Leveraging Machine-generated Rationales for Conversational Forecasting

Anonymous ACL submission

Abstract

Predicting outcomes in multi-turn dialogues is challenging due to the implicit nature of decision-making and the evolving dynamics between participants. In this work, we explore whether LLM-generated rationales can enhance the accuracy and generalizability of outcome prediction in task-oriented dialogues. We evaluate zero-shot in-context learning models on the Craigslist Bargain dataset, testing their ability to predict final sale prices at different dialogue checkpoints in absence and presence of rationales. Preliminary results with metrics such as RMSE and Pearson correlation suggest that rationale-augmented models better capture negotiation strategies and concession patterns, improving early-stage prediction accuracy.

1 Introduction

The task of conversational forecasting involves predicting the outcome of an unfolding dialogue (Sokolova et al., 2008; Zhang et al., 2018). We present a snippet of an ongoing negotiation scenario between a buyer and a seller in Figure 1 where the objective is to predict the final sales price at the end of the conversation, given the target prices that each party is trying to optimize. Beyond measuring success in task-oriented dialogues like (Chawla et al., 2021; Dutt et al., 2021; Reitter and Moore, 2007), conversational forecasting has also been adopted for content moderation in social media (Zhang et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019; Kementchedjhieva and Søgaard, 2021), predicting emotions (Wang et al., 2020; Matero and Schwartz, 2020) and even health codes (Cao et al., 2019) in dialogues.

Conversational forecasting is an inherently complex task due to the implicit and evolving nature of decision-making. For example, negotiation conversations involve an interplay of strategy, persuasion, and concession-making, often without any explicit cues until a conclusion is reached. Models that utilize only these dialogue sequences can miss the underlying intentions behind the participants' utterances that drive these interactions (Yamaguchi et al., 2021; Chan et al., 2024; Dutt et al., 2024). In this work, we investigate whether machine-generated "free-text" rationales, that capture the intentions behind each utterance, can serve as effective augmentations for conversational forecasting. Our hypothesis is that by making the implicit reasoning explicit, models can better capture the nuanced dynamics of conversations, particularly at early stages when only partial dialogue context is available (Hua et al., 2024).

2 Methodology

Dataset: For our preliminary experiments, we use the Craigslist Bargain dataset of He et al. (2018). The dataset comprises simulated multi-turn dialogues between a buyer and a seller as they negotiate the price of an item while trying to optimize their assigned target price. Our objective is to predict the final sales price for each negotiation at different stages of its completion, i.e. 25%, 32.5%, 50%, 62.5%, 75%, and 100% of the conversation.

Rationales: We leverage the rationale-generation framework of (Dutt et al., 2024) to obtain the speaker's intentions corresponding to each utterance in the negotiation dialogue. For example, in Figure 1 we showcase the intentions of the buyer and seller on the right. These intentions were generated on an utterance-by-utterance basis using GPT-4o as the backbone LLM. We investigate whether these rationales can aid forecasting when provided as additional inputs.

Prompting Experiments: We conduct zero-shot prompting experiments using two popular LLMs, i.e. GPT-3.5-turbo and GPT-4o to predict the final sales price of a given negotiation conversation. We explore two different prompting styles, i.e. (i) Simple Prompt that directly requests the final sales

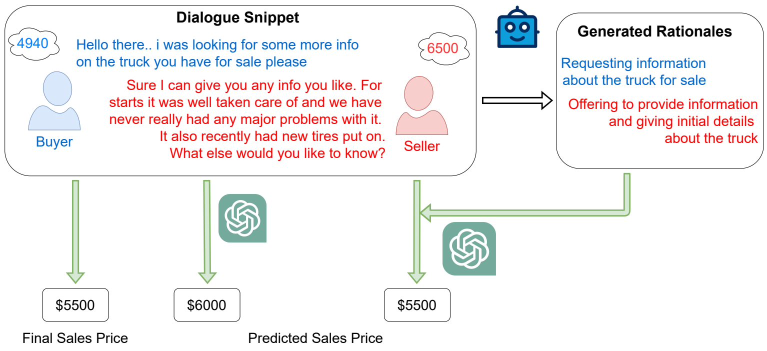

Figure 1: We present a snippet of a negotiation conversation between a buyer and a seller as they each try to match their targeted price. We observe that the sales price predicted by an LLM matches the final sales price when we augment the dialogue snippet with generated rationales.

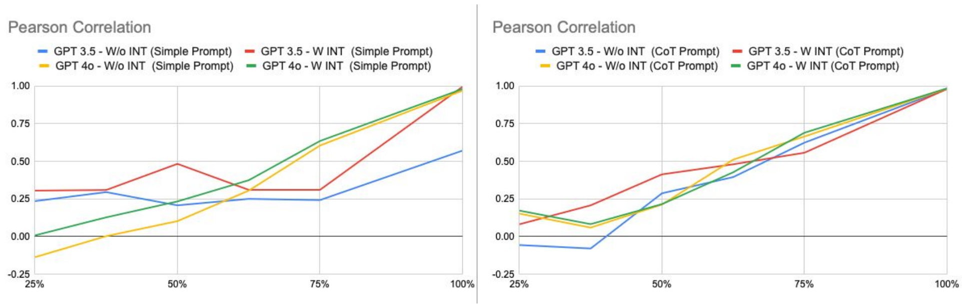

Figure 2: Pearson Correlation between the predicted success score and the true success score for different prompting styles in presence and absence of rationales.

price and (ii) Chain-of-Thought (CoT) Prompt that instructs the model to take into consideration additional information such as the persuasion techniques employed, power dynamics, and concession pace before predicting the final price.

In short the Chain-of-Thought (CoT) prompt is designed to summarize the global dialogue information, whereas the rationales are designed to capture the local information, i.e. at an utterance. The specific prompts used in our experiments without intentions are shown in Figures 4 and 5 in the Appendix. When including intentions in our experiments, the prompts are slightly modified with the dialogue formatted as [(u1, r1), (u2, r2)....(un, rn)] where, u1, u2... are the utterances and r1, r2... are the corresponding intention rationales.

Evaluation: For evaluation, we use the normalized success score below, which positions the predicted sale price relative to the buyer's and seller's target prices, providing a consistent metric across dialogues. (Dutt et al., 2021). We measure the correlation coefficient between the predicted success score where p_{SP} is the predicted sales price and the true success score where p_{SP} is the actual sales price for the conversation. We also measure the RMSE between the predicted and true sales price.

$$succ = \frac{p_{SP} - p_{BU}}{p_{SE} - p_{BU}} \quad (1)$$

3 Results and Future Work

Our results in Figures 3 and Figures 2 highlights that both GPT-3.5 Turbo and GPT-4o models achieve lower RMSE scores and higher correlation scores when augmented with rationales. This difference is particularly marked in the early stages of the conversation (<50% of dialogue as context), indicating enhanced prediction accuracy even with limited context. We also observe pronounced improvements from adding rationales even in the CoT based prompting style for the GPT-3.5-turbo model, but less for GPT-4o, thereby highlighting the efficacy of these prompts for less powerful models. Also unsurprisingly we see a strong monotonic increase in correlation and decrease in RMSE with more conversational turns. Our future work aims to explore the role of these rationales for instruction tuned models like FLAN-T5, other LLMs like LLama, and other forecasting tasks like conversational derailment (Zhang et al., 2018).

References

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn Rose. 2021. ResPer: Computationally modelling resisting strategies in persuasive conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 78–90, Online. Association for Computational Linguistics.

Ritam Dutt, Zhen Wu, Jiaxin Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Rose. 2024. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6901–6929, Bangkok, Thailand. Association for Computational Linguistics.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. How did we get here? summarizing conversation dynamics. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7452–7477, Mexico City, Mexico. Association for Computational Linguistics.

Yova Kementchedjhieva and Anders Søgaard. 2021. Dynamic forecasting of conversation derailment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Matero and H. Andrew Schwartz. 2020. Autoregressive affective language forecasting: A self-supervised task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2913–2923, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Marina Sokolova, Vivi Nastase, and Stan Szpakowicz. 2008. The telling tail: Signals of success in electronic negotiation texts. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Zhongqing Wang, Xiujun Zhu, Yue Zhang, Shoushan Li, and Guodong Zhou. 2020. Sentiment forecasting in dialog. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2448–2458, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita. 2021. Dialogue act-based breakdown detection in negotiation dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 745–757, Online. Association for Computational Linguistics.

Justine Zhang, Jonathan P Chang, Lucas Dixon, Yiqing Hua, Nithum Tahin, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Appendix

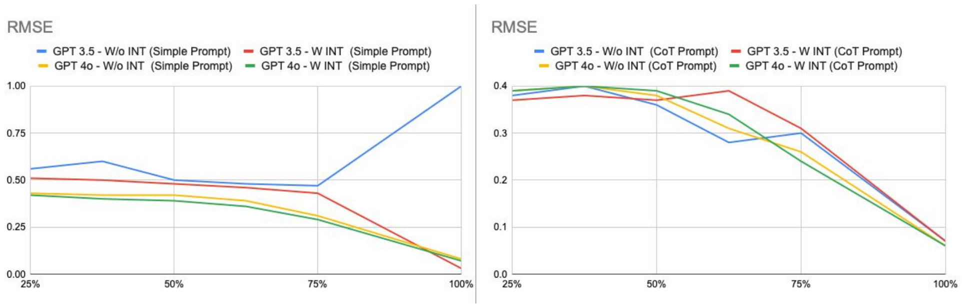

Figure 3: RMSE scores

Simple Prompt

Analyze this negotiation, given in the format <buyer target, seller target, [negotiation]> and predict the projected sale price. Provide only the final answer in the format ‘FINAL_PRICE: [number]’
INPUT: <{buyer_target}, {seller_target}, [{dialogue}]>

Figure 4: Simple prompt without rationales

Chain of Thought (CoT) Prompting

Analyze this negotiation and predict the final agreed price. Think through each step, then provide your final answer.
Context:
- Buyer’s Target: ${buyer_target}
- Seller’s Target: ${seller_target}
Dialogue:
{dialogue}
Consider:
1. Opening positions and target prices
2. Pace of concessions from both parties
3. Negotiation tactics and persuasion techniques used
4. Power dynamics and urgency signals
5. Number of turns and negotiation progression
Analyze the above factors, then print the output, the predicted final price, in this format, with no additional information:
FINAL_PRICE: [your prediction]

Figure 5: CoT prompt without rationales