CaresAI at SemEval-2024 Task 2: Improving Natural Language Inference in Clinical Trial Data using Model Ensemble and Data Explanation

Reem Abdel-Salam
Cairo University / Egypt
CaresAI/ Australia
reem.abdelsalam13@gmail.com

Mary Adetutu Adewunmi
University of Tasmania / Australia
CaresAI/ Australia
Mary.Adewunmi@utas.edu.au

Mercy Akinwale
Covenant University / Nigeria
CaresAI/ Australia
mercy.akowepgs@stu.cu.edu.ng

Abstract

Large language models (LLMs) have demonstrated state-of-the-art performance across multiple domains in various natural language tasks. Entailment tasks, however, are more difficult to achieve with a high-performance model. The task is to use safe natural language models to conclude biomedical clinical trial reports (CTRs). The Natural Language Inference for Clinical Trial Data (NLI4CT) task aims to define a given entailment and hypothesis based on CTRs. This paper aims to address the challenges of medical abbreviations and numerical data that can be logically inferred from one another due to acronyms, using different data pre-processing techniques to explain such data. This paper presents a model for NLI4CT SemEval 2024 task 2 that trains the data with DeBERTa, BioLink, BERT, GPT2, BioGPT, and Clinical BERT using the best training approaches, such as fine-tuning, prompt tuning, and contrastive learning. Furthermore, to validate these models, different experiments have been carried out. Our best system is built on an ensemble of different models with different training settings, which achieves an F1 score of 0.77, a faithfulness score of 0.76, and a consistency score of 0.75 and secures the sixth rank in the official leaderboard. In conclusion, this paper has addressed challenges in medical text analysis by exploring various NLP techniques, evaluating multiple advanced natural language models(NLM) models and achieving good results with the ensemble model. Additionally, this project has contributed to the advancement of safe and effective NLMs for analysing complex medical data in CTRs.

1 Introduction

Clinical trials play a crucial role in advancing medical knowledge, evaluating the safety and efficacy of new treatments, and improving patient care (Holford et al., 2010) which are essential for the development of new drugs, therapies, and medical interventions. Most importantly, they involve systematic investigations that aim to answer specific research questions and provide evidence-based guidance for medical decision-making (Tunis et al., 2003). Moreover, clinical trial reports (CTRs) have been published at an accelerated rate due to the rapid development of digital health. Currently, there are more than 10,000 CTRs just for breast cancer (Jullien et al., 2024; Bastian et al., 2010). Also, medical professionals have developed evidence-based clinical diagnoses through the increasing number of Clinical Trial Reports (CTRs) (Bastian et al., 2010), which serve as a broad source of factual and scientific information. Despite these CTRs, drawing valuable conclusions from these reports can be an uphill task due to the different medical domains and the unstructured nature of the report. Recent improvements in natural language processing (NLP) systems, on the other hand, have led to the idea of using multiple language models that have already been trained in the medical field to efficiently carry out medical NLP tasks. The growth of CTRs has also made it possible for a natural language inference (NLI) system to be created that can help with medical interpretation and finding evidence for individualized evidence-based therapy.(Agrawal et al., 2022) used InstructGPT with zero-shot and few-shot settings to extract information from clinical text. In addition, the authors introduced new datasets for benchmarking for few-shot clinical information extraction. The work in (Molinet et al., 2022) introduced a new tool, the ACTA automated tool, to support evidence-based clinical decision-making. The authors in (Yasunaga et al., 2022) proposed a new model, LinkBERT, that incorporates document link knowledge for medical domains. Despite substantial research on the use of advanced NLP approaches in the medical domain, evaluation benchmarks remain inadequate.

The SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1905–1911

(NLI4CT) task is proposed by (Jullien et al., 2024) by building an efficient evaluation benchmark using a set of statements, explanations, and CTRs for breast cancer. This task is an extension of the previous year's shared task Multi-evidence Natural Language Inference for Clinical Trials. The purpose of the NLI4CT task is to entail a statement based on one or multiple clinical reports. NLI4CT is challenging because hypothesis verification sometimes requires integrating multiple pieces of data from the premise. In some instances, validating a hypothesis necessitates a comparison of two distinct premise CTRs. Validating hypotheses based on each premise type demands varying levels of inference skills (textual, numerical, etc.).

This paper presents work done in the NLI4CT to address these challenges owing to the complexity of the medical domain and text structure. The objective of this task is to develop a system capable of deducing conclusions or implications about various CTRs. The system consists of an ensemble of different experiments using different training approaches. The rest of the papers go as follows: section 3 discusses the proposed methods, section 4 shows experimental results, and section 5 concludes the paper.

2 Background

The main goal of the task is to determine the validity of a claim (hypothesis) based on a single section from one or multiple clinical trial reports (CTRs) of breast cancer (premises). There are two possible inferential relations for each statement: entailment and contradiction. The dataset[1] used is provided by the task organizers and it is divided into two parts: The first part is derived from a compilation of CTRs, which is categorized into four sections.: a) eligibility criteria required for participation in the clinical trial; b) intervention detailing the treatment type, dosage, frequency, and duration; c) results showing participant numbers, outcome measures, units, and findings; and d) adverse effects observed in patients during the trial. The second part compromises the claim about the information contained in a single section, either in one or two CTRs and information about which CTRs are targeted and which section. The dataset consists of 1700 training samples and 200 validation samples. The dataset supplied has an equal distribution of labels.

[1] https://github.com/ai-systems/Task-2-SemEval-2024

3 System overview

This section presents the various approaches used while developing the final models. This includes techniques for preprocessing and ingesting data. Moreover, it includes the paradigms used for training as well as the experimental setup. DeBERTa and Bert-based models are fine-tuned using a weighted ensemble of refined iterations, as well as prompt-based fine-tuning (Lester et al., 2021) for DeBERTa final models.

3.1 Data Preparation

Large Models (LM) have challenges doing reasoning tasks when presented with acronyms, numerical, and quantitative information. Large language models (LLM) can be used to answer numerical entities and improve performance through a chain of thought, which is a methodical process of reasoning and explaining things step by step (Wang et al., 2023a). We have implemented a comparable pre-processing procedure for converting numeric quantities into textual representations and providing explanations in the presence of mathematical formulas. Furthermore, abbreviations are expanded to their original full names. The process goes as follows: first numbers, operations and abbreviations are converted into textual representation. For example **"Arm/Group Description: Phase 1 Cyclophosphamide Dose Escalation: Fludarabine 5x25mg/m2 + Cyclophosphamide 2x60mg/kg + Cells intravenous (IV) "** is converted to **"Arm/Group Description: Phase one Cyclophosphamide Dose Escalation: Fludarabine five times twenty-five milligram/ meterstwo plus Cyclophosphamide two times sixty milligram/ kilograms plus Cells intravenous (IV)."** The next steps involve using LLM as chatGPT or Mistral 7b to give an explanation for the converted text. The following prompt is feed to the LLM ***" Explain the following sentences in five sentences: text".*** . ChatGPT provides the following explanation: " **The given sentence describes a specific group or category within a clinical trial referred to as "Arm/Group Description" during the phase one stage. This particular arm involves a treatment protocol called "Cyclophosphamide Dose Escalation," which aims to determine the optimal and safe dosage of Cyclophosphamide, a chemotherapy drug. The treatment**

regimen includes the intravenous administration (IV) of Fludarabine at a dose of twenty-five milligrams per square meter of body surface area, repeated five times over a specific time period or schedule. Additionally, Cyclophosphamide is administered intravenously (IV) at a dose of sixty milligrams per kilogram of body weight, repeated twice over the same time period or schedule. The treatment also involves the administration of cells intravenously (IV), although the specific type of cells is not mentioned in the provided sentence.". Finally, Both Converted text and explanation are concatenated.

3.2 Data Representation

In types of reasoning tasks such as hypothesis and premise reasoning, the intuitive step is to feed both texts separated by a model-separating token ***("hypothesis [SEP] premise")***. However, it is not yet clear if it is an optimal representation or not. To address this, different representations were used to assess model performance and sensitivity to data representations.

The following illustrates several data representations used. The challenge included two sorts of reasoning. The first type of reasoning is to check whether a given premise is correct based on one single report. The second type is to check whether a given premise is correct based on two different reports. Therefore the model was two different representations were designed for each type.

- The first common design is to feed in each premise section separated by a separator token, then the hypothesis section.
  - ***"First premise [SEP] Second premise [SEP] hypothesis".***
- The second design was adding special token information to indicate the following sections:
  - **token_first** for the first premise
  - **token_second** for the second premise
  - **token_hypothesis** for the hypothesis.
  - ***"token_first_section First premise [SEP] token_second_section Second premise [SEP] token_hypothesis".***
- The third design was inverting order first feed hypothesis followed by premise.
- The remaining design explored the impact of adding different prompts to encourage model correct classification to each sentence and understanding of the current problem.
  - ***"First premise [SEP] Second premise [SEP] Is this statement correct based on previous CTR reports: hypothesis?".***
  - ***"First premise [SEP] Second premise [SEP] Question: Does this imply that: hypothesis?".***
  - ***"Task: Determine Claim Validity \n \n n CTR Report \n First premise [SEP] CTR Report \n Second premise [SEP] Evaluate the Claim: \n hypothesis".***

Also, since the organizers offered the specific lines that contributed to reasoning in a given section presented in both training and validation data, another crucial data-feeding option is whether to feed an entire section for the premise or choose selected lines from a premise section. Some models were trained on the whole section, while others were trained on chosen premise lines.

3.3 Model Selection, Design and training

Based on the following papers results (Wang et al., 2023b; Kanakarajan and Sankarasubbu, 2023; Zhou et al., 2023), experiments were conducted with a variety of different models, including 1) GPT2 (Lagler et al., 2013) 2) Deberta large (He et al., 2020) 3) BioLinkBERT (Yasunaga et al., 2022) 4) Clinical BERT (Alsentzer et al., 2019) 5) Scifive (Phan et al., 2021) 6) BioGPT (Luo et al., 2022).

3.3.1 Model architecture

It is important to modify the model architecture by deciding whether to simply use the last layer and input them into the Fully Connected (FC) layer, or to use the last n-layers from the model and implement average pooling before feeding them to the FC layer, or to direct the output to a convolutional or LSTM layer followed by the FC layer. Experiments were conducted with two alternative options. The first option is to apply mean pooling to the last layer of the model, while the second option is to use GeM pooling on the same layer.

3.3.2 Model Training

BioLinkBERT, Clinical BERT, BioGPT, DeBERTa-large and GPT2 models: Several training approaches have been investigated to improve the generalizability of the model and

its performance. The first approach involves fine-tuning the whole model while using cross-entropy loss. The second approach involves fine-tuning the whole model while using two losses. To improve model performance. The first loss is a cross-entropy loss so penalize the model for wrong prediction; the second loss is contrastive (Chen et al., 2020). The reason behind it is to improve model representation for both classes in the embedding space. The following weights were used: 0.7 for cross-entropy loss and 0.3 for contrastive loss. Following recent practices from the literature, parameter-efficient tuning methodologies as prompt-tuning, LoRA, have been shown to improve model performance over conventional fine-tuning (Fu et al., 2023; Ding et al., 2023). Therefore, the third approach leverages prompt-fine-tuning (Lester et al., 2021) for LM. In prompt-fine-tuning, the data is fed with a prompt to encourage the model to understand the task well, as well as the **"[MASK]"** token. The model task is to predict the correct class in the **"[MASK]"** token. The challenge in prompting lies in the design of the prompt and the model's output. The prompt we used was: ***"First premise [SEP]. Can we infer the hypothesis from the text above? [MASK]".*** The model's output is a binary prediction of either **"yes"** or **"no."**.

Scifive model training: Scifive is based on the T5 (Raffel et al., 2020) generator type, which is an encoder-decoder that transforms all tasks into text-to-text. Instruction fine-tuning has been conducted on the Scifive model with the following template: ***"Determine Claim Validity \n \n. First premise \n \n. Second premise \n\n. Evaluate the following Claim: hypothesis \n \n. Is the assertion accurate? Options: [yes, no]."*** For the loss function of the model BLEU score have been used. The model was constrained to predict either **"valid/invalid"**, or **"correct/incorrect"**, or **"yes/no"**.

3.3.3 Experimental setup

Table 2 shows the hyperparameter setup used during training.

4 Results

In this section, the performance of the proposed models is reported based on the official metric during the dev-phase and test-phase. Error analysis (Lu et al., 2023) was conducted to identify the weaknesses of the proposed models. For the task, the official metric is based on the F1 score and the average faithfulness [2] and consistency [3] scores.

4.1 Dev-phase results

Table 3 shows the results of the developed models on the dev-set, with their training settings. Clearly, the DeBERTa model with different settings showed superior performance compared to other models such as BioGPT, BioLinkBert, ClinicalBERT, GPT2 and Scifive. The Scifive model showed huge performance degradation when compared to BioLinkBert.

The first observation is that changing the pooling technique from mean pooling to GeM pooling, improved model performance by a magnitude of 3%. The second observation is that having two loss function contrastive loss with cross entropy loss improved performance by a magnitude of 3%. The third observation is that building two models for the different cases of reasoning (case single premise, hypothesis, and case of two premises and hypothesis) and including a task description in the data fed improved model performance by 2%. The fourth observation is that prompt-based fine-tuning is better than conventional fine-tuning by magnitude of 1-2%. Another key observation during training is that the model scores a similar f1-score for both classes in most of the settings. The fifth observation is that having data processing as converting numerical quantities to textual representation along with an explanation improves model performance over conventional ones by a magnitude of 1-2%.

4.2 Test-phase results

The results of the proposed system are presented in table 1. Our system ranks in sixth place, with a 0.77 F1-Score, a 0.76 Faithfulness score, and a 0.75 Consistency score. There are correlations between the dev-phase f1-score and the test-phase f1-score, which suggests that a greedy approach to choosing models and their weights is a good approach.

5 Conclusion

The study tested different ways to prepare and load data, as well as more advanced NlM models. It came to the conclusion that the ensemble

[2] Faithfulness measures the extent to which a given system arrives at the correct prediction for the correct reason (Li et al., 2022)

[3] Consistency is a measure of the extent to which a given system produces the same outputs for semantically equivalent problems (Fan et al., 2023)

Combination of selected Models	Leaderboard Results F1-score/ Consistency/Faithfulness	Dev f1-score
BioLinkBert DeBERTa (model 5,8,11,6,10 from table 3)	0.75/0.75/0.79	**0.8945**
DeBERTa (model 8,12,6,9,10 from table 3)	0.743/0.75/0.76	0.8899
DeBERTa (model 8,6,9,11, 10 from table 3)	0.754/0.74/0.75	0.88497
DeBERTa (model 5,8,12,6,9 from table 3)	0.744/0.76/0.80	0.88492
DeBERTa (model 8,6,9,10 from table 3)	**0.765/0.76/0.75**	0.8799
DeBERTa (model 8,11,6,10 from table 3)	0.73/0.75/0.75	0.8749
BioLinkBert DeBERTa (model 8,9,6 from table 3)	0.744/0.75/0.75	0.87474
DeBERTa (model 8,6,9 from table 3)	0.742/0.74/0.78	0.8746

Table 1: Performance of the submitted models on the leaderboard

Hyperparameter	Value
Learning-rate	4e-5 or 5e-6
Schedular	cosine-annealing
Weight decay	1e-3
Epochs	30
Optimizer	Adam
Metric	F1-macro on dev-set

Table 2: The full hyperparameter search space.

model worked well for medical text analysis. DeBERTA, ClinicalBERT, GPT2, Scifive, BioGPT, and BioLinkBert have been investigated and the results show that the DEBERTA model showed better performance compared to other models during the training phase. The final model submitted was an ensemble of various models and techniques. The best-performing model achieved an F1 score of 0.77, a faithfulness score of 0.76, and a consistency score of 0.75, securing the sixth rank in the official leaderboard. Overall, this study has enhanced safe and effective NLMs for complicated medical data analysis in clinical trial reports. Future recommendations could explore other large language models and training techniques, such as LoRA and prefix-tuning, for ingesting medical knowledge into CTRs.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. 2023. Large language models for software engineering: Survey and open problems. *arXiv preprint arXiv:2310.03533*.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12799–12807.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

N Holford, SC Ma, and BA Ploeger. 2010. Clinical trial simulation: a review. *Clinical Pharmacology & Therapeutics*, 88(2):166–182.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of*

Model	Training Paradigm	Data Ingestion Prompt	F1-Score
BioGPT	Architecture: Mean Pooling Loss function: Cross Entropy	"premise [SEP] hypothesis".	50
BioLinkBert	Architecture: Mean Pooling Loss function: Cross Entropy	"token_special hypothesis [SEP] token_special premise".	69.7
ClinicalBERT	Architecture: Mean Pooling Loss function: Cross Entropy	"token_special hypothesis [SEP] token_special premise".	63.5
ClinicalBERT	Architecture: Mean Pooling Loss function: Cross Entropy	"premise [SEP] hypothesis".	50
DeBERTa	Architecture: Mean Pooling Loss function: Cross Entropy	"token_special hypothesis [SEP] token_special premise".	80
DeBERTa	Architecture: Mean Pooling Loss function: Cross Entropy	Comparison type [SEP] token_special premise [SEP] premise".	77
DeBERTa	Architecture: Mean Pooling Loss function: Cross Entropy	" premise [SEP] hypothesis".	80
DeBERTa	Architecture: Mean Pooling Loss function: Cross-Entropy and Contrastive Learning	"premise [SEP] hypothesis".	**83**
DeBERTa	Architecture: GeM Pooling Loss function: Cross-Entropy Data preparation: Converted numeric values and abbreviation Two separate models for each comparison type	" premise [SEP] Is this statement correct based on previous CTR reports: hypothesis? ".	82
DeBERTa	Architecture: GeM Pooling Loss function: Cross-Entropy Data preparation: Converted numeric values and abbreviation	" premise [SEP] hypothesis".	82
DeBERTa	Prompting	" premise [SEP] Based on the paragraph above can we conclude that: hypothesis? [MASK] "	81
DeBERTa	Architecture: GeM Pooling Loss function: Cross-Entropy	" premise [SEP] hypothesis".	**83**
GPT-2	Architecture: GeM Pooling Loss function: Cross-Entropy	"premise [SEP] hypothesis".	60
Scifive		"premise [SEP] Question: Does this imply that: hypothesis? ".	50
Scifive		"Task: Determine Claim Validity\n\n CTR Report \n premise [SEP] premise [SEP] f'Evaluate the Claim:\n hypothesis. Options: [correct, incorrect] ".	63.9
Scifive		"Task: Determine Claim Validity\n\n CTR Report \ premise [SEP] f'Evaluate the Claim:\n hypothesis. Options: [valid, invalid] ".	63.73
Scifive		"Determine if a claim is correct based on the following reports.\n Report 1: premise. \n Claim: hypothesis Is the claim correct? \n Options: [yes, no]"	50

Table 3: Models and techniques developed during the experimental and F1-score based on dev-set.

the 18th International Workshop on Semantic Evaluation (SemEval-2024). Association for Computational Linguistics.

Kamal Raj Kanakarajan and Malaikannan Sankarasubbu. 2023. Saama ai research at semeval-2023 task 7: Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Benjamin Molinet, Santiago Marro, Elena Cabrio, Serena Villata, and Tobias Mayer. 2022. Acta 2.0: A modular architecture for multi-layer argumentative analysis of clinical trials. In *IJCAI 2022-Thirty-First International Joint Conference on Artificial Intelligence*.

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. *arXiv preprint arXiv:2106.03598*.

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.

Sean R Tunis, Daniel B Stryer, and Carolyn M Clancy. 2003. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Jama*, 290(12):1624–1632.

Chaojie Wang, Yishi Xu, Zhong Peng, Chenxi Zhang, Bo Chen, Xinrun Wang, Lei Feng, and Bo An. 2023a. keqing: knowledge-based question answering is a nature chain-of-thought mentor of llm. *arXiv preprint arXiv:2401.00426*.

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023b. Knowcomp at semeval-2023 task 7: Fine-tuning pre-trained language models for clinical trial entailment identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1–9.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Yuxuan Zhou, Ziyu Jin, Meiwei Li, Miao Li, Xien Liu, Xinxin You, and Ji Wu. 2023. Thifly research at semeval-2023 task 7: A multi-granularity system for ctr-based textual entailment and evidence retrieval. *arXiv preprint arXiv:2306.01245*.