TRACE: TRansformer-based Attribution using Contrastive Embeddings in LLMs

Anonymous ACL submission

Abstract

The rapid evolution of *large language models* (LLMs) represents a substantial leap forward in natural language understanding and generation. However, alongside these advancements come significant challenges related to the accountability and transparency of LLM outputs. Reliable source attribution is essential to adhering to stringent legal and regulatory standards, including those set forth by the General Data Protection Regulation. Despite the well-established methods in source attribution within the computer vision domain, the application of robust attribution frameworks to natural language processing remains underexplored. To bridge this gap, we propose a novel and versatile TRansformer-based Attribution framework using Contrastive Embeddings called TRACE that, in particular, exploits contrastive learning for source attribution. We perform an extensive empirical evaluation to demonstrate the performance and efficiency of TRACE in various settings and show that TRACE significantly improves the ability to attribute sources accurately, making it a valuable tool for enhancing the reliability and trustworthiness of LLMs.

1 Introduction

The recent era has seen a significant rise in the prevalence of *large language models* (LLMs) (Ouyang et al., 2022; Touvron et al., 2023) which have demonstrated an array of remarkable capabilities. However, studies (Huang et al., 2023; Liu et al., 2024; Wang et al., 2023a) have highlighted a critical concern on the accountability of LLMs. Considering the widespread usage and such a concern, it has brought to the forefront a critical need for source attribution that involves identifying the specific training data that contributes to generating part or all of an LLM's output, which is crucial for legal and regulatory compliance and enhances the reliability of LLMs. Various regulations mandate transparency and accountability in data usage, especially regarding intellectual property and privacy. For instance, the General Data Protection Regulation (GDPR) in the European Union requires that individuals have the right to be informed when their personal data is used. Proper source attribution ensures compliance with such legal frameworks, mitigating the risk of legal disputes and penalties.

A related topic would be that of *membership inference* (MI) (Mireshghallah et al., 2022) whose task is to determine whether a given piece of data was used during the training of a machine learning model. While MI and source attribution share some similarities, they differ significantly in their granularity: MI typically only involves a binary classification task and does not require identifying a specific data provider. In contrast, source attribution requires to identify one or more data providers.

Though there are some studies on source attribution (Marra et al., 2018; Yu et al., 2022), a majority of them are situated within the computer vision domain. Techniques developed for computer vision tasks cannot be directly applied to LLMs due to the fundamental differences in the data and model architectures. To the best of our knowledge, effective source attribution for LLMs still remains an open and underexplored problem.

While numerous properties are important to a source attribution framework, we identify **accuracy**, **scalability**, and **interpretability** as the most crucial components. These three attributes are fundamental to ensuring the effectiveness and applicability of the framework across various contexts. Accuracy is essential to guaranteeing that the framework consistently produces reliable results. Scalability ensures that the framework can handle increasing volumes of data and complexity without a significant performance degradation, making it suitable for large-scale applications. Interpretability is equally critical as it enables stakeholders to understand and trust the attribution outcomes, hence fostering transparency and facilitating informed decision making.

This paper presents a novel TRansformer-based Attribution framework using Contrastive Embeddings (TRACE) to achieve source attribution while satisfying the above three important properties. By detailing our methodology and presenting empirical results, we seek to demonstrate the accuracy, scalability, and interpretability of TRACE.

Our contributions can be summarized as follows:

- We propose the novel TRACE framework based on contrastive learning, which is designed to achieve effective source attribution. TRACE differs from traditional contrastive learning by using source information as the label. Fig. 1 illustrates the TRACE framework.
- We have performed an extensive empirical evaluation of TRACE to demonstrate its accuracy, scalability, and interpretability.

2 Preliminaries

Contrastive Learning and NT-Xent Loss. Contrastive learning is a conventional technique commonly used in representation learning (Arora et al., 2019; Hadsell et al., 2006). Its underlying idea is that similar objects should exhibit a closer distance in the embedding space while dissimilar objects should repel each other. This technique has been widely employed in computer vision tasks due to its convenient implementation to augment image input to form a self-supervised problem. Models using contrastive learning have achieved state-of-the-art performances (Cui et al., 2021; Tian et al., 2020). Apart from the attention it receives in computer vision, new approaches using contrastive learning in natural language processing (Meng et al., 2021; Wu et al., 2020) have also started gaining attention and showcasing great capabilities.

Our TRACE framework assigns the same label to all the data from the same source, hence naturally forming a *supervised contrastive learning* problem. In particular, TRACE utilizes NT-Xent Loss (Sohn, 2016) for supervised contrastive learning:

$$\mathcal{L} = \sum_{i \in I} \frac{-1}{|P_i|} \sum_{p \in P_i} \log \left(\frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_p / \tau\right)}{\sum_{a \in A_i} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_a / \tau\right)} \right)$$

where the set I ($P_i \subset I$) contains indices of the sentences in the given batch (sharing the same label as sentence i, but does not include i), $A_i = I \setminus \{i\}$, $\mathbf{z}_i$ denotes the embedding of sentence i, and $\tau \in \mathbb{R}^+$ is a temperature parameter. Minimizing $\mathcal{L}$ would maximize the similarity between embeddings (of sentences) from the same source while minimizing the similarity between embeddings from different sources.

Sentence Encoder. Similar to the concepts of Word2Vec (Mikolov et al., 2013) and GloVe (Brochier et al., 2019) which produce meaningful vector representations of words, such techniques can be applied to larger text units such as sentences. A straightforward way is to take the average of word embeddings within a sentence, but this often results in embeddings that lack semantic depth. Several models have been developed to address this issue, including InferSent (Conneau et al., 2018), Universal Sentence Encoder (Cer et al., 2018), and Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Given its superior performance and efficiency, SBERT is chosen to generate sentence embeddings in TRACE. SBERT leverages a pre-trained BERT network and utilizes Siamese and triplet network structures to produce semantically meaningful sentence embeddings.

3 TRACE Framework

3.1 Source-Specific Semantic Distillation

Projecting every piece of data from each provider into the embedding space is desirable but would incur considerable computational costs. Moreover, it is prudent to recognize that not all information carries equal importance: For example, sentences that occur less frequently typically tend to be more representative of the document. So, we propose to extract principal sentences from each source by leveraging the *Term Frequency-Inverse Document Frequency* (TF-IDF) which is effective for identifying significant sentences within documents. It is generally recommended to select 10-20% of the sentences, thereby striking a balance between complexity and performance; these sentences are subsequently defined as *principal sentences*. The length of these sentences is specified by a parameter called WINDOW_SIZE. Section 4.7 presents an ablation study examining the effect of different WINDOW_SIZEs on accuracy.

SBERT (Reimers and Gurevych, 2019) has proven effective in deriving high-quality sentence

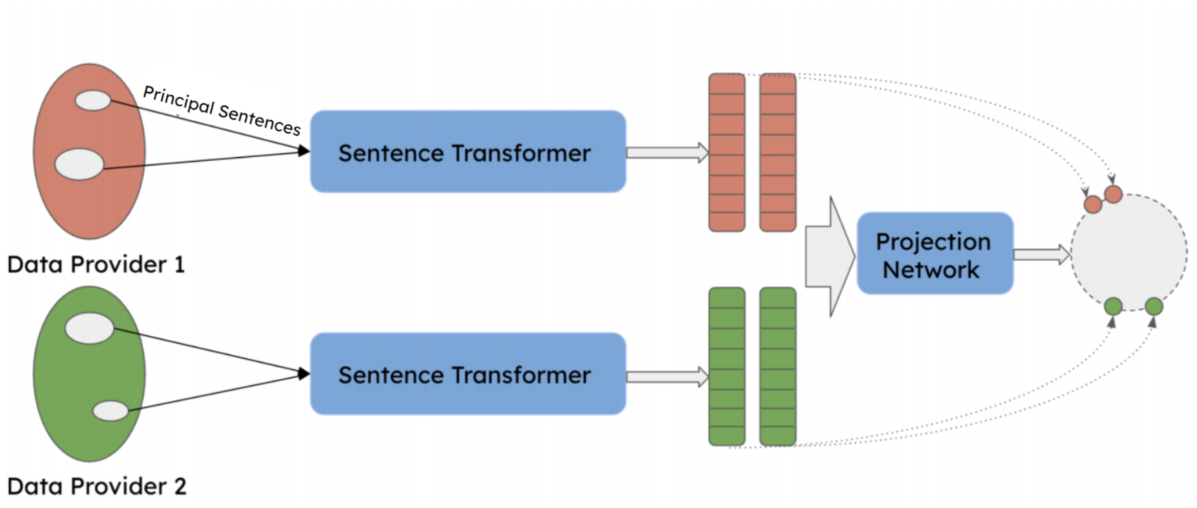

Figure 1: Illustration of TRACE framework.

representations. However, to enhance its suitability for TRACE, we propose several modifications inspired by the work of SimCLR (Chen et al., 2020). A key finding from SimCLR is that adding a nonlinear projection head significantly improves the representation quality. Following this insight, we incorporate a projection network at the end of the traditional SBERT architecture. This projection network is trained together with the base SBERT model, thus encouraging the learned representations to be more discriminative in the embedding space.

3.2 Supervised Contrastive Embedding Training for Source-Coherent Clustering

Unlike the other contrastive learning frameworks in computer vision whose tasks are typically defined to be auto-regressive due to the availability of various data augmentation techniques to generate positive samples, TRACE aims to achieve source-coherent clustering. In our case, we already possess the label of each sentence indicating its source. So, we can frame our task as a *supervised contrastive learning* problem. The supervision is derived from the label information which corresponds to the source. Contrastive learning aligns with our objective to form clusters based on these various sources.

SimCLR has demonstrated that NT-Xent Loss outperforms other contrastive loss functions such as logistic loss (Mikolov et al., 2013) and margin loss (Schroff et al., 2015). So, we employ NT-Xent Loss as the loss function for TRACE.

3.3 Proximity-based Inference

Once the training phase is completed, we transition to the inference stage where each data source is represented by its own set of contrastive embeddings. At this stage, when a language model generates a response, we employ the k-*Nearest Neighbor* (kNN) algorithm to assign the response to the closest data source in the embedding space, as demonstrated in Fig. 2. This ensures accurate source attribution by matching the generated response with its most similar source representation.

However, responses generated by language models may not always be exclusively influenced by a single data source: there could be instances where information from multiple sources contributes to the generated text. To consider this possibility, we introduce the concept of *multi-source attribution*. Multi-source attribution acknowledges and accounts for the potential influence of multiple data sources on the generated response.

We have developed three different implementations for single-source attribution and multi-source attribution, which allow users to select the most appropriate inference method based on time constraints and the number of sources. Section 4 provides a comparison of these methods.

Hard kNN (Single-Source Attribution). *Hard kNN* follows the traditional kNN algorithm closely. Here, the attribution is determined by considering the k embeddings that are closest in distance to the query. The source that appears most frequently among these k neighbors is assigned as the source of the query.

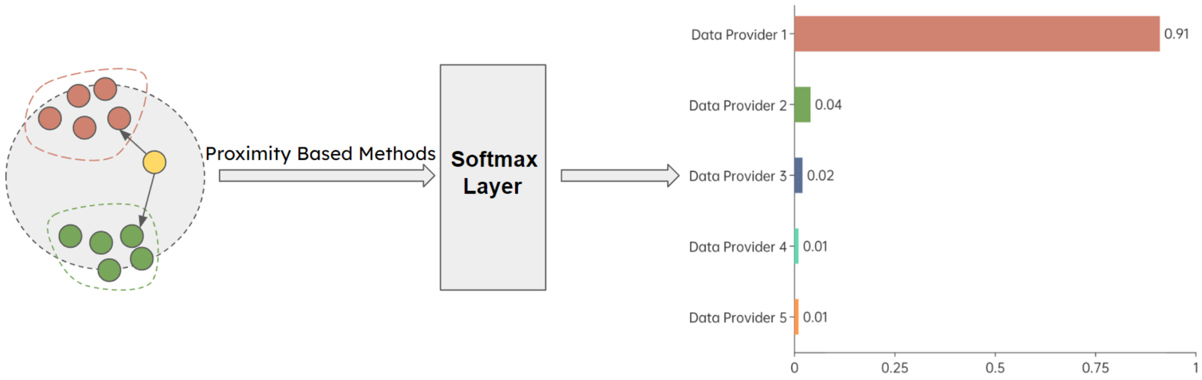

Figure 2: Illustration of the attribution step in TRACE framework.

Soft kNN (Multi-Source Attribution). To differentiate from traditional kNN where each query is assigned to a single source, we introduce *soft kNN*. Here, k represents the number of data sources rather than the number of closest neighbors. We rank the distances from the query to all other embeddings and select them in ascending order of distance until k distinct sources are covered.

Nearest Centroid (Single-Source Attribution). To reduce inference time, we employ the nearest centroid method. Here, the centroid of each cluster is calculated by adding the normalized embeddings within that cluster (i.e., corresponding to the sentences with the same label/source), as shown below. We then apply kNN using these centroids. This method significantly reduces inference time as it scales with the number of data providers rather than the volume of data from each source. We will demonstrate in the next section that this method maintains an impressively high accuracy.

Given a cluster of embeddings $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k$ with the same label/source, a good representative of the cluster would be the centroid $\bar{\boldsymbol{z}}$ that maximizes the sum of its cosine similarity with every normalized embedding $\boldsymbol{z}_i$ for $i = 1, \ldots, k$. Equivalently, $\bar{\boldsymbol{z}}$ minimizes the sum of its standard cosine distance with every normalized embedding:

$$\begin{aligned}
&\sum_{i=1}^{k}\left(1 - \frac{\boldsymbol{z}_i \cdot \bar{\boldsymbol{z}}}{\|\boldsymbol{z}_i\|\|\bar{\boldsymbol{z}}\|}\right) = k - \sum_{i=1}^{k} \frac{\boldsymbol{z}_i \cdot \bar{\boldsymbol{z}}}{\|\boldsymbol{z}_i\|\|\bar{\boldsymbol{z}}\|} \\
&= k - \left(\sum_{i=1}^{k} \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}\right) \cdot \frac{\bar{\boldsymbol{z}}}{\|\bar{\boldsymbol{z}}\|} \geq k - \left|\left(\sum_{i=1}^{k} \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}\right) \cdot \frac{\bar{\boldsymbol{z}}}{\|\bar{\boldsymbol{z}}\|}\right| \\
&\geq k - \left\|\sum_{i=1}^{k} \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}\right\|
\end{aligned}$$

by Cauchy-Schwarz inequality. The equality holds when there exists some $\lambda \in \mathbb{R}$ such that

$$\sum_{i=1}^{k} \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|} = \lambda \frac{\bar{\boldsymbol{z}}}{\|\bar{\boldsymbol{z}}\|} .$$

In other words, $\bar{\boldsymbol{z}}$ can be obtained by adding all normalized embeddings and setting $\lambda = 1$:

$$\bar{\boldsymbol{z}} = \sum_{i=1}^{k} \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|} .$$

4 Experiments

4.1 Experimental Setup

Data. We perform an extensive empirical evaluation of TRACE using three datasets: `booksum` (Kryściński et al., 2022), `dbpedia_14` (Zhang et al., 2015), and `cc_news` (Hamborg et al., 2017); a summary of these datasets can be found in Table 5 in the appendix. In the `booksum` dataset, we treat different books as distinct data providers and vary the number of data providers from 10, 25, 50, to 100 to demonstrate TRACE's scalability to a large number of data providers. Similarly, each class in `dbpedia_14` or each domain in `cc_news` is considered a separate data provider. In this section, we primarily present the experimental results on the `booksum` dataset with 25 data providers. Section 4.6 provides additional results.

Model. Focusing primarily on the `booksum` dataset, we evaluate the performance of TRACE using three different LLMs of varying sizes: `t5-small-booksum` (Raffel et al., 2020), `GPT-2` (Radford et al., 2019), and `Llama-2` (Touvron et al., 2023). The `t5-small-booksum` model is readily available on Hugging Face,[1] while `GPT-2`

[1] https://huggingface.co/cnicu/t5-small-booksum.

and `Llama-2` have been fine-tuned on a subset of the `booksum` dataset. This setup allows us to assess the performance of `TRACE` across LLMs of different scales. In App. A, we provide more details about our experiments.

4.2 Visualization of `TRACE`'s Embedding Space

After training for 150 epochs on `booksum`, a visualization tool such as `UMAP` (McInnes et al., 2020) can be used to view the distribution of principal sentences. Fig. 3 shows that after the contrastive learning step, the desired outcome has been achieved, i.e., data coming from the same source form clear and distinct clusters. This validates that our contrastive learning successfully groups different data providers. Supposing the responses from an LLM are projected into the embedding space without incorporating the contrastive learning step, the resulting neighborhood exhibits chaos and it is challenging to derive robust information. This further demonstrates the importance of the contrastive learning step.

4.3 Accuracy

Evaluating the accuracy of source attribution is particularly challenging due to the inherent difficulty in obtaining ground-truth test datasets. Even with a dataset, a language model, and specific inputs, pinpointing the exact parts of the training data that influence a particular output remains complex. Here are the key reasons:

1. **Lack of Explicit Traceability.** Language models like LLMs generate outputs based on patterns learned from vast amounts of data. However, these models do not provide explicit traceability back to the specific training data. This means we cannot directly observe which parts of the training data contribute to a given output.

2. **Intermixed Training Data.** The training data for LLMs is often a massive, intertwined collection of texts from various sources. Disentangling these sources to identify the precise contribution of each segment to the final output is nearly impossible due to the sheer volume and complexity.

3. **Influence of Pre-training Data.** It is also likely that the model generates outputs based on data encountered during the pre-training stage, which comprises a vast and diverse corpus. This pre-training data is often not fully documented or accessible, making it difficult to determine its influence on specific outputs during fine-tuning or evaluation.

Due to these challenges, obtaining ground-truth test datasets that accurately reflect the contribution of specific training data to the outputs of LLMs is exceedingly difficult. To address this issue, our approach involves using training data where the source is known. We then use this known source as the ground-truth label and evaluate whether `TRACE` can correctly determine the source. This allows us to approximate the evaluation of source attribution by leveraging the known origins of the specific training data.

Single-Source Attribution Accuracy. In this case, accuracy is simply defined as the number of correct source attributions divided by the total number of attributions evaluated, the latter of which is 250 in our experimental setup.

Multi-Source Attribution Accuracy. In certain settings, providing multiple sources and allowing the user to determine the justification of the attribution is acceptable. For a successful *soft kNN* attribution in such cases, the ground-truth source must appear among the *top-k* sources returned by `TRACE`. Using the same setup as that of single-source attribution, we have evaluated `TRACE` on 250 instances. Table 1 below shows the results:

Model	Soft kNN acc.	Soft kNN top-3 acc.	Soft kNN top-5 acc.	Hard kNN ($k = 10$, $k = 20$)	Nearest Centroid
t5	84.4%	95.3%	97.3%	84.4%	84.4%
GPT-2	81.3%	92.3%	94.0%	81.3%	81.3%
Llama-2	86.2%	96.1%	97.2%	86.2%	86.2%

Table 1: Source attribution accuracy for 25 data providers on `booksum` dataset using `TRACE`.

It can be observed that the accuracy for models of different sizes remains consistently high and significantly surpasses the random guess' accuracy of 4%. Another notable observation from the results is that varying the values of k in the hard kNN approach has minimal impact on accuracy and yields results identical to that of the nearest centroid method, which we attribute to the highly compact nature of the embeddings learned under the `TRACE` framework. When a query is projected into the embedding space, it becomes closely associated with its nearest neighbors regardless of

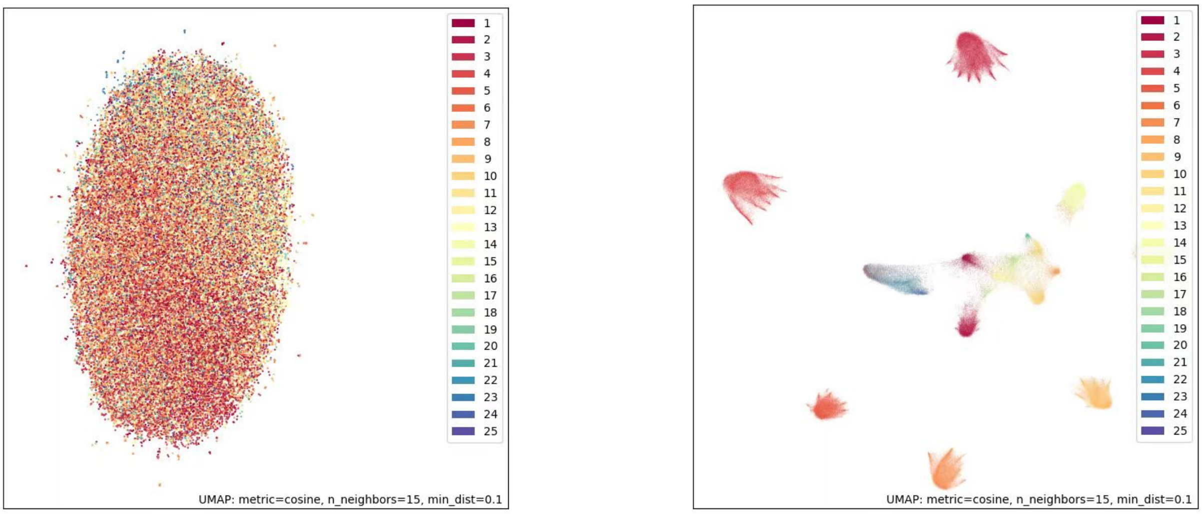

Figure 3: Visualization (using UMAP) of the embedding space before (left) and after (right) contrastive learning.

the specific value of k. This compactness suggests that the centroid of each cluster serves as an excellent representative of the entire cluster. Consequently, relying solely on these centroids can significantly reduce inference time. Even with 100 data providers as demonstrated in next subsection, the inference process remains almost instantaneous.

4.4 Scalability

Contemporary LLMs often necessitate substantial quantities of training data and the capability to manage a multitude of data providers. Hence, it is imperative to demonstrate the scalability of the TRACE framework under such settings. We assess the scalability of TRACE by selecting 10, 25, 50, and 100 distinct books from the booksum dataset, while maintaining a consistent experimental configuration. The results in Table 2 indicate a diminishing trend in accuracy with an increasing number of data providers, which is expected as the task complexity grows. However, despite this challenge, TRACE exhibits a relatively high level of accuracy across all settings, thus affirming its scalability.

4.5 Interpretability

The TRACE framework not only delivers accurate source attribution but also provides interpretability by offering additional insights into the attribution process. This interpretability is crucial for understanding the reasoning behind the model's decisions and gaining confidence in its outputs. We illustrate the interpretability of TRACE using responses from the `t5-small-booksum` model as a demonstration.

Table 3 shows a summary of correctly attributed single-source responses from the `t5-small-booksum` model. Each response is paired with the nearest principal sentence from the identified source. This pairing allows users to understand the specific evidence or context from the source text that influences the model's attribution decision.

Moreover, TRACE offers interpretability through the inclusion of different similarity scores. These scores provide insights into the model's confidence levels regarding the attribution outcomes. By examining the similarity scores, users can gauge the strength of the connection between the response and the identified source.

Overall, TRACE enhances interpretability by not only delivering the final attribution outcomes but also by providing supporting evidence from the source text and indicating the model's confidence levels through similarity scores. This transparency and insight into the attribution process empower users to trust and understand the model's outputs, which makes TRACE a valuable tool for source attribution tasks.

4.6 Additional Experimental Results

We conduct additional experiments to assess the performance of TRACE on alternative datasets, thereby evaluating its versatility. Table 4 summarizes the results. For a consistent comparison, we employ the same LLM across these datasets.

Our additional experiments affirm the adaptability of the TRACE framework across various datasets, thereby validating its applicability across various knowledge domains and settings.

n_books	t5 acc.	t5 top-3 acc.	t5 top-5 acc.	GPT2 acc.	GPT2 top-3 acc.	GPT2 top-5 acc.	Llama-2 acc.	Llama-2 top-3 acc.	Llama-2 top-5 acc.
10	87.5%	98.3%	99.4%	85.3%	96.8%	98.7%	88.2%	99.2%	99.5%
25	84.4%	95.3%	97.3%	81.3%	92.3%	94.0%	86.2%	96.1%	97.2%
50	73.1%	82.0%	84.0%	72.9%	82.9%	84.1%	70.3%	79.8%	82.2%
100	45.4%	74.8%	78.8%	49.0%	73.2%	77.7%	46.7%	76.8%	80.2%

Table 2: Source attribution accuracy for different no. of data providers on booksum dataset using TRACE.

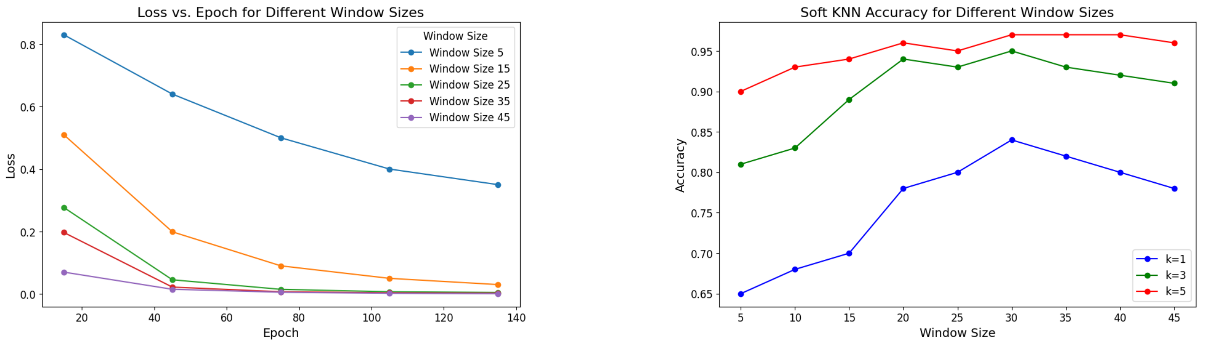

Figure 4: Contrastive loss (left) and soft kNN accuracy (right) with different WINDOW_SIZEs. Note that the results for hard kNN (regardless of the value of k) are identical to that of soft kNN when $k = 1$.

4.7 Ablation Study

The most important factor in TRACE is the user-defined WINDOW_SIZE. If the WINDOW_SIZE is too small, the principal sentences cannot capture sufficient contextual information, hence deteriorating the performance. However, an exceedingly large WINDOW_SIZE will not only require more computational resources and time to train but also the meaning will be diluted by other redundant information. This presents a natural trade-off between source attribution performance and computational efficiency. Therefore, in this subsection, we will analyze this trade-off and present the results in Fig. 4.

It can be observed that a larger WINDOW_SIZE facilitates faster model convergence. However, model loss alone is not a comprehensive indicator of the clustering quality. So, we evaluate the source attribution accuracy on the test dataset. When the WINDOW_SIZE is set to 30, our TRACE framework achieves its highest accuracy. We hypothesize that this is primarily because the WINDOW_SIZE of 30 is sufficient to capture essential contextual information without excessively diluting it.

5 Related Work

Source Attribution. Though source attribution remains relatively underexplored in the domain of natural language processing, WASA (Wang et al., 2023b) stands out as a notable framework.[2] Operating on the principle of watermarking, WASA embeds distinct source identifiers within the training data to ensure that responses convey pertinent data provider information. However, WASA necessitates extensive manipulation of training data and training the entire LLM from scratch, which is a time-consuming process given their sizes. In contrast, TRACE distinguishes itself by being model-agnostic, i.e., requiring no knowledge about the model. This characteristic enhances efficiency and adaptability.

In the context of identifying information sources for quotes, Quobert (Vaucher et al., 2021) is a minimally supervised framework designed for extracting and attributing quotations from extensive news corpora. Additionally, Spangher et al. (2023) have developed robust models for identifying and attributing information in news articles. However, these approaches are primarily focused on specific domains such as news. In contrast, TRACE is designed to handle knowledge across a wide range of domains and hence provides a more generalized and versatile solution for source attribution tasks.

Information Retrieval. A related topic to our work here is information retrieval. Traditional retrieval techniques like BM25 (Robertson et al., 1994) hinge heavily on frequency-based rules which prove to be inadequate when dealing with responses that share semantic similarities without significant lexical overlap. More contemporary

[2] Note that neither the source code nor comprehensive details of the experimental setup have been provided in (Wang et al., 2023b), making a fair comparison with WASA infeasible.

Response	**Nearest Principal Sentence**
Morel is in Sheffield, and he feels guilty towards Dawes, who is suffering and despairing, too. And besides, they had met in Nottingham in a way that is more or less responsible.	on his knees, feeling so awkward in presence of big trouble. Mrs. Morel did not change much. She stayed in Sheffield
But Emma thought at least it would turn out so. Mrs. Elton was first seen at church: but although devotion might be interrupted, curiosity could not be satisfied by a bride in. Pew, and it must be left for the visits in form which were then paid, to settle whether she was very pretty indeed, or only rather pretty at all.	or any thing just to keep my boot on." Mr. Elton looked all happiness at this proposition; and nothing could exceed
to marry Lord Warburton. Isabel enquired. "Your uncle's not an English nobleman," said Mrs. Touchett in her smallest, sparest voice. The girl asked if the correspondent of the Interviewer was to take the party to London under Ralph's escort. It was just the sort of plan, she said, that Miss Stackpole would be sure to suggest, and Isabel said that she did right to refuse him then.	he told Ralph he's engaged to be married." "Ah, to be married!" Isabel mildly exclaimed. "Unless he breaks it off. He seemed

Table 3: Sample responses with correct single-source attribution from `t5-small-booksum` model.

Dataset	Data Providers	Soft kNN $k=1$	Soft kNN $k=3$	Soft kNN $k=5$	Hard kNN $k=10$	Hard kNN $k=20$	Nearest Centroid
`booksum`	10	85.3%	96.8%	98.7%	85.3%		85.3%
`dbpedia_14`	10	88.2%	94.1%	97.2%	88.2%		88.2%
`booksum`	25	81.3%	92.3%	94.0%	81.3%		81.3%
`cc_news`	25	83.1%	90.8%	92.1%	83.1%		83.1%

Table 4: Source attribution accuracy on `dbpedia_14` and `cc_news` datasets using `TRACE`.

methods, such as ANCE (Xiong et al., 2020) and Contriever (Izacard et al., 2022), opt for generating compact, dense representations of documents rather than long, sparse ones. Thus, they tend to achieve better results.

While information retrieval and `TRACE` both use dense representations to measure text similarity, they differ in objectives and applications. Information retrieval aims to rank relevant documents for a user's query. In contrast, `TRACE` focuses on identifying and attributing the original source of specific information, hence ensuring accurate credit and authenticity.

Membership Inference Attack. The concept of membership inference attack was first introduced by Shokri et al. (2017). The primary objective of this attack is to ascertain whether a specific piece of information was part of the training data for a given machine learning model. Various assumptions about the available information lead to different attack models. For instance, some models assume access to hard labels (Li and Zhang, 2021), the model's confidence scores (Watson et al., 2022; Mattern et al., 2023), or the internal parameters of the model (Leino and Fredrikson, 2020). Wei et al. (2024) have achieved membership inference by inserting watermarks into data. Despite the variations, these attacks fundamentally seek to answer a binary question, i.e., whether the information was included in the training dataset or not.

In contrast, source attribution entails mapping the response to distinct and specific sources rather than simply determining the presence or absence of the data in the training set. Additionally, `TRACE` adheres to a black-box setting: It does not require access to internal information such as confidence scores or model parameters. Instead, `TRACE` only necessitates the output from a LLM.

6 Conclusion

This paper describes a novel `TRACE` framework which effectively achieves source attribution. By selecting principal sentences and projecting them into the embedding space via source-coherent contrastive learning, `TRACE` enhances the interpretability of responses generated by LLMs. This enhancement also conforms to regulations that aim to protect the privacy of users. After evaluating `TRACE` on various datasets, we have demonstrated the accuracy and effectiveness of our framework.

Limitations. Our experiments are subject to some limitations that can be addressed in the future work to ensure a comprehensive interpretation of results. Firstly, the balanced distribution of data across different sources may impact the final inference of TRACE given its reliance on the kNN algorithm. This uniformity in data volume may not be representative of real-world settings, which potentially limits the generalizability of our findings. Secondly, information within each source is quite distinct with no overlapping data. Future works can verify the setting where data sources contain similar information. These limitations underscore the importance of future research in addressing such challenges to enhance the robustness of TRACE across varied data environments.

Ethical Considerations. Our TRACE framework introduces a method for achieving source attribution. Utilizing this framework, a malicious actor may potentially identify the sources of data providers and reveal sensitive information about them. Therefore, the application of TRACE within this context necessitates meticulous handling to mitigate privacy concerns.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. arXiv:1902.09229.

Robin Brochier, Adrien Guille, and Julien Velcin. 2019. Global vectors for node representations. In *Proc. WWW*, pages 2587–2593.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. arXiv:1803.11175.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. arXiv:2002.05709.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised learning of universal sentence representations from natural language inference data. arXiv:1705.02364.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. 2021. Parametric contrastive learning. In *Proc. ICCV*, pages 715–724.

Raia Hadsell, Sumit Chopra, and Yann Lecun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proc. CVPR*, pages 1735–1742.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proc. ISI*, pages 218–223.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv:2311.05232.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. arXiv:2112.09118.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BookSum: A collection of datasets for long-form narrative summarization. In *Proc. EMNLP Findings*, pages 6536–6558.

Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proc. SEC*, pages 1605–1622.

Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. arXiv:2007.15528.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. arXiv:2308.05374.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2018. Do GANs leave artificial fingerprints? arXiv:1812.11842.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. arXiv:2305.18462.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. arXiv:2102.08473.

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

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. arXiv:2203.03929.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv:1908.10084.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proc. TREC*, pages 109–126.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, pages 815–823.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proc. IEEE S&P*, pages 3–18.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class N-pair loss objective. In *Proc. NIPS*.

Alexander Spangher, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2023. Identifying informational sources in news articles. In *Proc. EMNLP*, pages 3626–3639.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? arXiv:2005.10243.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. arXiv:2302.13971.

Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: A corpus of quotations from a decade of news. In *Proc. WSDM*, pages 328—336.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023a. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *Proc. NeurIPS*.

Jingtan Wang, Xinyang Lu, Zitong Zhao, Zhongxiang Dai, Chuan-Sheng Foo, See-Kiong Ng, and Bryan Kian Hsiang Low. 2023b. WASA: Watermark-based source attribution for large language model-generated data. arXiv:2310.00646.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. 2022. On the importance of difficulty calibration in membership inference attacks. arXiv:2111.08440.

Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. 2024. Proving membership in LLM pretraining data via data watermarks. arXiv:2402.10892.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive learning for sentence representation. arXiv:2012.15466.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv:2007.00808.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 2022. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. arXiv:2007.08457.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. NeurIPS*.

A Experimental Setup

Data Preparation. From `booksum`, we have randomly selected subsets of 10, 25, 50, and 100 books. From `dbpedia_14`, we chose 10 distinct classes. Additionally, we have extracted text samples from 25 diverse domains within the `cc_news` dataset.

Before proceeding with the analysis, we have performed standard preprocessing steps which include converting all text to lowercase and removing punctuation to ensure uniformity and cleanliness in the data.

Model. For sentence embedding, we have opted for `SBERT` (Reimers and Gurevych, 2019). Leveraging the pre-trained model `xlm-r-distilroberta-base-paraphrase-v1` that is readily accessible on Hugging Face, we have fine-tuned it within our `TRACE` framework. Moreover, we have augmented the model with additional feed-forward layers which serve as the projection network. The dimension for the embeddings is set as 64.

Training Details. The hyperparameters utilized in our experimental setup are configured as follows: the learning rate is 1×10^{-5}, the batch size is 64, the number of epochs is 150, and the temperature in the `NT-Xent Loss` is 0.1. Notably, all training procedures are conducted on a single NVIDIA L40 GPU, obviating model or data parallelism techniques. The results were obtained by averaging the outcomes of three executions, each with a different random seed.

Statistic	`booksum`	`dbpedia_14`	`cc_news`
Number of Documents	405 (books)	560,000	149,954,415
Languages Covered	English	English	English
Domains	Books	Encyclopedic	News

Table 5: Statistics of `booksum`, `dbpedia_14`, and `cc_news` datasets.