

# THiFLY Research at SemEval-2023 Task 7: A Multi-granularity System for CTR-based Textual Entailment and Evidence Retrieval

Yuxuan Zhou<sup>1\*</sup>, Ziyu Jin<sup>2,3\*</sup>, Meiwei Li<sup>2,3</sup>, Miao Li<sup>1</sup>, Xien Liu<sup>1</sup>, Xinxin You<sup>2,3</sup>, Ji Wu<sup>1</sup>

<sup>1</sup>Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

<sup>2</sup>THiFLY Research, Tsinghua University, Beijing 100084, China

<sup>3</sup>State Key Laboratory of Cognitive Intelligence, Hefei, Anhui 230088, China

## Abstract

The NLI4CT task aims to entail hypotheses based on Clinical Trial Reports (CTRs) and retrieve the corresponding evidence supporting the justification. This task poses a significant challenge, as verifying hypotheses in the NLI4CT task requires the **integration of multiple pieces of evidence from one or two CTR(s) and the application of diverse levels of reasoning, including textual and numerical.** To address these problems, we present a multi-granularity system for CTR-based textual entailment and evidence retrieval in this paper. Specifically, we construct a **Multi-granularity Inference Network (MGNet)** that exploits **sentence-level** and **token-level** encoding to handle both textual entailment and evidence retrieval tasks. Moreover, we enhance the numerical inference capability of the system **by leveraging a T5-based model, SciFive,** which is **pre-trained on the medical corpus.** **Model ensembling** and a **joint inference method** are further utilized in the system to increase the stability and consistency of inference. The system achieves f1-scores of **0.856** and **0.853** on textual entailment and evidence retrieval tasks, resulting in the best performance on both subtasks. The experimental results corroborate the effectiveness of our proposed method. Our code is publicly available at <https://github.com/THUMLP/NLI4CT>.

## 1 Introduction

In recent years, with the fast development of digital health, there has been a surge in the publication of Clinical Trial Reports (CTRs). Currently, there are over 10,000 CTRs for Breast Cancer alone<sup>1</sup>. This proliferation of CTRs has enabled the construction of a Natural Language Inference (NLI) system that can aid medical interpretation and evidence retrieval for personalized evidence-based care. The

\*: equal contribution

<sup>1</sup><https://sites.google.com/view/nli4ct/home>

Clinical Trial Report	
<b>Intervention</b>	
...	
<b>Eligibility</b>	
Inclusion Criteria:	
...	
Exclusion Criteria:	
<ul style="list-style-type: none"> <li>Prior treatment with a phosphatidylinositol 3-kinase (PI3K) inhibitor, mammalian target of rapamycin (mTOR) inhibitor (e.g. everolimus), or protein kinase B (AKT) inhibitor</li> </ul>	
<b>Results</b>	
...	
<b>Adverse Events</b>	
Adverse Events 1:	Adverse Events 2:
<ul style="list-style-type: none"> <li>Total: 19/213 (8.92%)</li> <li>Diarrhoea 0/213 (0.00%)</li> </ul>	<ul style="list-style-type: none"> <li>Total: 133/416 (31.97%)</li> <li>Diarrhoea 32/416 (7.69%)</li> </ul>
...	
<b>Hypotheses:</b>	
<ul style="list-style-type: none"> <li>A total of 32 patients in the primary trial had <b>Diarrhoea</b>.</li> <li>Prior treatment with fulvestrant or with a phosphatidylinositol 3-kinase (PI3K) inhibitor will not prevent a patient from participating in the primary trial</li> </ul>	

Figure 1: An Example of NLI4CT.

NLI4CT task (Jullien et al., 2023) focuses on constructing an explainable multi-evidence NLI system based on CTRs for Breast Cancer. It consists of two sub-tasks: textual entailment (Task A) and evidence retrieval (Task B). Task A requires the system to determine if the CTRs entail or contradict the given hypotheses. Task B requires the system to find evidence in the CTRs necessary to justify the prediction in Task A. The primary difficulty inherent in NLI4CT is that the process of hypothesis verification frequently involves the integration of several pieces of evidence from the premise. In certain cases, justifying a hypothesis requires a comparison of two separate premise CTRs. Moreover, unlike other clinical NLI benchmarks (e.g., MedNLI), NLI4CT presents a unique feature containing four distinct types of premises - Intervention, Eligibility, Results, and Adverse Events - that align with the four sections typically found in each CTR. Validating hypotheses based on each premise type requires different levels of inference skills (textual, numerical, etc.). Fig.1 presents an example of NLI4CT. The first hypothesis involves

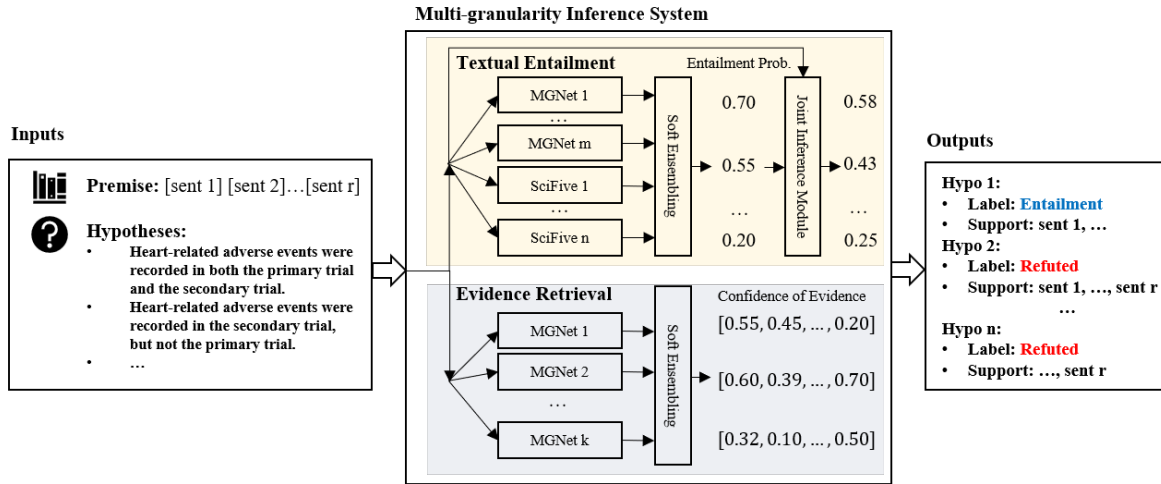


Figure 2: An overview of the proposed Multi-granularity System for NLI4CT task. MGNet refers to the proposed multi-granularity inference network.

summing the numbers of patients who experienced Diarrhoea in two distinct cohorts to conduct the verification. Alternatively, verifying the second hypothesis requires evaluating treatment modality against the criteria outlined in the Exclusion Criteria.

This paper proposes a multi-granularity inference system for the NLI4CT task to address the aforementioned challenges. An overview of the proposed system is presented in Fig.2. The system comprises a multi-granularity inference network that performs **joint semantics encoding of the hypothesis and premise, followed by multi-granularity inference through sentence-level and token-level encoding** of the encoded sequence to accomplish Task A and Task B. To handle the various levels of inference involved, **we utilize the SciFive model (Phan et al., 2021) to enhance the system’s performance on hypotheses requiring numerical inference skills**. Our system is presented as an ensemble of models to increase the stability and consistency of inference. A joint inference module is further proposed to refine the inference outcomes of hypotheses by detecting semantic consistency between hypothesis pairs sharing the same premise. Our system achieves micro f1-scores of 0.856 and 0.853 on Task A and Task B, respectively. The main contributions of this work are as follows:

- We present a multi-granularity inference network that performs both textual entailment and evidence retrieval tasks concurrently. We further incorporate the SciFive model to strengthen the system’s numerical inference capability.

- We introduce a joint inference module that improves the inference consistency by exploiting the polarity of hypotheses pairs that share the same premise.
- Our proposed framework achieves state-of-the-art f1-scores of 0.856 and 0.853 on Task A and B, respectively.

## 2 Related Work

**BiolinkBERT** BiolinkBERT (Yasunaga et al., 2022) is a pre-trained model that utilizes hyperlinks within documents. The model incorporates linked documents into the input of pre-training and is pre-trained using masked language modeling and document relation prediction tasks. BiolinkBERT has demonstrated state-of-the-art performance on various medical NLP benchmarks, such as BLURB (Gu et al., 2021) and BioASQ (Nentidis et al., 2020).

**SciFive** SciFive (Phan et al., 2021) is a T5-based model pre-trained on large biomedical corpora, namely PubMed Abstract and PubMed Central. This model has shown the potential of text-generating models in medical NLP tasks and has achieved state-of-the-art performance on MedNLI (Romanov and Shivade, 2018) and DDI (Herrero-Zazo et al., 2013).

## 3 Task Formulation

The Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) task comprises two sub-tasks: textual entailment and evidence

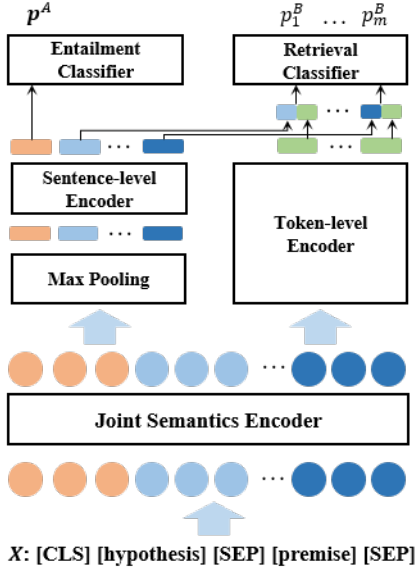


Figure 3: The structure of the proposed Multi-granularity Inference Network (MGNet).

retrieval. Both sub-tasks share a common input comprising a premise  $P$  and a hypothesis  $S$ . The hypothesis is a single sentence, while the premise comprises several sentences in the CTR(s), i.e.,  $P = [P^1, P^2, \dots, P^m]$ .

The textual entailment sub-task can be viewed as a binary classification problem, with the entailment module  $f_{\text{Ent}}$  generating a predicted result  $\hat{y} \in \{0, 1\}^n$  for all input hypotheses:  $f_{\text{Ent}}(S, P) = \hat{y}$ . In this context, the predicted value  $\hat{y} = 1$  indicates that the premise entails the hypothesis, while  $\hat{y} = 0$  denotes that the hypothesis contradicts the premise. On the other hand, the evidence retrieval sub-task can be seen as a binary classification problem targeting each sentence in the premise. The evidence retrieval module  $f_{\text{Ret}}$  produces a predicted result  $\hat{\mathbf{r}} \in \{0, 1\}^m$ :  $f_{\text{Ret}}(S, P) = \hat{\mathbf{r}}$ , where  $\hat{r}_j = 1$  indicates that the  $j^{\text{th}}$  sentence in the premise supports the justification of  $S$ ,  $\hat{r}_j = 0$  means the sentence is not related to the justification of  $S$ .

## 4 Models

In this section, we introduce models applied in our inference system: Multi-granularity Inference Network (MGNet), SciFive (Phan et al., 2021), and the joint inference network.

### 4.1 Multi-granularity Inference Network

The multi-granularity inference network aims to handle both the textual entailment and evidence retrieval subtasks using integrated token-level and

sentence-level representations. The network structure is illustrated in Figure 3. This section is organized as follows: Section 4.1.1 presents the joint semantics encoder, which encodes the contextual semantics between the hypothesis and the premise. Section 4.1.2 describes the sentence-level encoder, which learns the sentence-level contextual semantics of the hypothesis and the premise. Section 4.1.3 describes the token-level encoder, which extracts the token-level joint representation between the hypothesis and each sentence in the premise. Finally, Section 4.1.4 introduces the classifier module, which accomplishes Task A and Task B inference based on the learned multi-granularity semantics.

#### 4.1.1 Joint Semantics Encoder

The network initially employs a transformer-based (Vaswani et al., 2017) language model to learn the joint contextual representation of the hypothesis and premise, aiming to incorporate medical domain knowledge into the network. Specifically, we concatenate each hypothesis  $S$  with the premise  $P$  to form an input token sequence  $\mathbf{X} = [[\text{CLS}], \mathbf{S}, [\text{SEP}], \mathbf{P}, [\text{SEP}]]$ , where  $\mathbf{S}$  and  $\mathbf{P}$  refer to the tokenization of  $S$  and  $P$ , respectively. Further details regarding data preprocessing can be found at Sec.B in the appendix. The token sequence is subsequently embedded and encoded by the transformer-based language model as follows:

$$\mathbf{H}^0 = f_{\text{Emb}}(\mathbf{X}) \quad (1)$$

$$\mathbf{H}^l = f_{\text{Enc}}^l(\mathbf{H}^{l-1}) \quad (2)$$

where  $\mathbf{H}^0$  denotes the output of the embedding layer,  $\mathbf{H}^l \in \mathbb{R}^{N \times d}$  ( $l \in \{1, 2, \dots, L\}$ ) refers to the output representation of the  $l^{\text{th}}$  transformer encoder layer,  $N$  is the sequence length and  $d$  the dimension of the representation vector. Here,  $f_{\text{Emb}}$  and  $f_{\text{Enc}}^l$  refer to the embedding layer and the  $l^{\text{th}}$  encoding layer, respectively. Finally, the output representation of the final layer,  $\mathbf{H}^L$ , is further processed in the subsequent encoders.

#### 4.1.2 Sentence-level Encoder

The sentence-level encoder conducts contextual encoding of the hypothesis and the premise based on the learned token-level representation. Specifically, the encoder first conducts max-pooling<sup>2</sup> to extract

<sup>2</sup>We have tried other pooling methods (mean-pooling, attention pooling, etc.). However, they fail to achieve higher performance than max-pooling.

sentence-level representation from the token-level representation of each sentence, as follows:

$$\hat{\mathbf{H}}_i^s = \text{Max-pooling}(\{\mathbf{H}_j^L : j \in \mathbf{I}_i\}) \quad (3)$$

where  $\hat{\mathbf{H}}^s \in \mathbb{R}^{(m+1) \times d}$  refers to the pooled sentence-level representation and  $\mathbf{I}_i$  denotes the subscript set of  $\mathbf{X}$  that corresponds to the tokens in the  $i^{th}$  sentence. For example,  $\hat{\mathbf{H}}_1^s$  corresponds to the hypothesis, while  $\hat{\mathbf{H}}_{i+1}^s$  corresponds to the  $i^{th}$  sentence in the premise.

After that, the resulting  $\hat{\mathbf{H}}^s$  output is fed into an encoder designed to capture contextual semantics. The motivation for this encoder stems from the need to justify some hypotheses using multiple pieces of evidence. To achieve this, we implement the encoder by two alternative networks:

(1) BiLSTM (Hochreiter and Schmidhuber, 1997): the network utilizes two LSTMs to compute the representation of the given sequence from both directions:

$$\vec{\mathbf{H}}^s = \text{LSTM}_{for}^s(\hat{\mathbf{H}}^s) \quad (4)$$

$$\overleftarrow{\mathbf{H}}^s = \text{LSTM}_{rev}^s(\hat{\mathbf{H}}^s) \quad (5)$$

The final representation of each sentence  $\mathbf{H}^s \in \mathbb{R}^{(m+1) \times d}$  is obtained by concatenating the forward and backward representations and passing them to a linear layer:

$$\mathbf{H}_i^s = \mathbf{W}_1[\vec{\mathbf{H}}_i^s; \overleftarrow{\mathbf{H}}_i^s] + \mathbf{b}_1 \quad (6)$$

where  $\mathbf{W}_1 \in \mathbb{R}^{d \times 2d}$ ,  $\mathbf{b}_1 \in \mathbb{R}^d$  are trainable parameters. Finally, the global representation of the hypothesis-premise pair  $\tilde{\mathbf{h}}^s \in \mathbb{R}^d$  is computed by concatenating the last hidden state of the forward and backward encoding ( $\mathbf{W}_2 \in \mathbb{R}^{d \times 2d}$ ,  $\mathbf{b}_2 \in \mathbb{R}^d$  are also trainable parameters):

$$\tilde{\mathbf{h}}^s = \mathbf{W}_2[\vec{\mathbf{H}}_{m+1}^s; \overleftarrow{\mathbf{H}}_1^s] + \mathbf{b}_2 \quad (7)$$

(2) Transformer (Vaswani et al., 2017): a  $L^s$ -layer transformer encoder integrates reasoning-related evidence with the sentence-level attention mechanism. In this case, we take the last hidden state as the final representation of each sentence and the hypothesis's representation as the global representation:

$$\mathbf{H}^s = f_{Enc}^s(\hat{\mathbf{H}}^s) \quad (8)$$

$$\tilde{\mathbf{h}}^s = \mathbf{H}_i^s \quad (9)$$

#### 4.1.3 Token-level Encoder

The purpose of the token-level encoder is to capture the reasoning-related semantics of each sentence in the premise. Compared to the sentence-level encoder, it provides a fine-grained representation of individual sentences in premise, which has aided in evidence retrieval. To encode the  $i^{th}$  sentence in the premise, we concatenate it with the hypothesis and input it into the encoder. Two alternative networks can be used to implement the encoder:

(1) BiLSTM: A BiLSTM network is leveraged to encode the forward and backward representation of the given sequence:

$$\vec{\mathbf{H}}_i = \text{LSTM}_{for}^t([\mathbf{H}_1^L, \mathbf{H}_{i+1}^L]) \quad (10)$$

$$\overleftarrow{\mathbf{H}}_i = \text{LSTM}_{rev}^t([\mathbf{H}_1^L, \mathbf{H}_{i+1}^L]) \quad (11)$$

where  $\mathbf{H}_{I_j}^L \in \mathbb{R}^{N_j \times d}$  refers to the representation of tokens from the  $j^{th}$  sentence in  $\mathbf{X}$ ,  $N_j$  denotes the length of the  $j^{th}$  sentence,  $[\cdot, \cdot]$  denotes the concatenation of matrices on the first dimension,  $\vec{\mathbf{H}}_i, \overleftarrow{\mathbf{H}}_i \in \mathbb{R}^{(N_1+N_{i+1}) \times d}$  are the output representation of the LSTMs. Finally, the representation of the  $i^{th}$  sentence in the premise is also obtained by concatenating the last hidden state of LSTMs:

$$\mathbf{H}_i^t = \mathbf{W}_3[\vec{\mathbf{H}}_{N_1+N_{i+1}}^t; \overleftarrow{\mathbf{H}}_1^t] + \mathbf{b}_3 \quad (12)$$

where  $\mathbf{W}_3 \in \mathbb{R}^{d \times 2d}$  and  $\mathbf{b}_3 \in \mathbb{R}^d$  are trainable parameters as well.

(2) max-pooling: A simpler implementation of the encoder is a max-pooling layer that merges the representation of the sequence:

$$\mathbf{H}_i^t = \text{Max-pooling}([\vec{\mathbf{H}}_i^t, \overleftarrow{\mathbf{H}}_i^t]) \quad (13)$$

#### 4.1.4 Classifiers

We implement the classifiers with simple structures for both Task A and B. For Task A (textual entailment), we use a double-layer MLP that takes the sentence-level global representation  $\tilde{\mathbf{h}}^s$  as the input:

$$\mathbf{p}^A = \text{Softmax}(\mathbf{W}_2^A \sigma(\mathbf{W}_1^A \tilde{\mathbf{h}}^s + \mathbf{b}_1^A) + \mathbf{b}_2^A) \quad (14)$$

where  $\mathbf{W}_1^A \in \mathbb{R}^{d \times d}$ ,  $\mathbf{b}_1^A \in \mathbb{R}^d$ ,  $\mathbf{W}_2^A \in \mathbb{R}^{2 \times d}$ ,  $\mathbf{b}_2^A \in \mathbb{R}^2$  are MLP's parameters,  $\sigma$  denotes the GELU activation function (Hendrycks and Gimpel, 2016). The output  $\mathbf{p}^A$  suggests that the given hypothesis has a  $p_1^A$  probability of being false and a  $p_2^A$  probability of being true based on the premise. For Task



B (evidence retrieval), we project the concatenation of sentence-level and token-level representations into a scalar value to determine whether the  $i^{th}$  sentence in the premise supports the hypothesis:

$$p_i^B = \text{Sigmoid}(\mathbf{W}^B[\mathbf{H}_i^s; \mathbf{H}_i^t] + b^B) \quad (15)$$

where  $\mathbf{W}^B \in \mathbb{R}^{1 \times 2d}$ ,  $b^B \in \mathbb{R}$  are trainable parameters.  $p_i^B$  measures the probability that the  $i^{th}$  sentence in the premise supports the justification of the hypothesis.

## 4.2 SciFive

The proposed MGNet performs well on Task A and Task B through a unified multi-granularity encoding process. Nevertheless, we found in experiments that **MGNet struggles with hypotheses that require numerical inference skills**. Motivated, we implement SciFive to enhance the system’s numerical inference ability. Following the original paper, we preprocess the input sequence  $\mathbf{X}$  in the format of “nli hypothesis: [hypothesis] premise: [premise]”. SciFive then computes the probabilities of generating “entailment” ( $P(\text{“ent”}|\mathbf{X})$ ) and “contradiction” ( $P(\text{“con”}|\mathbf{X})$ ) given the input sequence. The prediction for Task A is then calculated using the following equations:

$$p_1^A = \frac{P(\text{“con”}|\mathbf{X})}{P(\text{“con”}|\mathbf{X}) + P(\text{“ent”}|\mathbf{X})} \quad (16)$$

$$p_2^A = \frac{P(\text{“ent”}|\mathbf{X})}{P(\text{“con”}|\mathbf{X}) + P(\text{“ent”}|\mathbf{X})} \quad (17)$$

## 4.3 Joint Inference Network

The inference models we implemented have demonstrated satisfactory performance in justifying single hypotheses. **However, these models tend to produce identical predictions for hypotheses that share the same premise but contradict each other**, leading to inaccurate results. For example, consider the hypotheses “Heart-related adverse events were recorded in **both** the primary trial **and** the secondary trial” and “Heart-related adverse events were recorded in the secondary trial, **but not** the primary trial,” which are semantically exclusive. Assigning the same label to both hypotheses is an unreasonable outcome. It is worth noting that such a problem will likely occur in real-world evidence-based care systems, where users may interact with the system multiple times and generate multiple hypotheses based on the same premise.

Motivated by this, we aim to enhance inference consistency and performance by leveraging

the mutual information among hypotheses that share the same premise. To achieve this, we consider a set of hypotheses  $S^1, S^2, \dots, S^n$  that share the same premise and design a joint inference network that determines whether a given pair of hypotheses has the same label for Task A under the given premise. We implement the network with a  $L$ -layer transformer encoder. The network initially takes the sequence  $\tilde{\mathbf{X}} = [[\text{CLS}], \mathbf{S}^i, [\text{SEP}], \mathbf{S}^j, [\text{SEP}], \mathbf{P}, [\text{SEP}]]$  as input:

$$\tilde{\mathbf{H}}^L = f_{Enc}^J(\tilde{\mathbf{X}}) \quad (18)$$

where  $f_{Enc}^J$  refers to the  $L$ -layer transformer encoder. A double-layer MLP is applied to make predictions based on the last hidden state of [CLS] that is considered as the representation vector of the entire sequence:

$$\mathbf{c}^{i,j} = \text{Softmax}(\mathbf{W}_2^J \sigma(\mathbf{W}_1^J \tilde{\mathbf{H}}^L + \mathbf{b}_1^J) + \mathbf{b}_2^J) \quad (19)$$

where  $\mathbf{W}_1^J \in \mathbb{R}^{d \times d}$ ,  $\mathbf{b}_1^J \in \mathbb{R}^d$ ,  $\mathbf{W}_2^J \in \mathbb{R}^{2 \times d}$ ,  $\mathbf{b}_2^J \in \mathbb{R}^2$  are trainable parameters and  $\sigma$  represents the Tanh activation function.  $\mathbf{c}_1^{i,j}$  measures the probability that  $S^i$  and  $S^j$  share the same label, and  $\mathbf{c}_2^{i,j}$  measures the probability of having different labels. The joint inference module rectifies the predictions of Task A with the prediction results of this network, and we will introduce the details in the next section.

## 5 System Details

**Model Setup** We initialized the joint semantics encoder in MGNet and the joint inference network with BiolinkBERT-Large-MNLI-SNLI, a language model proposed by Yasunaga et al. (2022) that is pre-trained on the PubMed dataset and further fine-tuned on the Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018) and Stanford Natural Language Inference (SNLI) datasets (Bowman et al., 2015). Following the setting of BiolinkBERT-Large, we set the number of layers  $L$  and the dimension of hidden states  $d$  to 24 and 1024, respectively. We apply Adam (Kingma and Ba, 2015) and Adafactor optimizers (Shazeer and Stern, 2018) for finetuning MGNet and SciFive, respectively.

To provide a comprehensive description of the inference models used in our system, we first outline the loss functions employed in model training:

(1) Cross-entropy loss for textual entailment:

$$\mathcal{L}_A = -(1 - y) \cdot \log p_1^A - y \cdot \log p_2^A \quad (20)$$

Name	Network	ML	Encoder		Loss
			Sentence	Token	
Task A					
M-512-Bi-Bi-mul	MGNet	512	BiLSTM	BiLSTM	$\mathcal{L}_{mul}$
M-512-Tf-Bi-cl	MGNet	512	Transformer	BiLSTM	$\mathcal{L}_{CL}$
M-1024-Tf-Bi-mul	MGNet	1024	Transformer	BiLSTM	$\mathcal{L}_{mul}$
SciFive	SciFive	1024	-	-	$\mathcal{L}_A^*$
Task B					
M-512-Tf-Bi	MGNet	512	Transformer	BiLSTM	$\mathcal{L}_B$
M-512-Bi-Bi	MGNet	512	BiLSTM	BiLSTM	$\mathcal{L}_B$
M-512-Bi-Max	MGNet	512	BiLSTM	Max-pooling	$\mathcal{L}_B$

Table 1: The inference models applied in the proposed system. ML: max length of the input sequence, Tf: transformer model, Bi: BiLSTM model, Max: Max-pooling.  $\mathcal{L}_A^*$  denotes the sequence cross-entropy loss proposed in Raffel et al. (2020).

Here,  $y \in \{0, 1\}$  represents the label of Task A.

(2) Cross-entropy loss for evidence retrieval:

$$\mathcal{L}_B = -\frac{1}{m} \sum_{i=1}^m [(1-r_i) \cdot \log(1-p_i^B) + r_i \cdot \log p_i^B] \quad (21)$$

where  $r_i \in \{0, 1\}$  is the label of the  $i^{th}$  sentence in the premise.

(3) Multitask learning loss: we use a multitask learning loss, which is a combination of the CE loss for Task A and Task B:

$$\mathcal{L}_{mul} = \mathcal{L}_A + \lambda \mathcal{L}_B \quad (22)$$

Here,  $\lambda$  is a hyperparameter that controls the pace of learning on both tasks.

(4) Contrastive learning loss: we use the contrastive learning loss proposed by Gunel et al. (2020), which is a combination of the CE loss and the supervised contrastive learning (SCL) loss:

$$\mathcal{L}_{CL} = \gamma \mathcal{L}_A + (1 - \gamma) \mathcal{L}_{SCL} \quad (23)$$

Here,  $\mathcal{L}_{SCL}$  is the SCL loss that takes the global representation  $\tilde{\mathbf{h}}_1^s, \tilde{\mathbf{h}}_2^s, \dots, \tilde{\mathbf{h}}_N^s$  and the labels  $y_1, y_2, \dots, y_N$  in the batch as input ( $N$  is the batch size). Since our dataset is relatively small, we apply the SCL loss as a regularization technique during training.

The inference models applied in the proposed system are listed in Table.1. To extend the maximum input length of the original BiolinkBERT model from 512 to 1024, we append the positional embeddings of BiolinkBERT with randomly initialized parameters. In Task A, we set the number of layers  $L^s$  for the transformer in the sentence-level encoder as one and the learning rates of the joint semantics encoder and other parts as  $2e-5$  and  $1e-4$ , respectively. The batch size, number of epochs,

and warmup ratio are set to 32, 100, and 0.3, respectively. For SciFive, we set the learning rate, batch size, number of epochs, and warmup steps as  $3e-5$ , 32, 100, and 500, respectively. The hyperparameter values  $\lambda = 0.01$  and  $\gamma = 0.5$  are chosen, and the temperature of the supervised contrastive learning (SCL) loss is set to 0.3. In Task B, we set the learning rate,  $L^s$ , batch size, number of epochs, and warmup ratio as  $5e-6$ , 2, 1, 50, and 0.05, respectively.

In the joint inference network, the learning rate, batch size, number of epochs, and warmup ratio are set to  $2e-5$ , 32, 100, and 0.3, respectively. Positive samples, i.e., hypothesis pairs with the same label, are created using back translation, while negative samples are obtained by identifying hypothesis pairs with different labels that appear in the datasets. Further training details of this network are presented in Section C in the appendix, owing to space limitations.

**Soft Ensembling** We apply a cross-validation-based soft ensembling method to summarize the inference results from different models. Specifically, we perform 10-fold validation for each model in Table.1 using all data in the train and dev sets. For Task A, we save the checkpoint that achieves the best f1-score on the corresponding dev set for each fold, while for Task B, we save two of the best checkpoints for each fold. Then, we average the predictions of  $4 \times 10 = 40$  models trained by cross-validation for Task A and average the predictions of  $3 \times 10 \times 2 + 3 = 63$  models trained by cross-validation (two checkpoints for each fold) and hold-out method (one checkpoint for each model) for soft ensembling. The ensemble predictions for the two tasks are denoted as  $\tilde{\mathbf{p}}^A$  and  $\tilde{\mathbf{p}}^B$ , respectively.

system	Individual Inference			Joint Inference		
	recall	precision	f1-score	recall	precision	f1-score
original system	0.816	0.739	0.776	0.856	0.856	<b>0.856</b>
- M-512-Bi-Bi-mul	0.820	0.732	0.774	0.840	0.837	0.838
- M-512-Tf-Bi-cl	0.796	0.748	0.771	0.848	0.845	0.846
- M-1024-Tf-Bi-mul	0.820	0.745	<b>0.781</b>	0.848	0.851	0.850
- SciFive	0.792	0.717	0.753	0.844	0.844	0.844
BiolinkBERT-Large	0.712	0.674	0.693	0.768	0.768	0.768
SciFive	0.784	0.713	0.747	0.804	0.798	0.801

Table 2: Performance of the proposed system on Task A with ablation experimental results compared with baseline models on the test set of NLI4CT.

system	recall	precision	f1-score
original system	0.898	0.811	<b>0.853</b>
- M-512-Tf-Bi	0.884	0.818	0.850
- M-512-Bi-Bi	0.882	0.801	0.840
- M-512-Bi-Max	0.895	0.811	0.852

Table 3: Performance of the proposed system on Task B with ablation experimental results.

**Joint Inference** As mentioned above, we utilize a joint inference module to rectify the ensembling results  $\tilde{\mathbf{p}}^A$  before performing Task A. This module corrects the prediction of hypotheses that share the same premise using the following equation:

$$\hat{\mathbf{p}}^{A,i} = \frac{1}{n} \sum_{j=1}^n (I_{ij} \tilde{\mathbf{p}}^{A,j} + (1 - I_{ij})(1 - \tilde{\mathbf{p}}^{A,j})) \quad (24)$$

Here,  $\tilde{\mathbf{p}}^{A,i}$  represents the prediction of the  $i^{th}$  hypothesis after ensembling, while  $c^{i,j}1$  denotes the prediction results of the joint inference network discussed in Sec.4.3. The function  $I_{ij} = \mathbb{I}[c^{i,j}1 > 0.5]$  employs an indicative function  $\mathbb{I}[\cdot]$ , which is equal to 1 if the condition in brackets holds, and 0 otherwise. We set  $I_{ii} = 1$  since two identical hypotheses must share the same labels. Notably, when  $n = 2$  and the two hypotheses are predicted to be contradicting each other ( $c_1^{1,2} < 0.5$ ), the hypothesis with a higher confidence level in its prediction retains its original prediction, while the hypothesis with a lower confidence level modifies its prediction. Therefore, the joint inference module improves the system’s prediction consistency.

**System Output** We use a threshold-based decision strategy in both tasks to obtain the ultimate predictions. Specifically, for Task A, hypotheses with  $\hat{p}_2^A > \eta_A$  are predicted as “entailment”, while the remaining hypotheses are predicted as “contradic-

tion.” For Task B, the system determines whether the  $i^{th}$  sentence in the premise supports the inference by checking if  $\tilde{p}_i^B > \eta_B$ . In our implementation, we set the threshold values to  $\eta_A = 0.57$  and  $\eta_B = 0.53$ .

## 6 Evaluation

### 6.1 Data, Metrics, and Baselines

We conduct all the evaluations on the NLI4CT dataset. The train, dev, and test sets contain 1,700, 200, and 500 hypotheses. The hypotheses can be divided by the number of CTRs involved: (1) Single, related to only one CTR; (2) Comparison, requiring the comparison between two CTRs for justification. The premise is categorized into four sections that correspond to the CTR sections: (1) Intervention; (2) Eligibility; (3) Results; (4) Adverse Events. Additional dataset details are available in Sec.A in the appendix. Recall, precision, and f1-score are chosen as the evaluation metrics for both Task A and Task B.

We compare our system with several baseline models (BiolinkBERT-Large, SciFive) that achieve SOTA performance on medical NLI tasks. For the BiolinkBERT-Large model, we feed the concatenated hypothesis and premise sequence into the language model and add an MLP that takes the representation of  $[\text{CLS}]$  as input for classification. Note that the performance of baseline models is obtained by ensembling 10 models through cross-validation to make the results comparable. We did not train baseline models for Task B because it is hard to find an existing implementation that can directly be applied to Task B.

### 6.2 Overall Performance

The results of the proposed system on Task A and B are presented in Table 2 and 3, respectively. Our

Model	Dev	Test
M-512-Bi-Bi	<b>0.896</b>	<b>0.829</b>
- token-level	0.878	0.816
- sentence-level	0.857	0.827

Table 4: Ablation study of the multi-granularity encoding in MGNet on Task B. All the models are trained on the train set and evaluated on the dev and test sets. The performance is reported by the f1-score.

system performs best on both tasks with f1-scores of 0.856 and 0.853 for Task A and Task B, respectively. Compared with the baseline models, our system shows a significant improvement of approximately 3% and 5% on the f1-score with and without the joint inference module, respectively, demonstrating the effectiveness of our approach. Note that the proposed multi-granularity encoding method greatly improves the performance of the original BiolinkBERT-Large model on Task A (-SciFive vs. BiolinkBERT-Large) by around 6% in f1-score (0.753 vs. 0.693, 0.844 vs. 0.768).

### 6.3 Ablation Study

We performed ablation experiments to investigate the effect of the proposed joint inference method and the contribution of each model in the soft ensembling. The results are presented in Table 2 and 3. The results show that the joint inference module leads to a significant improvement of 8% points on Task A, suggesting that the system is more likely to make the correct prediction when considering the mutual information between hypotheses, i.e., whether the hypotheses share the same label.

The experimental results suggest that the MGNet, which employs both sentence-level and token-level encoders implemented by BiLSTM (M-512-Bi-Bi/M-512-Bi-Bi-mul), plays a crucial role in both Task A and B, contributing to 1.8% and 1.3% improvements in f1-score, respectively. Furthermore, the individual inference case results reveal that SciFive provides the most significant improvement in the system’s performance, with a 2.3% increase in the f1-score.

To further investigate the impact of multi-granularity encoding in MGNet, we conduct an additional ablation experiment on Task B by removing either the sentence-level or token-level encoder (Task A only relies on the sentence-level encoder). The results in Table 4 show that both sentence-level and token-level encoders enhance the system’s performance, with the sentence-level

Model	Int	Elig	Res	AE
M-1024-Tf-Bi-mul	0.749	0.727	0.735	0.709
SciFive	0.748	0.735	0.712	<b>0.728</b>
M-512-Bi-Bi-mul	0.746	0.739	<b>0.755</b>	0.708
M-512-Tf-Bi-cl	<b>0.783</b>	<b>0.743</b>	0.724	0.709

Table 5: Performance (f1-score) of each model on different sections of CTR in the cross-validation of Task A. Int: Intervention, Elig: Eligibility, Res: Results, AE: Adverse Events.

encoder contributing more to the dev set (4% improvement) and the token-level encoder contributing more to the test set (1.3% improvement). The results indicate the effectiveness of the proposed multi-granularity encoding.

### 6.4 Performance on Different Sections

We further examine the behavior of each model that contributes to the ensembling by evaluating the performance of our system on different sections of CTR. The analysis results are presented in Table 5. It can be observed that MGNets demonstrate superior performance on hypotheses involving the Intervention, Eligibility, and Results sections. In contrast, SciFive achieves a higher f1-score on hypotheses related to the Adverse Events section. Note that hypotheses concerning the Adverse Events section generally require more numerical calculations than those related to other sections. These results demonstrate that SciFive enhances the system’s performance with its numerical inference ability.

## 7 Conclusion

We propose a Multi-granularity system for CTR-based textual entailment and evidence retrieval. The system comprises an ensembling of inference networks and a joint inference module to enhance prediction consistency. A Multi-granularity Inference Network (MGNet) is proposed to handle the two subtasks within a unified network. Another text-generation-based network SciFive is also leveraged to enhance the numerical inference of the system. We use model ensembling to integrate different models’ output and enhance the inference stability, while a joint inference method is further proposed to increase the consistency of inference. Our system achieves the best performance on both subtasks. The ablation experiments and analysis indicate the effectiveness of different parts of the system. In the future, we will try other large language



models and new inference methods (e.g, in-context learning) to inject more medical knowledge into the system and further improve the performance on this task.

## References

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.
- Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Mael Jullien, Marco Valentino, Hannah Frost, Paul O’Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. Results of the seventh edition of the bioasq challenge. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 553–568. Springer.
- Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. [Scifive: a text-to-text transformer model for biomedical literature](#).
- 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. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.
- Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. [LinkBERT: Pretraining language models with document links](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

## A Details of Data

Table 6 shows the basic statistics of NLI4CT. Avg Hypo denotes the average number of words in hypotheses, Avg CTR Sents refers to the average number of sentences in CTRs, and Avg CTR Words is the average number of words in CTRs.

Split	#Hypothesis	Avg Hypo	Avg CTR	
			Sents	Words
Train	1,700	19.7	16.9	165.8
Val	200	18.7	16.2	166.3
Test	500	21.6	14.1	161.7

Table 6: The basic statistics of NLI4CT.

Fig.4 depicts the distribution of the input sequences (concatenation of the hypothesis with premise) regarding the number of tokens in sequences. Most input sequences have less than 500 tokens, while there also exists input sequence longer than 1750 tokens.

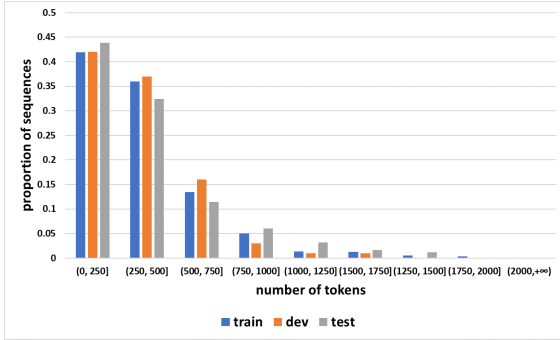


Figure 4: The proportion of input sequences regarding the number of tokens in sequences.

As for the CTR, each CTR contains a total of four sections:<sup>3</sup>

- **Intervention:** Information concerning the type, dosage, frequency, and duration of treatments being studied.
- **Eligibility:** A set of conditions for patients to be allowed to take part in the clinical trial
- **Results:** Number of participants in the trial, outcome measures, units, and the results.
- **Adverse Events:** These are signs and symptoms observed in patients during the clinical trial.

<sup>3</sup>The definition of each section is from <https://sites.google.com/view/nli4ct/home>.

It is worth to note that each hypothesis only involves a single section in CTR. For hypotheses that involve multiple CTRs, one is required to compare the same section of different CTRs.

## B Data Preprocessing

We process the evidence CTRs differently for Task A and B. For hypotheses that require single CTR, we simply concatenate all the sentences in the CTR to form the premise sequence for Task A and B. For hypotheses that require comparison between two CTRs, we concatenate the first and second CTR in a single sequence for Task A with the following format: "primary trial: [CTR 1]. secondary trial: [CTR 2]." Therefore, the joint semantics between two CTRs can be captured by the language model. Nevertheless, we find that encoding each CTR with the hypothesis achieves higher performance for Task B. Therefore, we conduct evidence retrieval on each CTR individually.

## C Details of Training Joint Inference Network

It is hard to train the joint inference network on NLI4CT, because most of the premise corresponds to a pair of hypotheses that contradicts with each other and we can't find positive samples (hypotheses that share the same label). To address this problem, we use the back-translate method to generate positive samples. For each hypothesis  $S$  in the train and dev set, we translate it to Chinese and back translate to English to get a new hypothesis  $S'$ . Then  $S$  and  $S'$  share the same label for Task A. For a pair of conflicting hypotheses  $S_1, S_2$ , we generate four input sequences for the joint inference network:

1.  $[[CLS], S_1, [SEP], S'_1, [SEP], P, [SEP]]$
2.  $[[CLS], S_2, [SEP], S'_2, [SEP], P, [SEP]]$
3.  $[[CLS], S_1, [SEP], S_2, [SEP], P, [SEP]]$
4.  $[[CLS], S_2, [SEP], S_1, [SEP], P, [SEP]]$

After the data generation, we train the joint inference network on the train set and choose the best model on the dev set. We set the learning rate, warmup ratio, and batch size as  $2e-5$ , 0.1 and 32 respectively.