Identifying Power Relations in Conversations using Multi-Agent Social Reasoning

Zhaoqing Wu
Purdue University
wu1828@purdue.edu

Dan Goldwasser
Purdue University
dgoldwas@purdue.edu

Maria Leonor Pacheco
University of Colorado Boulder
maria.pacheco@colorado.edu

Leora Morgenstern
SRI International
leora.morgenstern@sri.com

Abstract

Large language models (LLMs) struggle in social science domains, where critical thinking and human-level inference are crucial. In this work, we propose a multi-agent social reasoning framework that leverages the generative and reasoning capabilities of LLMs to generate and evaluate reasons from multiple perspectives grounded in social science theories, and construct a factor graph for inference. Experimental results on understanding power dynamics in conversations show that our method outperforms standard prompting baselines, demonstrating its potential for tackling hard Computational Social Science (CSS) tasks.

1 Introduction

Understanding conversational dynamics is a multi-faceted problem, which requires situating the interlocutors' utterances in a specific social context such that the intent behind them, and the reaction to them, could be revealed. Past social science work has studied the interaction between conversations and social relationships (Evans and Aceves, 2016), language use in different social situations (Snyder and Stukas Jr, 1999; Gibbs, 2000) and social identities (Tracy and Robles, 2013). This paper focuses on the connection between a specific social indicator, *power relations*, and several aspects of language use, namely *style* (e.g., apologetic, assertive), *content* (e.g.,judgments over dialog acts), *coordination* (e.g., steering and setting the tone) and *engagement* (e.g., active participation).

Identifying power relationships in conversations, taking place in different settings such as organizational emails, online forums and chats, has been studied extensively in the NLP literature (Bramsen et al., 2011; Danescu-Niculescu-Mizil et al., 2012a; Biran et al., 2012; Prabhakaran and Rambow, 2013, 2014; Lam et al., 2018) and was typically formulated as a supervised learning problem focusing on different aspects such as lexical features (Bramsen et al., 2011), linguistic coordination (Danescu-Niculescu-Mizil et al., 2012a) or conversational structure (Prabhakaran and Rambow, 2013). The recent paradigm shift in NLP, moving away from task-specific supervised learning and towards broader-purpose LLMs, raises an open question – **Can LLMs understand such social dynamics, without dedicated training?** Initial results for conversation analysis tasks (including power-relation prediction) were mixed (Ziems et al., 2024) motivating further research in this area.

In this paper we argue this question should be studied with more nuance. Instead of directly accounting for the complex interactions between social settings and conversational behaviors via LLM autoregressive (i.e., greedy) decoding, we argue that LLMs can demonstrate their ability to understand conversational data by focusing on different *aspects* of conversational behavior and raising hypotheses on how they provide evidence for the power relationship between interlocutors. Specifically, building on prior work in social science, we identify *style*, *content*, *coordination*, and *engagement* as key aspects that capture the implicit dynamics of conversations, including speakers' social status (Irvine, 1985), power relations (Danescu-Niculescu-Mizil et al., 2012a), and the overall conversation flow (Liu et al., 2020). We formulate the problem as a multi-agent social reasoning task (see Guo et al.,2024 for an overview), in which each interlocutor is associated with an LLM-based agent advocating for their high power status in the conversation, by providing aspect-specific *reasons* and *rebuttals* in response to the other side's reasons. We define LLM-based assessment functions for scoring the strength of these claims (Sec. 2.1) and organize them based on their argumentation structure; we then compile this structure into a factor-graph (Sec. 2.2) and perform probabilistic reasoning over that structure (Jung et al., 2022; Kassner et al., 2023) to find the most probable power-relation con-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 855–865
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

sistent with that structure. Figure 1 provides an illustration of our overall framework.

We conduct our experiments over the ICSI Meeting Corpus (Janin et al., 2003) by sampling conversation snippets and applying our multi-agent reasoning architecture over them.[1] These are very challenging settings, as each snippets captures only a handful of relevant behaviors, which are often misleading as the data consists of informal work-related interactions between students, postdocs and faculty. This is reflected in our experimental results, showing that the performance of both human and direct LLM prompting is worse than random. Augmenting the LLM prompts with the generated reasons and rebuttals leads to even worse performance, as the model is not able to effectively prioritize between them. However, when applying our argumentation-based reasoning framework, the model can detect inconsistencies and prefer reasons that uniquely identify one of the sides, leading to an 8 points improvement compared to human performance and 5 compared to direct LLM prompting.

2 Multi-Agent Social Reasoning

Our framework incorporates a generation step and an inference step. In the generation step, we use two LLM agents to produce reasons and rebuttals for each prediction candidate. For each reason and rebuttal, a separate LLM assigns a score indicating its likelihood of validity and grounding. In addition, the LLM also evaluates the reasons' relationships with the other generated reasons. Based on these scores and interdependencies, we construct and resolve a factor graph to finalize the prediction.

2.1 Text and Score Generation

We prompt an LLM to generate text that explains the input conversation and analyzes the speakers' behaviors, then score the generated text with probability potentials that support inference.

Given an input conversation C, we decompose C into n distinct aspects and use an LLM to generate aspect-specific summaries, S_{ASP}, for both speakers, A and B, based on the predefined definitions of each aspect. This process is formally denoted as

$$S_{ASP} = P_{SUMMARY}(C, ASP)$$

where $P_{SUMMARY}$ is a natural language prompt used to generate the summaries.

[1] https://github.com/kandicew/social-reasoning

User:
Summarize the style of the conversation.
<conversation>
ChatGPT:
style of A:
Informal, slightly hesitant and fragmented language, focusing on short and direct responses with some repetition.
style of B:
Technical and focused on the task at hand, using informal language with some interruptions and repeating points.
User:
Provide reasons for A has more power based on the summary.
<style of A>
<style of B>
ChatGPT:
Direct response is more influential in decision-making.

Table 1: An example of LLM-generated summary on the *style* aspect is shown as the first interaction in the table. This summary is then used in prompts to generate reasons supporting that Speaker A holds higher power (the second interaction in the table). Summaries for other aspects and additional reasons for both speakers follow a similar process. Further details, including additional textual generations and the conversation snippet used for this example, can be found in the Appendix E.

Conditioned on the aspect-specific summaries, we prompt the LLM with P_{REASON} to generate sets of supporting reasons R_s, for each speaker s. In our case, given that the speakers are restricted to A and B, the model produces R_A and R_B to support A and B respectively. Table 1 shows an example of such process.

$$R_s = P_{REASON}(s, S_{ASP})$$

To incorporate critical thinking, we use a separate LLM with the prompt $P_{REBUTTAL}$ and utilize both the original conversation C and the reason R_s, where s is A or B, as the context to generate a rebuttal R_s^b for each reason.

$$R_s^b = P_{REBUTTAL}(C, R_s)$$

All reasons and rebuttals are scored using a scoring function, f_{score}, following the approach of Kassner et al. (2023) to evaluate a statement. A reason is accessed on whether it qualifies as a strong

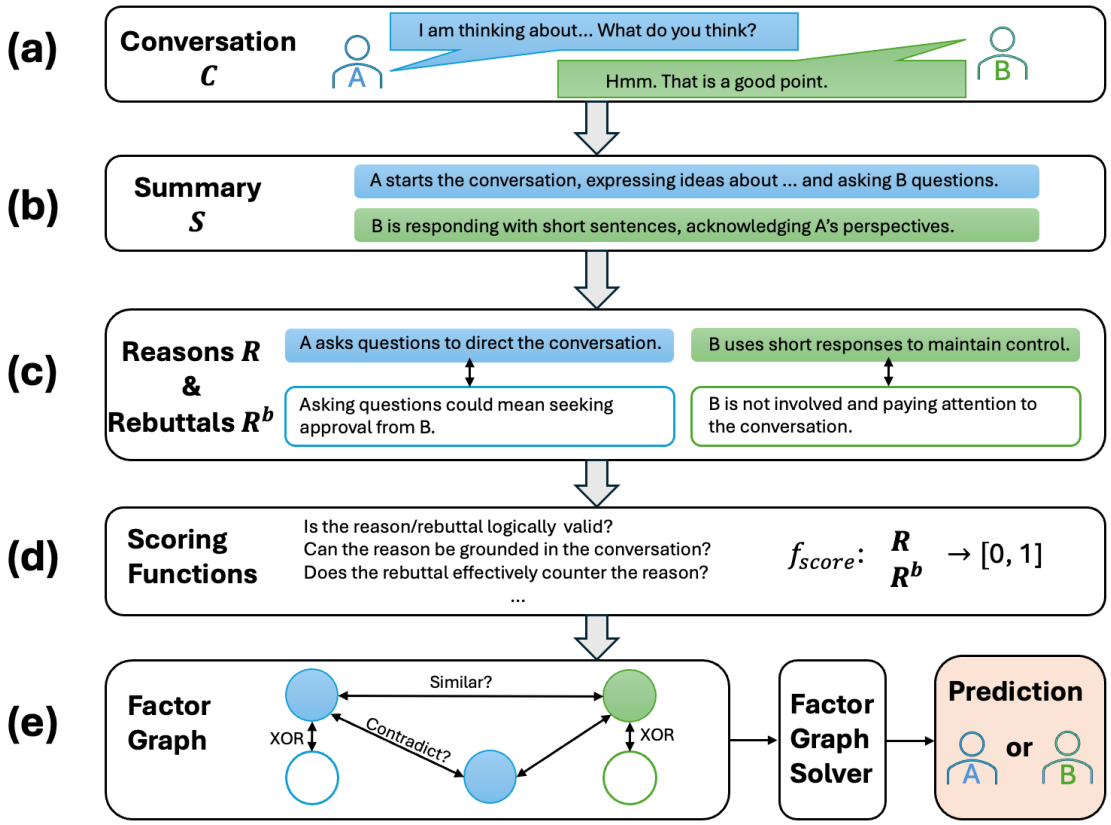

Figure 1: A high-level illustration of each of the steps in our social reasoning framework. In step (a), a conversation snippet C is provided as the input. Step (b) generates the aspect-specific summaries S_{ASP} for each speaker. Step (c) further, based on the summaries, generates supporting reasons R_s for s in higher power, along with rebuttals R_s^b to challenge those reasons. Step (d) employs scoring functions to evaluate the probabilities or strength of both reasons and rebuttals, resulting in a score between 0 and 1. Step (e) builds a factor graph using the generated texts and their corresponding scores. The final prediction is derived by solving the factor graph, assigning approximate probabilities to each speaker's level of power in the conversation.

reason (valid) and whether it can be directly supported by the conversation (grounding). As for a rebuttal, it is scored on whether it directly challenges the corresponding reason and makes it less convincing, and whether it is grounded in the conversation. This process results in a score between 0 and 1, and is shown in Figure 1(d).

$$f_{score} : R_s, R_s^b \rightarrow [0, 1]$$

Additionally, we assign scores to the relationships between reasons. For each pair of reasons that supports the same speaker, we find a contradictory score indicating whether they are in conflict. For each pair of reasons that supports different speakers, a similar score is assigned. To quantify these relationship, we prompt the LLM to use a Likert scale as in Appendix A for scoring.

2.2 Factor Graph Inference

We construct a factor graph with the generated text and scores described in 2.1, and solve the factor graph with AD3 (Martins et al., 2011). AD3 relaxes the input factor graph to a Linear Programming (LP) problem, providing an efficient approximation of probability assignments for each variable, enabling fast inference in our case. An example of subgraph with variables and factors is shown as Figure 1(e).

The variables in the graph include the reasons, rebuttals, and the relationships, similar or contradictory. The potentials of these variables are the weighted scores, details in Appendix D. We define two variables, P_A and P_B, initially set to 0, representing the probability that each speaker holds the power in the conversation. We consider the following factors for constructing the graph: 1) Only one speaker can hold power; 2) At least one

reason must support the speaker in power; 3) A reason and its corresponding rebuttal cannot be valid simultaneously; 4) A high similarity score between reasons supporting opposing speakers suggests weaker decision-making confidence; 5) A high contradiction score between reasons supporting the same speaker implies that only one of them can be valid.

AD3 assigns a probability between 0 and 1 to each of the variables after solving the factor graph. We compare the probabilities assigned to P_A and P_B, selecting the higher value as our final prediction of which speaker holds greater power in the conversation.

3 Experiments

3.1 Setup

For all LLM interactions, we utilize GPT-3.5-Turbo in a zero-shot prompting setup. We break down the conversations into four aspects, details defined in Appendix B. For each aspect, we generate three reasons to support each of the two speaker' positions, resulting in 12 reasons supporting each speaker. Each reason is then challenged with a rebuttal.

As this is a binary classification task, we evaluate performance using exact match accuracy based on the number of correct classifications.

3.2 Dataset

We use the the transcripts of ICSI Meeting Corpus (Janin et al., 2003), which consists of natural meetings. These meetings involve multiple participants such as undergraduates, graduate students, postdocs, and professors, which contains nuanced interaction in an academic setting. We assume that the professors are the ones with the highest power among all participants. For our analysis, we focus on conversations that are limited to six alternating turns between two speakers. We specifically filter the data to include only interactions between a professor and a student. 80% of the filtered data is used to train a BERT (Devlin et al., 2019) classifier, and the remaining data is used for testing, resulting in a test set of 151 such conversations snippets.

3.3 Baselines

3.3.1 Direct Prompting

We prompt GPT-3.5-Turbo directly to predict which one of the two speakers holds more power in a given conversation. The answer is limited to either 'A' or 'B'. We also include generated reasons and rebuttals in the prompt to experiment whether providing more information about power dynamics affects the prediction. All of this is done using a zero-shot approach, without providing in-context examples.

3.3.2 Trained Classifiers

We trained a BERT (Devlin et al., 2019) classifier using 80% of the filtered conversations snippets from the ICSI Corpus (Janin et al., 2003) as mentioned in 3.2 and evaluated its performance on the test set.

Additionally, Danescu-Niculescu-Mizil et al. (2012b) introduces a dataset of Supreme Court conversations between justices and lawyers, where the power dynamics are clearly defined. Both in-domain and out-of-domain predictions demonstrate that this dataset can be utilized for learning about power dynamics in conversations. We train a separate BERT classifier using this dataset and apply it to the test dataset.

3.3.3 Human Judges

To better understand human performance on this task, we conduct a human evaluation on the same test dataset with six PhD students as judges. Each data point is decided by two human judges with an agreement of 63%, and a third judge resolves any disagreements.

3.4 Our Model

We construct three variants of factor graphs using the generated potentials described in 3.1: 1) only reason potentials are considered; 2) all reason and rebuttal potentials, along with conflicting relation between each reason-rebuttal pair, are considered; 3) all the reason and rebuttal potentials as well as all relation potentials are considered.

4 Results

Table 2 shows the main results. While individuals perceive power dynamics in conversations differently due to their diverse backgrounds, the suboptimal accuracy of human performance suggests that this predicting power relations in such setting is a challenging task. In zero-shot direct prompting, the accuracy decreases with the increasing context provided to the LLM, indicating that incorporating conflicting viewpoints complicates the decision-making process. All variants of our models show improved performance. The increasing

Model	Accuracy%
Human Judges	46.3
0-shot Conversation Only	49.0
0-shot w/Reasons	48.3
0-shot w/Reasons+Rebuttals	44.3
Bert In-Domain	55.0
Bert Out-of-Domain	51.7
Our Model Reasons Only	50.9
Our Model Reasons+Rebuttals	52.9
Our Model All Relations	54.3

Table 2: Experiment Results

Aspect	Top Reasons
Style	Conversational style enhances authority and influence.
Content	Expression of concern or hesitation suggests power and control.
Coordination	Initiating topics, steering discussions, and setting the tone reflect assertiveness and authority.
Engagement	Active participation, contribution, and engagement in a conversation indicate power.

Table 3: Summaries Reason Clusters Based on Aspects

performance with complete relations between variables suggests the model's ability to utilize all information into reasoning and predicting. The best performance comes from the classifier that trained on in-domain data, Ziems et al. (2024) argues that LLMs fail to outperform finetuned models in complicated social tasks, so the goal of our model is to reach this benchmark.

4.1 Analysis

Table 3 presents summaries of the top reasons clustered using BERTopic (Grootendorst, 2022). To identify weak reasons, we define them as those exhibiting high similarity to reasons supporting the opposing speaker. Table 4 reports the proportion of weak reasons conditioned on the four predefined aspects. Additionally, an example of a weak reason accompanied by a strong rebuttal is provided in the Appendix E.3.

For a more in-depth understanding of the results, we conduct statistical analysis to assess the performance distribution of our framework against human evaluation. The findings are presented in Appendix G.

5 Discussion and Summary

This paper presents a multi-agent probabilistic reasoning framework for analyzing conversations. We intentionally structure the agents' interactions to create an argumentation structure based on aspect-based reason-rebuttal pairs and capture global consistency between them, using LLM judgments. Our results demonstrate that each aspect of the model enhances performance, highlighting the potential of LLMs to transform social analysis tasks—provided they are leveraged through careful, structured problem decomposition.

Aspect	Weak Reason%
Style	25.4
Content	30.0
Coordination	44.5
Engagement	38.7

Table 4: Weak Reason Percentage Based on Aspects

Looking forward, we believe that this paper is only a first step in this direction, motivating several future research directions. First, our framework can be generalized to a broader range of social reasoning tasks. Second, we aim to explore the connection between our system and Formal Theories of Argumentation (FTA) (Dung, 1995; Dung et al., 2009; Prakken, 2010; Prakken et al., 2017). Our conjecture is that our structure can be mapped to a subset of FTA (i.e., our rules, such as reason-rebuttal, naturally align with the concept of defeaters in FTA). This connection has the potential to bridge LLM-based reasoning with theoretically grounded argumentation frameworks.

Acknowledgments

We thank the reviewers for their insightful comments that helped improve the paper. This work was partially supported by NSF CAREER award IIS-2048001 and the DARPA CCU program. The contents of this paper reflect the perspectives of the author(s) and do not necessarily represent the official views of, nor an endorsement by, DARPA, or the US Government.

Limitations

We believe our contributions algin with the scope of a short paper, and our findings align with prior work, highlighting a promising direction for further exploration. However, we recognize that additional research is needed to fully realize this framework's potential. In particular, cultural considerations must be addressed, as our judgments are based primarily on US-centric interactions. More work is also needed to evaluate the ability of Large Language Models to capture social interactions across diverse settings.

As an initial study, we acknowledge the need for further validation of the generated reasons and rebuttals, particularly their alignment with human judgments. Given the current performance of our framework, we do not consider it ready for casual use and recommend its application strictly for academic research purposes.

Ethics Statement

We acknowledge the risk that readers might be led to believe AI systems are capable of social reasoning. Such claims should be carefully evaluated, which is beyond the scope of this paper. Our human evaluations were collected voluntarily and required less than 30 minutes work.

References

Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen Mckeown, and Owen Rambow. 2012. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012a. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012b. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Phan Minh Dung, Robert A Kowalski, and Francesca Toni. 2009. Assumption-based argumentation. *Argumentation in artificial intelligence*, pages 199–218.

James A Evans and Pedro Aceves. 2016. Machine translation: Mining text for social theory. *Annual review of sociology*, 42(1):21–50.

Raymond W Gibbs. 2000. Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

T Guo, X Chen, Y Wang, R Chang, S Pei, NV Chawla, O Wiest, and X Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*. IJCAI; Cornell arxiv.

Judith T. Irvine. 1985. Status and style in language. *Annual Review of Anthropology*, 14:557–581.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore. Association for Computational Linguistics.

Michelle Lam, Catherina Xu, and Vinodkumar Prabhakaran. 2018. Power networks: A novel neural architecture to predict power relations. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 97–102, Santa Fe, New Mexico. Association for Computational Linguistics.

Yafei Liu, Hongjin Qian, Hengpeng Xu, and Jinmao Wei. 2020. Speaker or listener? the role of a dialog agent. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4861–4869, Online. Association for Computational Linguistics.

André Martins, Mário Figueiredo, Pedro Aguiar, Noah Smith, and Eric Xing. 2011. An augmented lagrangian approach to constrained map inference. pages 169–176.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 216–224, Nagoya, Japan. Asian Federation of Natural Language Processing.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344, Baltimore, Maryland. Association for Computational Linguistics.

Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124.

Henry Prakken et al. 2017. Historical overview of formal argumentation. *IfCoLog Journal of Logics and their Applications*, 4(8):2183–2262.

Mark Snyder and Arthur A Stukas Jr. 1999. Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in social interaction. *Annual review of psychology*, 50(1):273–303.

Karen Tracy and Jessica S Robles. 2013. *Everyday talk: Building and reflecting identities*. Guilford Press.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Likert Scale

This section presents the Likert scale used in prompts for assessing the similarity and contraction score between reasons.

A.1 Similarity

1: The reasons mention different behaviors of the speakers, and provide different reasoning of why they could be the one with higher power in the conversation.
2: The reasons mention somewhat similar behaviors of the speakers, but provide different reasoning of why they could be the one with higher power in the conversation.
3: The reasons mention somewhat similar behaviors of the speakers, and provide similar reasoning of why such behaviors could indicate higher power in the conversation.
4: The reasons mention similar behaviors of the speakers, and provide similar reasoning of why such behaviors could indicate higher power in the conversation.
5: The reasons mention the same behavior of the speakers, and provide very similar reasoning of why such behaviors could indicate higher power in the conversation.

A.2 Contradiction

1: The reasons mention somewhat similar behavior of the speaker, while provide different reasoning on how such behavior could indicate higher power in the conversation.
2: The reasons mention different behaviors of the speaker, and provide different reasoning of why such behaviors could indicate higher power in the conversation.
3: The reasons mention somewhat contradictory behaviors of the speaker, and provide different reasoning of why such behaviors could indicate higher power in the conversation.
4: The reasons mention somewhat contradictory behaviors of the speaker, but provide somewhat similar reasoning of why such behaviors could indicate higher power in the conversation.
5: The reasons mention contradictory behaviors of the speaker, and provide somewhat similar reasoning of why such behaviors could indicate higher power in the conversation.

B Aspect Definitions

We define the conversation aspects as the following:

Style: Style encompasses the tone, manner, and language used during the conversation. It can range from formal to informal, polite to blunt, friendly to hostile, etc.

Content: Content is the substance or subject matter of the conversation. It includes the topics being discussed, the information exchanged, and the sentence type used.

Coordination: Coordination is how participants manage turn-taking, interruptions, and transitions between topics. It involves maintaining a balance between speaking and listening, ensuring everyone has a chance to contribute.

Engagement: Engagement is the level of interest and involvement of participants in the conversation. Engaging conversations often involve asking

questions, sharing personal experiences, and expressing empathy.

C Prompts

In this section, we present all the prompts we use in the framework. For prompts with variables <high> and <low>, <high> designates the speaker assigned a high-power role by the LLM agent, while <low> represents the speaker assigned a low-power role.

C.1 Summary Prompt

```
In this task, you will summarize the <aspect> of the conversation based on the definition of <aspect> for each participant, A and B.

Definition:
<aspect definition>

Conversation:
<conversation>

Please provide the <aspect> summary of A and B separately. Provide the <aspect> summary of A on the first line, starting with "<aspect> of A: "; then provide the <aspect> summary of B on the next line, starting with "<aspect> of B: ".

<aspect> of A:
```

C.2 Reason Prompt

```
In this task, you will need to come up with the reasons for <high> has more power than <low> based on the conversation summaries.

Summaries:
<summary>

Please list three reasons to support <high> has more than <low>, one in a line, start with "-" and surrounded by quotes.

The reasons for <high> has more power than <low> are:
- "
```

C.3 Rebuttal Prompt

```
In this task, you are given a conversation and reason that supports <high> has more power than <low>. You will need to provide a rebuttal against this reason for <low> has more power than <high>.

Conversation:
<conversation>

Reason:
<reason>

Please provide the rebuttal, start with "-" and surrounded by quotes.

- "
```

C.4 Evaluation Prompt

C.4.1 Reason Validation

```
In this task, you will need to decide whether the reason is valid to indicate <high> has more power than <low> in a conversation between A and B. Respond with Yes or No. When uncertain, output No.

Reason:
<reason>

output:
```

C.4.2 Reason Grounding

```
In this task, you are given a conversation and a reason for why <high> has more power than <low> based on the conversation. You will need to decide whether this reason can be grounded through the conversation. Respond with Yes or No. When uncertain, output No.

Conversation:
<conversation>

Reason:
<reason>

output:
```

C.4.3 Rebuttal Validation

In this task, you will need to decide whether the Rebuttal is valid to counter the Reason to indicate <high> has more power than <low> in a conversation between A and B. Respond with Yes or No. When uncertain, output No.

Reason:
<reason>

Rebuttal:
<rebuttal>

output:

C.4.4 Rebuttal Grounding

In this task, you are given a conversation and a reason. You will need to decide whether this reason can be grounded through the conversation. Respond with Yes or No. When uncertain, output No.

Conversation:
<conversation>

Reason:
<reason>

output:

C.5 Relation Assessment

C.5.1 Similarity

In this task you are given two descriptions [1] and [2] about the power dynamics of the the same conversation between two speakers, A and B. Give a similarity score of these two descriptions based on the following rubrics.

Rubrics:
1: Description [1] and [2] mention different behaviors of A and B, and provide different reasoning of why they could be the one with higher power in the conversation.
2: Description [1] and [2] mention somewhat similar behaviors of A and B, but provide different reasoning of why they could be the one with higher power in the conversation.
3: Description [1] and [2] mention somewhat similar behaviors of A and B, and provide similar reasoning of why such behaviors could indicate higher power in the conversation.
4: Description [1] and [2] mention similar behaviors of A and B, and provide similar reasoning of why such behaviors could indicate higher power in the conversation.
5: Description [1] and [2] mention the same behavior of A and B, and provide very similar reasoning of why such behaviors could indicate higher power in the conversation.

Descriptions:
<description1>
<description2>

In your response, provide the similarity score of [1] and [2]. Only print '1', '2', '3', '4' or '5'.

Score:

C.5.2 Contradiction

In this task you are given two descriptions, [1] and [2], about the power dynamics of the same conversation between two speakers, A and B. Both descriptions support the same speaker, A or B, for holding higher power in the conversation. Give a score on how contradicting the descriptions are based on the following rubrics.

Rubrics:
1: Description [1] and [2] mention the somewhat similar behavior of the speaker, while provide different reasoning on how such behavior could indicate higher power in the conversation.
2: Description [1] and [2] mention different behaviors of the speaker, and provide different reasoning of why such behaviors could indicate higher power in the conversation.
3: Description [1] and [2] mention somewhat contradictory behaviors of the speaker, and provide different reasoning of why such behaviors could indicate

```
higher power in the conversation.
4: Description [1] and [2] mention somewhat contradictory behaviors of the speaker, but provide somewhat similar reasoning of why such behaviors could indicate higher power in the conversation.
5: Description [1] and [2] mention contradictory behaviors of the speaker, and provide somewhat similar reasoning of why such behaviors could indicate higher power in the conversation.

Descriptions:
<description1>
<description2>

In your response, provide the similarity score of [1] and [2]. Only print '1', '2', '3', '4' or '5'.

Score:
```

D Variable Potential Calculation

We prompt GPT-3.5-Turbo with yes/no answer for whether a reason is valid or grounded. We use the average of generated token probabilities for validity and grounding as the potentials for a reason or rebuttal. We optionally apply a logit function over the probability to penalize lower probability potentials.

Reason variables, rebuttal variables, and relation variables are associated with optional weights, (default set to 1), that can scale all reasons, or rebuttals, or same/contradict relations, and is used as a hyperparameter to adjust its influence in solving the factor graph. Each aspect is also associate with an optional weight, (default set to 1), that can be applied to all reasons from that aspect.

E Examples

This section provides an example of an input conversation snippets and LLM generated content following the framework.

E.1 Conversation

A: Right.
B: And then do some segmenting and recognition - initial recognition would be interesting to do. Yeah, although it - it - it - it may be separating out these numbers from the rest. Yeah.
A: That's what I mean.
B: And then
A: Yeah just doing a digits on it - uh, connected digits.
B: Yeah and uh

E.2 Summaries

E.2.1 Content

Content of A: A mentioned the need for segmenting and recognizing, specifically mentioning separating out numbers from the rest and focusing on connected digits.
Content of B: B also discussed segmenting and recognizing, and emphasized the importance of initial recognition and separating numbers from the rest.

E.2.2 Coordination

Coordination of A: A maintains a balanced turn-taking pattern with B, allowing them both to contribute equally to the conversation. A listens actively to B's suggestions and adds their own input, ensuring a smooth transition between topics and avoiding interruptions.
Coordination of B: B utilizes turn-taking effectively with A, allowing for a back-and-forth exchange of ideas. B responds to A's comments and builds upon them, demonstrating active listening and contributing to the conversation in a collaborative manner. The conversation flows smoothly between B and A, with both participants having the opportunity to share their thoughts.

E.2.3 Engagement

Engagement of A: A seems to be actively engaged in the conversation by demonstrating understanding of the points made by B and adding to the discussion by suggesting specific ideas like segmenting and recognizing numbers.
Engagement of B: B also seems engaged in the conversation by sharing their thoughts on segmenting and recognizing numbers, showing interest in the topic and actively participating in the exchange of ideas with A.

E.3 Pairs of reasons and rebuttals

R_A **(reasons that support Speaker A):** A is able to communicate in a more concise and direct manner, which can be more influential in decision-making processes.

Variable/Aspect	Weight
Reason	20
Rebuttal	18
Same	1
Contradict	1
Style	1
Content	1
Coordination	1
Engagement	1

Table 5: Factor Graph Hyperparameters

R_A^b **(rebuttals that counter** R_A**):** A may communicate in a more concise manner, but that does not necessarily equate to having more power. B's ability to have a thorough understanding and analysis of the situation can also be influential in decision-making processes. Just because A's communication style is more direct does not automatically mean they hold more power.
R_B **(reasons that support Speaker B):** B demonstrates a greater level of technical expertise and focus on the task at hand compared to A.
R_B^b **(rebuttals that counter** R_B**):** Technical expertise and focus on the task at hand do not necessarily equate to having more power in a conversation. Power dynamics are influenced by various factors such as communication style, assertiveness, and persuasiveness, which may vary between individuals regardless of technical expertise.

F Model Parameters

F.1 Factor Graph

One instance of a full factor graph contains 98 variables and 75 factors. The weights used for variables and aspects for the best model are shown in Table 5.

F.2 BERT Classifier

We trained 'bert-base-uncased' model for 3 epochs with learning rate $2e-5$ for both in-domain and out-of-domain training dataset.

G Statistical Analysis

We perform a t-test using the results of human judgment and our best-performing model, yielding $t = -1.87$ and a p-value of 0.062. Additionally, a McNemar test results in a p-value of 0.059.

While both tests fail to reject the null hypothesis, the p-values are close to the 0.05 threshold. This suggests that further investigation using larger datasets may provide deeper insights into the approach.