AICCC www.aicccnet Kyoto, Japan (http://www.aiccc.net/)

# Notification of Acceptance of the AICCC 2023

**Kyoto, 16-18 December, 2023**

**http://www.aiccc.net/**

Paper ID: CA0047

Paper Title: CoT-STS: A Zero Shot Chain-of-Thought Prompting for Semantic Textual Similarity

Dear Musarrat Hussain, Ubaid Ur Rehman, Tri D.T. Nguyen and Sungyoung Lee,

First of all, thank you for your concern. 2023 6th Artificial Intelligence and Cloud Computing Conference (AICCC 2023) review procedure has been finished. Your paper was tripling blind-reviewed and, based on the evaluations. The comments are enclosed.

The selected papers could be published in the international conference proceeding with high quality. According to the recommendations from reviewers and technical program committees, we are glad to inform you that your paper identified above have been selected for publication and oral presentation. You are invited to present your paper and studies during our AICCC conference that would be held during December 16-18, 2023.

All peer reviewed and accepted papers can be published in the Conference Proceedings by ACM (ISBN: 979-8-4007-1622-5) and be indexed by Ei Compendex and Scopus, and submitted to be reviewed by Thomson Reuters Conference Proceedings Citation Index (ISI Web of Science).

## (Important Steps for your registration): Please do finish all the 4 steps on time to guarantee the paper published in the proceedings successfully:

**1. Revise your paper according to the Review Comments in the attachment carefully.**

**2. Format your paper according to the Formatting Instructions carefully**.
http://www.aiccc.net/template.docx (DOC Format)

**3. Register through the link directly:**

https://www.zmeeting.org/register/aiccc2023

**4. Add the CCS Concepts part and XML code in your paper.**
http://www.aiccc.net/CCS&XML.docx

**Note: To register through the link above, you are required to sign in the system first. Please do fill the registration information required, upload the final paper, payment proof in the system before November 27, 2023 to guarantee the registration successfully.**

AICCC 2023 will check the format of all the registered papers and send e-mail confirmation to the authors. After the registration, we will send all qualified papers to the publish house and index organization for publishing directly.

We are looking forward to meeting all the authors in our conference.

Please strictly adhere to the format specified in the conference template while preparing your final paper. If you have any problem, please feel free to contact us via **aiccc.contact@gmail.com.** For the most updated information on the conference, please check the conference website at **http://www.aiccc.net/.** The Conference Program will be available at the website in **the early of December, 2023**.

Yours sincerely,
www.aiccc.net
AICCC 2023

# AICCC&ADIP 2023
## KYOTO CONFERENCE ABSTRACT

Kyoto, Japan
December 16-18, 2023

2023 6TH ARTIFICIAL INTELLIGENCE AND CLOUD COMPUTING CONFERENCE **AICCC**

2023 5TH ASIA DIGITAL IMAGE PROCESSING CONFERENCE **ADIP**

# Registration and Zoom test (Online)

## Online Zoom Test-December 16, 2023

## Japan Local Time (GMT+9)

*Online test is for testing the Internet connection and helping participants get familiar with software Zoom, please make sure that you will attend online test.

**Zoom link:** https://us02web.zoom.us/j/83229506745

**Zoom Room ID:** 83229506745

| | | |
|---|---|---|
| 10:10-10:20 | Keynote speaker | Prof. Wenbing Zhao |
| 10:20-10:30 | Keynote speaker | Prof. Yan Li |
| 10:30-11:00 | Session 6 (Online) | CA0036, CA0030, CA5004-A, CA5007-A, CA0029, CA5001 |
| 11:00-11:30 | Session 7 (Online) | CA0053, CA0068, CA0080, CA0079 |
| 11:30-12:00 | Session 8 (Online) | CA0031, CA0052, CA0059, CA2003, CA0034, CA0070 |
| 16:00-16:05 | Invited Speaker | Prof. Narendra D. Londhe |
| 16:05-16:10 | Invited Speaker | Prof. Umesh C. Pati |
| 16:10-16:15 | Invited Speaker | Prof. Pascal Lorenz |
| 16:15-16:20 | Invited Speaker | Prof. Tossapon Boongoen |

# Schedule Simple Version

***Tips:* The time in the schedule is according to Japan Standard Time (GMT+9)**

*Regular Oral Presentation: about **15 Minutes** including Presentation and **2-3 Minutes** for Question and Answer.

## December 17, 2023- Keynote Speeches, Invited Speeches and Onsite Sessions

## Japan Local Time (GMT+9)

**Zoom link:** https://us02web.zoom.us/j/83229506745

**Zoom Room ID:** 83229506745

| Time | Each talk includes Q&A time | Presenter | Room |
|---|---|---|---|
| 9:00-9:05 | Opening Remark | Prof. Seiichi Ozawa, Kobe University, Japan | 9F<br>Shiho(紫峰)<br>Zoom ID:<br>83229506745 |
| 9:05-9:45 | Keynote Speech 1<br>Speech Title: Digital Economy Demands 5G/6G Networks, Generative AI, and Intelligent Twin City | Prof. Kai Hwang, The Chinese University of Hong Kong (Shenzhen), China (IEEE Life Fellow) | |
| 9:45-10:25 | Keynote Speech 2<br>Speech Title: Fine-Grained Activity Recognition in Professional Sports: A Preliminary Study | Prof. Wenbing Zhao, Cleveland State University, USA | |
| 10:25-10:40 | Group Photo & Coffee Break | | |
| 10:40-12:10 | Session 1<br>Topic: Neural Network Models and Data Computing<br>Session Chair: Prof. Dmitriy K. Levonevskiy, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia | CA0050, CA0056-A, CA0060, CA0076-A, CA0027, CA0055 | |
| 2:10-13:00 | Lunch Time | | 9F<br>Hakusui(白水) |
| :00-13:40 | Keynote Speech 3<br>Speech Title: Small-data Deep Learning for AI-Aided Diagnosis and AI Medical Imaging | Prof. Kenji Suzuki, Tokyo Institute of Technology, Japan | 9F<br>Hakusui(白水)<br>Zoom ID:<br>83229506745 |
| 40-14:20 | Keynote Speech 4<br>Speech Title: Sleep Stages Classification Using Deep Learning Methods | Prof. Yan Li, University of Southern Queensland, Australia | |

| | | | |
|---|---|---|---|
| 14:20-14:40 | **Invited Speech 1**<br>**Speech Title: Video-based Intrusion Detection System using Deep Learning Techniques** | Prof. Umesh C. Pati, National Institute of Technology (NIT), India | |
| 14:40-16:10 | **Parallel Session 2**<br>**Topic: Intelligent system monitoring and information management based on IoT**<br>**Session Chair: Prof. Herlina Abdul Rahim, University of Technology Malaysia, Malaysia** | CA0004, CA0021, CA0024, CA0032, CA0046, CA0047 | 9F Hakusui( |
| | **Parallel Session 3**<br>**Topic: Next generation artificial intelligence technology and applications**<br>**Session Chair: Prof. Ahmad Al-Qerem, Zarqa University, Jordan** | CA0005-A, CA0010, CA0048, CA0057, CA2002, CA0049 | 9F Benihana( |
| 16:10-16:25 | Coffee Break | | 9F Hakusui Benihana (白水+紅 |
| 16:25-18:10 | **Parallel Session 4**<br>**Topic: Machine Learning and Prediction Models**<br>**Session Chair: Assoc. Prof. Swee Chuan Tan, Singapore University of Social Sciences, Singapore** | CA0002, CA0033-A, CA5016, CA0054, CA5011, CA0016, CA0082-A | 9F Hakusui(白 |
| | **Parallel Session 5**<br>**Topic: Intelligent Image Analysis and Processing Methods**<br>**Session Chair: Prof. Kenji Suzuki, Tokyo Institute of Technology, Japan** | CA0065-A, CA5014, CA5002-A, CA5010, CA5003-A, CA5013, CA5005-A, CA0058(Poster) | 9F Benihana(紅 |
| 18:10-20:00 | Dinner Time | | 9F Hakaku(八 |

# December 18, 2023- Invited Speeches and Online Sessions

## Japan Local Time (GMT+9)

**Zoom link:** https://us02web.zoom.us/j/83229506745

**Zoom Room ID:** 83229506745

| Time | Each talk includes Q&A time | Presenter |
|---|---|---|
| 10:30-12:00 | **Session 6**<br>Topic: Image detection and speech recognition<br>Session chair: Assoc. Prof. Ts. Dr Hamimah binti Ujir, Universiti Malaysia Sarawak, Malaysia | CA0036, CA0030, CA5004-A, CA5007-A, CA0029, CA5001 |
| 12:00-13:30 | Lunch Time | |
| 13:30-13:50 | **Invited Speech 2**<br>Speech Title: Multiparametric Behavioral Machine Based Pain Estimation | Prof. Narendra D. Londhe, National Institute of Technology Raipur, India |
| 13:50-14:50 | **Session 7**<br>Topic: Information Privacy and Security in Data Networks<br>Session chair: Prof. Anand Nayyar, Duy Tan University, Viet Nam | CA0053, CA0068, CA0080, CA0079 |
| 14:50-15:00 | Break time | |
| 15:00-15:20 | **Invited Speech 3**<br>Speech Title: Advanced architectures of Next Generation Wireless Networks | Prof. Pascal Lorenz, University of Haute-Alsace, France |
| 15:20-15:40 | **Invited Speech 4**<br>Speech Title: Imperfect data, the curse for any data science project. | Prof. Tossapon Boongoen, Aberystwyth University, UK |
| 15:40-17:10 | **Session 8**<br>Topic: AI based data computing and recommendation system<br>Session Chair: | CA0031, CA0052, CA0059, CA2003, CA0034, CA0070 |

CA0047 (15:55-16:10 GMT+9)

CoT-STS: A Zero-Shot Chain-of-Thought Prompting for Semantic Textual Similarity

**Musarrat Hussain**, Ubaid Ur Rehman, Tri D.T. Nguyen and Sungyoung Lee
Kyung Hee University, South Korea

Abstract-The emergence of Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) by changing the focus of technical development from features engineering, architecture engineering, and objective engineering to prompt engineering. The main goal of the prompt engineering is to craft clear and concise instructions, known as input prompts, for LLMs to effectively perform the targeted NLP task. Semantic Textual Similarity (STS) is one such significant NLP task, which aims to assess the similarity between the semantic meanings of two input sentences. Numerous approaches have been proposed in the literature, including syntactic similarity evaluations, word-embedding based methods, and dedicated model training. However, these approaches require substantial effort, such as creating extensive annotated datasets and training dedicated STS models.

This research introduces CoT-STS, which aims to customize the use of the chain-of-thought prompting with LLMs for the STS task. We proposed four influential factors as part of the Chain-of-Thought approach, including theme similarity, participating object similarity, similarity of the activity being carried out, and the evaluation of other factors before arriving at the final similarity assessments. The application of the proposed CoT-STS on the BIOSSES dataset achieved a Pearson's correlation of 0.72, surpassing the 0.45 correlation achieved by the standard prompting and the correlation of 0.71 achieved by the existing zero-shot CoT methodology. The result achieved demonstrates the potential of LLMs with an appropriate prompting strategy to significantly improve the performance of the STS task.

# CoT-STS: A Zero-Shot Chain-of-Thought Prompting for Semantic Textual Similarity

MUSARRAT HUSSAIN

Kyung Hee University (Global Campus), Yongin-si, Republic of Korea.

musarrat.hussain@oslab.khu.ac.kr

UBAID UR REHMAN

Kyung Hee University (Global Campus), Yongin-si, Republic of Korea.

ubaid.rehman@khu.ac.kr

TRI D.T. NGUYEN

Kyung Hee University (Global Campus), Yongin-si, Republic of Korea.

tringuyendt@khu.ac.kr

SUNGYOUNG LEE

Kyung Hee University (Global Campus), Yongin-si, Republic of Korea.

sylee@oslab.khu.ac.kr

The emergence of Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) by changing the focus of technical development from features engineering, architecture engineering, and objective engineering to prompt engineering. The main goal of the prompt engineering is to craft clear and concise instructions, known as input prompts, for LLMs to effectively perform the targeted NLP task. Semantic Textual Similarity (STS) is one such significant NLP task, which aims to assess the similarity between the semantic meanings of two input sentences. Numerous approaches have been proposed in the literature, including syntactic similarity evaluations, word-embedding based methods, and dedicated model training. However, these approaches require substantial effort, such as creating extensive annotated datasets and training dedicated STS models.

This research introduces CoT-STS, which aims to customize the use of the chain-of-thought prompting with LLMs for the STS task. We proposed four influential factors as part of the Chain-of-Thought approach, including theme similarity, participating object similarity, similarity of the activity being carried out, and the evaluation of other factors before arriving at the final similarity assessments. The application of the proposed CoT-STS on the BIOSSES dataset achieved a Pearson's correlation of 0.72, surpassing the 0.45 correlation achieved by the standard prompting and the correlation of 0.71 achieved by the existing zero-shot CoT methodology. The result achieved demonstrates the potential of LLMs with an appropriate prompting strategy to significantly improve the performance of the STS task.

**CCS CONCEPTS** • Information Systems • Information Retrieval • Evaluation of Retrieval Results • Relevance Assessment

**Additional Keywords and Phrases:** Chain-of-Thought Prompting, Large Language Models, Semantic Textual Similarity

**ACM Reference Format:**

Musarrat Hussain, Ubaid Ur Rehman, Tri D.T. Nguyen, and Sungyoung Lee. 2023. CoT-STS: A Zero-Shot Chain-of-Thought Prompting for Semantic Textual Similarity. In 2023 6th Artificial Intelligence and Cloud Computing Conference (AICCC) (AICCC 2023), December 16–18, 2023, Kyoto, Japan. ACM, New York, NY, USA, 8 pages.

## 1 INTRODUCTION

Semantic textual similarity (STS) is a natural language processing (NLP) task that seeks to assess the level of semantic similarity between a pair of text [1]. It empowers NLP systems to comprehend and process human language more effectively, leading to improved responses and enhancing various applications, including information retrieval, question-answering, paraphrase detection, language translation, and information extraction [2]. Over time, researchers have proposed diverse methodologies for STS, spanning from syntactic and structural evaluation, word-embedding methods to deep learning-based approaches, with the latest focus being on Large Language Models (LLMs) [2], [3].

The emergence of LLMs have revolutionized the field of NLP due to their ability to perform various language-related tasks, such as text processing, text generation, translation, sentiment analysis, question-answering, and more [4]. These models have shown remarkable performance in understanding context, grammar, and semantics, allowing them to generate coherent and contextually appropriate responses. Therefore, it has transformed the focus of the NLP research and technical development from features engineering, architecture engineering, and objective engineering to prompt engineering.

Prompt engineering is the process of formulating precise and efficient instructions, referred to as prompts, which guide LLMs in carrying out specific NLP tasks. These prompts enforce constraints, rules, automate processes, and ensure desired qualities and quantities of the generated output [5]. Previously, various prompting methodologies have been proposed and applied, including role prompting, zero-shot prompting, few-shot prompting, and Chain-of-Thought (CoT) prompting [6], [7]. The CoT prompting method enhances the reasoning capabilities of LLMs by breaking down the target NLP task into intermediate sub-steps and seeking their solutions before generating the final results. This methodology can help in improving various NLP tasks including STS. However, the CoT methodology requires a detailed sub-steps example so that LLMs can mimic the reasoning processing of dividing a problem into sub-steps before arriving at final solution. While in case of sematic similarity evaluation, as highlighted by Deshpande et al. [1], the task of STS is inherently ill-defined, as the similarity between a pair of text can fluctuate significantly depending on various attributes and factors being considered. Therefore, identifying the most appropriate and influential factors and sub-steps of the Chain-of-Thought can increase the overall performance of the STS task.

This research proposes CoT-STS, a novel semantic textual similarity approach, which is founded on the innovative chain-of-thought prompting methodology. The CoT-STS comprehensively assesses and constraints the similarity between two input texts through four key factors as chain-of-thought aspects. These factors encompass theme similarity, the similarity of objects involved, the similarity of activities described within both texts, as well as other potential factors that influence similarity. The identified influential factors remain the same throughout the evaluation and across various datasets. In the evaluation phase, the effectiveness of the CoT-STS is demonstrated through its ability to aggregate individual similarity scores from these chain-of-thought aspects to produce a final similarity score. This innovative approach was tested on the BIOSSES dataset, boasting a Pearson's correlation coefficient of 0.72. This result is significantly higher than the 0.45 correlation achieved with conventional prompting methodology as well as 0.71 correlation score of existing zero-shot CoT methodology [8]. The result achieved reveals the substantial potential of leveraging LLMs in combination with a well-suited prompting methodology like CoT-STS to significantly enhance the performance of the STS task, offering new avenues for advancing the field.

## 2 RELATED WORK

The objective of STS is to evaluate the semantic similarity of two text snippets. Researchers have proposed various methodologies in the past to achieve this objective. The proposed methodologies can be broadly divided into three categories, including syntactic or string based similarity, structural, and semantic similarity measures [3], [9]. Syntactic similarity measures mainly focus on the tokens of the text and evaluate word overlap between two texts for similarity evaluation; however, they suffer from token synonyms and polysemy, as the same content can be represented in diverse textual forms using different terminologies. Commonly used syntactic similarity evaluation methodologies include bags of words overlap, Jaccard similarity, windows of words overlap, the ratio of shared skipped bigrams, edit distance, and others, [3], [10].

On the other hand, the structural similarity measures evaluate the lexical structure and taxonomical relationships among various words using diverse parsing methodologies [11], [12]. These methods help in decomposing various clauses and their associated structure to ensure the similarity between text pairs. The primary drawback of these methods roots back to the basic hypothesis that similar structured text tends to be semantically similar, which is not always true; diverse text can have similar structure, and the same structured text can have a different meaning. To overcome this drawback, researchers moved on to embedding-based similarity evaluations [13]. The most commonly employed techniques for assessing similarity rely on embeddings. In this approach, both input texts are converted into vector representations (embeddings), and the vectors are evaluated using similarity measures such as Cosine similarity for their semantic similarity evaluation. Despite wide usage, the correctness of embedding-based methods heavily depends on the text-to-vector transformation methodology; accurate representation leads to more robust performance, and, vice versa.

In recent years, the advancement of generative AI and LLMs have caused a paradigm shift in all NLP-related tasks, including the STS. The NLP research of the current era is more focused on the utilization of LLMs with effective prompt methodologies, which are the driving force behind the LLMs [7], [14]. Some of the most common prompting methodologies include, shot-prompting, chain-of-thought (CoT) [7], zero-shot CoT [8], Auto CoT [15], Least-to-Most [16], DecomP [17], and plan-and-solve prompting [18]. All of these prompting methodologies aim to provide LLMs with appropriate context of handling various NLP tasks. In particular the widely used prompting methodology Chain-of-Thought (CoT) [7], which aims to break down a task into sub-steps to make LLMs better understand the problem and increase its reasoning capabilities. However, building sub-steps examples could be difficult in some satiation, therefore, Kojima et al. [8] proposed a zero-shot version of this methodology hereafter zero-shot CoT. The authors were able to produced comparable results in zero-shot settings by providing LLMs with a string of “*lets think step by step*”. The provided string enables LLMs to identify sub-steps required to solve a problem and apparently produce appropriate results for complex problems. LLMs with effective prompt methodologies have already proven results in various domains, including clinical text de-identification [19], among many others [20]. However, to the best of our knowledge, this is the first study exploring the application of prompt engineering with LLMs for STS evaluation. As highlighted by Deshpande et al. [1], for appropriate semantic similarity evaluation, we need to provide the model with explicit targeted aspects, because a text may be semantically equivalent in one aspect while may have completely different sematic meanings in other. Therefore, taking inspiration from Deshpande et al. [1] and Kojima et al. [8] who proposed a zero-shot CoT, the proposed methodology utilizes zero-shot Chain-of-Thought and conditions the similarity evaluation on four major factors, including theme, participating objects, activities, and others. We believe conditioning the STS task on the provided factors help LLMs to better evaluate semantic similarity of a text pair. Therefore, this research can be a stepping stone for the applications of effective prompt engineering and LLMs for evaluating the semantic similarity of text pairs.

## 3 PROPSOED METHODLOGY

The similarity between a pair of texts can be influenced by multiple factors, as mentioned in Deshpande et al.'s work [1]. Consequently, it is necessary to break down the task of evaluating similarity into sub-tasks to enhance the evaluation process. Thus, the proposed CoT-STS methodology breaks down the STS task into four sub-factors as part of the Chain-of-Thought framework. In our opinion the most important factors impacting textual similarity includes; theme of the text, participating objects, activities being carried out, and any additional factors described in the sentences. The resultant proposed CoT-STS prompt, compared to the standard prompt, is shown in Figure 1.

In the context of textual analysis, theme pertains to the central subject matter addressed within a given text. Therefore, it stands as a pivotal determinant, significantly impacting the level of textual similarity observed between two text. When two texts share a common theme, they often exhibit a higher degree of similarity in their content and language, resulting in a substantial overlap of semantic meaning. Conversely, when text excerpts delve into disparate thematic domains, the semantic correlation between them becomes notably minimal, even if they employ similar lexical tokens or words. Thus, within the framework of our CoT prompt, we have prioritized theme as the primary factor influencing text similarity. As such, our initial consideration point "*Similarity between the themes of the sentences*" focuses on evaluating the similarity between the themes encapsulated within the sentences under examination.

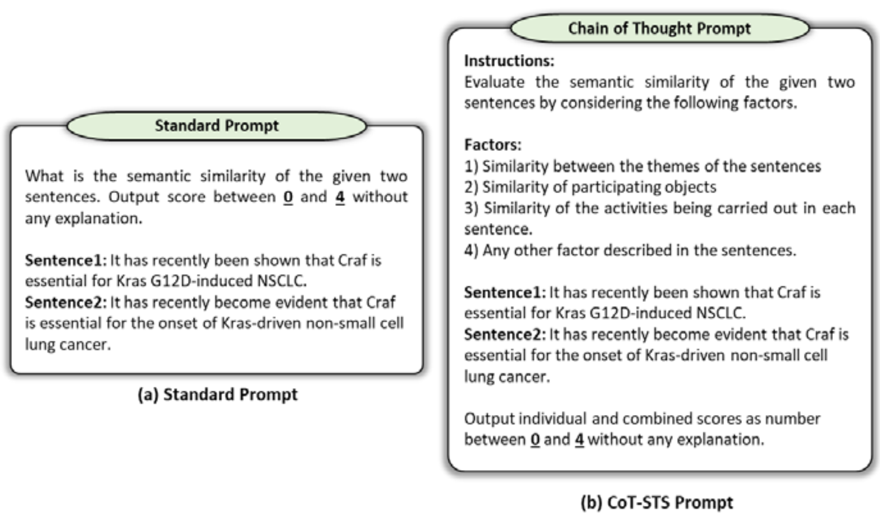


Figure 1: Standard vs proposed zero-shot CoT-STS prompts used for STS evaluation.

The second important factor that plays a crucial role in determining textual similarity pertains to the participating objects mentioned within the text. When the objects in two texts are similar, it tends to result in a higher degree of semantic similarity between them. Conversely, if the objects in the texts are diverse or dissimilar, it can have a negative impact on the similarity between pairs of texts. Hence, we can categorize this as the second factor of the prompt, which we refer to as the "*Similarity of participating objects*". However, it is also important to note that a text might feature similar objects engaged in various activities. In light of this, we need to consider the activities being carried out within the text as another factor influencing its semantics. This can be described as the "*Similarity of the activities being carried out in each sentence*" making it our third factor to examine in the context of text similarity analysis.

Additionally, there may be some additional factors that can modify textual meaning. Therefore, we include the text "*Any other factor described in the sentence*" as an additional factor of the Chain-of-Thought for evaluating textual similarity. In our opinion, the aforementioned four factors are good enough to evaluate the semantic similarity of a text pair. The final score of a text snippet is achieved by averaging the aforementioned factors individual scores as shown in the Equation 1. where $S$ is final similarity score and $s_i$ represent individual factor similarity. As we have utilized only four factors as part of the Chain-of-Thought, therefore $N=4$ in this special scenario, theme, participating objects, activities, and other factors similarities, respectively. In this study, we have considered each factor with similar importance; however, diverse weightage can be defined for various factors to give more emphasis to some aspects compared to others. This holistic approach allows for a more comprehensive assessment of the semantic aspects of the text and contributes to a deeper understanding of its similarity to other texts.

$$S = \frac{1}{N}\sum_{i=1}^{N}(s_i) \qquad (1)$$

## 4 EXPERIMENTAL SETUP

The proposed methodology, presented in Section 4, represents a theoretical framework for semantic similarity evaluation. To construct a robust implementation and assess the effectiveness of the proposed framework, we utilize the ChatGPT version 3.5 interface[1] on the Biomedical Semantic Similarity Estimation System (BIOSSES) dataset [21]. For each sentence evaluation, we refresh the ChatGPT interface to reset its context before each sentence pair. The BIOSSES dataset comprises a total of 100 sentences from the biomedical domain, each annotated by five independent human experts with scores ranging from 0 to 4. Therefore, we instructed ChatGPT to provide a score within the same range of 0 to 4. A score of zero indicates a significant difference in semantic meaning between the two sentences, while a score of four indicates a perfect semantic match.

The Pearson's correlation coefficient is used as an evaluation metric, measuring the linear relationship between scores. The Pearson score varies between -1 and +1, where 0 indicates no correlation, while -1 or +1 imply a perfect linear relationship. The negative and positive signs indicate the direction of the correlation in terms of negative and positive correlations, respectively. However, in our setting, we treat both correlations as equivalent, regardless of their direction.

[1] https://chat.openai.com

## 5 RESULTS AND DISCUSSION

The prompts depicted in Figure 1 serve as essential instructions guiding ChatGPT's evaluation for similarity scores. The prompts are evaluated by individually inserting all the sentences from the BIOSSES dataset into the standard (base-line), zero-shot CoT (state-of-the-art related methodology) and CoT-STS prompts (proposed methodology), resulting in similarity scores ranging from 0 to 4. The assessment outcomes, as measured by Pearson's correlation, are then presented in Figure 2, highlighting the comparison between the standard prompt, zero-shot CoT and the innovative CoT-STS prompt. These results are compared with those obtained from human annotators (Annotator A, Annotator B, Annotator C, Annotator D, Annotator E, and their average)

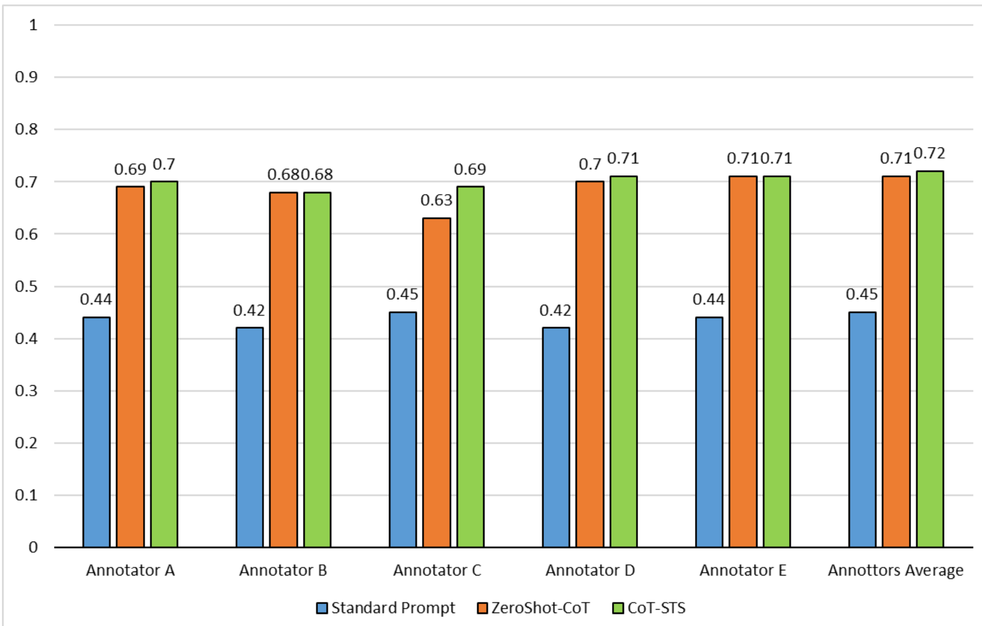


Figure 2: Pearson correlation comparison among the standard prompt, zero-shot CoT and the proposed CoT-STS prompt results in contrast with human annotators.

The performance of various prompts within the context of similarity evaluation is illustrated in Figure 2. The initial prompt, commonly referred to as the standard prompt, produces a correlation score of 0.45. In contrast, a state-of-the-art approach known as zero-shot CoT [8] attains a higher correlation score of 0.71. Conversely, our proposed CoT-STS methodology achieves an even better correlation score of 0.72, as compared to the average scores assigned by human annotators. To obtain results on the BIOSSES dataset, we follow the step-by-step methodology described in [9] for zero-shot CoT.

The proposed CoT-STS prompt demonstrates a significant improvement of 0.27 points in correlation when compared to the standard prompt. Furthermore, it exhibits a slight 0.01-point improvement compared to the zero-shot CoT. In addition to the marginal improvement in correlation score compared to the zero-shot CoT, our methodology necessitates only a single request per text pair, whereas the zero-shot CoT requires two requests per text pair. The first request is used for extracting reasoning steps, followed by another request for result extraction. In contrast, our proposed methodology maintains the identified influential factors consistently throughout the task, allowing us to extract results with just a single request to ChatGPT.

Delving deeper into the specifics of our evaluation, when we assess the CoT-STS prompt against individual annotators, we observe its superior performance with Annotator A, showcasing a remarkable correlation score of 0.70. This impressive score represents a significant increase of 0.26 points compared to the standard prompt's performance and a marginal 0.01-point improvement over the zero-shot CoT in correlation with this particular annotator. The trend of substantial gains extends across the spectrum, as Annotators B, C, D, and E all exhibit substantial enhancements in their correlation scores. Specifically, we observe an improvement of 0.26, 0.24, 0.29, and 0.27 points, respectively, when comparing the performance of the CoT-STS prompt to that of the standard prompt. However, when we compare it to the zero-shot CoT, we find that similar results were obtained with Annotators B and E. There was a minor 0.06-point improvement in correlation with Annotator C and a slight 0.01-point improvement with Annotator D. As highlighted earlier, in addition to the improved correlation scores, the proposed prompting methodology, CoT-STS, also excels in terms of execution time efficiency. Unlike the zero-shot CoT, which requires two requests per text pair, the CoT-STS prompt necessitates only a single request for evaluating the similarity of a text pair. Consequently, the results obtained underscore a consistent and meaningful advancement in the CoT-STS prompt's ability to capture semantic similarity. These findings disclose the potential of conditioning similarity evaluation on factors such as theme, actors, and activities. By incorporating these elements into the prompt, the CoT-STS prompt clearly demonstrates its advantage in helping language models better comprehend the semantic meanings embedded within the provided text.

As previously mentioned, in our approach to determine the final similarity score of a text, we follow a specific procedure. This procedure involves the averaging of individual factor scores, as outlined in the proposed CoT-STS equation shown in Equation 1. It's worth noting that in some instances, ChatGPT did not identify any additional factors beyond the core elements of theme, objects, and activities. Conversely, there were also cases where certain sentence pairs contained more than one external factor, leading to the generation of more than four distinct scores.

In all scenarios, regardless of the number of individual factors detected, we consistently computed the final similarity score by summing up these individual factor scores and then dividing the sum by four. This standardization approach ensures that the final similarity score remains consistent and comparable across different text pairs. Furthermore, it is important to highlight that despite our explicit instructions to generate output scores without textual explanations, ChatGPT occasionally provided results accompanied by explanatory text. Interestingly, this phenomenon was more visible in the context of the CoT-STS prompt when compared to the standard prompt. This variation in behavior between the two prompts warrants further investigation and analysis.

## 6 CONCLUSION

The emergence of generative AI and Large Language Models (LLMs) have drastically improved various NLP tasks, including question-answering, information retrieval, text summarization, and others. The focus of NLP research has shifted to the utilization of LLMs through an efficient and effective prompt engineering methodology. Therefore, this paper presents a zero-shot Chain-of-Thought prompting methodology (CoT-

STS) for evaluating semantic similarity between text pairs using ChatGPT. We considered four fundamental factors, including theme, objects, activities, and other external factors, as Chain-of-Thought elements affecting the similarity measurements between text pairs. The evaluation of the proposed CoT-STS, compared to the standard prompt, and zero-shot CoT [8], on the Biomedical Semantic Similarity Estimation System (BIOSSES) dataset resulted in a 0.27-point and 0.01-point improvement in terms of Pearson correlation, respectively. The results reveal the effectiveness of breaking down the STS task into sub-factors. In the future, we will include an ablation study of the targeted four factors and will also consider replacing the targeted factor with a diverse set of factors to identify the most influential Chain-of-Thought factors.

## ACKNOWLEGEMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2022-2020-0-01489), the ITRC (Information Technology Research Center) support program (RS-2023-00259004) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation), (IITP-2017-0-00655, Lean UX core technology and platform for any digital artifacts UX evaluation).

## REFERENCES

[1] A. Deshpande *et al.*, “CSTS: Conditional Semantic Textual Similarity.” arXiv, May 24, 2023. Accessed: Oct. 20, 2023. [Online]. Available: http://arxiv.org/abs/2305.15093
[2] “Measurement of Text Similarity: A Survey”, doi: 10.3390/info11090421.
[3] H. Hassanzadeh, T. Groza, A. Nguyen, and J. Hunter, “A Supervised Approach to Quantifying Sentence Similarity: With Application to Evidence Based Medicine,” *PLOS ONE*, 2015.
[4] J. Wang *et al.*, “Prompt Engineering for Healthcare: Methodologies and Applications.” arXiv, Apr. 28, 2023. Accessed: Oct. 20, 2023. [Online]. Available: http://arxiv.org/abs/2304.14670
[5] J. White *et al.*, “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.” arXiv, Feb. 21, 2023. Accessed: Oct. 20, 2023. [Online]. Available: http://arxiv.org/abs/2302.11382
[6] S. Sivarajkumar and Y. Wang, “HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing”.
[7] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”.
[8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners”.
[9] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, “Semantic Textual Similarity Methods, Tools, and Applications: A Survey,” *Comput. Sist.*, vol. 20, no. 4, 2016.
[10] N. Gali, “Framework for syntactic string similarity measures,” *Expert Syst. Appl.*, 2019.
[11] M. Farouk, “Measuring text similarity based on structure and word embedding,” *Cogn. Syst. Res.*, 2020.
[12] M. Alian and A. Awajan, “Syntactic-Semantic Similarity Based on Dependency Tree Kernel,” *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 10937–10948, Aug. 2023, doi: 10.1007/s13369-023-07694-z.
[13] T. Ranasinghe, C. Orasan, and R. Mitkov, “Semantic Textual Similarity with Siamese Neural Networks”.
[14] J. D. Zamfrescu-Pereira, R. Wong, B. Hartmann, and Q. Yang, “Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts,” 2023.
[15] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic Chain of Thought Prompting in Large Language Models.” arXiv, Oct. 07, 2022. Accessed: Oct. 23, 2023. [Online]. Available: http://arxiv.org/abs/2210.03493
[16] D. Zhou *et al.*, “Least-to-Most Prompting Enables Complex Reasoning in Large Language Models.” arXiv, Apr. 16, 2023. Accessed: Oct. 23, 2023. [Online]. Available: http://arxiv.org/abs/2205.10625
[17] T. Khot *et al.*, “Decomposed Prompting: A Modular Approach for Solving Complex Tasks.” arXiv, Apr. 11, 2023. Accessed: Oct. 23, 2023. [Online]. Available: http://arxiv.org/abs/2210.02406
[18] L. Wang *et al.*, “Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models.” arXiv, May 26, 2023. Accessed: Oct. 23, 2023. [Online]. Available: http://arxiv.org/abs/2305.04091
[19] Z. Liu *et al.*, “DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4.” arXiv, Mar. 20, 2023. Accessed: Oct. 20, 2023. [Online]. Available: http://arxiv.org/abs/2303.11032
[20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *ACM Comput. Surv.*, vol. 55, no. 9.
[21] G. Sog, “BIOSSES: a semantic sentence similarity estimation system for the biomedical domain”.