Discourse Relation Recognition with Language Models Under Different Data Availability

Shuhaib Mehri **Chuyuan Li** **Giuseppe Carenini**
Department of Computer Science, University of British Columbia
V6T 1Z4, Vancouver, BC, Canada
shuhaibm@student.ubc.ca, chuyuan.li@ubc.ca, carenini@cs.ubc.ca

Abstract

Large Language Models (LLMs) have demonstrated remarkable performance across various NLP tasks, yet they continue to face challenges in discourse relation recognition (DRR). Current state-of-the-art methods for DRR primarily rely on smaller pre-trained language models (PLMs). In this study, we conduct a comprehensive analysis of different approaches using both PLMs and LLMs, evaluating their effectiveness for DRR at multiple granularities and under different data availability settings. Our findings indicate that no single approach consistently outperforms the others, and we offer a general comparison framework to guide the selection of the most appropriate model based on specific DRR requirements and data conditions.

1 Introduction

Discourse parsing automatically extracts the underlying discourse structure of a text, playing a pivotal role in various natural language processing (NLP) tasks. Its utility has been demonstrated in applications such as machine translation (Chen et al., 2020), summarization (Xu et al., 2020; Chen and Yang, 2021; Rennard et al., 2024), and question-answering (Jansen et al., 2014). Discourse parsing is particularly useful in scenarios that involve handling complex or large-scale text, such as in multi-document summarization (Chen et al., 2021; Li et al., 2020; Liu and Lapata, 2019).

A fundamental task in discourse parsing is discourse relation recognition (DRR), which aims to identify the relation sense between argument pairs. Typically, argument pairs are made up of text spans known as elementary discourse units (EDUs). When connectives are present between argument pairs (explicit DRR), training a simple classifier on the connectives can achieve a classification accuracy close to 95% (Xiang and Wang, 2023; Pitler and Nenkova, 2009; Varia et al., 2019). On the other hand, the task becomes more difficult when

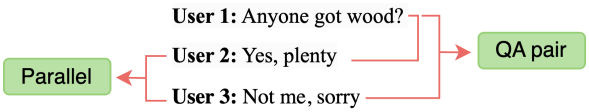

Figure 1: An example of discourse relation parsing in dialogue, taken from STAC corpus (Asher et al., 2016).

connectives are not present (implicit DRR), and current approaches for this task struggle to achieve an accuracy above 80% (Xiang et al., 2023; Zhou et al., 2022; Chan et al., 2023). To address this challenge, we explore relation recognition in dialogue discourse parsing (see Figure 1), where connectives play a less prominent role, alongside implicit discourse relation recognition (IDRR) in monologues. In dialogue discourse parsing, the argument pairs are made up of user utterances, and in IDRR, the argument pairs are made up EDUs.

Recent large language models (LLMs) (e.g., ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI et al., 2024)) have demonstrated remarkable performance on many NLP benchmarks, and display advanced reasoning and understanding capabilities. They also exhibit impressive abilities in zero-shot and few-shot settings (Wei et al., 2022), and can sometimes be competitive with prior state-of-the-art fine-tuning approaches (Brown et al., 2020). At the same time, many studies suggest that LLMs do not perform as well as small encoder-only models fine-tuned on specific-tasks (Qin et al., 2023; Lu et al., 2023).

This is the case for the DRR task on which LLMs seem to struggle with (Fan et al., 2024; Chan et al., 2024). Many of the current top-performing approaches rely on fine-tuning relatively smaller encoder-based pre-trained language models (PLMs) like RoBERTa (Zhou et al., 2022; Wu et al., 2023; Xiang et al., 2023, 2022; Li et al., 2023, 2024a,b).

In spite of these established approaches, it is still

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 148–156

unclear when it is more effective to use LLMs or PLMs for DRR. With this in mind, we conduct a comprehensive analysis of different approaches for the DRR task, focusing on comparing PLMs and LLMs under different data availability settings. For PLMs, we use the data for fine-tuning. For LLMs, we employ zero-shot prompting, in-context learning, and a new self-reflection technique we call confusion-matrix prompting. We explore these techniques using both monologues and dialogues with different relation types and granularities.

Confusion-matrix prompting is a novel technique that uses information from a confusion matrix to inform an LLM about the errors it tends to make, enabling it to self-reflect and adjust its predictions accordingly. This is inspired by the many studies that have shown how LLMs benefit from self-reflecting on and improving their initial generation (Madaan et al., 2023; Fernando et al., 2023; Welleck et al., 2023; Shinn et al., 2023), as well as learning from their mistakes (Zhang et al., 2024).

Our work advances the understanding of fine-tuning PLMs and various prompting techniques with LLMs in the context of DRR, across different dataset sizes and multiple relation sense granularities. Key takeaways include: **(1)** Zero-shot prompting leverages inherent knowledge embedded in LLMs and performs better than other techniques when there is little available data; **(2)** Confusion-matrix prompting achieves optimal performance when there is insufficient data for fine-tuning, but enough to surpass zero-shot performance; **(3)** Fine-tuned PLMs excel in scenarios with increased data, and is robust across various datasets regardless of complexity or number of relation senses.

2 Methodology

To simulate different data availability settings, we extract seven subsets of training datasets, each with different sizes. For each subset, we randomly select a certain number of examples for each relation sense. We start with a single example per relation sense, and increment the number up to 250 examples per relation sense. When a specific relation sense does not enough examples available, we randomly select the remaining examples from other relation senses to satisfy the target example count.

Next, we employ fine-tuning and prompting techniques that leverage these subsets for the DRR task, assessing how each performs across different data volumes.

- **Fine-tuning (FT).** We fine-tune an encoder-only PLM to encode the representation of argument pairs and predict a relation sense. Our representation of argument pairs follows the template from Zhou et al. (2022): `Arg1: <Arg1>. Arg2: <Arg2>. In summary, the discourse relation between Arg1 and Arg2 is`
- **Zero-shot (ZS).** Without using any annotated data, we frame the problem as a zero-shot fill-in-the-blank prompt to a LLM. Our prompt follows the same format as FT.
- **In-context learning (ICL).** Input-label pairs from the dataset are incorporated directly into the LLM's prompt to leverage in-context learning. Typically, in-context learning approaches manually select the input-label pairs to ensure high-quality examples. However, in our approach, we use randomly selected pairs from the dataset to maintain consistency with our other techniques. Due to the limited input context length of the early GPT-3.5 version, we cannot include examples for all data availability settings, particularly for larger numbers of examples per relation sense, e.g., >25 examples per relation sense for PDTB top-level experiment.
- **Confusion-Matrix Prompting (CMP).** Using the dataset, we collect zero-shot performance of the LLM in the form of a confusion matrix, recording the model's predictions against the true labels. This confusion matrix allows us to determine how often the model correctly predicts a relation, and how often it confuses it with another relation. During inference, we first let the model make its initial prediction. Based on this prediction we formulate a follow-up prompt using the confusion matrix, informing the model of its prediction accuracy and common mistakes. We provided this prompt as a follow-up, giving the model a chance to self-reflect and correct it's initial prediction (see Appendix A for an example).

3 Experimental Setup

Monologue Data: The Penn Discourse Treebank 3.0 (PDTB 3.0). PDTB 3.0 is an annotated corpus of discourse relations that come from Wall Street Journal articles (Webber et al., 2019). It uses 3 hierarchies of relation senses, and contains both implicit and explicit relation types. For our

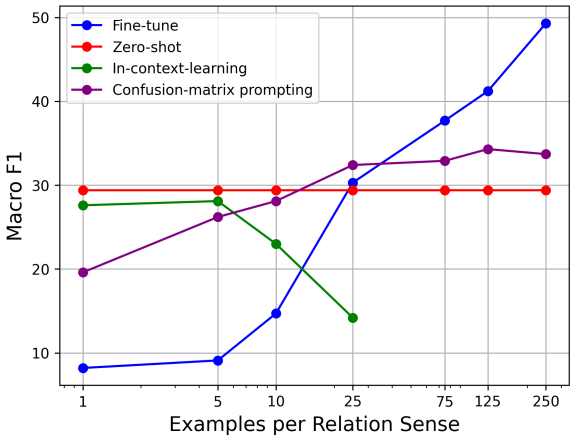

Figure 2: Comparisons of Macro F1 scores of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB top-level. The number of ICL examples is restricted to up to 25 due to the input context length of GPT-3.5.[1]

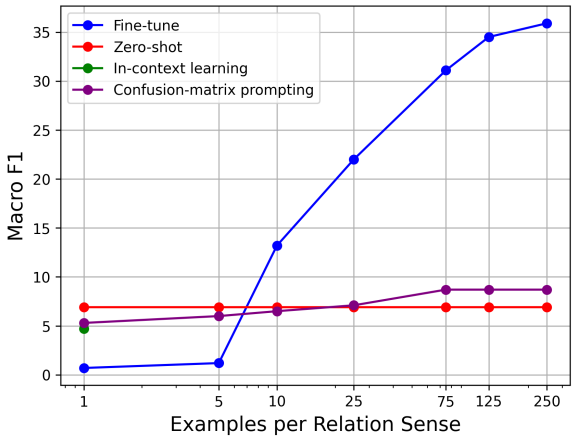

Figure 3: Comparisons of Macro F1 scores of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB second-level. ICL examples is restricted to 1 example per relation sense. [1]

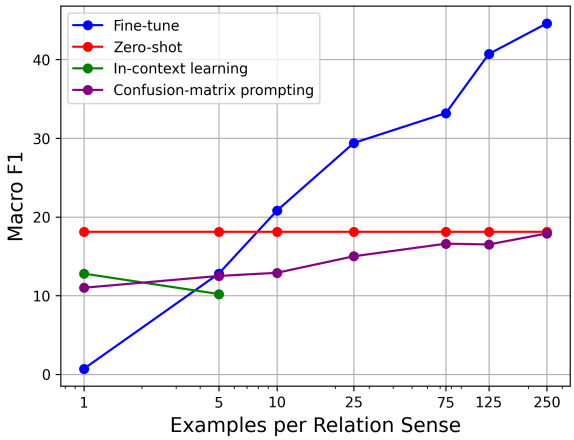

Figure 4: Comparisons of Macro F1 scores of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on STAC. ICL examples is restricted to up to 5 examples per relation sense. [1]

experiments, we use the four top-level and twenty second-level implicit relation senses, with a test set of 1,538 examples.

Dialogue Data: STAC. STAC is a corpus of multi-party dialogues collected from an online game called *The Settlers of Catan* (Asher et al., 2016). The dialogues are annotated in the style of Segmented Discourse Representation Theory (SDRT), which uses sixteen relation senses (Asher and Lascarides, 2003). The test set consists of 1,128 examples.

Implementation Details. The fine-tuning experiments were conducted using RoBERTa-base (Liu et al., 2019). We selected this lightweight model for its strong performance on the DRR task. We employ a learning rate of $1e-5$ and trained the model for 20 epochs with early stopping based on performance on the development set. To ensure robustness of our results, we repeat each experiment over 10 random seeds and report the average score.

ZS, ICL and CMP experiments were done using GPT-3.5 Turbo. We report the average of our results over 5 random seeds. Additionally, preliminary experiments were performed on Mistral 7B (Jiang et al., 2023), and indicated a similar trend of improvement in performance.

4 Results and Analysis

The results of our experiments are displayed in Table 1 and illustrated in Figures 2, 3, and 4 for PDTB top-level, second-level, and STAC corpus, respectively. Our analysis primarily uses Macro F1 scores, though similar trends are observed for accuracy (relevant figures are included in Appendix B).

In general, ZS is consistently the better technique for lesser amounts of data. As the number of examples per relation sense increases, fine-tuning (FT) demonstrates constant improvement and soon surpasses ZS, underlining the benefits of the technique. While CMP starts out with lower performance, it has shown to improve and eventually surpass ZS at higher data volumes. ICL, on the other hand, exhibits underwhelming performance across all datasets.

The observed trends indicate that ZS, which relies on inherent discourse knowledge embedded in

[1]Logarithmic scale is used for the x-axis

Number of Training Examples Per Relation Sense (always zero for ZS)														
	1		5		10		25		75		125		250	
Technique	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
PDTB Top-Level (4 relation senses)														
FT	20.9	8.2	22.7	9.1	22.9	14.7	34.5	30.3	39.9	**37.7**	**43.1**	**41.2**	**51.7**	**49.3**
ZS	**42.8**	**29.4**	**42.8**	**29.4**	**42.8**	**29.4**	**42.8**	29.4	**42.8**	29.4	42.8	29.4	42.8	29.4
ICL	40.8	27.6	39.6	28.1	33.1	23.0	18.5	14.2	-	-	-	-	-	-
CMP	22.5	19.6	33.1	26.2	35.4	28.1	39.5	**32.4**	42.0	32.9	42.1	34.3	42.3	33.7
PDTB Second-Level (20 relation senses)														
FT	5.6	0.7	6.2	1.2	15.0	**13.2**	**25.4**	**22.0**	**38.4**	**31.1**	**42.6**	**34.5**	**50.1**	**35.9**
ZS	**17.8**	**6.9**	**17.8**	**6.9**	**17.8**	6.9	17.8	6.9	17.8	6.9	17.8	6.9	17.8	6.9
ICL	9.1	4.7	-	-	-	-	-	-	-	-	-	-	-	-
CMP	11.5	5.3	13.9	6.0	13.4	6.5	11.7	7.1	19.3	8.7	19.5	8.7	20.5	8.7
STAC (16 relation senses)														
FT	4.3	0.7	20.0	12.8	**31.0**	**20.8**	**39.9**	**29.4**	**46.5**	**33.2**	**53.0**	**40.7**	**59.1**	**44.6**
ZS	**24.9**	**18.1**	**24.9**	**18.1**	24.9	18.1	24.9	18.1	24.9	18.1	24.9	18.1	24.9	18.1
ICL	20.6	12.8	17.2	10.2	-	-	-	-	-	-	-	-	-	-
CMP	16.2	11.0	14.9	12.5	16.2	12.9	19.3	15.0	21.6	16.6	22.1	16.5	26.8	17.9

Table 1: Accuracy (Acc) and Macro *F1* (F1) scores of FT, ZS, ICL, and CMP techniques on different numbers of examples per relation sense. The best results for each technique are bolded. The - values indicate that we were unable to experiments due to input length limitations.

the model, is the highest performing technique for DRR in low data availability scenarios. When considering ZS performance across the datasets, the performance diminishes for more complex problems where there are greater numbers of relation senses. ZS achieves higher performance on top-level PDTB, with 4 relations senses, and lower performance on STAC, with 16 relation senses, and even lower performance in second-level PDTB, with 20 relation senses. This increased difficulty highlights the limitations of relying solely on pre-trained knowledge.

FT scales very well with the data and always emerges as the most effective technique as data availability increases. Notably, the accuracy and F1 scores achieved by FT are relatively consistent across the different datasets. Unlike in ZS, we do not see a similar drop in performance as the number of relation senses increases. From this, we can gather that a more complex task does not proportionally impact the performance of PLMs the same way it does for LLMs.

CMP begins to outperform ZS as dataset sizes increase, showing that it is optimal when the amount of data is insufficient for fine-tuning, or if fine-tuning is not a viable option. In scenarios involving smaller datasets, the volatility of the confusion matrix is less representative of model performance, often causing a drop in performance. However, as we use larger datasets, the confusion matrix provides a more accurate depiction of the model's errors and overall performance. This allows CMP to help the LLM learn from its past performance and start outperforming ZS.

Furthermore, CMP proves to be more effective in the more complex datasets with larger numbers of relation senses. This effectiveness is attributed to the technique being beneficial when there are more potential mistakes that the LLM can make.

The results observed from ICL gives poor results, which is likely due to the random selection of examples and context length limitations. It never outperformed ZS, and the performance decreases as the datasets get larger, as if adding more data into the prompt makes it more difficult for the LLM to effectively process.

5 Conclusion

In order to identify the optimal techniques for DRR under different data availability settings, we perform an analysis on how these techniques perform with varying amounts of data. The techniques we explore include fine-tuning for PLMs, and various prompting techniques with LLMs. In our experiments, we find that in low data availability scenarios, zero-shot prompting performs best. CMP achieves the best performance when there is more data available, but not enough for effective fine-tuning. When we have more data, fine-tuning PLMs dominates, and performance is not affected

by more complex relation sense granularities. Unexpectedly, ICL is always dominated by ZS.

In future work, we plan to further investigate the trade-off between PLMs and LLMs for discourse processing tasks. We would like to extend this work by conducting further experiments on more powerful LLMs, more specific ICL techniques such as similarity-based selection, as well as more complex tasks such as discourse parsing. Additionally, we would like to explore self-reflection learning techniques for LLMs as we have found quite promising results. The methodology and experimental framework we have designed and implemented [1] will be critical in facilitating these further investigations by us and other researchers.

6 Limitations

In our experiments, we consider GPT-3.5 Turbo and RoBERTa-base as representatives for LLMs and PLMs, respectively. While these models serve as good representatives, exploring more powerful models would further strengthen our study. Furthermore, due to the lack of annotated data, our experiments were limited to two English datasets: PDTB 3.0 and STAC. These limitations highlight areas for future research to provide a more comprehensive understanding of discourse relation recognition

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback and suggestions. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chunkit Chan, Cheng Jiayang, 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*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021. SgSum:transforming multi-document summarization into sub-graph selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4063–4074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. *Preprint*, arXiv:2305.08391.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *Preprint*, arXiv:2309.16797.

[1] The code will be made public after publication

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Chuyuan Li, Maxime Amblard, and Chloé Braud. 2023. A semi-supervised dialogue discourse parsing pipeline. In *Journées Scientifiques du GDR Lift (LIFT 2023)*.

Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. 2024a. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176, St. Julians, Malta. Association for Computational Linguistics.

Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024b. Dialogue discourse parsing as generation: A sequence-to-sequence LLM-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–14, Kyoto, Japan. Association for Computational Linguistics.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6232–6243, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

Yuxiang Lu, Yu Hong, Zhipang Wang, and Guodong Zhou. 2023. Enhancing reasoning capabilities by instruction learning and chain-of-thoughts for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5634–5640, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David

Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Virgile Rennard, Guokan Shang, Michalis Vazirgiannis, and Julie Hunter. 2024. Leveraging discourse structure for extractive meeting summarization. *Preprint*, arXiv:2405.11055.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

Wei Xiang, Chao Liang, and Bang Wang. 2023. TEPrompt: Task enlightenment prompt learning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12403–12414, Toronto, Canada. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Comput. Surv.*, 55(12).

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Tianjun Zhang, Aman Madaan, Luyu Gao, Steven Zheng, Swaroop Mishra, Yiming Yang, Niket Tandon, and Uri Alon. 2024. In-context principle learning from mistakes. *Preprint*, arXiv:2402.05403.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Confusion-Matrix Prompting Examples

User: Arg1: Coupons and a newsletter will be mailed. Arg2: And the sponsor will be able to gather a list of desirable potential customers.
In summary, the discourse relation between Arg1 and Arg2 is

Model: Expansion

User: Arg1: Coupons and a newsletter will be mailed. Arg2: And the sponsor will be able to gather a list of desirable potential customers.
The initial prediction for the discourse relation between Arg1 and Arg2 was Expansion. 27% of the time when Expansion was predicted, the correct answer was Expansion. 27% of the time when Expansion was predicted, the correct answer was Contingency. 24% of the time when Expansion was predicted, the correct answer was Temporal. 20% of the time when Expansion was predicted, the correct answer was Comparison. Considering this information, what is the relation sense?

Model: Contingency

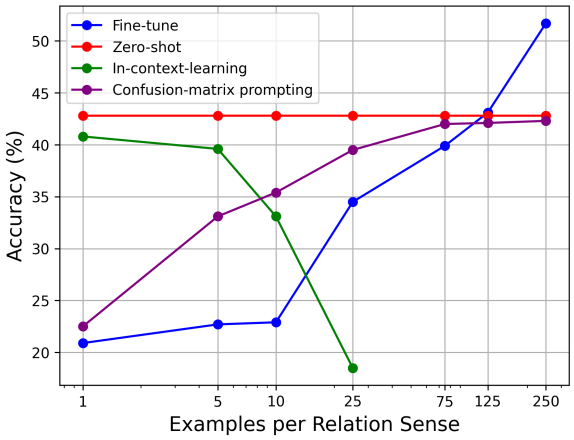

Figure 5: Comparisons of accuracy of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB top-level. [1]

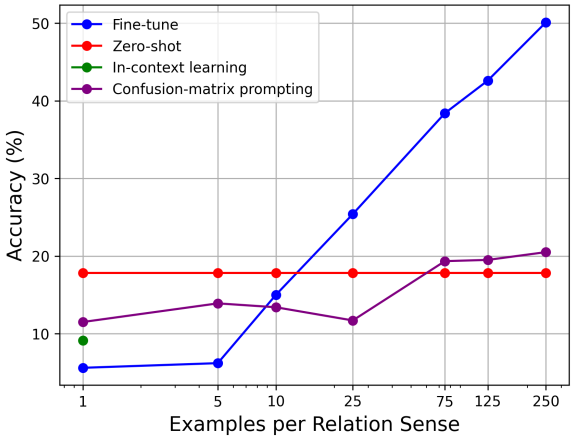

Figure 6: Comparisons of accuracy of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB second-level. [1]

B Accuracy Comparisons of Techniques

[1]Logarithmic scale is used for the x-axis

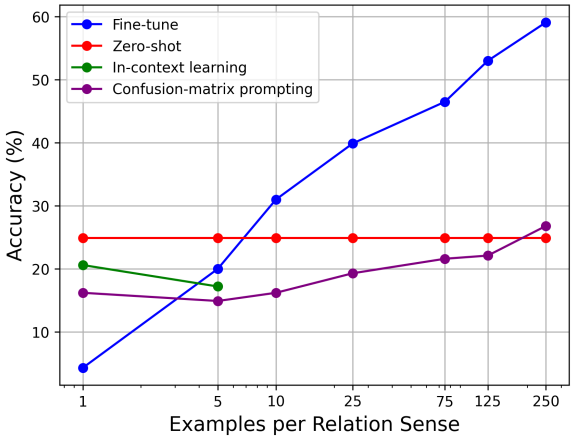

Figure 7: Comparisons of accuracy of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on STAC. [1]