KnowComp at DialAM-2024: Fine-tuning Pre-trained Language Models for Dialogical Argument Mining with Inference Anchoring Theory

Yuetong Wu*, **Yukai Zhou***, **Baixuan Xu, Weiqi Wang, Yangqiu Song**
Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
{ywufe, yzhoueg, bxuan}@connect.ust.hk

Abstract

In this paper, we present our framework for DialAM-2024 Task A: Identification of Propositional Relations and Task B: Identification of Illocutionary Relations. The goal of Task A is to detect argumentative relations between propositions in an argumentative dialogue (Inference, Conflict, Rephrase), while Task B while Task B aims to detect illocutionary relations between locutions and argumentative propositions in a dialogue, e.g.,, Asserting, Agreeing, Arguing, Disagreeing. Noticing the definition of the relations are strict and professional under the context of IAT framework, we meticulously curate prompts which not only incorporate formal definition of the relations, but also exhibit the subtle differences between them. The PTLMs are then fine-tuned on the human-designed prompts to enhance its discrimination capability in classifying different theoretical relations by learning from the human instruction and the ground truth samples. After extensive experiments, a fine-tuned DeBERTa-v3-base model exhibits the best performance among all PTLMs with an F1 score of 78.90% on Task B. It is worth noticing that our framework ranks #2 in the ILO - General official leaderboard.

1 Introduction

Dialogical argument mining is an emerging field that aims to bridge the gap between the analysis of argumentation and dialogue (Budzynska et al., 2014b; Ruiz-Dolz et al., 2024; Kawarada et al., 2024). Traditional argument mining approaches have often focused on opinion mining within monological texts (Lawrence and Reed, 2019; Arugam, 2022) or document form contents (Ruosch et al., 2022; Sazid and Mercer, 2022; Khondoker and Yousuf, 2022). However, real-world argumentation frequently occurs in dialogical contexts, where multiple participants engage in a dynamic exchange of viewpoints (Feger and Dietze, 2024; Lai et al., 2024; Alsinet et al., 2022). This complexity necessitates a more holistic approach that considers both the argumentative structures and the dialogical interactions.

*Equal Contributions

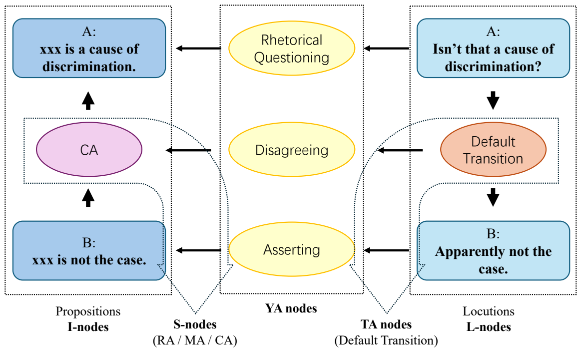

Figure 1: Inference Anchoring dialogical map example.

Apart from the dialogical information extraction paradigms explored by previous works (Dutta et al., 2022; Mestre et al., 2021), A generic modelling formalism for extracting dialogical information is the Inference Anchoring Theory (IAT) introduced by Budzynska and Reed (2011). It offers a systematic approach to decomposing text speech into distinct units (ADUs), while also anchoring and categorising logical inferences between propositions and locutions. As such, IAT provides a comprehensive methodology for analyzing the maneuvers of dialogues within a given theoretical framework, thus building an explicit scaffolding for language models to handle semantics analysis tasks (Budzynska et al., 2014a).

Based on this theory, DialAM-2024 workshop (Ruiz-Dolz et al., 2024) introduces the first shared task in dialogical argument mining, aimed at modeling argumentation and dialogue information together within a domain-independent framework. The proposed tasks of DialAM-2024 involves classification of the three-way argumentative relations between locutions and corresponding propositions,

Proceedings of the 11th Workshop on Argument Mining, pages 103–109
August 15, 2024 ©2024 Association for Computational Linguistics

detection of relevant dialogical components and completion of the inference anchoring map.

Due to the in-context learning ability of LLMs on unconventional tasks with demonstrated examples (Sun et al., 2023), our initial attempt was to use Large Language Models (LLMs) as the classifier for illocutionary relations (Chan et al., 2024; Wang et al., 2023b,a, 2024a,b; Wang and Song, 2024). A combination of zero-shot and few-shot (Brown et al., 2020) prompts integrated with Chain-of-thought (Wei et al., 2022) were tested. However, we observed that popular LLMs, such as gpt3.5-turbo (OpenAI, 2023), fail to show significant understanding of the task and yield relatively low performance after exhaustive experiment.

Notably, recent developments in Pre-Trained Language Models (PTLMs) on text classification tasks (Howard and Ruder, 2018) have empowered us to build our system the other way round. After the compilation of paired ADUs of propositions and locutions nodes embedded in a meticulously designed textual prompt, we fine-tuned our PTLMs on the reconstituted dataset as that of a traditional text classification task (Wang et al., 2023c; Peng et al., 2024; Yan et al., 2024). Using this method, we were able to achieve relatively high accuracy in the identification of illocutionary relations. The classification results of Task B were then used as textual information to assist the identification of propositional relations.

An extensive ablation study was also conducted to test the effectiveness and generalizability of our proposed system. A maximal F1 score of 78.90% and precision of 82.35% on Task B was achieved using a fine-tuned DeBERTa-v3-base model (Howard and Ruder, 2018). It is also noted that DeBERTa-v3-large underperforms its base version, with a precision difference of -0.2%. The proposed explanation is that the model already converges on the given dataset, provided the base version parameters. Several other PTLMs, including RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020) are also trialed using identical prompt design, which yield results inferior to DeBERTa-v3.

However, fine-tuned PTLMs converges inconsistently for Task A, with a recall of only 33.79%. We suspect that besides text from adjacent propositions and locutions, the system will need more in-context information (e.g., a dialogue 2-3 nodes away) to assist the process of relation identification according to recent works on reasoning under contexts (Dong et al., 2024; Zhang et al., 2024; Li et al., 2024).

As such, our proposed system provides valuable insight for dialogical argument mining using PTLMs on a IAT layout, and future works should be more focused on the revamp of methodology in in-context training information extraction. Our code and results are publicaly available at Arwenwutietie/DialAM-2024

2 Problem Definition

In this section, we would introduce the dataset format and elaborate on the formal definition of the shared task in DialAM-2024.

2.1 Dataset Description

In the DialAM-2024 dataset, all input texts are categorized into two primary types: locutions (L-nodes) and propositions (I-nodes). Locutions represent the original sentence segments within a complete dialogue, typically featuring speakers and timestamps. Conversely, propositions are reconstructed locutions, where linguistic elements such as anaphora, pronouns, and deixis have been resolved. These two text types are then structured into a navigable graph based on IAT, with corresponding L-nodes and I-nodes connected by three distinct relation types: (i) relations between locutions in a dialogue, known as transitions (TA-nodes); (ii) relations between propositions and locutions (YA-nodes); and (iii) illocutionary connections that link locutions with their semantic content (S-nodes).

We use QT30 corpus (Hautli-Janisz et al., 2022) as our dataset. QT30 is a collection of 30 episodes of Question Time aired between June 2020 and November 2021, with a total of more than 29 hours of transcribed broadcast material and comprising 19,842 locutions by more than 400 participants. The QT30 dataset contains 10,818 propositional relations that include Default Inferences, Default Conflicts, and Default Rephrases, and 32,303 illocutionary relations divided into Asserting, Agreeing, Arguing, Disagreeing, Restating, Questioning, and Default Illocuting.

2.2 Task Definition

The DialAM-2024 challenge comprises two distinct sub-tasks. Task A aims to detect the argumentative relations that exist between the propositions identified and segmented within the argumentative dialogue. More specifically, the objective is to use two connected I-nodes to predict the S-nodes between them. Task B, on the other hand, seeks to

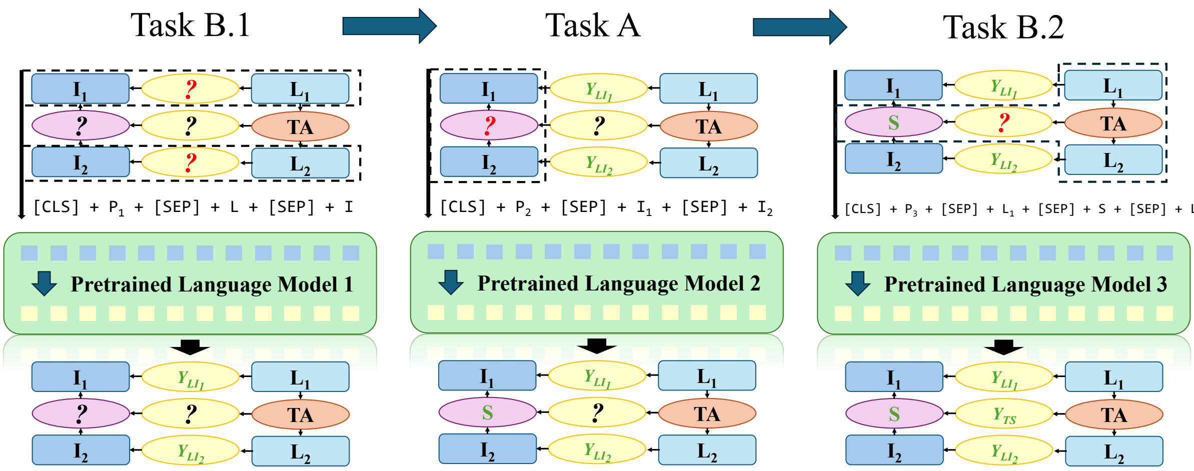

Figure 2: PTLM pipeline for DialAM-2024 dialogical argument mining tasks. Three PTLMs are fine-tuned in sequence to cope with Task B.1, Task A and Task B.2 respectively. The IAT map structure is optimally utilized for propositional & illocutionary relation classification.

identify the illocutionary relations that exist between the locutions uttered in the dialogue and the argumentative propositions associated with them. In other words, given a set of locutions (L-nodes) and propositions (I-nodes), the goal is to uncover the Illocutionary connections (YA-nodes) that link them.

To allow us to establish a clear and formal framework for analyzing the relationships. Formally, let us denote two coherent locutions as L_1 and L_2, their corresponding propositions as I_1 and I_2, the intermediate TA-nodes between L_1 and L_2 as T, the YA-nodes connecting L_1 and I_1 as Y_{LI_1}, the YA-nodes connecting L_2 and I_2 as Y_{LI_2}, the intermediary S-nodes between I_1 and I_2 as S, and the YA-nodes connecting T and S as Y_{TS}. We denote LLMs as F and the curated prompt as P_1, P_2 respectively for Task A and Task B. By these notations, the Task A and Task B could be reformatted formally as:

$$\textbf{Task A:}\;\; S = \max_i F(S_i|I_1, I_2, P_1);$$

$$\textbf{Task B:}\;\; Y_{LI_1} = \max_i F(Y_{LI_i}|I_1, L_1, P_2),$$

where S_i and Y_{LI_1} denote the output of PTLMs.

3 System Overview

In this section, we will introduce our proposed system. Our method conducts sequential inferences where we predict Y_{LI_1}, Y_{LI_2} and S in the first stage, then infer Y_{TS} with the predicted S in the previous stage.

3.1 Prompt Design

With the rapid advancement exhibited in prompt engineering technique (Chang et al., 2024; Qiao et al., 2023; Xu et al., 2024) it has been pointed out that prompting makes better use of the pre-trained data of PTLMs, allowing the model to perform better on fewer training examples, which can be helpful when classifying classes with smaller examples in this task. Being aware of this, since this text classification task is highly specified and targeted, we meticulously curated descriptive prompting for both sub-tasks. The prompt is then aggregated with given texts as the inputs for large model. Predefined special tokens like [SEP], [CLS] and [EOS] are also added to the final input texts to assist the model to understand the relationship between the different parts of the input. Totally, three different prompts have been used for Task A and B: P_1 (prompt used to predict Y_{LI_1} and Y_{LI_2}), P_2 (prompt used to predict S) and P_3 (prompt used to predict Y_{TS}).

3.2 Sequential inference and model training

Recently, decomposing complex problems into several simple one has become a fashion in LLM reasoning field (Bueno et al., 2024; Besta et al., 2024). Following this trend, in this project, the training of PTLMs is divided into three sequential stages, as shown in figure 2.

3.2.1 Stage 1: Direct Illocutionary Relation Detection (Task B.1)

In Stage 1, we instruct the model to predict Y_{LI_1} and Y_{LI_2} separately, since the illocutionary rela-

Model/Epoch	1-epoch			2-epoch		
	Y_{LI_1}+Y_{LI_2}	Y_{TS}	S	Y_{LI_1}+Y_{LI_2}	Y_{TS}	S
$DeBERTa_{base}$	0.9423	0.6137	0.5198	0.9450	0.6486	0.5676
$DeBERTa_{large}$	0.9428	0.6056	0.513	0.9359	0.6322	0.5681
RoBERTa	0.901	0.5481	0.4388	0.9234	0.5745	0.4503
ALBERT	0.8906	0.5364	0.4637	0.8906	0.5891	0.498
ChatGPT	0.72	-	-	0.72	-	-

Table 1: The experiment result for three stage inference. The result is evaluated on the validation set manually seperated by the author to demonstrate the model performance comparison.

tions between L-nodes (L_1 and L_2) and I-nodes (I_1 and I_2) is more intuitive and requires less information to classify. The raw textual prompt used is $('[CLS]' + P_1 +' [SEP]' + L_1 +' [SEP]' + I_1)$ and $('[CLS]' + P_1 +' [SEP]' + L_2 +' [SEP]' + I_2)$.

3.2.2 Stage 2: Propositional Relation Detection (Task A)

Then, in Stage 2 we subsequently classify S-nodes with textual prompt $('[CLS]' + P_2 +' [SEP]' + I_1 +' [SEP]' + I_2)$.

3.2.3 Stage 3: Indirect Illocutionary Relation Detection (Task B.2)

Finally, motivated by our observation that S and Y_{TS} are highly related, we incorporate the information yield through the previous two stages. Specifically, we leverage L_1, L_2 and the already predicted S for the prediction of Y_{TS}. The prompt we used is $('[CLS]' + P_3 +' [SEP]' + L_1 +' [SEP]' + S +' [SEP]' + L_2 +' [SEP]')$.

3.2.4 Training Objective

All models are trained with cross-entropy loss. Denote each input as x_i, its token length as $|x_i|$. Our models are denoted by p, and thus $p(x_i)$ represents the prediction made by the corresponding node, with $q(x_i)$ as its true label.

$$L(x_i, q) = -\sum_{i=1}^{|x|} p(x_i) \log(q(x_i)) \qquad (1)$$

4 Experimental Setup

We followed a standard approach to partition our input data into training and validation sets. Please refer to Appendix C for more details.

5 Results and Analysis

In this section, we demonstrate our experiment results and conduct analysis on the issue we encountered through the experiments.

Our overall result is shown in Table 1. From the data we can observe that both DeBERTa-base and DeBERTa-large can achieve a relatively high accuracy on the prediction of Y_{LI_1}+Y_{LI_2}, Y_{TS} and S. However, ChatGPT's results were clearly not satisfactory, and it achieved the lowest accuracy rate on all 3 tasks. The reason could be that this text classification task is highly specialized and targeted where related resources rarely occur in ChatGPT's training data. Consequently, ChatGPT would fall short in relevant reasoning tasks. In the classification of Y_{LI_1}+Y_{LI_2}, we realize that the most numerous type in Y_{LI_1}+Y_{LI_2}, Asserting, accounts for 90% of the total number of Y_{LI_1}+Y_{LI_2}. We suspect that this may affect the final performance of the model, making it more inclined to split a new Y_{LI_1}orY_{LI_2}node into the Asserting class. Based on this, we tried to reduce the number of Asserting classes in the training set to train a more comprehensive model. However, the final results demonstrated that this actually led to a decrease in the overall accuracy. This implies that the model is scarcely affected by the imbalance of the dataset.

Further experiments indicate that the accuracy of S-node classification is greatly affected by the size of the training set. According to our observation, when 60% of the data is sampled for training, the accuracy on the test set reaches the highest (65.73%), and when all data is used for training, the accuracy decreases to 56.76%. We suspect that this may be due to model's overfitting to the training data.

6 Conclusion

In this paper, we present our system for the DialAM-2024 dialogical argument mining task, focusing on the identification of propositional and illocutionary relations within dialogues. By leveraging the IAT framework, we developed a methodology that integrates human-defined prompts to stimulate PTLMs' reasoning. Our approach features commendable results in the identification of illocutionary relations with concise preprocessing procedures, as evidenced by our high F1 score and precision in Task B. Despite the notable success in Task B, our system encountered challenges in Task A, particularly in achieving consistent recall rates. This indicates that additional context beyond adjacent propositions and locutions may be necessary for enhancing the identification of argumentative relations. Our findings contribute valuable insights into the application of PTLMs in dialogical argument mining. The results underscore the importance of designing effective prompts and highlight the need for ongoing methodological advancements to fully harness the capabilities of PTLMs in complex argumentation analysis tasks.

Ethics Consideration

The authors believe that this paper does not yield additional ethics concerns. All models and datasets accessed are freely accessible for research purposes.

References

Teresa Alsinet, Josep Argelich, Ramón Béjar, Daniel Gibert, and Jordi Planes. 2022. Argumentation reasoning with graph isomorphism networks for reddit conversation analysis. *Int. J. Comput. Intell. Syst.*, 15(1):86.

S. S. Arumugam. 2022. Development of argument based opinion mining model with sentimental data analysis from twitter content. *Concurr. Comput. Pract. Exp.*, 34(15).

Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwasniewski, Jürgen Müller, Lukas Gianinazzi, Ales Kubicek, Hubert Niewiadomski, Onur Mutlu, and Torsten Hoefler. 2024. Topologies of reasoning: Demystifying chains, trees, and graphs of thoughts. *CoRR*, abs/2401.14295.

Tom Brown, Benjamin F Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Thomas Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey C.S. Wu, Clemens Winter, Christopher Hesse, Mark I-Cheng Chen, Eric J Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack A Clark, Christopher Berner, Samuel McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv (Cornell University)*.

Kasia Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yakorska. 2014a. A model for processing illocutionary structures and argumentation in debates.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014b. Towards argument mining from dialogue. *HAL (Le Centre pour la Communication Scientifique Directe)*.

Katarzyna Budzynska and Chris Reed. 2011. Speech acts of argumentation: Inference anchors and peripheral cues in dialogue.

Mirelle Bueno, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2024. Lissard: Long and simple sequential reasoning datasets. *CoRR*, abs/2402.07859.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *CoRR*, abs/2404.01077.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning.

Subhabrata Dutta, Jeevesh Juneja, Dipankar Das, and Tanmoy Chakraborty. 2022. Can unsupervised knowledge transfer from social discussions help argument mining? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7774–7786. Association for Computational Linguistics.

Marc Feger and Stefan Dietze. 2024. TACO - twitter arguments from conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15522–15529. ELRA and ICCL.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument mining as a text-to-text generation task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2002–2014. Association for Computational Linguistics.

Md Yasin Arafat Khondoker and Mohammad Abu Yousuf. 2022. Argument mining on clinical trial abstracts on lung cancer patients. In *TCCE*, pages 49–60.

Viet Dac Lai, Duy Ngoc Pham, Jonathan Steinberg, Jamie Mikeska, and Thien Huu Nguyen. 2024. CAMAL: A novel dataset for multi-label conversational argument move analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2673–2682. ELRA and ICCL.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *arXiv:1909.11942 [cs]*.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45:1–55.

Yun Li, Zhe Liu, Hang Chen, and Lina Yao. 2024. Context-based and diversity-driven specificity in compositional zero-shot learning. *CoRR*, abs/2402.17251.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Rafael Mestre, Razvan Milicin, Stuart Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining, ArgMining@EMNLP 2021, Punta Cana, Dominican Republic, November 10-11, 2021*, pages 78–88. Association for Computational Linguistics.

OpenAI. 2023. Gpt-3.5 turbo.

Yinbin Peng, Wei Wu, Jiansi Ren, and Xiang Yu. 2024. Novel GCN model using dense connection and attention mechanism for text classification. *Neural Process. Lett.*, 56(2):144.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5368–5393. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Chr-Jr Chiu, Chung-Chi Chen, Noriko Kando, and Hsin-Hsi Chen. 2024. Learning strategies for robust argument mining: An analysis of variations in language and domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10286–10292. ELRA and ICCL.

Ramon Ruiz-Dolz, John Lawrence, Schad Schad, and Chris Reed. 2024. Overview of DialAM-2024: Argument Mining in Natural Language Dialogues. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand.

Florian Ruosch, Cristina Sarasua, and Abraham Bernstein. 2022. BAM: benchmarking argument mining on scientific documents. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 36th AAAI Conference on Artificial Inteligence, SDU@AAAI 2022, Virtual Event, March 1, 2022*, volume 3164 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Muhammad Tawsif Sazid and Robert E. Mercer. 2022. A unified representation and a decoupled deep learning architecture for argumentation mining of students' persuasive essays. In *Proceedings of the 9th Workshop on Argument Mining, ArgMining@COLING 2022, Online and in Gyeongju, Republic of Korea, October 12 - 17, 2022*, pages 74–83. International Conference on Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024a. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Haochen Shi, Baixuan Xu, Wenxuan Ding, Liyu Zhang, Wei Fan, Jiaxin Bai, Haoran Li, Xin Liu, et al. 2024b. On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions. *arXiv preprint arXiv:2406.10885*.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Weiqi Wang and Yangqiu Song. 2024. Mars: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *arXiv preprint arXiv:2406.02106*.

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023c. Knowcomp at semeval-2023 task 7: Fine-tuning pre-trained language models for clinical trial entailment identification. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 1–9. Association for Computational Linguistics.

Jason Zhanshun Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Long Chen, and Yangqiu Song. 2024. Mind: Multimodal shopping intention distillation from large vision-language models for e-commerce purchase understanding.

Xueming Yan, Han Huang, Yaochu Jin, Liang Chen, Zhanning Liang, and Zhifeng Hao. 2024. Neural architecture search via multi-hashing embedding and graph tensor networks for multilingual text classification. *IEEE Trans. Emerg. Top. Comput. Intell.*, 8(1):350–363.

Kaiyi Zhang, Ang Lv, Yuhan Chen, Hansen Ha, Tao Xu, and Rui Yan. 2024. Batch-icl: Effective, efficient, and order-agnostic in-context learning. *CoRR*, abs/2401.06469.

A Inference Anchoring Theory Glossary

Refer to *A Quick Start Guide to Inference Anchoring Theory (IAT)* and *Inference Anchoring Theory* for details.

B Prompt design

P1="Illocutionary relations include 0:Asserting, 1:Pure Questioning, 2:Challenging, 3:Assertive Questioning, 4:Rhetorical Questioning, 5:Agreeing, 6:Default Illocuting, 7:Arguing, 8:Restating, 9:Disagreeing.The illocutionary relation between the two sentences is [mask].".

P2="Illocutionary relations include 0:Default Inference, 1:Default Rephrase, 2:Default Conflict.The illocutionary relation between the two sentences is [mask].".

P3="Illocutionary relations include 0:Asserting, 1:Pure Questioning, 2:Challenging, 3:Assertive Questioning, 4:Rhetorical Questioning", 5:"Agreeing", 6:"Default Illocuting", 7:"Arguing", 8:"Restating", 9:"Disagreeing".The illocutionary relation between the two sentences is [mask].".

C Experiment Setup

We allocated 80% of the data to the training set, while the remaining 20% was assigned to the validation set. Prior to training, the datasets were tokenized and then fed into language models for fine-tuning. The learning rate was set to 2e-5, and the model underwent training for 2 epochs. To update the model's parameters, we employed the AdamW optimizer.

During the evaluation phase, we assessed the model's performance on the validation using accuracy as the metric. This metric takes the model's predictions and the ground-truth label as input and returns the portion of the correct predications. Every epoch, we printed out the achieved accuracy. To ensure optimal model performance, we conducted experiments with various input sizes and epochs, aiming to strike a balance between underfitting and overfitting.

To support our computations, we leveraged a single NVIDIA RTX A6000 card as our computational infrastructure. The best checkpoint, determined by our experiments, was utilized to generate the submitted maps.