DOI: 10.19884/j.1672-5220.202407003

LTDDA: Large Language Model-Enhanced Text Truth Discovery with Dual Attention

FANG Xiu, CUI Zhihong, SUN Guohao*, LU Jinhu
School of Computer Science and Technology, Donghua University, Shanghai 201620, China

Abstract: Existing text truth discovery methods fail to address two challenges: the inherent long-distance dependencies and thematic diversity of long texts; the inherent subjective sentiment that obscures objective evaluation of source reliability. To address these challenges, a novel truth discovery method named large language model (LLM)-enhanced text truth discovery with dual attention (LTDDA) is proposed. First, LLMs generate embedded representations of text claims, and enhance the feature space to tackle long-distance dependencies and thematic diversity. Then, the complex relationship between source reliability and claim credibility is captured by integrating semantic and sentiment features. Finally, dual-layer attention is applied to extract key semantic information and assign consistent weights to similar sources, resulting in accurate truth outputs. Extensive experiments on three real-world datasets demonstrate that the effectiveness of LTDDA outperforms that of state-of-the-art methods, providing new insights for building more reliable and accurate text truth discovery systems.

Keywords: large language model (LLM); truth discovery; attention mechanism
CLC number: TP389.1 **Document code**: A
Article ID: 1672-5220(2025)06-0699-12

Open Science Identity (OSID)

0 Introduction

Nowadays, there are trillions of online resources on the Internet, which contain a large amount of valuable information, and a lot of errors and disinformation. Information on the same data item of interest can often be collected from different sources. These sources may have differences in the data quality, format, language and so on, resulting in a large number of data conflicts. How to automatically infer trustworthy information from these conflicting data is a challenging problem.

To solve this challenge, truth discovery methods have been proposed[1-2]. These methods identify the truth of each data while estimating source reliability. Though different models and scenarios are applied, they all follow the same general principle: a source is highly reliable if it frequently provides many trustworthy claims, and a claim is more likely to be true if many reliable sources support it. Noteworthy, most existing truth discovery methods are designed for structured data[3-4]. It is difficult to apply them directly to unstructured text data due to the complex and unique characteristics of natural language.

In recent years, some text truth discovery methods[5-8] have emerged. Some researches aim to address challenges in the field of truth discovery by examining the issues of semantic diversity and answer factors in text claims[5-6], while the convolutional neural network-long short-term memory (CNN-LSTM) employs pattern matching to extract triples from text data for truth discovery[8]. However, existing text truth discovery methods cannot well address the following two challenges.

First, the inherent long-distance dependencies and thematic diversity of long texts present challenges to truth discovery, resulting in information loss and difficulties in modeling relationships. As depicted in Fig. 1, considering the query "Does wearing less in cold weather make you gain weight?" in Claim 1, information extends beyond the local vocabulary, where the term "the issue" reaches far enough to require significant time or space to ascertain its referent. Additionally, Claim 1 is not confined to a single theme; it encompasses various related thematic, evident in the presence of diverse thematic keywords such as "Caloric Expenditure", "Behavioral Changes" and "Thermogenesis". Existing text truth discovery methods encounter challenges of information loss when attempting to learn dependencies from distant positions in the text. Furthermore, they struggle to effectively capture relationships between different themes in multi-theme texts, resulting in suboptimal modeling performance.

Second, the subjective sentiment of a source influences its expression of a particular viewpoint, potentially introducing bias toward the subjectivity, which affects our interpretation of the information, thus leading to incorrect reliability assessments. The subjective sentiment of a source may impact the accurate expression of claims. Emotional language may render claims

Received date: 2024-07-09
* Correspondence should be addressed to SUN Guohao, email: ghsun@dhu.edu.cn
Citation: FANG X, CUI Z H, SUN G H, et al. LTDDA: large language model-enhanced text truth discovery with dual attention[J]. *Journal of Donghua University (English Edition)*, 2025, 42(6): 699-710.

ambiguous, obscuring the objective viewpoint and making it challenging to extract accurately. As shown in Fig. 1, based on the content of Claim 2, it is evident that Source 2 initially employs sentiments of sarcasm and is surprised to emphasize disagreement with the viewpoint, followed by a gradual shift toward anger and expressing strong resentment. Current truth discovery methods may misinterpret messages due to their inability to accurately capture dynamic sentiment in texts, thereby reducing the reliability of Source 2.

Fig. 1 An example response to "Does wearing less in cold weather make you gain weight?"

To address the above challenges, this paper proposes the large language model (LLM)-enhanced text truth discovery with dual attention (LTDDA). The method leverages recent advancements in LLMs, and could exhibit strong language comprehension and the ability to handle longer contexts through zero-shot learning, in-context learning, fine-tuning and instruction tuning. Specifically, LTDDA leverages the transformer decoder architecture, commonly used in LLMs, to better capture and generate coherent and contextually relevant texts. During the pre-training stage, the model is exposed to a vast amount of text data, enabling it to learn richer language features and diverse topics, which enhances its ability to understand long-distance dependencies. Additionally, LTDDA incorporates factors such as subjective sentiment that could influence the reliability of sources. These factors, along with the embedded representations of text claims, are fed into a dual attention layer. This dual attention mechanism allows LTDDA to effectively evaluate the credibility of claims.

Experiments are done on three real-world datasets to compare LTDDA and the state-of-the-art truth discovery methods. Two ablation experiments are conducted to verify the hypothesis on the impact of source subjective sentiment on source reliability modeling, as well as to verify the effectiveness of LTDDA in alleviating long-distance dependencies and thematic diversity issues. The major contributions of this paper are as follows.

1) We propose a novel text truth discovery model that utilizes robust contextual capabilities of LLMs and a dual attention mechanism. This model effectively addresses challenges posed by the inherent long-distance dependencies and the thematic diversity of long texts.

2) We leverage the influence of the subjective sentiment on source reliability. By integrating semantic and sentiment features, we capture the intricate relationship between source reliability and claim credibility.

3) Extensive experiments on three real-world datasets show that LTDDA outperforms the state-of-the-art truth discovery methods.

1 Related Work

1.1 Truth discovery

Veracity is one of the features of big data, which is of great significance in the field of the big data analysis. In the past decade, truth discovery has become a hot topic in the database community[1-2, 9]. To resolve conflicts in multi-source data and reveal the underlying truth, extensive research efforts have been conducted to develop multiple truth discovery methods across different application scenarios[3-4, 10-11].

Truth discovery methods can be broadly categorized into the following four types: iterative methods[12] that involve repetitive truth calculation and source reliability

assessment until convergence; optimization-based methods[1-3, 13] that define a distance function to measure the difference between the claims provided by sources and the ground truth; probabilistic graphical model-based methods[14] in which the observed values are generated based on two parameters, namely truth and source reliability; neural network-based methods[15-16] that utilize neural networks to model the complex dependencies between sources and claims.

Among these, only a few methods are relevant to unstructured text data and can be classified into two groups. The first group extracts structured triples from the raw text and then performs truth discovery on structured data. Dong et al.[17] proposed a confidence-aware source reliability estimation approach which performed truth discovery based on the subject-verb-object (SVO) triples extracted from webpages in the process of knowledge base construction. Ye et al.[8] utilized a hybrid model of CNN and LSTM for extracting triple information from text data by using the pattern-based fact extraction method. This method not only emphasizes the credibility of the triples but also takes into account the reliability of the patterns. However, these methods overlook the natural language features of text data, resulting in suboptimal truth discovery results. Additionally, the use of extractors and patterns inevitably introduces additional noise.

The second group conducts truth discovery directly on raw texts. Zhang et al.[5] incorporated semantic meanings into the truth discovery procedure and proposed a method to identify trustworthy medical diagnoses from crowdsourcing users. Li et al.[18] combined the keywords extracted from the answers to specific questions into multiple interpretable factors and used the method based on the probabilistic graphical model to perform truth discovery to find trustworthy answers. Unfortunately, these methods can only handle short texts and cannot capture the long-distance dependencies and multi-theme features of long text sequences, so they are not suitable for long texts.

1.2 LLMs

LLMs have made significant progress in the field of natural language processing (NLP) in recent years. These models, such as GPT-3[19] and LLaMA[20], demonstrate powerful language generation and understanding capabilities through large-scale pre-training and fine-tuning. The current mainstream LLM uses the task of predicting the next token during pre-training, which helps to make better use of contextual information.

Currently, LLMs are mainly built upon the transformer architecture where multi-head attention layers are stacked in a very deep neural network. Existing LLMs adopt similar transformer architectures and pre-training objectives (e.g., language modeling) as small language models. The emergent ability of LLMs is one of the most significant features and distinguishes them from previous natural language models. In Ref. [21], emergent abilities are defined as abilities that are not present in small models but are present in large models. When the parameter scale of a language model reaches a certain level, its performance improves significantly. Furthermore, through instruction tuning, LLMs can follow task instructions[22] for new tasks without using explicit examples, thereby exhibiting better generalization capabilities. By fine-tuning on a multi-task dataset described in natural language (called instruction tuning), LLMs perform well on unseen tasks that are also described in the form of instructions.

1.3 Parameter-efficient fine-tuning

Fine-tuning pre-trained LLMs on downstream datasets can bring huge performance gains when they are compared to pre-trained LLMs out-of-the-box. However, as LLMs get larger and larger, full fine-tuning becomes very expensive in terms of computational cost and memory requirements. In addition, massive models might not be data efficient, and overfitting issues might be observed, yielding suboptimal generalization. To address these issues, parameter-efficient fine-tuning (PEFT) approaches have been proposed. PEFT approaches only fine-tune a small number of (extra) model parameters and freeze most parameters of the pre-trained LLMs, thereby greatly decreasing the computational and storage costs. PEFT approaches are better than fine-tuning in the low-data regimes and generalize better to out-of-domain scenarios. Existing PEFT approaches include LoRA[23], Prefix Tuning[24], Soft Prompt Tuning[25] and P-Tuning[26]. In this work, we use LoRA to improve the performance of LLMs in truth discovery.

As shown in Fig. 2, the basic concept of LoRA involves adding an auxiliary branch to the original pre-trained language model. Here, $\boldsymbol{x}$ denotes the input vector, and $\boldsymbol{h}$ represents the intermediate hidden state of the model. During fine-tuning, the weight matrix $\boldsymbol{W} \in \mathbf{R}^{d\times d}$ of the original pre-trained model remains unchanged, while the dimension expansion matrix $\boldsymbol{A} \in \mathbf{R}^{d\times r}$ and the dimension reduction matrix $\boldsymbol{B} \in \mathbf{R}^{r\times d}$ in the auxiliary branch are fine-tuned. The matrix product $\boldsymbol{BA}$ is then added to the original pre-trained model $\boldsymbol{W}$ to obtain a new pre-trained model. Since the magnitude order r is much smaller than d, the number of trainable parameters is greatly reduced.

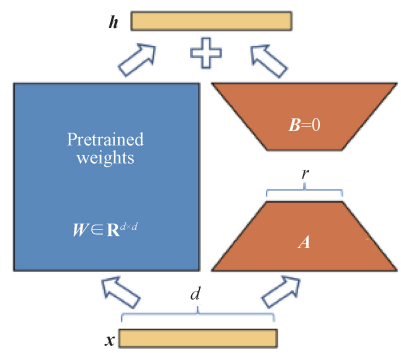


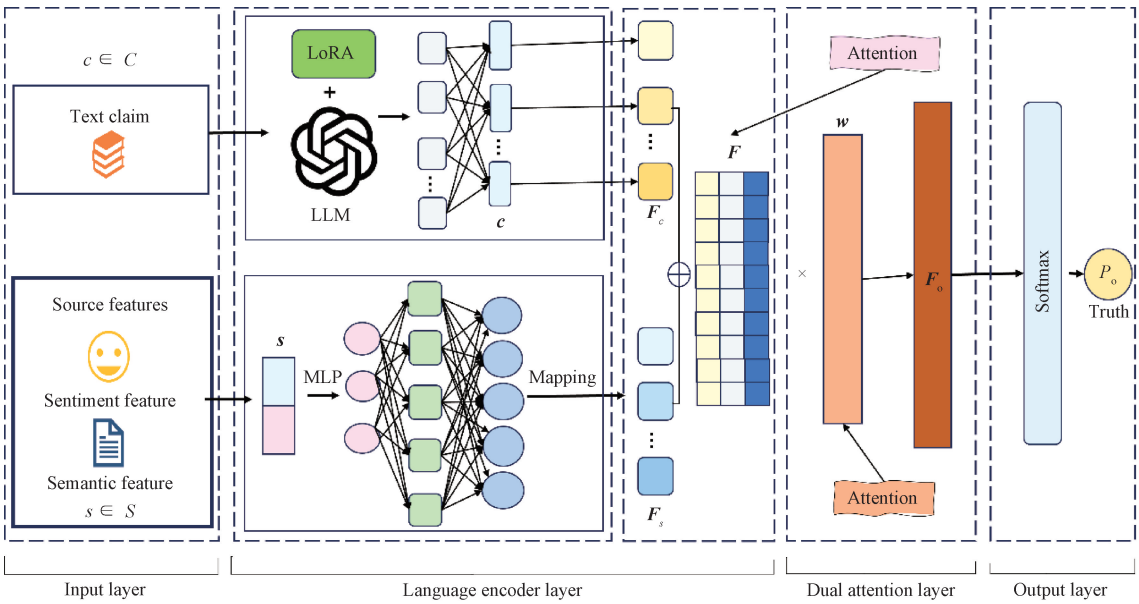
Fig. 2 Visualizing essence of LoRA

2 Methodology

2.1 Overview

Figure 3 illustrates the LTDDA framework, which consists of four key layers: an input layer, a language encoder layer, a dual attention layer and an output layer.

For each question $q \in Q$, the following process is carried out for truth reasoning. In the input layer, we use the LLM with a linear layer on the top as a regression head to embed representations of text claims and employ methods such as term frequency-inverse document frequency (TF-IDF) to generate semantic and sentiment features for sources. In the language encoder layer, we further extract features and concatenate the source features with the embedded representations of text claims. Subsequently, in the dual attention layer, we focus on capturing the reliability of similar sources and key information from text claims.



MLP—multilayer perceptron; s— data source item; S— data source set; $\boldsymbol{s}$— source vector; c—claim item; C—claim set; $\boldsymbol{c}$— claim vector; $\boldsymbol{F}_c$— claim's embedding representation; $\boldsymbol{F}_s$— source's embedding representation; $\boldsymbol{F}$—ground truth discovery feature space; $\boldsymbol{F}_o$— overall vector that determines whether the text claim is true; $\boldsymbol{w}$—text claim reliability weight matrix; P_o— probability of claim being true.

Fig. 3 Architecture of LTDDA framework

2.2 Input layer

When processing text claims, to capture the relationship between questions and text claims, each claim $c \in C$ is first spliced with its corresponding question q to provide a more complete representation of the claim input. Then, to correctly allocate attention weights and not be affected by differences in the sequence length, we use dynamic padding and truncation to unify the length of the text claim, which is set to the maximum length that the LLM can handle. Finally, the claim input is segmented into words, a word embedding vector with a dimension of d_{model} is created for each word, and a multi-dimensional vector $\boldsymbol{c}$ of the claim is constructed.

When dealing with data sources, simple one-hot encoding cannot incorporate the reliability of the data source into the vector representation. To accurately assess the reliability of data sources, we consider both semantic features and sentiment features when encoding data sources into vector representations. For each data source $s \in S$, we use the TF-IDF method to extract semantic features. Specifically, we first calculate the TF-IDF value for each word across all claims and treat these values as the semantic features of the words. Then, we concatenate the semantic features of all words to obtain the semantic feature vector $\boldsymbol{s}_1$ for the data source. For sentiment features, we leverage an LLM and a sentiment analysis head. The data source text is input into the LLM, and the sentiment analysis head predicts the sentiment polarity of the text (e.g., positive, negative, or neutral). Finally, the predicted sentiment polarity is used as the sentiment feature vector $\boldsymbol{s}_e$ of the data source. We concatenate $\boldsymbol{s}_1$ and $\boldsymbol{s}_e$ to form $\boldsymbol{s}$ of the data source.

2.3 Language encoder layer

We first describe the generation process of the pre-trained models. Suppose that we have a pre-trained model f for the text truth discovery. The output can be represented as a sequence of tokens from the vocabulary V. Let V^* be the space of sequences of tokens. Suppose the logits of f on $v \in V$ are $\bar{f}(v \mid x)$. The likelihood of the

next token following x being v is defined as

$$f(v \mid x) = \frac{\exp(\bar{f}(v \mid x))}{\sum_{v' \in V} \exp(\bar{f}(v \mid x))}. \tag{1}$$

The likelihood of generating $\hat{\boldsymbol{y}} \in V^*$ given x is defined as

$$f(\hat{\boldsymbol{y}} \mid x) = \prod_{i=1}^{|\hat{\boldsymbol{y}}|} f(\hat{y}_i \mid x, \hat{y}_{1:i-1}), \tag{2}$$

where $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{|\hat{\boldsymbol{y}}|})$ and $|\hat{\boldsymbol{y}}|$ is the length of $\hat{\boldsymbol{y}}$. The likelihood can be very small when $|\hat{\boldsymbol{y}}|$ is very large. To address this issue, we define the normalized likelihood $f_{\text{norm}}(\hat{\boldsymbol{y}} \mid x)$ as

$$f_{\text{norm}}(\hat{\boldsymbol{y}} \mid x) = f(\hat{\boldsymbol{y}} \mid x)^{\frac{1}{|\hat{\boldsymbol{y}}|}}. \tag{3}$$

We use f to generate the output sequence $\hat{\boldsymbol{y}}^*$ for the given input x by solving the following objective:

$$\hat{\boldsymbol{y}}^* = \arg\max \log f(\hat{\boldsymbol{y}} \mid x). \tag{4}$$

We first input $\boldsymbol{c}$ into a pre-trained model, using the LLM architecture to generate $\boldsymbol{F}_c$. Specifically, the pre-trained model breaks the text claim into a sequence of tokens and generates an embedding vector for each token. These vectors are concatenated into a long vector, representing the features of the entire text claim. After applying regularization and normalization, $\boldsymbol{s}$ is fed into a multi-layer perceptron to extract $\boldsymbol{F}_s$:

$$\boldsymbol{F}_s = \text{MLP}(\boldsymbol{w}_s \cdot \boldsymbol{s} + \boldsymbol{b}), \tag{5}$$

where $\text{MLP}(\cdot)$ represents the multi-layer perceptron; $\boldsymbol{w}_s$ represents the weight matrix in the trained model; $\boldsymbol{b}$ represents the bias term. Assuming that the embedding vector dimension of each word is d_{model}, the dimension of $\boldsymbol{F}_c$ of the text statement is $d_{\text{model}} \times l$, where l is the length of the text claim. We use the TF-IDF method and sentiment analysis to extract the semantic and sentiment features of the data source. Assuming that the embedding vector dimension of each word is d_{model}, the dimension of $\boldsymbol{F}_s$ is $d_{\text{model}} \times 2$.

$\text{MLP}(\cdot)$ can not only capture the implicit relationships between features but also map the data source features into the same dimensions as the declared features. Next, we concatenate and fuse $\boldsymbol{F}_c$ and $\boldsymbol{F}_s$ to obtain $\boldsymbol{F}$.

2.4 Dual attention layer

Our model incorporates a sophisticated dual-layer attention mechanism, designed to tackle the intricacies of truth discovery tasks effectively.

Our focus is on fostering robust embeddings among sources that exhibit similarity. By allocating similar weights to these embeddings, we aim to improve reliability and proximity within clusters of similar sources. This approach serves as a foundational step toward enhancing the model's ability to discern truth amidst a myriad of conflicting information.

Moving beyond the first attention layer, we introduce the second attention layer to delve deeper into the nuances of semantic information embedded within each text claim. This layer is pivotal in capturing the varying contributions of different words to the truth discovery task. Through the meticulous attention to the semantic detail, we strive to extract the essence of each claim, thus enriching the model's understanding of complex textual contexts.

2.4.1 First attention layer

In the first attention layer, when different data sources $\boldsymbol{s}$ have similar semantic features and sentiment features and the text claims $\boldsymbol{c}$ are also similar, their reliability matrices $\boldsymbol{w}$ also should be similar. This assumption helps to strengthen the model's focus on similar information, thereby improving the representation of relevant embeddings in the feature space.

Assuming n is the number of data sources, taking $\boldsymbol{s}_p$, $\boldsymbol{s}_q$, $\boldsymbol{c}_p$, and $\boldsymbol{c}_q$ ($p, q = 1, 2, \cdots, n$) as examples, in $\boldsymbol{F}$, we utilize the first attention layer to aggregate source embeddings with similarity and assign similar weights to these embeddings:

$$\text{Score}^s_{pq} = \text{sim}(\boldsymbol{s}_p, \boldsymbol{s}_q) = \frac{\boldsymbol{s}_p \cdot \boldsymbol{s}_q}{\|\boldsymbol{s}_p\| \cdot \|\boldsymbol{s}_q\|}, \tag{6}$$

$$\text{Score}^c_{pq} = \text{sim}(\boldsymbol{c}_p, \boldsymbol{c}_q) = \frac{\boldsymbol{c}_p \cdot \boldsymbol{c}_q}{\|\boldsymbol{c}_p\| \cdot \|\boldsymbol{c}_q\|}, \tag{7}$$

$$\text{Score}_{pn} = \{(\text{Score}^s_{p1} + \text{Score}^c_{p1}), (\text{Score}^s_{p2} + \text{Score}^c_{p2}), \cdots, (\text{Score}^s_{pn} + \text{Score}^c_{pn})\}, \tag{8}$$

$$\boldsymbol{w} = \text{Softmax}(\text{Score}_{pn}), \tag{9}$$

where $\text{sim}(\cdot)$ is the cosine similarity function.

2.4.2 Second attention layer

In the second attention layer, we are committed to deeply mining the key semantic information in the text to better understand the context and internal logical relationships of the text.

By accurately capturing important semantic features in each textual statement, we can better grasp the gist and key points of the text, thereby improving the efficiency and accuracy of the model in processing long texts. This nuanced focus on semantic information not only helps reduce the information loss but also effectively addresses the challenges posed by long texts, including long-distance dependencies and topic diversity complexity.

Therefore, the second attention layer plays a crucial role in enhancing the model's understanding of long texts:

$$\boldsymbol{u} = \tanh(\boldsymbol{W} \cdot \boldsymbol{F} + \boldsymbol{b}), \tag{10}$$

where $\boldsymbol{W}$ is the $\boldsymbol{F}$-weight matrix; $\boldsymbol{u}$ is the intermediate hidden layer vector of $\boldsymbol{F}$. We specifically use the final

hidden layer because it captures the most abstract and high-level features of the input, which are crucial for making accurate predictions. $\boldsymbol{u}_w$ is a randomly initialized vector that is optimized as a model parameter during the training process. This ensures that the model learns to align the final hidden layer representation with the target task.

We can calculate the probability distribution of $\boldsymbol{u}$ and $\boldsymbol{u}_w$ through the Softmax function to determine the weight α of the output $\boldsymbol{F}_o$:

$$\alpha = \mathrm{Softmax}(\boldsymbol{u}^T \boldsymbol{u}_w), \tag{11}$$

$$\boldsymbol{F}_o = \alpha \cdot \boldsymbol{F}. \tag{12}$$

Pass $\boldsymbol{F}_o$ to the Softmax function to obtain the probability of the claim being true, resulting in the final truth discovery result:

$$P_o = \mathrm{Softmax}(\boldsymbol{F}_o). \tag{13}$$

3 Experiments

In this section, we conduct experiments to evaluate the LTDDA performance for truth discovery. Our experimental evaluation is focused on the following questions.

Question 1: how does our proposed LTDDA method perform compared to existing truth discovery methods?

Answer 1: compared to existing truth discovery methods, LTDDA has superior performance by combining the excellent contextual capabilities of LLMs and source text features.

Question 2: what LLMs can be used with LTDDA for text truth discovery?

Answer 2: in our experiments, we focus on small models such as Qwen2-0.5B to demonstrate that efficient parameter fine-tuning can achieve better accuracy than larger models, and the experimental results verify our hypothesis.

Question 3: can the truth discovery method using a dual attention mechanism enhanced by LLMs capture long-distance dependencies and thematic diversity?

Answer 3: by leveraging the powerful contextual capabilities of LLMs, we can significantly improve the ability of LTDDA to handle long-distance dependencies and thematic diversity.

3.1 Experimental setup

3.1.1 Datasets

We focus on the free-from question answering task on the datasets SQuAD[27], ARC-Easy[28] and CommonsenseQA[29].

1) SQuAD. SQuAD is a reading comprehension dataset and consists of questions posed by crowd-workers on a set of Wikipedia articles, where the answer to every question is a segment of a text or span, from the corresponding reading passage. We use the SQuAD 1.1 version that contains more than 100 000 question-answer pairs on more than 500 articles. The training set contains 86 821 question queries, while the test set contains 5 928 question queries. We use the following template to construct prompt queries:

[*Please help me answer the Question based on the Context content.* \n *Context*: {*The provided Context paragraph*} \n *Question*: {*Provided Question*}].

In addition, in order to make the dataset suitable for the truth discovery scenario, we also sample answers for each piece of data and ultimately generate more than 60 000 sampled data.

2) ARC-Easy. ARC-Easy is a new dataset of 7 787 genuine grade-school levels, contains multiple-choice science questions, and is assembled to encourage research in advanced question-answering. The dataset is partitioned into a challenge set and an easy set, where the former contains only questions answered incorrectly by both a retrieval-based algorithm and a word co-occurrence algorithm. We use the following template to construct prompt queries:

[*Please answer the Question by choosing the most appropriate answer from the four Options below.* \n *Question*: {*Provided Question*} \n *Choices*: {*The provided Choices*}].

3) CommonsenseQA. CommonsenseQA is a new multiple-choice question answering dataset that requires different types of commonsense knowledge to predict the correct answers. It contains 12 102 questions with one correct answer and four distractor answers. We use the following template to construct prompt queries:

[*Please answer the Question by choosing the most appropriate answer from the six Options below.* \n *Question*: {*Provided Question*} \n *Choices*: {*The provided Choices*}].

3.1.2 LLMs

We use Mistral[30], LLaMA3 and Qwen2 models of various sizes. The specific details are shown in Table 1.

1) Mistral. Mistral is known for its efficiency and versatility across both general-purpose and specialized tasks. Mistral 7B, one of its standout models, offers strong performance despite its relatively small size. It excels in multilingual tasks, supporting numerous languages such as English, French, German and Chinese.

2) LLaMA3. Meta's LLaMA3 is the successor to LLaMA2 and offers significant improvements in multilingual support and natural language understanding. It is available in sizes ranging from 8 billion to 70 billion parameters, making it versatile for different applications. LLaMA3-8B represents the model with 8 billion parameters.

3) Qwen2. Developed by Alibaba, Qwen2 focuses heavily on multilingual capabilities and large-scale applications. Qwen models perform competitively in natural language generation and understanding tasks.

Table 1 Specific details of LLMs

Model	Number of parameters	d_{model}	Context length	Vocabulary size
Mistral-v0.2	7×10^9	4 096	3.2×10^4	32 000
Mistral-v0.3	7×10^9	4 096	4×10^3	32 768
LLaMA3-8B	8×10^9	4 096	8×10^3	128 256
Qwen2-0.5B	5×10^8	896	3.2×10^4	151 936
Qwen2-7B	7×10^9	3 584	1.28×10^5	152 064

3.1.3 Baseline methods

We compare LTDDA with state-of-the-art truth discovery methods. For those methods designed for structured data, we revise them to be suitable for text data.

1) MV. This approach chooses a claim as the identified truth when it has the highest support among all sources.

2) CRH + Topic Dist. CRH adopts an optimization-based truth discovery framework to handle heterogeneous data, which aims to minimize the weighted loss of the aggregation results[13]. We employ the latent Dirichlet distribution to extract a 100-dimensional topic representation for each question and its corresponding answer. Answers are ordered based on the cosine similarity with the question. In experiments, we use the topic distribution as a representation for each claim, which is then input into CRH.

3) CRH + Word2Vec. This baseline approach shares similarities with CRH + Topic Dist, with the only difference being a modification in the inputs to the average word vectors of answers.

4) CATD + Topic Dist. CATD is an optimization-based truth discovery framework that addresses the long-tail phenomenon in data by employing a confidence interval-based approach for assessing source reliability[31]. Similar to CRH + Topic Dist, we use the topic distributions as the representations of the whole answers to be fed to CATD.

5) CATD + Word2Vec. This baseline approach is similar to CATD + Topic Dist, with the difference being the change in inputs to the average word vectors of answers. The word vector representation is the same as that used in CRH + Word2Vec.

6) CNN-LSTM[8]. The pattern-based extraction method using CNN-LSTM employs a hybrid model of CNN and LSTM to capture complex patterns and dependencies between facts that are challenging for traditional learning approaches to uncover.

7) UTD[32]. UTD formulates the truth discovery task as a joint maximum likelihood estimation problem of unknown true claims and source reliability. It follows and applies the profile likelihood procedure to derive the joint maximum likelihood estimators.

3.1.4 Evaluation metrics

For the SQuAD dataset, we use expectation-maximization (EM) and $F1$ scores as evaluation metrics; for the other two datasets (ARC-Easy and CommonsenseQA), we use accuracy as the evaluation metric. The larger the value of all these evaluation metrics, the better the task performance.

3.1.5 Training hyperparameters

We use the LoRA[23] method to fine-tune the LLMs, setting the parameters as follows: lora_target is all, lora_rank is 64, lora_alpha is 128, and lora_dropout is 0.05. We train the LoRA for three epochs by using the AdamW optimizer with a batch size of 32, a learning rate of 0.001, a weight decay of 0.01, and a cosine learning rate scheduling with 10% warmup steps. Gradient clipping is set to 1.0. The maximum sequence length is set to 1 024. Figure 4 shows the convergence of some LLMs on different datasets during fine-tuning. We run all experiments using Python 3.10 and the HuggingFace API on 80 GB NVIDIA A100 GPUs in the Debian GNU/Linux 10 system. We use the Mistral, Qwen2, LLaMA2 and LLaMA3 models via the HuggingFace transformers library which can be easily adapted for reproducibility. We modify the trainer class provided by the HuggingFace API for LoRA tuning. We use the generate(·) function of the HuggingFace API to generate answers. Unless specified, we use default parameters of the generate(·) function.

We analyze the complexity of the LTDDA method, focusing on the number of forward passes required by the LLM. Since the LLM produced the output sequence in an auto-regressive way, the number of forward passes is proportional to the length of the generated output sequence. Assume that the maximum length of the output sequence is denoted as $l_{\max}$. For each input sequence, one forward pass is needed to encode the input, and up to $l_{\max}$ forward passes are required to generate the output sequence. Consequently, the upper bound for the number of forward passes is $1 + l_{\max}$, giving the complexity as $O(l_{\max})$.

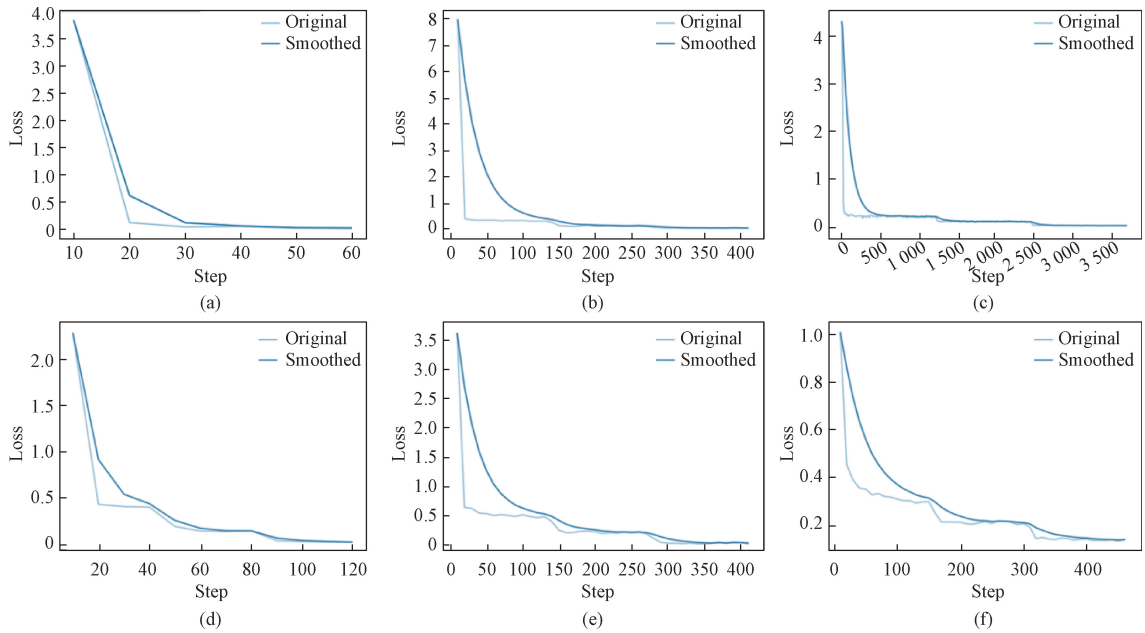


Fig. 4 Convergence of LLaMA3-8B and Qwen2-0.5B during fine-tuning: (a) LLaMA3-8B on ARC-Easy dataset; (b) LLaMA3-8B on CommonsenseQA dataset; (c) LLaMA3-8B on SQuAD dataset; (d) Qwen2-0.5B on ARC-Easy dataset; (e) Qwen2-0.5B on CommonsenseQA dataset; (f) Qwen2-0.5B on SQuAD dataset

3.2 Experimental results

We first evaluate the accuracy of different LLMs. The results in Table 2 show that after LoRA tuning, the accuracy of LLMs is improved significantly.

Table 2 Performance of different LLMs

Method	SQuAD		ARC-Easy accuracy	CommonsenseQA accuracy
	EM	*F*1		
Pre-trained Qwen2-0.5B	0.117 4	0.327 4	0.388 9	0.351 4
Adapted Qwen2-0.5B with LoRA	0.664 3	0.803 9	0.704 8	0.637 5
Pre-trained Qwen2-7B	0.495 3	0.699 7	0.907 8	0.774 0
Adapted Qwen2-7B with LoRA	0.737 0	0.873 0	0.945 3	0.846 0
Pre-trained Mistral-v0.2	0.118 9	0.385 4	0.810 6	0.662 6
Adapted Mistral-v0.2 with LoRA	0.739 8	0.8737	0.898 6	0.822 3
Pre-trained Mistral-v0.3-7B	0.273 2	0.535 3	0.849 3	0.701 9
Adapted Mistral-v0.3-7B with LoRA	0.738 5	0.873 3	0.894 8	0.805 1
Pre-trained LLaMA3-8B	0.666 5	0.830 3	0.824 5	0.674 0
Adapted LLaMA3-8B with LoRA	0.737 9	0.873 0	0.931 3	0.818 2

On the SQuAD dataset, the adapted Qwen2-0.5B can even outperform the pre-trained Qwen2-7B and Mistral-v0.2, which demonstrates that it is possible to adapt a smaller LLM to achieve higher accuracy than a much larger LLM. We then report the results of LTDDA as well as baselines on the real-world datasets. As shown in Table 3, LTDDA consistently outperforms almost all the baselines on the three real-world datasets. LTDDA surpasses state-of-the-art methods in text veracity detection.

The reasons for the superior performance of LTDDA compared to the retrieval-based approaches and state-of-the-art truth discovery methods are analyzed as follows.

Table 3 Performance of different truth discovery methods

Method	SQuAD		ARC-Easy	CommonsenseQA
	EM	*F*1	accuracy	accuracy
LTDDA-Qwen2-0. 5B	0. 732 2	0. 815 3	0. 902 4	0. 753 8
LTDDA-Qwen2-7B	0. 785 3	0. 905 4	0. 978 9	0. 867 6
LTDDA-LLaMA3-8B	0. 782 2	0. 898 7	0. 962 2	0. 844 2
LTDDA-Mistral-v0. 2	0. 756 2	0. 882 3	0. 928 6	0. 846 0
LTDDA-Mistral-v0. 3	0. 763 2	0. 901 2	0. 944 5	0. 858 2
MV	0. 3251	0. 348 1	0. 401 3	0. 341 1
CRH + Topic Dist	0. 431 2	0. 440 2	0. 587 2	0. 541 2
CRH + Word2Vec	0. 443 2	0. 449 8	0. 595 2	0. 552 3
CATD + Word2Vec	0. 523 1	0. 671 2	0. 614 3	0. 632 7
CATD + Topic Dist	0. 510 1	0. 668 0	0. 605 4	0. 625 6
CNN-LSTM	0. 678 2	0. 778 3	0. 895 6	0. 783 2
UTD	0. 665 4	0. 757 6	0. 842 1	0. 654 3

First, compared to the existing truth discovery method, although CRH aims to capture source reliability, its performance is not ideal. On the one hand, CRH assumes that a priori assumed functions can represent the relationship between the source reliability and the credibility of claims. This assumption leads to suboptimal results in text truth discovery, as the actual dependencies between sources and claims are often complex and a priori unknown. On the other hand, CRH does not take into account unique features of the natural language, including the impact of the source's subjective sentiment on its reliability, as well as challenges posed by long-distance dependencies and thematic diversity in long text sequences. Additionally, while CATD considers the difficulties introduced by the long-tail phenomenon in assessing the source reliability, it also fails to account for the unique features of natural language fully.

Second, compared to our method, the CNN-LSTM, which also utilizes neural networks, does not perform well. On the one hand, the pattern-based fact extraction method employed by CNN-LSTM does not directly apply to raw text data, introducing additional noise when the patterns used are not suitable for the respective dataset. On the other hand, the architecture of CNN and LSTM, where information is transmitted through a state vector, is limited in capacity and struggles to capture long-distance dependencies. This limitation may lead to issues such as gradient explosions, resulting in poorer performance on the disaster tweets dataset with larger average lengths. UTD treats the truth discovery problem as a joint maximum likelihood estimation problem, aiming to estimate the reliability and truth value of unknown sources. Numerical solutions are provided by iteratively calculating joint maximum likelihood estimates of the true value and source reliability. However, UTD does not directly process original text data, but processes preprocessed text classification data, which limits its application scenarios. Furthermore, modeling the problem as a joint maximum likelihood estimation problem may not resolve complex relationships in the text well.

3. 3 Ablation study

3. 3. 1 Alleviating long-distance dependencies and thematic diversity

To evaluate the effectiveness of LTDDA in capturing long-distance dependencies and thematic diversity, experiments are done by comparing two groups of popular pre-trained word embedding methods used in NLP: contextual embeddings including Transformer and BERT, and non-contextual embeddings including Word2Vec. In the experiment, we utilized BERT, RoBERTa and Word2Vec as word embeddings for the LTDDA model while keeping the rest of the structure unchanged.

Table 4 presents the performance comparison results of different word embedding methods across three datasets. As expected, contextual embeddings significantly outperform non-contextual ones in our model. This is largely due to models like BERT and RoBERTa that leverage the Transformer architecture. The key advantage of the Transformer lies in its attention mechanism, which excels at capturing long-distance dependencies and extracting richer semantic features. In contrast, non-contextual embeddings assign the same representation to all meanings of a word, limiting their capacity to handle semantic nuances. Furthermore, in LTDDA, the LLM outperforms BERT and RoBERTa overall. This is primarily because modern LLMs are trained on much larger datasets and have substantially more parameters, enhancing their generalization ability and making them particularly effective when processing longer texts, where the attention mechanism further proves advantageous in modeling long texts.

Table 4 Performance comparison for different word embedding methods

Word embedding method	Dimension	SQuAD		ARC-Easy	CommonsenseQA
		EM	*F*1	accuracy	accuracy
BERT	768	0.692 3	0.732 1	0.753 4	0.751 4
RoBERTa	768	0.743 3	0.795 6	0.863 2	0.756 1
LTDDA-Qwen2-0.5B	896	0.732 2	0.815 3	0.902 4	0.753 8
Word2Vec	100	0.533 1	0.698 4	0.608 9	0.651 4

3.3.2 Source feature ablation experiment

To evaluate the contributions of two types of source features (i.e., source subjective sentiment features and semantic features) to LTDDA, we conduct source feature ablation experiments by using the complete set of statistical features and two subsets of statistical features.

The results of the source feature ablation experiments are depicted in Fig. 5. It can be observed that the complete set of statistical features performs the best in the truth discovery process, indicating the crucial role of these two features in distinguishing between credible and non-credible claims. Furthermore, LTDDA Qwen2-0.5B performs better when using semantic features compared to using sentiment features, highlighting the essential role of semantic information in the complex task of truth discovery from text data. Lastly, the accuracy is higher when using sentiment features compared to not using sentiment features, confirming our hypothesis that source subjective sentiment significantly influences people's perception of source reliability.

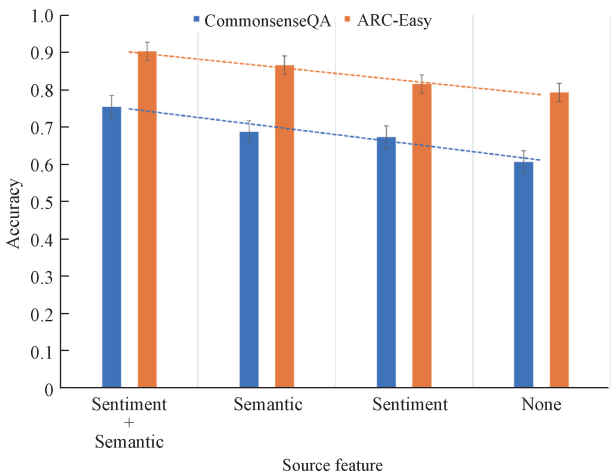


Fig. 5 Accuracy of source feature ablation experiments on CommonsenseQA and ARC-Easy by using LTDDA-Qwen2-0.5B

4 Conclusions

In this paper, we propose a text truth discovery model called LTDDA. Experimental results on three real datasets demonstrate the effectiveness of the proposed LTDDA model. In the future, we plan to enhance the performance of truth discovery by introducing additional features about the source and adopting more powerful pre-trained models to model the complex relationship between sources and claims more comprehensively.

References

[1] LI Y L, LI Q, GAO J, et al. On the discovery of evolving truth[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 675-684.

[2] MA F L, LI Y L, LI Q, et al. FaitCrowd: fine grained truth discovery for crowdsourced data aggregation[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 745-754.

[3] LI Q, LI Y L, GAO J, et al. A confidence-aware approach for truth discovery on long-tail data[J]. *Proceedings of the VLDB Endowment*, 2014, 8(4): 425-436.

[4] LI Q, LI Y L, GAO J, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2014: 1187-1198.

[5] ZHANG H T, LI Y L, MA F L, et al. TextTruth: an unsupervised approach to discover trustworthy information from multi-sourced text data[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2018: 2729-2737.

[6] CHANG C, CAO J J, ZHENG Q B, et al. An unsupervised approach of truth discovery from multi-sourced text data[J]. *IEEE Access*, 2019, 7: 143479-143489.

[7] LIU J C, TANG F L, HUANG J L. Truth inference with bipartite attention graph neural network from a comprehensive view[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2021: 1-6.

[8] YE C, WANG H Z, LU W B, et al. Deep truth discovery for pattern-based fact extraction[J]. *Information Sciences*, 2021, 580: 478-494.

[9] LI X, DONG X L, LYONS K B, et al. Truth finding on the deep web[J]. *Proceedings of the VLDB Endowment*, 2012, 6(2): 97-108.

[10] LI X, DONG X L, LYONS K B, et al. Scaling up copy detection [C]//2015 IEEE 31st International Conference on Data Engineering. New York: IEEE, 2015: 89-100.
[11] MENG C S, JIANG W J, LI Y L, et al. Truth discovery on crowd sensing of correlated entities [C]//Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems. New York: ACM, 2015: 169-182.
[12] PASTERNACK J, ROTH D. Knowing what to believe (when you already know something) [C]//International Conference on Computational Linguistics. Beijing: COLING, 2010, 2: 877-885.
[13] LI Y L, LI Q, GAO J, et al. Conflicts to harmony: a framework for resolving conflicts in heterogeneous data by truth discovery[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(8): 1986-1999.
[14] YANG Y, BAI Q, LIU Q. A probabilistic model for truth discovery with object correlations[J]. *Knowledge-Based Systems*, 2019, 165: 360-373.
[15] LI L Y, QIN B, REN W J, et al. Truth discovery with memory network[J]. *Tsinghua Science and Technology*, 2017, 22(6): 609-618.
[16] MARSHALL J, ARGUETA A, WANG D. A neural network approach for truth discovery in social sensing[C]//2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). New York: IEEE, 2017: 343-347.
[17] DONG X L, GABRILOVICH E, MURPHY K, et al. Knowledge-based trust[J]. *Proceedings of the VLDB Endowment*, 2015, 8(9): 938-949.
[18] LI Y L, DU N, LIU C C, et al. Reliable medical diagnosis from crowdsourcing: discover trustworthy answers from non-experts [C]// Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. New York: ACM, 2017: 253-261.
[19] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 1877-1901.
[20] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models [EB/OL]. (2023-02-17) [2024-07-09]. https://arxiv.org/abs/2302.13971.
[21] WEU J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models [EB/OL]. (2022-10-26) [2024-07-09]. https://arxiv.org/abs/2206.07682.
[22] SANH V, WEBSON A, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization [EB/OL]. (2022-03-17) [2024-07-09]. https://arxiv.org/abs/2110.08207.
[23] HU J E, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models [EB/OL]. (2021-10-16) [2024-07-09]. https://arxiv.org/abs/2106.09685.
[24] LIU X, JI K X, FU Y C, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks [EB/OL]. (2022-03-20) [2024-07-09]. https://arxiv.org/abs/2110.07602.
[25] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning [EB/OL]. (2021-09-02) [2024-07-09]. https://arxiv.org/abs/2104.08691.
[26] LIU X, ZHENG Y N, DU Z X, et al. GPT understands, too [EB/OL]. (2023-10-25) [2024-07-09]. https://arxiv.org/abs/2103.10385.
[27] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100 000+ questions for machine comprehension of text [EB/OL]. (2016-10-11) [2024-07-09]. https://arxiv.org/abs/1606.05250.
[28] CLARK P, COWHEY I, ETZIONI O, et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge [EB/OL]. (2018-03-14) [2024-07-09]. https://arxiv.org/abs/1803.05457.
[29] TALMOR A, HERZIG J, LOURIE N, et al. CommonsenseQA: a question answering challenge targeting commonsense knowledge [EB/OL]. (2019-03-15) [2024-07-09]. https://arxiv.org/abs/1811.00937.
[30] JIANG A Q, ABLAYROLLES A, MENSCH A, et al. Mistral 7[EB/OL]. (2023-10-10) [2024-07-09]. https://arxiv.org/abs/2310.06825.
[31] TOUVRON H, MARTIN L, STONE K R, et al. LLaMA 2: open foundation and fine-tuned chat model[EB/OL]. (2023-07-19) [2024-07-09]. https://arxiv.org/abs/2307.09288.
[32] XIAO H P, WANG S Y. A joint maximum likelihood estimation framework for truth discovery: a unified perspective [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 35(6): 5521-5533.

LTDDA：基于大语言模型增强的双重注意力文本真值发现

方　秀，崔志宏，孙国豪*，陆金虎
东华大学 计算机科学与技术学院，上海 201620

摘　要：现有文本真值发现方法无法解决两大挑战：其一，长文本中固有的长距离依赖性和主题多样性；其二，固有的主观情感使源可靠性的客观评估更加复杂。为此，提出了大语言模型（large language model，LLM）增强双重注意力真值发现（large language model-enhanced text truth discovery with dual attention，LTDDA）的新方法。首先，LTDDA 使用大语言模型生成文本声明的嵌入表示，增强特征空间以解决长距离依赖性和主题多样性问题。其次，通过融合语义和情感特征，捕捉源可靠性和声明可信度之间的复杂关系。最后，应用双重注意力层来提取关键语义信息，并为相似来源分配一致权重，从而实现准确的真值输出。在三个真实数据集上的广泛实验表明，LTDDA 在有效性方面优于现有的最佳方法，为构建更可靠、准确的文本真值发现系统提供了新思路。

关键词：大语言模型（LLM）；真值发现；注意力机制