of baselines for both sub-tasks.

[2] Downloaded from https://github.com/google-research-datasets/QuoteSum/tree/main. The dataset is licensed under CC-BY-SA-4.0 license.

[3] Downloaded from https://tinyurl.com/verifiability. The dataset is licensed under MIT license.

[4] https://github.com/nltk/nltk

[5] https://github.com/google/diff-match-patch

5.3.1 Baselines for Sub-task 1

For the first sub-task, the identification of copied spans, it is worth noting that, to our best knowledge, no existing system can perform this task without being explicitly trained which alters the quality of the answers. Common methods such as few-shot prompting and fine-tuning tend to modify the answer. Thus, we resort to a modified version of the SEMQA prompt (Schuster et al., 2023), whereby instead of prompting the system to generate an answer via few-shot prompting, we instruct it to identify spans within a provided answer (prompt included in Appendix A.7, Figure 13). We prompt GPT 3.5[6] and GPT-4[7] using this modified prompt and they serve as our baselines for the first sub-task.

5.3.2 Baselines for Sub-task 2

GPT 3.5 and GPT-4 prompted with the modified SEMQA prompt (prompt included in Appendix A.7, Figure 14) are used as baseline for the second sub-task as well, i.e., attributing the identified spans to their original source paragraph. In addition, we also employ retrieval methods such as BM25 (sparse) (Robertson et al., 2009), GTR (dense) (Ni et al., 2021), and MonoT5 (Nogueira et al., 2020) as baselines for this task.

By comparing the performance of our system against these baselines, we aim to evaluate its effectiveness in tracing the origin of the spans.

6 Results and Analysis

We perform experiments for our approach (§ 4.2) with the following LLMs: Llama-7b (Touvron et al., 2023), Llama-70b (Touvron et al., 2023), Mistral-7b (Jiang et al., 2023), Yi-6b (Young et al., 2024) and OPT-350m (Zhang et al., 2022). Note that OPT-350m has been excluded from comparision on VERIFIABILITY-GRANULAR dataset due to its limitation of 2048 token context length and the dataset contains samples upto 4096 tokens. All of our experiments were run on a A100 machine with 4 80GB GPUs.

6.1 Results for Sub-task 1

The performance of our method and the baselines on sub-task 1, i.e., identifying the copied spans, is summarized in Table 1 and Figure 3 for QuoteSum and VERIFIABILITY-GRANULAR datasets respectively. We observe that our proposed method performs better than GPT models on the QuoteSum dataset and performs on par with GPT-4 on VERIFIABILITY-GRANULAR.

As seen in Table 1, while trying to optimize for F1 on the VERIFIABILITY-GRANULAR dataset the scores we obtain indicate that the optimal solution is to mark all tokens as copied. Therefore, for a better analysis of model performance we plot the PR curves as shown in Figure 3.

The consistency in performance metrics across all models for this sub-task suggests that the capability to identify extracted tokens via hidden representations is not confined to any particular model family or size. This observation underscores the broader applicability and versatility of our proposed methodology, affirming its effectiveness regardless of the underlying architecture or capacity of the language model being used.

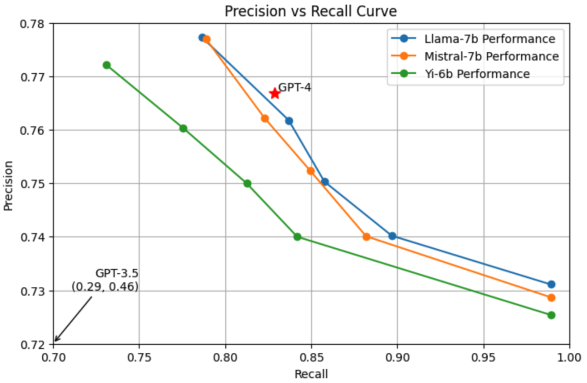


Figure 3: Performance of our method and baselines on the VERI-GRAN test set illustrated using the Precision-Recall curve.

The selection of the best hyperparameters (layer, threshold) for each of our models was based on the F1 performance metric. We share more details in the Appendix (Section A.3). The results presented in Table 1 and Figure 3 are derived from the test set.

The results indicate a considerable shortfall in the performance of GPT-3.5, particularly in terms of recall. This suggests that the model is unable to accurately identify all tokens considered by the annotators as being extracted from the document. A representative example of this failure case is illustrated in the Appendix (Figure 6, Section A.4).

As can be seen from the example, GPT-3.5 tends to identify entities rather than the specific spans that have been directly copied from the document. Even GPT-4, which demonstrates a better understanding of the task, occasionally overlooks certain

[6] https://openai.com/blog/chatgpt
[7] https://openai.com/research/gpt-4

Model	QuoteSum			Verifiability-Granular		
	P	R	F1	P	R	F1
GPT-3.5	0.92	0.46	0.56	0.46	0.29	0.36
GPT-4	**0.96**	0.87	0.90	**0.76**	0.83	0.79
Llama-7b (Ours)	0.96	0.97	**0.96**	0.73	**0.99**	**0.84**
Mistral-7b (Ours)	0.94	0.98	**0.96**	0.73	0.99	0.84
Yi-6b (Ours)	0.94	**0.99**	0.96	0.73	0.99	0.84
OPT-350m (Ours)	0.94	0.99	0.96	-	-	-

Table 1: Token level P, R & F1 scores for identifying output tokens extracted from the document on QuoteSum and Verifiability test sets. For PR plot on VERIFIABILITY-GRANULAR dataset, please refer to Figure 3

components like "The" that are parts of the directly copied spans.

An additional challenge with the GPT models is their tendency to hallucinate, leading to the introduction or elimination of content that doesn't exist in the original answer. We try to recover from the failure by taking spans marked on the modified response and superimposing it on the original response.

To elaborate our observations, we delve into the performance variation across different model families and layers in this sub-task, as illustrated in Figure 4. For more precise comparisons among closely located numbers, the later layers from the Yi-6b model have been excluded due to their substantially lower performance compared to others.

The smoothest graph among all the model families belongs to OPT-350m which shows a gradual increase in performance until the middle layer, followed by a similarly gradual decrease. This performance peak at the middle layers aligns with the findings of (Zou et al., 2023), suggesting that earlier layers are often dedicated to low-level tasks, while later layers tend to be excessively focused on next token prediction.

Interestingly, the larger models, despite their varied architectures, exhibit a shared performance trend, with a peak at the earlier layers followed by a subsequent decrease. This pattern suggests that for these larger models, the sub-task of identifying extracted tokens tends to be a relatively low-level or straightforward task. Consequently, the best performance for these models is typically observed in the earlier layers.

6.2 Results for Sub-task 2

Table 2 presents the performance results for the second sub-task. For each model utilizing our methodology in this task, the sole hyperparameter is the layer. The results provided in the table are derived from evaluations conducted on the test set of QuoteSum and VERIFIABILITY-GRANULAR datasets.

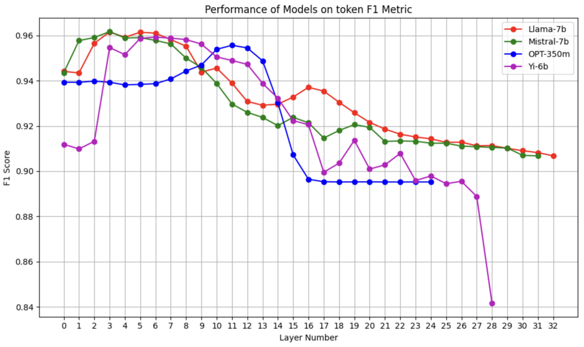


Figure 4: Comparison of model token F1 performance across layers of different models, for identifying output tokens extracted from the document on QuoteSum train set.

Model	Accuracy (%)	
	QUOTESUM	VERI-GRAN
GPT-3.5	90.18	26.40
GPT-4	**90.59**	62.11
BM25	75.72	68.20
GTR	72.57	53.15
MT5	89.24	67.43
Llama-7b (Ours)	87.51	77.33
Mistral-7b (Ours)	89.95	**77.71**
Yi-6b (Ours)	89.24	77.61
OPT-350m (Ours)	75.29	–

Table 2: Paragraph-level accuracy for attributing extractive spans on QuoteSum and VERI-GRAN test sets.

The results indicate that our methods perform on par with GPT-4 ($\sim$ 90%) on the QuoteSum dataset, while they outperform GPT-4 by $\sim$ 15% on VERIFIABILITY-GRANULAR. Note that the human performance on VERI-GRAN is 92.04%, calculated as the mean performance of 4 NLP practitioners on the dataset.

The reason for this is the much harder nature of the later dataset due to the larger number of passages (3.38 vs 69.1 passages on average). Prompting GPT to attribute already generated answers is essentially a retrieval task. It has been observed by (Liu et al., 2023a) that LLM retrieval performance degrades with increasing context length. Our method benefits from the contextual representations of LLM embeddings without relying on the LLM for retrieval and matching representations directly.

Two noteworthy observations from the dataset and results are:

1. The specific nature of the datasets necessitate a delicate balance between two competing tasks: directly matching the substring in the span to be attributed, and disambiguating among multiple instances of the substring. The former is a low-level task, while the latter requires a more nuanced, high-level understanding of context-based matching.

2. The successful performance of models across different families and capacities suggests that approach of matching based on hidden representations is broadly applicable.

Similar to sub-task 1, we carry out a detailed analysis of performance on sub-task 2 across various model configurations and layers, as depicted in Figure 5.

A fundamental difference between the OPT-350m and larger models lies in the influence of token positions on the layer 0 embedding. For OPT-350m, the token's position directly impacts its layer 0 representation. This differs significantly from larger models, where the layer 0 representation depends solely on the token ID like Word2Vec (Mikolov et al., 2013), with the token's position not having a direct influence. Instead, in larger models, the token position is a part of the transformer's learning process (Su et al., 2023).

In light of this distinction, our methodology, observed in Table 2, excludes layer 0 for the larger models while choosing the best layer. Because its representation, which is solely dependent on the token ID, tends to prioritize exact substring matches.

Therefore, OPT-350m begins with lower performance that gradually at the middle and the final layers. On the other hand, all larger models across different families exhibit a similar trend where they start with the highest performance at layer 0. This high initial performance is largely attributable to the dataset's characteristics, which contain numerous examples that can be unambiguously matched using exact substring match.

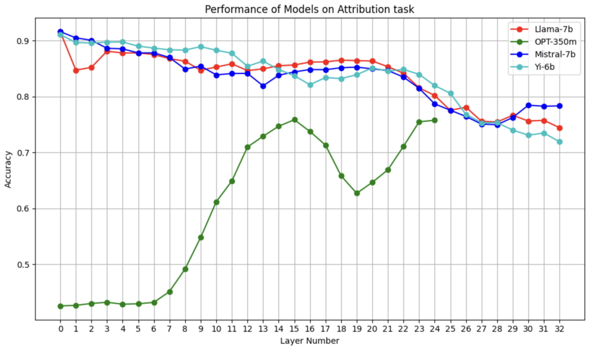


Figure 5: Comparison of model accuracy across layers for attributing extractive spans on QuoteSum train set.

However, the dataset also comprises examples with multiple similar substrings, necessitating disambiguation. This requirement for contextual understanding is well catered to by the middle and later layers, as our ablation study in the Appendix (Section A.2) reveals. Thus, the slight upturns in performance experienced by the larger models in the middle and later layers indicates the trade-off between exact and contextual matching.

7 Discussion & Conclusion

Through this study, we introduce a novel, efficient solution to the challenge of attribution in contextual question answering. Our method leverages the hidden state representations of LLMs, providing detailed, granular attributions without requiring extensive model retraining. Our approach's ability to perform token-level attribution easily lends itself to end-user applications, and we have included an actual system input-output in the Appendix (Section A.6) depicting the same.

Our experimental results, across various model configurations and layers, demonstrate the efficacy of our approach. Notably, our method performs on par or better than training-free baselines in both identifying tokens in an answer that are directly copied from the context and attributing these tokens to their original positions in the document. Additionally we make available VERIFIABILITY-GRANULAR dataset, which contains token-level attributions for LLM generations in the contextual question answering setup.

Perhaps one of the most interesting findings from our study is that the ability to identify extracted tokens and attribute them back to the source is not exclusive to a specific model family. This suggests that our approach can be applied broadly across different LLM architectures, underscoring its versatility.

8 Limitations & Future Work

One limitation of our work stems from the specific nature of the QuoteSum and VERIFIABILITY-GRANULAR datasets we utilized for our experiments. The datasets are constructed such that only verbatim spans from the document are annotated and attributed in the answers. Our method, thus, is primarily evaluated on verbatim spans. However, our method is not specifically designed for verbatim spans only; it could potentially work with paraphrased information as indicated in our experiments with a synthetically paraphrased version of Quotesum presented in the Appendix (Section A.5).

In the future, we intend to employ datasets that include paraphrased spans. Testing on such data will allow us to assess the method's effectiveness in attributing paraphrased information, identify potential challenges, and adapt our methodology accordingly.

We also see potential for our method to be useful beyond the realm of contextual question answering. We tried mapping an LLM generated text span in Spanish to its source document in English (Figure 12).

Interestingly, our method shows potential for this use case too. We envision applying our approach to other tasks that involve information extraction and attribution to a given context. This will not only enhance our understanding of the generation process of LLMs but also extend the applicability and value of our method. We are currently limited by the availability of datasets to perform detailed analyses of other applications.

References

Marco Ancona, Enea Ceolini, A. Cengiz Öztireli, and Markus H. Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Brooke Borel. 2023. *The Chicago guide to fact-checking*. University of Chicago Press.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: developing and governing AI that does not lie. *CoRR*, abs/2110.06674.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1163–1168. The Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023a. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. 2021. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Inf. Fusion*, 71:28–37.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 11–20.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *ArXiv*, abs/2311.03731.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. A survey on truth discovery. *SIGKDD Explor.*, 17(2):1–16.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023b. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *ArXiv*, abs/2309.07852.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick S. H. Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. 2022. Improving wikipedia verifiability with AI. *CoRR*, abs/2207.06220.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake news challenge*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–64.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Gabriele Sarti, Grzegorz Chrupala, Malvina Nissim, and Arianna Bisazza. 2023a. Quantifying the plausibility of context reliance in neural machine translation. *CoRR*, abs/2310.01188.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023b. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Tal Schuster, Adam D Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William W Cohen, and Donald Metzler. 2023. Semqa: Semi-extractive multi-source question answering. *arXiv preprint arXiv:2311.04886*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, page 127063.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (FEVER) shared task. *CoRR*, abs/1811.10971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pre-training data. *arXiv preprint arXiv:2305.13252*.

Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Inference with reference: Lossless acceleration of large language models.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Do Xuan Long, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4736–4756. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Performance over target span position

In this section, we analyze if the position of the target span (the span that needs to be attributed) within the answer influences the efficacy of our methodology. The start character of each span is normalized with respect to the length of the answer and the accuracy is measured for all the models on Sub-task2. As shown in Figure 8, the accuracy improves with increasing span position for all models. In decoder-only models each step can access information from previously generated or processed tokens only. As the model progresses through the text, it accumulates more information and context about the text it is analyzing. We believe that this behaviour is the reason why accuracy increases when we attribute later text spans.

A.2 Performance of layers in the disambiguation task setting

Given a span that needs to be attributed, there are instances in the dataset where ground truth attribution sub-string is present multiple times over different sources. During inference, our approach has to both identify the correct attribution sub-string and also disambiguate between the multiple occurances of the sub-string. Disambiguating over these multiple occurences, requires contextual understanding of the span to be attributed in order to choose between the occurences. Data points that require disambiguation are collected from the test set of the QuoteSum dataset.

We measure performance across all layers for different models on this subset (Figure 9). We compute the performance when one of the multiple occurrences of the sub-string is chosen randomly. This is highlighted as the random baseline in the figure.

Figure 9 shows an interesting trend, where early layers perform worse than the random baseline on the disambiguation task. Early layers capture low level, basic features of the input text, making them less suitable for a complex task such as disambiguation. Performance improves as we use later layers, with the middle layers yielding the best performance for Llama-70b and Mistral-7b, and later layers yielding the best performance for Llama-7b and OPT-350m.

A.3 Effect of hyper-parameters on Sub-task 1 performance

We present graphs for precision, recall and F1 in Figure 7 generated during our experiments on the QuoteSum train set, which show the impact of the threshold θ, and the choice of layer on the performance of our method (Llama-7b) for Sub-task 1. θ was chosen for each layer-model setup to achieve the best f1 score. Interestingly we note that in the earlier layers, higher thresholds are preferred. We hypothesize that there's a higher chance of low-level information overlapping, necessitating more stringent filtering at the earlier layers.

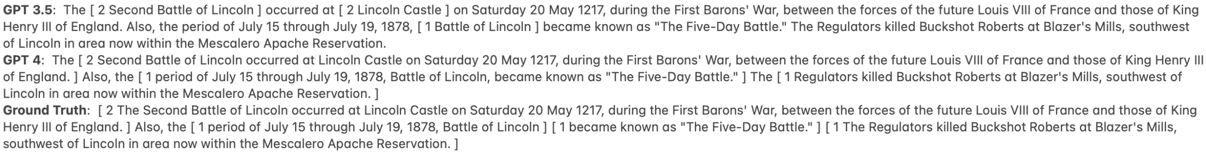
A.4 Limitations of GPT models for Sub-task 1

Figure 6 depicts a qualitative example where GPT models fail in identifying directly copied spans from a source document on the QuoteSum dataset. In this example GPT-3.5 is only attributing entities like Lincoln Castle, Battle of Lincoln and the Second Battle of Lincoln. GPT-4 has a better understanding of the task but leaves out tokens like 'The' which were copied verbatim from the source.

A.5 Synthetic Paraphrasing of QuoteSum

We paraphrase the QuoteSum dataset by prompting GPT-4 as shown below. A sample from the dataset is shown in Figure 10.

```
Document:
{document}
Based on the information contained in
    the document, answer the question
    with details to the best of your
    abilities. Think step by step and
    explain your answer if that will
    help better understand the answer.
Q: {question} A:
{answer}
Given the above source passages, a
    question and an answer. The answer
    summarizes the given sources while
    explicitly copying spans from the
    sources. Paraphrase the non-entity
```



GPT 3.5: The [2 Second Battle of Lincoln] occurred at [2 Lincoln Castle] on Saturday 20 May 1217, during the First Barons' War, between the forces of the future Louis VIII of France and those of King Henry III of England. Also, the period of July 15 through July 19, 1878, [1 Battle of Lincoln] became known as "The Five-Day Battle." The Regulators killed Buckshot Roberts at Blazer's Mills, southwest of Lincoln in area now within the Mescalero Apache Reservation.
GPT 4: The [2 Second Battle of Lincoln occurred at Lincoln Castle on Saturday 20 May 1217, during the First Barons' War, between the forces of the future Louis VIII of France and those of King Henry III of England.] Also, the [1 period of July 15 through July 19, 1878, Battle of Lincoln, became known as "The Five-Day Battle."] The [1 Regulators killed Buckshot Roberts at Blazer's Mills, southwest of Lincoln in area now within the Mescalero Apache Reservation.]
Ground Truth: [2 The Second Battle of Lincoln occurred at Lincoln Castle on Saturday 20 May 1217, during the First Barons' War, between the forces of the future Louis VIII of France and those of King Henry III of England.] Also, the [1 period of July 15 through July 19, 1878, Battle of Lincoln] [1 became known as "The Five-Day Battle."] [1 The Regulators killed Buckshot Roberts at Blazer's Mills, southwest of Lincoln in area now within the Mescalero Apache Reservation.]

Figure 6: Limitations of GPT models in identifying directly copied spans from a source document

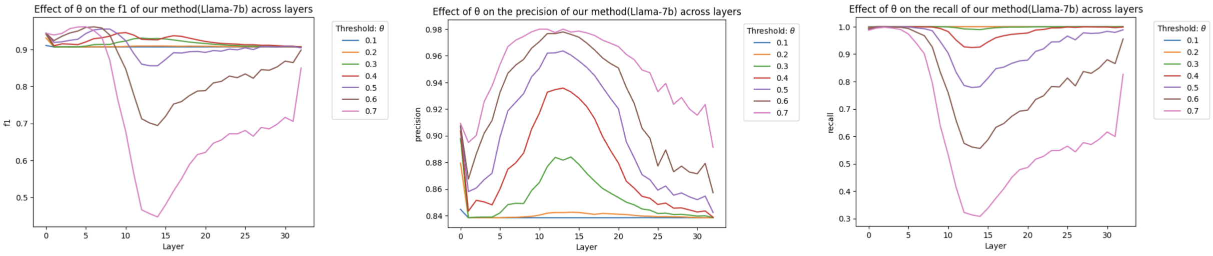


Figure 7: Effect of the threshold θ on the performance of our method across different layers of Llama-7b on QuoteSum train set.

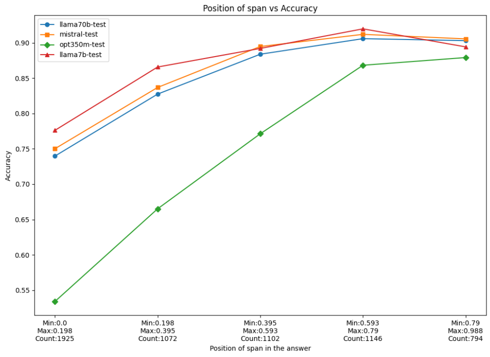


Figure 8: Performance of different models over span position on Sub-task 2.

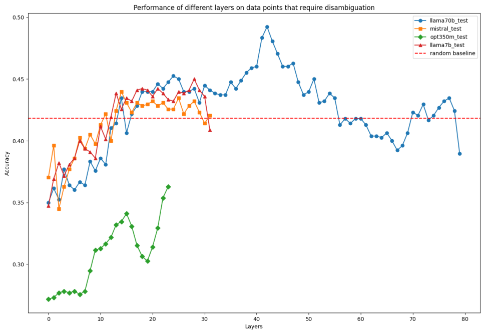


Figure 9: Performance of different models over the disambiguation task.

```
    parts of the answer within [] while
    keeping entities intact and rewrite
    in the same format as original
    answer.
Paraphrased Answer:
```

We computed the sub-task 2 performance of the various methods on the paraphrased test set and summarize the results in Table 3.

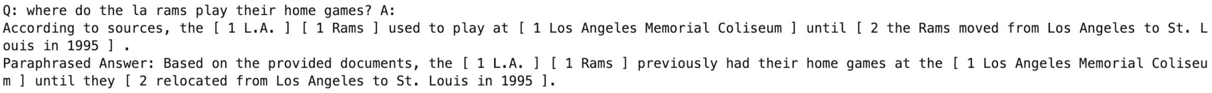
We observe that the performance of all the methods drop on the paraphrased version of QuoteSum by nearly the same amount(1-3%). This probably implying that synthetic paraphrasing setup while harder than the original dataset is not difficult enough to offer additional insights.

A.6 Qualitative Example for Sub-task 1 and Sub-task 2

To illustrate the utility of our method, a qualitative example is presented in Figure 11.

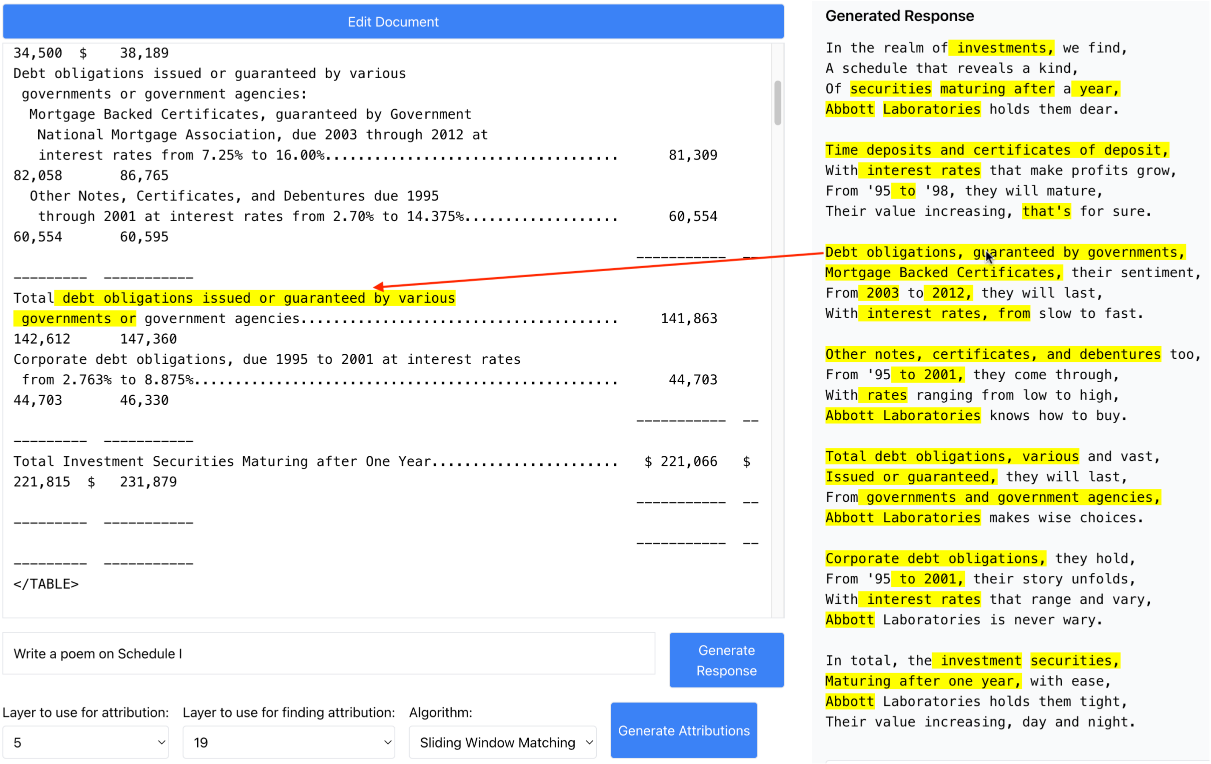
A.7 Prompts for GPT baselines

We present the prompts used for our GPT-based baselines for Sub-task 1 and Sub-task 2 in Figure 13 and Figure 14 respectively. For Sub-task 1, the LLM has to respond with copied spans along with the source paragraph number. For Sub-task 2, we mark the verbatim copied spans in "[]" and prompt the LLM to identify the source paragraph only.



```
Q: where do the la rams play their home games? A:
According to sources, the [ 1 L.A. ] [ 1 Rams ] used to play at [ 1 Los Angeles Memorial Coliseum ] until [ 2 the Rams moved from Los Angeles to St. L
ouis in 1995 ] .
Paraphrased Answer: Based on the provided documents, the [ 1 L.A. ] [ 1 Rams ] previously had their home games at the [ 1 Los Angeles Memorial Coliseu
m ] until they [ 2 relocated from Los Angeles to St. Louis in 1995 ].
```

Figure 10: Sample datapoint from the paraphrased QuoteSum dataset



```
34,500  $    38,189
Debt obligations issued or guaranteed by various
 governments or government agencies:
  Mortgage Backed Certificates, guaranteed by Government
   National Mortgage Association, due 2003 through 2012 at
   interest rates from 7.25% to 16.00%..................................      81,309
82,058       86,765
  Other Notes, Certificates, and Debentures due 1995
   through 2001 at interest rates from 2.70% to 14.375%..................      60,554
60,554       60,595
                                                                          -----------  --
---------  -----------
Total debt obligations issued or guaranteed by various
 governments or government agencies......................................     141,863
142,612      147,360
Corporate debt obligations, due 1995 to 2001 at interest rates
 from 2.763% to 8.875%...................................................      44,703
44,703       46,330
                                                                          -----------  --
---------  -----------
Total Investment Securities Maturing after One Year......................   $ 221,066   $
221,815  $   231,879
                                                                          -----------  --
---------  -----------
                                                                          -----------  --
---------  -----------
</TABLE>
```

Write a poem on Schedule I

Layer to use for attribution: 5

Layer to use for finding attribution: 19

Algorithm: Sliding Window Matching

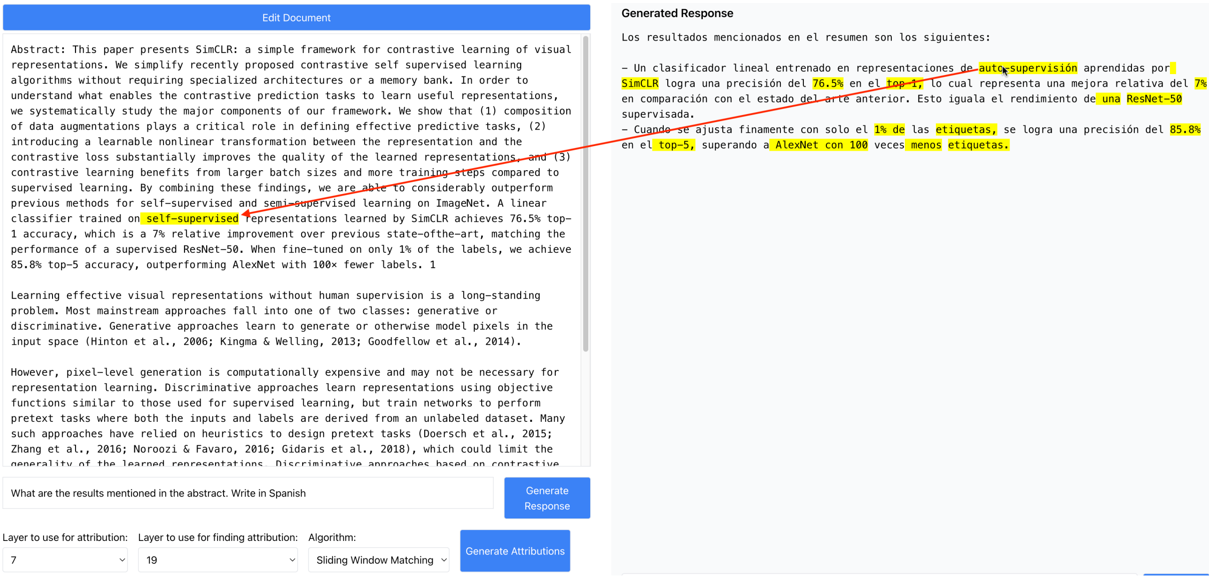
Generated Response

```
In the realm of investments, we find,
A schedule that reveals a kind,
Of securities maturing after a year,
Abbott Laboratories holds them dear.

Time deposits and certificates of deposit,
With interest rates that make profits grow,
From '95 to '98, they will mature,
Their value increasing, that's for sure.

Debt obligations, guaranteed by governments,
Mortgage Backed Certificates, their sentiment,
From 2003 to 2012, they will last,
With interest rates, from slow to fast.

Other notes, certificates, and debentures too,
From '95 to 2001, they come through,
With rates ranging from low to high,
Abbott Laboratories knows how to buy.

Total debt obligations, various and vast,
Issued or guaranteed, they will last,
From governments and government agencies,
Abbott Laboratories makes wise choices.

Corporate debt obligations, they hold,
From '95 to 2001, their story unfolds,
With interest rates that range and vary,
Abbott Laboratories is never wary.

In total, the investment securities,
Maturing after one year, with ease,
Abbott Laboratories holds them tight,
Their value increasing, day and night.
```

Figure 11: Given a document and a question, the user selects generate response to get the answer. Once the answer is generated, to identify extractive answer tokens, the user selects the Generate attributions button. Answer tokens copied from the document are highlighted and are clickable. When the user selects 'Debt Obligations..' in the generated response, the corresponding attribution is shown on the left panel.

Model	Accuracy
BM25	73.99
GTR	71.79
MT5	**88.46**
Llama-7b (Ours)	85.74
Mistral-7b (Ours)	87.06
Yi-6b (Ours)	86.99
OPT-350m (Ours)	71.18

Table 3: Paragraph-level accuracy for attributing extractive spans on paraphrased QuoteSum test set.



Abstract: This paper presents SimCLR: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-ofthe-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100× fewer labels. 1

Learning effective visual representations without human supervision is a long-standing problem. Most mainstream approaches fall into one of two classes: generative or discriminative. Generative approaches learn to generate or otherwise model pixels in the input space (Hinton et al., 2006; Kingma & Welling, 2013; Goodfellow et al., 2014).

However, pixel-level generation is computationally expensive and may not be necessary for representation learning. Discriminative approaches learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset. Many such approaches have relied on heuristics to design pretext tasks (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018), which could limit the generality of the learned representations. Discriminative approaches based on contrastive

What are the results mentioned in the abstract. Write in Spanish

Layer to use for attribution: 7

Layer to use for finding attribution: 19

Algorithm: Sliding Window Matching

Generated Response

Los resultados mencionados en el resumen son los siguientes:

- Un clasificador lineal entrenado en representaciones de auto-supervisión aprendidas por SimCLR logra una precisión del 76.5% en el top-1, lo cual representa una mejora relativa del 7% en comparación con el estado del arte anterior. Esto iguala el rendimiento de una ResNet-50 supervisada.
- Cuando se ajusta finamente con solo el 1% de las etiquetas, se logra una precisión del 85.8% en el top-5, superando a AlexNet con 100 veces menos etiquetas.

Figure 12: Given a document and a question, the user selects "Generate response" to get the answer. Once the answer is generated, to identify extractive answer tokens, the user selects the "Generate Attributions" button. Answer tokens copied from the document are highlighted and are clickable. When the user selects "auto super.." in the generated response, the corresponding attribution is shown on the left panel.

```
Given source passages, a question and an answer. The answer summarizes the given sources while explicitly copying spans from the
sources. Identify the copied spans using brackets and the respective source number to indicate that this span was copied. Make
sure to use all relevant sources and properly quote them. Here are some examples:

Question: Who was the bond girl in you only live twice?
[1]You Only Live Twice (film) : Blofeld then orders Osato to kill Bond. Bond then meets ...
[2]Aki (James Bond) : Aki is a fictional character created for the 1967 "James Bond" film "You Only Live Twice” ...
[3]Aki (James Bond) : Following her death, Aki is almost immediately replaced by Kissy Suzuki and is not ... [4]You Only Live
Twice (film) : The tall set could be seen from away, and attracted many people from ...
[5]Aki (James Bond) : Aki does not appear in Ian Fleming's 1964 novel. She was originally named Suki ...
Answer: Aki is the Bond girl in You Only Live Twice , although Kissy Suzuki is considered the 'main' Bond girl in this film, but
Aki has a bigger role and is more memorable .
Quoted answer: [ 3 Aki ] is [ 3 the ] [ 3 Bond ] girl in [ 1 You Only Live Twice ] , although [ 3 Kissy Suzuki is considered the
'main' Bond girl in this film, but Aki has a bigger role and is more memorable ] .

Question: who are the new cast members in american horror story?
[1]American Horror Story: Roanoke : American Horror Story: Roanoke is the sixth season of the FX horror anthology television  ...
[2]American Horror Story: Coven : Returning cast members from the previous seasons of the series include: Sarah Paulson ...
[3]American Horror Story: Freak Show : Returning cast members from the of the series include: Sarah Paulson, Evan Peters ...
[4]American Horror Story: Cult : Returning cast members from previous seasons include Sarah Paulson, Evan Peters, Cheyenne ...
[5]American Horror Story : cast. In September 2016, the full main cast was announced after the first episode with the ...
Answer: According to sources, the new cast members in Roanoke are Cuba Gooding Jr. and André Holland . In Coven, sources say the
new cast members are Emma Roberts and Kathy Bates. In Freak Show, the new cast members are Michael Chiklis and Finn
Wittrock. In Cult, the new cast members are Billie Lourd and Alison Pill.
Quoted answer: According to sources, the new [ 1 cast members ] in [ 1 Roanoke ] are [ 1 Cuba Gooding Jr. and André Holland ] . In
Coven, sources say the [ 2 new cast members ] are [ 2 Emma Roberts and Kathy Bates. ] In Freak Show, the [ 3 new cast members ]
are [ 3 Michael Chiklis and Finn Wittrock. ] In Cult, the [ 4 new cast members ] are [ 4 Billie Lourd and Alison Pill. ]

Question: {question}
{document}
Answer: {clean_answer}
Quoted answer:
```

Figure 13: Prompt for GPT-based baselines for Task-1.

```
Given source passages, a question and an answer. The answer summarizes the given sources while explicitly copying spans from the
sources. Identify the copied spans using brackets and the respective source number to indicate that this span was copied. Make
sure to use all relevant sources and properly quote them. Here are some examples:

Question: Who was the bond girl in you only live twice?
[1]You Only Live Twice (film) : Blofeld then orders Osato to kill Bond. Bond then meets ...
[2]Aki (James Bond) : Aki is a fictional character created for the 1967 "James Bond" film "You Only Live Twice” ...
[3]Aki (James Bond) : Following her death, Aki is almost immediately replaced by Kissy Suzuki and is not ... [4]You Only Live
Twice (film) : The tall set could be seen from away, and attracted many people from ...
[5]Aki (James Bond) : Aki does not appear in Ian Fleming's 1964 novel. She was originally named Suki ...
Marked answer: [ Aki ] is [ the ] [ Bond ] girl in [ You Only Live Twice ] , although [ Kissy Suzuki is considered the 'main' Bond
girl in this film, but Aki has a bigger role and is more memorable ] .
Annotated answer: [ 3 Aki ] is [ 3 the ] [ 3 Bond ] girl in [ 1 You Only Live Twice ] , although [ 3 Kissy Suzuki is considered
the 'main' Bond girl in this film, but Aki has a bigger role and is more memorable ] .

Question: who are the new cast members in american horror story?
[1]American Horror Story: Roanoke : American Horror Story: Roanoke is the sixth season of the FX horror anthology television  ...
[2]American Horror Story: Coven : Returning cast members from the previous seasons of the series include: Sarah Paulson ...
[3]American Horror Story: Freak Show : Returning cast members from the of the series include: Sarah Paulson, Evan Peters ...
[4]American Horror Story: Cult : Returning cast members from previous seasons include Sarah Paulson, Evan Peters, Cheyenne ...
[5]American Horror Story : cast. In September 2016, the full main cast was announced after the first episode with the ...
Marked answer: According to sources, the new [ cast members ] in [ Roanoke ] are [ Cuba Gooding Jr. and André Holland ] . In
Coven, sources say the [ new cast members ] are [ Emma Roberts and Kathy Bates. ] In Freak Show, the [ new cast members ] are [
Michael Chiklis and Finn Wittrock. ] In Cult, the [ new cast members ] are [ Billie Lourd and Alison Pill. ]
Annotated answer: According to sources, the new [ 1 cast members ] in [ 1 Roanoke ] are [ 1 Cuba Gooding Jr. and André Holland ] .
In Coven, sources say the [ 2 new cast members ] are [ 2 Emma Roberts and Kathy Bates. ] In Freak Show, the [ 3 new cast members ]
are [ 3 Michael Chiklis and Finn Wittrock. ] In Cult, the [ 4 new cast members ] are [ 4 Billie Lourd and Alison Pill. ]

Question: {question}
{document}
Answer: {clean_answer}
Quoted answer:
```

Figure 14: Prompt for GPT-based baselines for Task-2.