How did we get here? Summarizing conversation dynamics

Anonymous ACL submission

Abstract

Throughout a conversation, the way participants interact with each other is in constant flux: their tones may change, they may resort to different strategies to convey their points, or they might alter their interaction patterns. An understanding of these *dynamics* can complement that of the actual facts and opinions discussed, offering a more holistic view of the trajectory of the conversation: how it arrived at its current state and where it is likely heading.

In this work, we introduce the task of summarizing the dynamics of conversations, by constructing a dataset of human-written summaries, and exploring several automated baselines. We evaluate whether such summaries can capture the trajectory of conversations via an established downstream task: forecasting whether an ongoing conversation will eventually derail into toxic behavior. We show that they help both humans and automated systems with this forecasting task. Humans make predictions three times faster, and with greater confidence, when reading the summaries than when reading the transcripts. Furthermore, automated forecasting systems are more accurate when constructing, and then predicting based on, summaries of conversation dynamics, compared to directly predicting on the transcripts.

1 Introduction

Conversations take place on at least two different levels (Tannen, 2005). On one level, participants directly communicate ideas, facts, and opinions, providing the topical context of the discussion. On the other level, the *dynamics* of their interactions reveal how they feel about each other, through changes in their tone—e.g., polite (Lakoff, 1973; Brown, 2015), condescending (Huckin, 2002), or sarcastic (Jorgensen, 1996)—, conversational strategies they employ—e.g., rhetorical questions (Han, 2002)—and even the patterns of their exchanges (Sacks et al., 1974; Silverstein, 1984).

Traditional summary: In this online conversation, participants discuss capitalism and its sustainability. Speaker1 argues that profit always trumps morals and ethics in business. Speaker2 disagrees, stating that unchecked capitalism is economically unsustainable and detrimental to human dignity. Speaker4 seeks clarification on the consequences of capitalism's unsustainability, and Speaker2 suggests it could lead to economic collapse, mass unemployment, and toxic environments. They emphasize the need for checks on capitalism to prevent these issues. The conversation highlights concerns about unregulated capitalism and its impact on society.

Summary of conversation dynamics (human-written): Several users discuss regulation of capitalism. Speaker2 shares an opinion which Speaker4 questions. Speaker2 addresses Speaker4's question in a sincere manner. Then, the overall tone shifts to aggressive and confrontational when Speaker4 rudely invalidates Speaker2's response. Speaker2 sarcastically criticizes Speaker4's attitude which aggravates Speaker4 more. Then, Speaker2 genuinely apologizes and elaborates on their point and Speaker4 gratefully acknowledges this and clarifies their intentions. They continue the discussion in a calmed down, civil tone.

Summary of conversation dynamics (generated): Four speakers engage in a discussion about the nature of capitalism and its consequences. Speaker2 expresses concerns about unchecked capitalism and argues for the need to consider human dignity. Speaker4 challenges Speaker2 to explain the consequences of an economically unsustainable capitalist system. The conversation becomes heated, with Speaker2 perceiving Speaker4's questions as confrontational. Speaker2 defends their views and provides examples to support their argument. The overall tone of the conversation remains argumentative, but civil.

Figure 1: Traditional and dynamics summaries for the same conversation (transcript in Appendix F). Elements of conversational dynamics are colored in blue.

A holistic description of a conversation and its trajectory requires accounting for both of these communication levels. We complement prior work that has largely focused on summarizing the topical context of the discussion (Yang and Zhu, 2023), by introducing the task of generating summaries that instead capture the dynamics of the interaction between the participants. As shown in Figure 1, these cover aspects lost in a traditional summary.

Summaries of conversation dynamics (or SCDs for short) provide a way for humans to quickly understand the trajectory of a discussion: what type of interactions lead to its current state, and how these are likely to develop? This type of understanding can benefit various applications, including supervision of conversations in time-sensitive domains (e.g., online community moderation and mental health crisis counseling), providing context to users (re)joining an online conversation, or identifying and reviewing common problems in human-human or human-AI conversations (Section 8).

However, generating SCDs that effectively capture conversation trajectories presents several new challenges. While prior computational work introduced models for separately capturing individual aspects of conversation dynamics (Section 6), an effective and concise summary must *select* those that are most relevant for understanding the trajectory of the conversation. Additionally, an informative summary must not simply identify these aspects separately, but should also describe how they evolve and interrelate throughout a conversation: for example, a conversation that transitions from an aggressive tone to a calmer one has a completely different trajectory than one that proceeds in reverse order. Thus, to provide an understanding of the trajectory of a conversation, an SCD must *synthesize* different aspects of its dynamics across multiple utterances and participants.

As a first step, we devise a multi-step procedure for human annotators to collaboratively write SCDs. Importantly, this procedure is designed to address the selection and synthesis challenges described above. Building on this procedure, we develop a large language model prompt for generating SCDs and compare them with summaries generated by other baselines, including traditional summaries.[1]

Specifically, in this paper we evaluate the usefulness of SCDs for conversation trajectory understanding via an established task: forecasting whether an ongoing conversation will eventually derail into toxic behavior (Zhang et al., 2018a; Liu et al., 2018). While prior attempts at this task started directly from the transcript (Section 6), we explore generating SCDs as an intermediate step. This approach has the potential advantage of adding interpretability to automated forecasting systems and improving efficiency for humans (such as moderators) trying to make such judgments (Schluger et al., 2022).

Our findings reveal the potential of SCDs to help both humans and automated systems understand a conversation's trajectory, motivating further work on this new task. In the downstream task of forecasting the future derailment of a conversation, humans make predictions three times faster, and with greater confidence, when reading the SCDs than when reading the transcript. Furthermore, automated systems are more accurate when constructing, and then predicting based on, SCDs compared to state-of-the-art systems that base their forecast directly on the transcript. Finally, by comparing human-written and machine-generated summaries, we reveal a quality gap that motivates further computational work on this new task.

In summary, this work:

1. introduces the task of summarizing conversation dynamics, together with a collection of human-written summaries;

2. proposes a downstream evaluation method that allows for comparison between methods for generating them;

3. shows the usefulness of dynamics summaries, motivating further work on this new task.

2 Human-written Summaries

To start, we introduce a procedure for writing SCDs and a collection of such summaries for an existing dataset of online conversations.

Procedure for writing summaries. To construct the first collection of SCDs, we iteratively designed a writing procedure that addresses the selection and synthesis challenges described in the introduction. In early iterations in which we asked a single annotator to both read the transcript and write its SCD, we observed that they consistently omitted key information that they take for granted, perhaps because some aspects of the dynamics are often processed non-consciously (Tannen, 2005). To address this issue, we devise a procedure that uses interaction between two annotators to surface key elements of the conversation dynamics that readers *who can not see the transcript* would consider relevant. Thus, we settle on a procedure that has two parts—one in which an annotator works individually and one in which they interact with another

[1]We include the collection of human-written summaries and sample model outputs as supplemental material and will make them and our code publicly available upon acceptance to encourage further work on this problem.

annotator—which we briefly outline here (and detail in Appendix A).

For the individual work, Annotator A will draft several summaries for a transcript in 4 steps:

1. skim over the transcript to have an overview of the topic and of the role of each speaker;
2. read the transcript utterance by utterance and write a comprehensive summary, including opinions and arguments expressed within most utterances, turning points, and elements of conversation dynamics;
3. condense the summary by selecting key points and dynamics and replacing specific opinions and arguments with high-level descriptions;
4. write a brief summary for each of the main speaker, focusing on (the changes in their) tone and on their conversational strategies.

In the interactive part, Annotator B will write the SCD, by interacting with Annotator A with a goal of understanding the conversation trajectory. In this process, Annotator B may read the summaries written in the previous steps by Annotator A, but not the transcript, and may make inquiries on details they deem important to understand the trajectory, such as 'was this said neutrally, or is there something about the tone that I should note?' or 'is the comment overtly rude, or is it just passive-aggressive or blunt?', surfacing key aspects that were not explicitly mentioned in Annotator A's summaries.

Conversation transcripts data. We apply this procedure to summarize conversations from the Conversations Gone Awry (CGA) dataset (Chang and Danescu-Niculescu-Mizil, 2019),[2] a conversation corpus collected from the 'Change My View' subreddit, where people actively seek to have others challenge their views on controversial topics. This community has been studied extensively in part because of their explicit norms against toxic behavior, and corresponding labels inferred from the moderators' interventions.

In the CGA corpus, conversations are paired such that every conversation that derailed—i.e., ended in a toxic comment removed by moderators—is matched with another conversation on the same topic that did not. For us, these labels provide an opportunity to test the extent to which SCDs provide an intuition about the future trajectory of the conversation (i.e., will it derail or not). To focus on the *future* trajectory, we remove the last 3 utterances from every conversation (in addition to the toxic comment, if there was one). Since our interest is in summarization, we only keep pairs where both conversations are longer than 10 utterances.

Collection of human-written summaries. We produce human-written summaries for 50 conversations from the train split of CGA. The summary writing process took roughly 240 annotator-hours.[3] Summaries are on average 66 words long (annotators are instructed to keep them under 80); for comparison, the transcripts are on average 984 words long. An example summary is shown in Figure 1, and a qualitative analysis is provided in Section 5.

Informativeness check. Before we proceed, we check whether the summaries are actually informative. Given their highly abstractive nature, there is a risk that they become so general as to not distinguish between different conversations (e.g., 'Speaker1 disagreed with Speaker2.' would apply to most of the conversations in the data). We devise a procedure for systematically checking whether this is the case, avoiding subjective interpretations that are prone to apophenia.

We ask new annotators to read a transcript, and then present them with a multiple-choice question. Each choice corresponds to a summary segment involving two speakers. One of the choices is from the actual summary of the provided transcript, while the other two are distractors: one from the summary of the paired conversation (thus, on the same topic, but with the opposite derailment label) and the other from the summary of another conversation with the same label as the transcript, but on a different topic. This way, neither the topic nor the label fully reveals the answer: to be identified correctly, the segment must contain information that matches the transcript better than the distractors.

For example, for our introductory example, three choices could be: "SpeakerX sarcastically criticizes SpeakerY's attitude which aggravates SpeakerY more." (actual segment), "SpeakerX poses a rhetorical question, which SpeakerY contradicts sarcastically, raising the tension and causing SpeakerX to disagree rudely." (same-pair distractor), "SpeakerX first shares their opinion and later poses rhetorical questions, and SpeakerY disagrees in a matter-of-fact manner." (same-label distractor).

[2] Accessed via the ConvoKit library (Chang et al., 2020).

[3] For each conversation transcript, the individual part takes about 2 hours and the interactive part takes about 20 minutes.

Though we designed this procedure to avoid excessive workload when evaluating informativeness,[4] each question still requires reading one transcript and carefully checking the segment choices against it. Therefore, we limit our total number of questions to 10, covering 30 conversations through distractors. (Further details in Appendix B.)

Two annotators completed the task, one of which answered 10 out of 10 questions correctly and the other answered 8 of them correctly (noting low confidence on the 2 answers they got wrong), suggesting that our summaries indeed pass this basic informativeness check.

3 Machine-generated Summaries

We now turn to explore several simple baselines for generating SCDs, setting the stage for developing more specialized methods in future work. The GPT-family models have shown remarkable results in various summarization benchmarks (Zhang et al., 2023; Yang et al., 2023). Among them, ChatGPT is particularly suitable for adapting to new tasks like ours without demanding a sizable train set. Thus, for the first group of baselines, we query OpenAI's ChatGPT (GPT-3.5-turbo-0613) API with default parameters using different prompts, from the most common prompt for traditional summarization tasks to prompts inspired by the procedure we developed for humans:[5]

Traditional prompt. After experimenting with several prompts on a development set, we use a concise prompt for our traditional summarization baseline: 'briefly summarize the following online conversation in 80 words.' Figure 1 includes a traditional summary generated by this prompt.

Zeroshot prompt. We devise a prompt that explicitly integrates our goal of generating summaries that can help people understand the conversation trajectory. After experimenting with several word choices for referring to trajectory, dynamics, and specific dynamics elements, we settle on a concise prompt, 'write a short summary capturing the trajectory of the online conversation' with additional constraints such as not including specific arguments and capturing elements of tone and conversation strategies (Figure 3 in the Appendix).

Procedural prompt. We build on the insights we gathered from developing the procedure for human annotators (Section 2) to construct a more elaborate prompt. This prompt (Figure 3 in the Appendix) thus includes instructions adapted from those provided to the annotators, together with examples that they found useful for understanding the instructions. Because we only include segments of summary examples instead of complete transcript and summary pairs, the procedural prompt can be positioned in-between zeroshot and few-shot in-context learning. Figure 1 shows the procedural prompt summary for our introductory example.

We also experimented with few-shot in-context learning on a small subset of the train set, but manual inspection did not reveal an increase in quality. Thus, due to significantly higher API costs, we did not pursue this path. Appendix D includes more discussion on our prompt engineering.

Finetuning. Finally, we experimented with finetuning GPT-3.5-turbo as well as with smaller dialog summarization systems (BART-large and DialogLED) (Lewis et al., 2020; Zhong et al., 2022) using the 50 human-written summaries and the corresponding transcripts. We provide details in Appendix D.

4 Downstream Evaluation: Forecasting Derailment

Popular metrics for summarization—e.g., ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and QA-based metrics—are notoriously unreliable when evaluating LLM-generated summaries or summaries of long documents (Goyal et al., 2023; Koh et al., 2022). We thus follow recommendations of Deutsch et al. (2021) and perform a downstream evaluation, in which we quantify the extent to which SCDs provide an understanding of the conversation trajectory.

Specifically, we choose the task of forecasting whether a conversation will eventually derail into toxic behavior (Zhang et al., 2018a). Unlike previous work in which the prediction was made based on a truncated transcript of the conversation (for a comprehensive discussion of prior models see Section 6), here we aim to make the prediction directly

[4] An equivalent check could be implemented by providing one summary segment and three transcripts to pick from. This method corresponds to the existing literature in communication-based evaluations for natural language generation (Newman et al., 2019), and specifically the idea that an informative summary should capture the salient information that makes the source text stand out with respect to other related texts (Zhang et al., 2018c). However, this equivalent method would require substantially longer time due to the lengths of the transcripts.

[5] We prompt the model to generate summaries of at most 80 words and set the max new token limit to 128 (corresponding to approximately 96 words) as a hard limit .

on the SCD of that truncated transcript. In addition to providing means to evaluate and compare current and future models for generating SCDs, this derailment forecasting task is also important in itself, as it was shown to enable important practical applications: automated forecasts can be used to inform users during ongoing discussions (Chang et al., 2022) while human forecasts are made by moderators in their everyday workflow (Schluger et al., 2022) (see Section 8 for practical and ethical considerations of real-world deployment).

We first compare the usefulness of SCDs for automated forecasting systems. Then we devise an experiment to estimate their usefulness for human forecasts. Throughout, the forecasts are done on a balanced dataset of derailing and non-derailing conversations paired by topic, following the setup of the CGA dataset (Section 2); thus the overall topic of the discussion plays a minimal role and the random baseline is 50%. To leave room for future work and avoid polluting the available data, we leave the original CGA test set untouched. Using truncated transcripts from the original train split, we construct a new train set (234 conversations), a new dev set (100), and a new test set (100); the new test set includes the 50 conversations for which we also have human-written summaries (Section 2).

4.1 Useful for automated forecasts?

We train classifiers to predict if a conversation will eventually derail based on the various types of summaries of the truncated transcripts. We adopt GPT-3.5-turbo to develop few-shot classifiers for each summary type, using examples from outside the test split. We also train BART (Lewis et al., 2020) and longformer (Beltagy et al., 2020) as finetuned classifiers. To provide more robust estimates, for each summarization method we generate 4 different summaries for each conversation, and average the classifiers' performance on them (details in Appendix D.3). We find GPT-3.5 few-shot classifier gives the best performance across all types of summaries, so we use it for our main analysis here. While the performance of other classifiers is substantially lower (Appendix D.3), the comparisons discussed below still hold.

Comparison of summaries. As shown in Table 1, the classifier based on the procedural prompt achieves the best accuracy, significantly outperforming the other types of summaries ($p < 0.05$, throughout we use the Wilcoxon signed-rank test significance testing). In particular, the information conveyed by the SCDs generated with the procedural prompt appears to be more useful for the automatic derailment forecaster than that included in traditional summaries. Other metrics (Macro-F1, precision, recall) support the same conclusion (Appendix F.5).

Based on...	**Accuracy**
transcripts (CRAFT classifier)	56.2
transcripts (GPT-16k classifier)	60.0
traditional prompt summaries	58.3 (5.85)
zeroshot prompt summaries	58.8 (6.24)
procedural prompt summaries	67.3* (2.63)

Table 1: Derailment forecasting results for systems based on truncated transcripts and on different types of machine-generated summaries. For summary-based systems, we report standard deviation across 4 summary-generation trials, and indicate with * the highest performance (p<0.05, Wilcoxon signed-rank test). Results for the GPT-3.5-turbo few-shot classifiers are shown unless otherwise noted.

The finetuned models—finetuned on the 50 human-written examples and evaluated on the remaining of the test set—perform worse than the procedural prompt on the same set (Appendix D). This could be due to the relatively small collection of human-written summaries, as well as the generic fine-tuning methodology.

Summary vs transcript. For reference, we also include two baselines operating directly on the truncated transcripts. The first baseline, CRAFT, was introduced before the advent of the LLM era and remained the state-of-the art system for this task (Chang and Danescu-Niculescu-Mizil, 2019).[6] The second baseline is a few-shot GPT-3.5-turbo-16k classifier, which can take up to 16k tokens to cope with the greater input lengths of the transcripts.[7]

As shown by Table 1, predictions based on procedural prompt outperform those based directly on the transcripts. This suggests that SCDs are effective in distilling from the transcripts information that is useful for the forecasting task. Perhaps more importantly, the feasibility of this 'summarize-then-

[6]For a fair comparison, we modify the ConvoKit implementation of CRAFT (Chang et al., 2020) to trigger forecasts exactly 3 utterances before the end of the conversation. This setup is harder than the original setup in which the system could make predictions all the way up to right before the attack or the end of the conversation, potentially having access to more explicit signals of upcoming toxic behavior.

[7]Both baseline systems might have an advantage in that they might have accessed the full untruncated transcripts during pre-training.

forecast' approach points out a promising future direction for improving the interpretability of the user-facing forecasting systems, where the summary could be presented as an easily digestible rationale for the prediction. In fact, users of such systems have identified the lack of explanations as one of their most important drawbacks (Chang et al., 2022).

4.2 Useful for human forecasts?

We now switch to the other main motivation: can SCDs help *humans* quickly grasp the trajectory of a conversation? To answer this question we devise an experiment in which subjects are asked to guess whether a conversation will eventually derail based either on a transcript or its SCD. We compare both their accuracy and efficiency, in terms of the time they spend to make their guess, as well as their confidence in their guess.

To better focus our resources, we use a subset of 20 paired conversations out of those for which we created human summaries. In addition to the transcripts and the human-written summaries, we also consider the corresponding procedural prompt summaries (since those were shown to fare best in the automatic prediction task).

We recruit 20 university students fluent in English as participants. A subset of participants make their guesses based on the transcripts only, while another subset will make guesses based on summaries only. Each participant in the latter subset will see a mix of human-written and machine-generated summaries (without being aware that these are produced differently) such that any observed differences between them cannot be attributed to participant idiosyncrasies. In addition to providing a guess of whether the conversation will derail or not, each participant is asked to rate their confidence in their guess (on a scale from 1 to 5). We also record the time it took for the participants to make their guess (starting from the time they see the transcript or summary until the time they select their guess), and instruct them to work on each question without pausing. The specific instructions and details about how participants are grouped are in Appendix C.

Unlike in the automatic evaluation in Section 4, we adopt a zero-shot prediction setting, in which humans do not have labeled examples of summaries (or transcripts) to assist their guessing. This way, we can better test if the summaries are immediately intuitive to humans rather than testing the participants' ability to learn patterns that might not be visible to untrained individuals. This means, however, that the accuracies of the human participants are not directly comparable with those of the automated system.

Summaries vs transcripts. As shown in Table 2, participants can make guesses 3-4 times faster based on SCDs while maintaining similar accuracy. This improvement in efficiency is critical for applications such as proactive online moderation, as earlier work has found that moderators are faced with "too many [potentially at-risk conversations] to proactively monitor (Schluger et al., 2022)."

Human vs generated summary. Participants are significantly more confident when making predictions based on human-written summaries than on machine-generated summaries (and even on the transcript).[8] This gap is important for applications where summaries are used for decision-making (e.g., moderation) and motivates future work on improving summarization models. Another noticeable difference is that machine-generated summaries provide a better understanding of the topical content of the discussion, perhaps to the detriment of better coverage of aspects of conversation dynamics. In Section 5 we further explore this trade-off via a qualitative analysis of the summaries.[9]

Based on...	Acc	Conf	Topic	Time
transcripts	60	3.5	-	132
gen. summ.	59	3.6	3.9	45*
human summ.	62	4.0*†	3.4†	31*

Table 2: Results on the human forecasting experiment. "gen. summ." refers to the summaries generated using the procedural prompt. Time is measured in seconds. * indicates a significant difference when compared with transcripts ($p<0.05$, Wilcoxon signed rank test), † indicates a significant difference when comparing human-written with automated summaries ($p<0.05$).

5 Qualitative Analysis

To complement our quantitative evaluation and understand what might drive the differences between human and machine-generated summaries, we now

[8] This difference continues to hold when only considering correct guesses. Also, reassuringly, confidence in correct guesses is higher than in incorrect ones throughout.

[9] We also experimented with directly asking participants to report their understanding of the trajectory of the conversation, on a scale from 1 to 5. There was no significant difference between human and machine-written summaries (4 and 3.9 respectively), perhaps due to the difficulty of briefly explaining what a trajectory is and how it differs from the guess, a confusion that surfaced during debriefing.

turn to the actual content of the SCDs. Through a close reading of the 20 human-written and 20 machine-generated summaries used in the experiment described above, we identify, annotate, and compare several aspects that were shown to provide clues about the conversation trajectories.

Tone. Whether 'polite,' 'rude,' 'aggressive,' 'condescending,' or 'sarcastic' (Tannen, 2005; Brown and Levinson, 1987), the tone employed by the participants is a prominent feature of the SCDs. Tone can be explicitly stated, as in 'Speaker1 disagrees [...] **in a somewhat passive-aggressive tone**.' Other times, especially in human-written summaries, it is expressed as modifying a speech act, as in 'contradicts **sarcastically**', 'disagrees **rudely**,' and '**admantly** defends.' Overall, tone is indicated less frequently in the machine-generated summaries (75% of them mention tone at least once) than in the human-written summaries of the same conversations (all mention tone at least once), suggesting a potential path for improvement.

Changes in tone. Tone can evolve throughout a conversation, and changes in tone can provide an intuition about its trajectory (Niculae et al., 2015). When participants use an '**increasingly** passive aggressive tone,' or when the 'tension **rises**' the conversation seems more likely to be getting out of hand than when a 'slight tension [...] is **maintained but doesn't escalate**' or when the 'tone **remains** argumentative but civil'. The latter quote is an example of an overall assessment of tone dynamics that both humans and (more commonly) automated systems sometimes include at the end of the summary, even though neither is explicitly instructed to do so. Overall, 75% of the human summaries feature phrases explicitly mentioning changes in tone whereas only 50% of the machine-generated counterparts do so.

Patterns of interaction. Beyond the content of the messages, the structural properties of the interactions were shown to be indicative of future trajectories (Backstrom et al., 2013; Zhang et al., 2018b). Two participants can have a '**brief exchange**' or an extended '**back-and-forth**', which can be interrupted when another participant '**jumps in**'. While explicit mentions of such patterns are relatively rare (found in 45% of the human summaries and 31% of the machine-generated summaries), they can often be inferred by following the sequence of speakers mentioned in the summaries.

Conversation strategies. Interlocutors employ strategies that can put the conversation on various trajectories. For example, 'pos[ing] a rhetorical question' or 'questioning each other's logic', can often lead to personal attacks (Habernal et al., 2018), whereas expressing uncertainty about one's own view (e.g., via hedging), would soften an impending disagreement and prevent the escalation of tension (Zhang et al., 2018a). 'Supporting [a] point with evidence', 'justifying objective claims with personal experiences', 'draw[ing] a comparison' or 'question[ing] the importance of specific details' are classic persuasion strategies (Zeng et al., 2020; Li et al., 2020a). A list of strategies considered in this analysis is included in Appendix E. Overall, we find that mentions of strategies are evenly distributed among human-written (80%) and machine-generated summaries (85%).

Topical context. Finally, these dynamics can only exist in the context of the content being discussed. Though not the primary focus of SCDs, a small amount of topical context is needed to provide a scaffolding for the phenomena discussed above. Both human and machine-generated summaries generally start with a sentence about the general topic of the discussion. Beyond that, machine-generated summaries include substantially more topical context to the detriment of actual aspects of conversation dynamics, despite the explicit instruction and in-context-learning examples against this behavior. This echoes the subjective ratings of the participants in the human forecasting experiment (Table 2). This phenomenon suggests that in-context learning is not sufficient to 'untrain' LLMs from the traditional summary examples seen in pre-training. This motivates developing models that are specifically designed to select and synthesize aspects of conversation dynamics, perhaps inspired by the interactive human-writing procedure.

6 Further Related Work

Our work falls at the intersection of three broad areas of NLP: studies of conversation dynamics, summarization, and conversation forecasting.

Conversation dynamics. We are primarily inspired by extensive computational work on modeling various aspects of conversation dynamics. Some studies have focused on identifying specific aspects, such as such as politeness (Burke and Kraut, 2008; Danescu-Niculescu-Mizil et al., 2013; Li et al., 2020b), formality (Pavlick and Tetreault, 2016; Krishnan and Eisenstein, 2015), passive-aggressiveness (Chhaya et al., 2018), condescen-

sion (Wang and Potts, 2019) or sarcasm (Oraby et al., 2017). Others have studied changes along these dimensions during the discussion (Wang and Cardie, 2014; Niculae et al., 2015; Niculae and Danescu-Niculescu-Mizil, 2016). A separate but related thrust focused on persuasive strategies interlocutors employ in a conversation, mostly in the context of debates (see Lawrence and Reed (2020) for a survey). Unlike these studies, the goal of SCDs is not to exhaustively identify occurrences of either one of these phenomena, but to convey how such key aspects combine towards an understanding of the conversation's trajectory.

Dialogue summarization. The vast majority of dialogue summarization systems focus on the content of the utterances, rather than on the more subtle non-topical dynamics. Early approaches to dialogue summarization focused on using external tools to explicitly model dialogue structures, such as topic segmentation and conversation stages (Li et al., 2019; Chen and Yang, 2020), dialogue acts (Goo and Chen, 2018), along with discourse dependency and speaker-action relations (Chen and Yang, 2021), which are processed into features that can help language models. Later, pretraining on dialogue corpora also attracted increasing research interest and achieved state-of-the-art results on many datasets (Zhong et al., 2022). Most recently, extensively pretrained, instruction-tuned, LLMs, such as the GPT-family models, have achieved superior results on various summarization leaderboards (Goyal et al., 2023; Zhang et al., 2023; Yang et al., 2023). In dialogue summarization, these instruction-tuned LLMs possess strong in-context-learning capabilities (Wu et al., 2023), making them strong candidates for solving new summarization tasks that have limited training data.

Conversation forecasting. We motivate and evaluate dynamics summaries with applications requiring an understanding of a conversation's trajectory. Beyond forecasting derailment (Zhang et al., 2018a; Liu et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019), other tasks include forecasting thread growth (Backstrom et al., 2013), prosocial outcomes (Bao et al., 2021), editorial decisions (Mayfield and Black, 2019), the outcomes of negotiations (Chawla et al., 2020) or team resilience (Whiting et al., 2019). It would be interesting to consider the extent to which SCDs can aid with these other forecasting tasks, and what modifications might be needed to summaries specifically dedicated to these tasks.

To the best of our knowledge, all conversational forecasting work operate directly on conversation transcripts. The early work by Chang and Danescu-Niculescu-Mizil (2019) adopts a recurrent network and applies unsupervised training to learn a representation of conversation dynamics. More recently, Kementchedjhieva and Søgaard (2021) explores pretraining and various training paradigms for this task, Altarawneh et al. (2023) applies a graph convolutional network, and Yuan and Singh (2023) uses a hierarchical transformer-based framework to combine utterance-level and conversation-level information. However, since it aims to guess the future, this task remains challenging and none of the mentioned approaches achieved an accuracy beyond 65% on the CGA balanced dataset.

Unlike detecting the toxic language after the fact (Wulczyn et al., 2017; Breitfeller et al., 2019), the signs of future derailment are subtle and require a more thorough understanding of the conversation trajectory. Our results suggest that SCDs can provide this information concisely and effectively, suggesting a new summarize-then-forecast approach to conversational forecasting tasks. This inspires future work that integrates SCDs in real-time forecasting systems, which would require tackling shorter conversations where summaries might not be appropriate, as well as the 'unknown horizon' problem: not knowing when to trigger the prediction (Chang and Danescu-Niculescu-Mizil, 2019).

7 Conclusions

In this work, we introduce the task of summarizing the dynamics of interaction between participants in a text-based conversation. By introducing human and automated procedures for writing such summaries, we show that they can capture information that is mostly missing from traditional summaries, such as the tone in which the participants write and how it changes throughout a conversation. Summaries of these dynamics are useful to both humans and automated systems for understanding the overall trajectory of the conversation, as shown through the downstream evaluation task of forecasting whether a conversation will eventually derail or not. Humans can make similarly accurate forecasts three to four times faster by starting from SCDs than by reading the transcripts. Automated systems can surpass the performance of state-of-the-art forecasting systems when generating SCDs as an intermediate step for forecasting.

8 Limitations

This work, however, only takes the first steps towards solving and evaluating this task automatically. In fact, we show that there is a substantial gap remaining between human-written summaries and machine-generated ones. Since in this work we focus on defining the task and demonstrating its feasibility, we only employ simple prompting and standard fine-tuning procedures. This sets the stage for the future development of more specialized models and training regimes. These models could be more tightly integrated with the downstream task, learning to select aspects of the dynamics that are most relevant as well as to determine the right level of abstraction.

To continue improving on dynamics generation models, more diverse automated evaluation methods are required. Given the highly abstractive nature of the task, traditional metrics based on token overlap or semantic similarity are not immediately applicable (Goyal et al., 2023). Our informativeness check provides an avenue for evaluation that could potentially be scaled up through automation. Furthermore, considering other downstream applications, such as forecasting prosocial outcomes (Bao et al., 2021) or how likely it is for participants to change their mind (Tan et al., 2016; Hovy and Yang, 2021), could further help evaluate the usefulness of dynamics summaries.

While the current work is restricted to summaries of text-based conversations, important dynamics can be encoded in vocal features (e.g., intonation, or pitch) or gestures (laughter, body positioning). A multimodal approach could enable applications that go beyond text-based conversations, and could even aid conversational analysis researchers explore the intricate ways in which conversations develop (Sidnell, 2011).

Additionally, while we tested how useful summaries are for humans in a small-scale control setting, further work could test this more comprehensively through user studies, for example by integrating these summaries into conversational assistance tools (Chang et al., 2022) or moderation assistance tools (Schluger et al., 2022). From a technical perspective, a real-time deployment would require iteratively generating summaries in real-time, as the conversation progresses, rather than at a set moment in the conversation as we do in this work for the sake of scalability.

Ethical concerns surrounding fairness and bias should necessarily take center stage in any deployment of summarization systems, especially since SCDs may include mentions of emotions and affect of the people involved in the conversation (Zhou and Tan, 2023). Any broad usage scenario should undergo rigorous scrutiny of potential for unintended consequences (Weidinger et al., 2022). For example, SCDs and automated forecasts relying on them should not be used to make automated censoring or moderation decisions, to avoid propagating biases embedded in the underlying large language models. If future developments will result in summaries that are reliable enough to inform human decisions (e.g., helping moderators decide whether to closely monitor an ongoing conversation), the users should be informed about systematic mistakes the summary is likely to make in that respective setting.

References

Enas Altarawneh, Ammeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. Conversation Derailment Forecasting with Graph Convolutional Networks. ArXiv:2306.12982 [cs].

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2013. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry. In *Proceedings of WSDM*, WSDM.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of WWW*, pages 1134–1145.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Penelope Brown. 2015. Politeness and Language. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 326–330. Elsevier.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Moira Burke and Robert Kraut. 2008. Mind Your Ps and Qs: The Impact of Politeness and Rudeness in Online Communities. In *Proceedings of CSCW*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 28–39, Berlin, Heidelberg. Springer.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. 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.

Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):545:1–545:37.

Kushal Chawla, Gale Lucas, Jonathan Gratch, and Jonathan May. 2020. BERT in Negotiations: Early Prediction of Buyer-Seller Negotiation Outcomes. *arXiv:2004.02363 [cs]*. ArXiv: 2004.02363.

Jiaao Chen and Diyi Yang. 2020. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. 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.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, Polite, or Formal: Quantifying Feelings and Tone in Email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146. Place: Cambridge, MA Publisher: MIT Press.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742, Athens, Greece. IEEE.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News Summarization and Evaluation in the Era of GPT-3. ArXiv:2209.12356 [cs].

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of NAACL*, pages 386–396.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 112(3):201–229.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of NAACL*, page 15.

T. Huckin. 2002. Critical Discourse Analysis and the Discourse of Condescension.

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. ISSN: 1520-6149.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of Pragmatics*, 26(5):613–634.

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.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *ACM Computing Surveys*, 55(8):154:1–154:35.

Vinodh Krishnan and Jacob Eisenstein. 2015. "You're Mr. Lebowski, I'm the Dude": Inducing Address Term Formality in Signed Social Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626, Denver, Colorado. Association for Computational Linguistics.

Robin T. Lakoff. 1973. *The Logic of Politeness: Minding Your P's and Q's*. Chicago Linguistic Society. Google-Books-ID: DfWfNAAACAAJ.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jialu Li, Esin Durmus, and Claire Cardie. 2020a. Exploring the Role of Argument Structure in Online Debate Persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020b. Studying Politeness across Cultures Using English Twitter and Mandarin Weibo. *arXiv:2008.02449 [cs]*. ArXiv: 2008.02449.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of ICWSM*.

Elijah Mayfield and Alan W. Black. 2019. Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

Benjamin Newman, Reuben Cohn-Gordon, and Christopher Potts. 2019. Communication-based Evaluation for Natural Language Generation. ArXiv:1909.07290 [cs].

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *Proceedings of NAACL*.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of ACL*.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical Questions and Sarcasm in Social Media Dialog. pages 310–319, SaarbrÃ¼cken, Germany. Association for Computational Linguistics.

Ellie Pavlick and Joel Tetreault. 2016. An Empirical Analysis of Formality in Online Communication. *TACL*.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735. Publisher: Linguistic Society of America.

Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):370:1–370:27.

Jack Sidnell. 2011. *Conversation Analysis: An Introduction*. John Wiley & Sons.

Michael Silverstein. 1984. On the pragmatic 'poetry'of prose: Parallelism, repetition, and cohesive structure in the time course of dyadic conversation. *Meaning, form, and use in context: Linguistic applications*, pages 181–99. Publisher: Georgetown University Press Washington, DC.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*.

Deborah Tannen. 2005. *Conversational style : analyzing talk among friends*. Oxford University Press, New York.

Lu Wang and Claire Cardie. 2014. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. pages 693–699.

Zijian Wang and Christopher Potts. 2019. TalkDown: A Corpus for Condescension Detection in Context. *arXiv:1909.11272 [cs]*. ArXiv: 1909.11272 version: 1.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne

Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery.

Mark E. Whiting, Allie Blaising, Chloe Barreau, Laura Fiuza, Nik Marda, Melissa Valentine, and Michael S. Bernstein. 2019. Did It Have To End This Way?: Understanding The Consistency of Team Fracture. In *Proceedings of CHI*, volume 3, pages 1–23.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of WWW*.

Diyi Yang and Chenguang Zhu. 2023. Summarization of Dialogues and Conversations At Scale. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 13–18, Dubrovnik, Croatia. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization. ArXiv:2302.08081 [cs].

Jiaqing Yuan and Munindar P. Singh. 2023. Conversation Modeling to Predict Derailment. ArXiv:2303.11184 [cs].

Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2020. What Changed Your Mind: The Roles of Dynamic Topics and Discourse in Argumentation Process. *arXiv:2002.03536 [cs]*. ArXiv: 2002.03536.

Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. 2018a. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J. Taylor. 2018b. Characterizing Online Public Discussions Through Patterns of Participant Interactions. In *Proceedings of CSCW*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. ArXiv:2301.13848 [cs].

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018c. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. ArXiv:2109.02492 [cs].

Karen Zhou and Chenhao Tan. 2023. Characterizing Political Bias in Automatic Summaries: A Case Study of Trump and Biden. ArXiv:2305.02321 [cs].

A Instructions for Writing Summaries

In this section we explain our annotation procedure and provide definitions for the terminologies in our instructions along the way. As described in Section 2, the procedure is divided into two parts: one in which an annotator works individually and the other in which they interact with another annotator.

A.1 Individual Work

Instructions for an individual annotator:

1. Depending on the *complexity* of the conversations, either 1) thoroughly read the whole conversation or 2) skim through the conversation to understand the general idea
    - *Complexity*: number of speakers, familiarity of the topic to the annotators, length. For shorter conversations, it is easier to read through the whole conversation before moving on to summarizing, while for really longer ones, annotators would read a few comments at a time, summarize, read the next few, etc.
2. Go through the conversation *comment-by-comment* and write a comprehensive summary that captures the content of each *comment* and any *key points*.
    - *comment*: all speakers' utterances are in the form of reddit comments.
    - *key points/moments*: also referred to as "turning points" are where the tension of the conversation or the speakers' opinions notably change. Annotators should highlight them in both the original transcript and the summary in the following way: increase in tension (red), decrease in tension (blue), change in opinions towards disagreement (yellow), change in opinions towards agreement (green)
3. Then revise the comprehensive summary to
    (a) change any wording that's confusing (not accurately describing the original comment)
    (b) review if the summary reflects the conversation accurately (specifically the conversation dynamics and tension) and add any tone indicators that might be missing
        i. Indicate changing tension (e.g. curse words, all-caps, rhetorical questions, polite words) and indicate sentiments with phrases like "sarcastically," "passive-aggressively," "politely," etc. Use direct quotes (no need to explicitly describe the emotion) if they are concise and hard to capture in a summary. Focus on the highlighted elements of the conversation when adding indicators in order to capture changes in tension.
    (c) Condense the summary to 150 words while trying to preserve the turning points from step 2 and tone indicators indicated during revision. Omit the parts of the conversation that didn't contribute much to the overall trajectory and otherwise reword for brevity. For example,
        i. Condense lengthy or redundant back-and-forth conversation that doesn't introduce new points (but may impact tension) into fewer sentences summarizing the main developments
        ii. Omit irrelevant comments (e.g. brief interjections by a new user that did not have any substantial follow-ups)
        iii. Change a few direct quotes/details to more concise sentiment words (ex. "calling this blatant racism" $\rightarrow$ "... with condemnation")
        iv. Other editorial changes
4. After comprehensive summary, write the speaker summary by
    (a) First identifying the key speakers based on the comprehensive summary.
        - Usually whichever speakers spoke the most, but also considering those contributed to the *key moments*
    (b) Then for each key speaker, rereading only their comments in the original conversation, in order to describe their specific changes in tone/stance/conversation strategies and interactions/response to other key speakers

A.2 Interactive Work

Annotators start the interaction from the following setup:

- Annotator A: having completed the individual work for the conversation, i.e., read the original conversation and wrote the comprehensive summary and the speaker summary

- Annotator B: didn't read the original conversation, now writes the summary of conversation dynamics.

Collaboratively, they follow these steps, which we describe from a third-person perspective for better clarity.

1. Annotator B reads the comprehensive summary and speaker summary out loud. They ask initial questions to Annotator A confirming the order of speaker comments and key speakers ("Speaker1 then Speaker2 then Speaker1 again?", "Speaker1 spoke the most?"), the overall stance/speaker relationship of the argument ("Speaker1 and 3 agreed, and both disagreed with Speaker2?")

2. Annotator B begins writing the SCD by first copying the first sentence of the comprehensive summary, which often describes the overall topic of the conversation in a few words.

3. Annotator B identifies the first *section* of the comprehensive summary, highlighting the summary sentences on the document so that Annotator A can also reference.

    - *section* – usually 1-3 *comments* that fall before/in between any *key moments*. These *comments* should have a similar impact on the overall conversation dynamics, so that it makes sense to condense them into one sentence in the SCD
    - Annotator A may disagree with condensing the section if they think important information from within the section would be lost (e.g. different tone/rhetorical elements, argumentative stances)

4. For each *section*, Annotator B writes a corresponding summary capturing the dialogue acts, conversation strategies, and tonal elements, without any topical details.

    - dialogue acts and conversation strategies examples: disagreement, agreement, counterargument, criticism, accusation, providing sources, requesting sources, insulting, defending, acknowledging, conceding, rhetorical questions, invalidating, repetition, using long comment
    - tonal elements example: sarcasm, passive-aggressiveness, bluntness, rudeness, civility, neutrality, passion, harshness, strength, assertiveness, politeness, friendliness, objectivity, annoyance, frustration, tension, provocation, skepticism, demanding
    - If the indicator of tone is missing or not clear, Annotator B asks Annotator A questions such as the ones below, and Annotator A often goes back to the original conversation to reread comments and provide accurate answers to the questions or even read aloud whole phrases of a comment if needed to give proper context
        - B: "Was this said neutrally, or is there something about the tone that I should note?"
        - B: "Is the comment overtly rude, or is it just passive-aggressive or blunt?"
    - Annotator A reviews the work done on this *section* and makes corrections or suggestions if they think the conversation dynamics summary isn't an accurate representation of the conversation. And, Annotator A and B would revise the sentences together.
    - They repeat this process for each *section*.
    - Annotator A rereads the whole conversation dynamics summary, noting if any part does not seem to accurately reflect the original conversation/comprehensive summary. Both people work together to correct any such cases with the question-asking method above.
        - If needed, annotators would condense the summaries to be under 80 words, but usually they were already within range.

B Informativeness Check

Conversations covered in the check. We first sample 10 conversations on 10 different topics. 5 of the conversations are 'derailing' and 5 are 'non-derailing'. Each of these conversations makes one

question, where this conversation offers its transcript and a segment from its summary as the correct choice. The paired conversations of these 10 conversations offer the first type of distractors as discussed in the main text (same topic but opposite derailment label). Then, for the second distractor of each question, we use a conversation that has a different topic but the same derailment label as the correct choice. We also ensure that each conversation is used only once across all questions (either offering a transcript and correct choice or offering a distractor). This way, each choice in the question represents a unique conversation and we maximize the coverage of our check, covering a total of 30 conversations.

Extracting and processing segments. For this basic check, we define a segment as a sentence that has 2 speakers. Each summary would have multiple segments and we always randomly select one. For the three segments (choices) of a question, we rename them in such a way that the speakers in all three segments have the same pseudonames. The speakers in the transcript is also renamed accordingly to be consistent with the correct segment. This effort prevents a question from being trivial when, for example, "Speaker5" appears in a distractor but never appears in the transcript, which immediately rules out this distractor. With this renaming, annotators have to carefully read all 3 choices against the transcript to answer a question correctly.[10]

C Human Forecasting Experiment

We now discuss our design for evaluating human forecasting on conversation summaries. To design experiments that respect the annotators' attention span, we divide the 20 conversations into two batches of 10 conversations for 2 rounds of exercises with the same procedure. All annotators participate in both rounds.

In each round, we have 10 subjects divided into 2 groups (A and B), each completing one version of our questionnaires (each containing 10 summaries). The questions are designed such that the i-th question in either questionnaire presents a summary for the same conversation but the summaries are created differently (one is human-written and the other is machine-generated). For example, if the i-th question in Questionnaire A is a human-written summary for a conversation, then the i-th question in Questionnaire B is a procedural prompt summary for the same conversation. This way, each participant has an equal weight on the results for human-written and machine-generated summaries, and thus any difference between these results can not be attributed to a single annotator (e.g., that is exceptionally good at the forecasting task).

For each conversation, the annotator sees the conversation summary and is asked to guess whether the conversation will derail in the future and give scores for their confidence in their guess, topic understanding, and conversation trajectory understanding. We briefly define conversation trajectory at the start of the questionnaire, as how the interaction between speakers evolves during the discussion, independent of the actual topics discussed. Additionally, we also record the time between the subjects seeing the summary and submitting the forecast. Figure 2 presents an example question. After the experiment, we also debriefed the subjects to understand how they understood the questions; one observation that stood out was confusion regarding the trajectory scale and how that relates to the guess they are making.

For evaluating human forecasting on transcripts, we follow a similar design with some modifications. First, we have a different group of 10 participants, such that there is no pollution between the two experiments. Since reading a transcript requires much longer time than reading a summary, each participant only reads 10 transcripts, with the exception of 2 participants who volunteered to read all 20 transcripts. This results in 6 guesses for each transcript.

D Details of Summarization Models

D.1 Generating Multiple Summaries For a Conversation

For every summary type (e.g., traditional prompt, procedural prompt, finetuned BART), we repeat the process of generating summaries and running the downstream evaluation in 4 trials, each trial generating a different summary for a conversation. For a summary type based on a finetuned model, in each trial we finetune the model with a different random seed for summary generation. For a summary type based on prompting GPT-3.5-turbo, we simply utilizes its stochasticity, using its default sampling parameters to generate a new summary for each conversation.

[10] Before this renaming, we've already anonymized the speakers' usernames with Speaker1, Speaker2, etc., to respect the their identity.

D.2 Finetuned Summarization Systems

For finetuned summarization systems, we use 40 transcript-summary pairs from our human summary dataset for finetuning, 10 pairs for development, and generate summaries for the remaining 50 test set conversations that do not have human summaries. The generated summaries are then evaluated with our downstream task in Section 4.

We first experimented with the SOTA conversation summarization systems, BART-large and DialogLED (Lewis et al., 2020; Zhong et al., 2022). Both systems previously showed strong performance on long dialogue summarization datasets with small train sets, such as AMI (train size 97) (Carletta et al., 2006) and ICSI (train size 43) (Janin et al., 2003), as reported in Zhong et al. (2022). Table 3 reports the performance brought by summaries from finetuned BART and DialogLED in our downstream task. We find that these models finetuned on the 40 human written summaries, do not produce summaries that lead to better forecasting results than procedural prompt summaries.

Additionally, we attempted to finetune GPT-3.5-turbo using OpenAI's API. Due to the high cost OpenAI charges for finetuning and inferencing on finetuned checkpoints, we find adequate hyperparameter search unfeasible and stopped after obtaining one checkpoint with reasonable summary quality. The summaries by this checkpoint led to an accuracy of 61.9% in the downstream task, substantially lower than the accuracy brought by the procedural prompt summaries (Table 3).

Based on...	**Accuracy**
transcripts	56.0
procedural prompt summ.	**71.5** (2.52)
finetuned BART summ.	57.5 (2.52)
finetuned DialogLED summ.	54.5 (5.26)

Table 3: Few-shot GPT derailment forecaster performance based on finetuned models summaries (for the 50 test set conversations that do not have human summaries). We include results on transcripts and procedural prompt summaries of the same 50 conversations for reference.

D.3 Other Forecasting Systems

For using GPT-3.5-turbo as few-shot classifiers, we set the sampling temperature to 0 for deterministic behaviors (because we could not set a random seed with OpenAI's API).

Additionally, we also experimented with other classifiers using supervised training to forecast conversation derailment. We use the transcripts or the generated summaries of the train (234 conversations) and dev (100 conversations) splits of our dataset to obtain trained classifiers and run inference on the transcripts or generated summaries of the test split (100 conversations). We examine two strong baseline models for text classification for this supervised setting, namely BART and Longformer. Although these supervised models are consistently outperformed by the GPT few-shot classifier (Table 4), when comparing their performances on the generated summaries, we still find that procedural prompt summaries best help the downstream forecasting of conversation derailment, indicating that our conversation dynamics summary task indeed helps automatic systems to forecast conversation derailment.

Based on...	Acc. by Forecasting Model		
	GPT	BART	Longformer
transcripts	60	46	50
convent. summ.	58	57	58
proced. summ.	**67**	**63**	**62**
BART summ.	58	54	54
LED summ.	55	52	58

Table 4: Comparing different classifier architecture for derailment forecasting. Similar to Table 1, "convent. summ." and "procedu. summ" refer to the GPT-3.5 generated traditional prompt and procedural prompt summaries; "BART summ." and "LED summ." refer to summaries by finetuning BART and DialogLED. Additionally, the column name "GPT" refers to GPT-3.5 few-shot classifier.

D.4 Prompt Engineering

When developing our zeroshot and procedural prompts for dynamics summaries, we tried different synonyms for conversation dynamics and specific dynamics elements, as well as changing the phrasing of their definitions and examples. For example, instead of simply prompting the model to summarize 'conversation dynamics', which might appear as a novel jargon to the model's parametric knowledge, we instruct the model to write a summary that captures the trajectory of the conversation, especially focusing on how elements like tone, sentiment, conversation strategies may change or remain the same throughout the conver-

sation. We then manually examine the quality of generated summaries for a small prompt engineering dataset (size 10) that's disjoint with our dev and test splits. For the procedural prompts, in particular, we manually wrote example summary segments to contrast different aspects of traditional summaries with those of SCDs, and included these examples in the procedural prompt. Figure 3 shows the two prompts we eventually chose as the zeroshot and procedural prompts for SCDs.

E Qualitative Analysis

Inspired by prior literature (Habernal et al., 2018; Zeng et al., 2020; Zhang et al., 2018a; Li et al., 2020a), we focus on a set of conversation strategies related to conversation trajectories for our qualitative analysis in Section 5. Here, we present the list in Table 11.

F Miscellaneous

F.1 Transcript of the Introductory Example

We provide the transcript of our introductory example in Figure 4 to 7. The last 3 utterances of the transcript are omitted as how it appears in our dataset.

F.2 Data Collection

Annonymization. We collect human summaries for conversation transcripts from the published dataset CGA, which we accessed through ConvoKit 2.5.3. The dataset contains the usernames of the conversation participants, which we replace with 'Speaker1', 'Speaker2', and etc. to respect the users' identity.

Annotators. All annotators for our evaluations are recruited as volunteers from university students in the US. The two annotators who wrote the summaries of conversation dynamics are co-authors of this paper. The data collection was approved by an Institutional Review Board at the authors' institution. All annotators were informed that their data would be used for an NLP research and eventually a published paper before they gave consent.

Disclaimer of Risks. All annotators are informed that "some of the conversations presented in the annotation task can be extremely biased and offensive and speak of sensitive topics." All annotators gave their consent to participate.

F.3 Implementation Details

For our finetuned models, we conducted hyperparameter search over learning rates [3e-6, 5e-6, 1e-5, 2e-5, 3e-5, 5e-5, 1e-4] and warmup steps ([40, 80] for summarizers and [234, 468] for classifiers), and used the default values from their original implementation for other hyperparameters. For the DialogLED and BART summarizers, we eventually used a learning rate of 3e-5 and 80 warmup steps. For the BART classifier, we used a learning rate of 3e-6 and 468 warmup steps. For the Longformer classifier, we used a learning rate of 5e-6 and 468 warmup steps. The finetuning experiments in total took about 150 GPU hrs on an Nvidia A40 GPU.

F.4 Used Artifacts

We include a list of existing artifacts we used. Some of them have been cited in the main sections of this paper above. We have closely followed their intended use.

- GPT-3.5-turbo-0613:
  a snapshot of GPT-3.5-turbo from June 13th, 2023. Closed-source but accessible at a low cost via OpenAI's API, https://platform.openai.com/docs/
- ConvoKit 2.5.3:
  https://convokit.cornell.edu/, MIT License
- PyTorch 1.8:
  https://pytorch.org, BSD-3 License
- Transformers 4.25:
  https://github.com/huggingface/transformers, Apache License 2.0
- Scikit-learn 1.3.2:
  https://scikit-learn.org, BSD-3 License

F.5 Additional Evaluation Metrics

Here, we provide additional performance metrics (precision, recall, macro-averaged F1) for different summary types, when they are evaluated with our derailment forecasting task. Each summary type is evaluated with its respective GPT-3.5 few-shot derailment forecasting model as described in Section 4.1.

Derailing?	prec.	rec.	F1
False	72.7	32.0	44.4
True	56.4	88.0	68.8
macro avg	64.6	60.0	56.6

Table 5: Additional metrics for derailment forecasting on **transcripts**

Derailing?	prec.	rec.	F1
False	57.1	66.5	61.4
True	59.9	50.0	54.5
macro avg	58.5	58.3	58.0

Table 6: Additional metrics for derailment forecasting on **traditional summaries**

Derailing?	prec.	rec.	F1
False	56.1	80.0	66.0
True	65.2	37.5	47.6
macro avg	60.7	58.8	56.8

Table 7: Additional metrics for derailment forecasting on **zeroshot prompt summaries**

Derailing?	prec.	rec.	F1
False	62.9	84.0	72.0
True	75.9	50.5	60.7
macro avg	69.4	67.3	66.3

Table 8: Additional metrics for derailment forecasting on **procedural prompt summaries**

Derailing?	prec.	rec.	F1
False	56.1	69.0	61.9
True	59.7	46.0	52.0
macro avg	57.9	57.5	56.3

Table 9: Additional metrics for derailment forecasting on **finetuned BART summaries**

Derailing?	prec.	rec.	F1
False	55.3	47.0	50.8
True	53.9	62.0	57.7
macro avg	54.6	54.5	54.2

Table 10: Additional metrics for derailment forecasting on **finetuned DialogLED summaries**

[Conversation Summary]
Speakers discuss the responsibilities of caregivers of autistic children. One Speaker opens up the discussion using strong language. Speaker3 and Speaker4 begin to argue in a passive-aggressive manner, which then transitions into sarcasm, accusations, and questioning each other's logic. Speaker4 supports their point with a personal experience, which Speaker3 refutes rudely.
Will the conversation go awry (derail)?

◯ Yes

◯ No

Confidence of your answer (1 for least confident and 5 for most confident)

◯ 1

◯ 2

◯ 3

◯ 4

◯ 5

To what extent did the summary help you understand the topic of the conversation (on a scale of 1 to 5)?

◯ 1: I don't even know the general topic.

◯ 2

◯ 3: I know the general topic of the discussion.

◯ 4

◯ 5: I know how each Speaker is related to the topic.

To what extent did the summary help you understand the conversation trajectory (on a scale of 1 to 5)?

◯ 1: I don't have any idea of the trajectory of the conversation.

◯ 2

◯ 3: I have a general understanding of the trajectory.

◯ 4

◯ 5: I have a thorough understanding of how each Speaker contributed to the trajectory.

Figure 2: Example question for derailment forecasting based on summaries

Zeroshot Prompt:
Write a short summary capturing the trajectory of an online conversation. Do not include specific topics, claims, or arguments from the conversation. Instead, try to capture how the speakers' sentiments, intentions, and conversational/persuasive strategies change or persist throughout the conversation. Limit the trajectory summary to 80 words.

Procedural Prompt:
Write a short summary capturing the trajectory of an online conversation. Do not include specific topics, claims, or arguments from the conversation. The style you should avoid:
Example Sentence 1: "Speaker1, who is Asian, defended Asians and pointed out that a study found that whites, Hispanics, and blacks were accepted into universities in that order, with Asians being accepted the least. Speaker2 acknowledged that Asians have high household income, but argued that this could be a plausible explanation for the study's findings. Speaker1 disagreed and stated that the study did not take wealth into consideration." This style mentions specific claims and topics, which are not needed.
Instead, do include indicators of sentiments (e.g., sarcasm, passive-aggressive, polite, frustration, attack, blame), individual intentions (e.g., agreement, disagreement, persistent-agreement, persistent-disagreement, rebuttal, defense, concession, confusion, clarification, neutral, accusation) and conversational strategies (if any) such as 'rhetorical questions', 'straw man fallacy', 'identify fallacies', and 'appealing to emotions.' The following sentences demonstrate the style you should follow:
Example Sentence 2: "Both speakers have differing opinions and appeared defensive. Speaker1 attacks Speaker2 by diminishing the importance of his argument and Speaker2 blames Speaker1 for using profane words. Both speakers accuse each other of being overly judgemental of their personal qualities rather than arguments."
Example Sentence 3: "The two speakers refuted each other with back and forth accusations. Throughout the conversation, they kept harshly fault-finding with overly critical viewpoints, creating an intense and inefficient discussion."
Example Sentence 4: "Speaker1 attacks Speaker2 by questioning the relevance of his premise and Speaker2 blames Speaker1 for using profane words. Both speakers accuse each other of being overly judgemental of their personal qualities rather than arguments."
Overall, the trajectory summary should capture the key moments where the tension of the conversation notably changes. Here is an example of a complete trajectory summary.
Trajectory summary:
Multiple users discuss minimum wage. Four speakers express their different points of view subsequently, building off of each other's arguments. Speaker1 disagrees with a specific point from Speaker2's argument, triggering Speaker2 to contradict Speaker1 in response. Then, Speaker3 jumps into the conversation to support Speaker1's argument, which leads Speaker2 to adamantly defend their argument. Speaker2 then quotes a deleted comment, giving an extensive counterargument. The overall tone remains civil.
Now, provide the trajectory summary for the following conversation.
Conversation Transcript: [...]

Figure 3: Zero-shot prompt and procedural prompt for SCDs

Strategies	**How they can be mentioned in dynamics summaries**
Rhetorical questions	"poses a rheotrical question", "rhetorically asks"
Attacking logic	"point out flaws in [the other speaker]'s arguments", "accuses [the other speaker] of their logical fallacy"
Anecdotal experience	"shares a personal story", "uses an anecdotal example"
Evidence	"cites statistics and data to support their viewpoint", 'uses external sources to support"
Juxtaposition	"makes a comparison between", "provides a detailed explanation of the differences between"
Analogy	"uses an analogy to support"
Pointing at missing or unsupported evidence	"asks for evidence", "criticizes the lack of evidence"
Accusing of not correctly treating their argument	"accuses [the other speaker] of not reading their arguments", "accuses [the other speaker] of reinterpreting their positions"
Questioning one's knowledge or attacking one's lack of knowledge	"insulting [the other speaker]'s knowledge of [the subject]", "accusing [the other speaker] of lacking the knowledge of [the subject]"
Hypothetical example	"proposing another hypothetical scenario"
Counterexample	"presents counterexamples"

Table 11: List of conversation strategies a speaker may use. For our qualitative analysis, we consider an SCD 'mentions' a conversation strategy only if it explicitly identifies what the strategy is, as shown in the examples listed in the second column of this table. For example, if a summary simply paraphrases the exact rhetorical question or the cited evidence, then we do not count it as 'mentioning' a conversation strategy.

Transcript:
Speaker1: Businesses aren't charities. They exist to make a profit. "Morals" and "ethics" are always trumped by profit in the business world.
Speaker2: That's.... Kind of the inherent problem.
Speaker3: For you. Whenever there is a "problem", it is usually some party wanting to further their interests. Remember that morals do not exist out there, they are a construct of society. If a majority is disadvantaged, they may use "morals" to push for their interests.
Speaker2: Unchecked capitalism is economically unsustainable. So yes, it is a problem for me, and for everyone participating in the economy. Business involves more than just profits. It involves human beings investing their labor. These are not machines. The idea that profit alone should drive our economic decisions is morally bankrupt. Human beings deserve a modicum of dignity. If you can't agree to that, and you think that slavery is OK and justified as long as the business is profitable, then I would posit that you too are morally bankrupt.
Speaker4: "Unchecked capitalism is economically unsustainable." what do you mean "unsustainable"? what happens?
edit: [MFW I get my daily reddit downvotes for questioning a socialist](http://i.imgur.com/QoGM3.gif)
Speaker2: Well, let's have an economic thought experiment.
Rule 1: Businesses only care about profit (and typically the short term profit at that, not longer term returns).
Rule 2: Businesses have nothing to check them from abusive practices.
Let's think of some common things people believe are good about this scenario:
1) Everyone seeks to further their own means so with supply and demand everything works out in the end!
2) If a company has poor practices, the consumer will migrate to other options.
3) If a company has poor practices, another company will be developed to compete against it, forcing it to rid itself of those poor practices.
Now let's think about whether or not those thoughts can sustain themselves in the thought experiment.
1) Why wouldn't companies seek to form cartels? We already have evidence that they do this, even among competitors. You might say that they'll be seeking to further their own means, so this is only temporary. But I would ask you to consider that if Company A and Company B can make larger profit margins colluding than they can competing, why wouldn't they? Sure, Company A might be able to make slightly higher profits down the road if they were able to beat Company B outright, but that takes short term investment (and therefore cost) and effort. With collusion, we can maximize profits with minimal effort and cost.
2) But surely if A&B collude with one another, C will come out of the woodwork and offer a better product/service at a better price, right? But will they? If A&B catch wind of C, what's to stop them from using potentially coercive means from stopping them from competing. For example: hostile takeovers. A&B are a cartel now, and they are established, so they certainly have the capital necessary to absorb C as a fledgling company.
3) If A&B are colluding, what's to stop them from slowly degrading the conditions of their workforce? Like the employees of office space only working hard enough not to get fired, A&B are only going to create conditions good enough as to not lose profit margins.
4) If A&B are able to block competitors from the industry, what motivation do they have to innovate and improve their products?
5) What's to stop A&B from doing environmentally disastrous things (pollution for example)? If short term profits are all that matter, why would they bother to care?
[continues on the next page..]

Figure 4: Example Transcript, Part 1 of 4. Its conversation ID is cz2r8ig in the CGA corpus.

[continues from the previous page]
All of this isn't to say that Government must be the check on businesses. Government can be the tool of business to enshrine their power as well. But capitalism by itself can have some pretty negative consequences when it is extrapolated out. The problem is the fundamentals of capitalism are built on a foundation of assumption. That assumption is that when people seek to fulfill their individual desires, with supply and demand and perfect competition at play, that everything reaches an equilibrium. The problem here is that perfect competition is never really at play. Like most things "perfect" it's an idea, not a reality. You can't enforce perfect competition. So the next closest thing you can do is create an environment where competition is maintained and our perpetually imperfect competition doesn't get too out of control.
Personally, I'm a believer in social democracy - I believe that capitalism can be good when it is juxtaposed by the idea of solidarity with all of mankind (as opposed to purely selfish means). That's not communism, mind you, it just means that we shouldn't lose sight of humanity in the process. When we think about that as a counter to selfish greed, we start thinking about longer term returns, we start thinking about societal benefit, we start humanizing the capitalistic process (they aren't "capital" or "human resources" or "employees" they're people investing their efforts in making these companies perform).
I hope that addresses some of your question. It was a bit open ended, so I tried not to ramble on too much.
Speaker4: that addressed ZERO of my question. you completely failed to address what happens when a capitalist system reaches its alleged "sustainability" threshold.
my question was not open-ended. what occurs when it is no longer sustainable? you said it was unsustainable. your hypothetical "what ifs" don't support your assertion.
Speaker2: Ok. Thanks for the attitude. Now I'm just going to give you a curt response.
Economic collapse. Mass unemployment. Overly polluted and toxic environments. Did you just want a parade of horribles?
Speaker4: attitude? i just explained that you didn't answer my question. you gave me an opinionated rant instead of an explanation.
what am i supposed to do, just say "oh thanks"?
"economic collapse" yeah, you said that already. why? how?
"mass unemployment" how? everybody just gradually becomes unemployed "because capitalism"?
"toxic environments" why? is there some aspect of socialism that prevents toxicity in products? does socialism provide some sort of waste-disposal service unavailable in a capitalist system? you just keep throwing out matter-of-fact assertions, but i don't see how you are arriving at your conclusions.
and apparently you interpret scrutiny as "attitude" that you take offense to... i dunno. I'm not convinced.
edit: annnnd im downvoted instead of having any of those valid questions answered.
Speaker2: "attitude? i just explained that you didn't answer my question. you gave me an opinionated rant instead of an explanation." Sorry, perhaps I'm just reading into the emphasis from "ZERO" and "completely". The tone of your response and the one that followed seemed to be ... dickish, for lack of a better word. If I'm reading into it, my apologies.
Economic collapse. Increasingly volatile market behaviors as a result of increasingly risky investments. Essentially: short term profit at any cost. The system fails to account for long term sustainability.
[continues on the next page..]

Figure 5: Example Transcript, Part 2 of 4.

[continues from the previous page]
Mass unemployment. Actually, we're already on our way to mass unemployment due to automation. I just read an article about us losing 5 million (net) jobs by 2020 to automation and AI (that includes the 2mil that will pick up new jobs from the new tech). You also have an increasing population who can be provided for but not necessarily enough jobs to have them making a worthy contribution.
Toxic environments. I actually didn't mention socialism, you did. I said social democracy. The primary economic driver of social democracy is capitalism. It's just checked. But even still: yes. There is something inherent in those systems that greatly reduces the risk of toxicity. The people who make the products and and invest the labor are the ones who realize the gains. They are also the ones to decide whether or not those gains are worthwhile given the risks.
In a raw capitalism system: the capitalist realizes the gains and, due to the increased capital, can largely insulate themselves from the associated risks. Drinking water pollution is an example of this. The rich typically don't have this problem as they can afford to purchase potable water. The poor and working class may not always have their luxury (Flint, MI for example). So the people who have to live with the realities and consequences of their production are the ones making the decisions as opposed to someone who doesn't have to live with them or who can use their financially superior position to avoid them.
Other examples include the short term profits associated with taking unnecessary risks: see countless oil spills and deepwater horizon. Also see: covering up of global warming by oil companies or covering up of health hazards by cigarette companies.
The capitalist system *encourages* this behavior. It is financially beneficial (short term profits are encouraged over long term profits and investment - this is legally supported through cases dating back to Dodge v. Ford). A checked capitalist economy disincentivizes that behavior through regulation and social welfare. A socialist economic system would probably do the same through the reality that people who would be causing the harm are the people who have to live with the harm and are less likely to remove themselves from it.
If we want to debate the core tenets of a capitalist economy: that's fine. But I operate on the presumption that we both agree the goal of a capitalist economy is to generate profit (with a weight towards short term profit especially with larger firms having to provide quarterly earnings and financial benefit for shareholders).
To get back to OP - the minimum wage is *one* check on the default capitalist economic framework (to drive costs as low as possible). OSHA is another. The FDA is another. The SEC is another. And so on and so forth.
Speaker4: thank you for that thoughtful response. i do type in a very tonal, speech-based style, so people actually accuse me of being "a dick" or "irate" pretty frequently. i guess i will just have to get used to it.
i just mentioned socialism as an alternative to compare to in the pollution argument, i didn't mean to presume to say you were, necessarily. my apologies if it seemed that way.
So anyway, yes. I agree that capitalism creates volatile markets. However, I don't consider it to be as threatening as you seem to. I can't understand any situation that would cause a cataclysmic death or cessation of those markets. There will always simply be peaks and valleys.
"The system fails to account for long term sustainability."
[continues on the next page..]

Figure 6: Example Transcript, Part 3 of 4.

[continues from the previous page]
This is really vague. What happens when it no longer becomes sustainable? Anarchy? Does the wealth stop existing? Where does it go? Does it become impossible to make purchases or be paid for your work? Does everyone die? I am really trying to pin down what exactly you mean by a system "sustaining" itself. It seems more like a system that would have to be electively given up, like a language that dies out, instead of something that can break and has to be discarded, like a broken dish or a burnt-out light bulb.
In this sense, I don't see how any other ideology or economic system would succeed or fail in any of these areas, without the "unchecked" qualifier I see so commonly applied to critique of capitalism.
"unchecked" *is* the problem. greed is the problem. overpopulation is the problem. creation of toxic byproducts and harmful processes is the problem. a sheltered ruling-class is the problem. none of these things will disappear by switching ideologies. no other economic system is better, or worse, -equipped to eradicate these things. similarly, blaming "unchecked" application of any other system for the existence of these problems would be just as silly as blaming capitalism.
Speaker2: That was my point in the beginning though. That unchecked capitalism is bad. I am merely arguing for a check on it. Something to keep it from becoming too volatile, too disruptive, too detrimental. Something to keep it contained. I view capitalism like a nuclear reactor. If you keep the reaction going and have the right containment - you've got a nice source of clean energy. But if you fail to keep it contained - you've got bad news.
To your question of what happens. It can really mean any number of things. It could mean a very slow dystopian degradation of society wherein the rich get progressively richer and the poor become bottom feeders with short lifespans. It could mean everyone dies (perhaps through experimentation gone awry - particularly in the energy or medical fields). I'm not sure about wealth ceasing to exist outright, I don't know what that would look like or how that would come about. Anarchy is certainly possible, so is revolution (as the rich get richer, the poor get angrier - extrapolate that and we have a recipe for repeating history).
Like I said, some of those problems can be avoided when the people profiting from the production are the people doing the production. It connects them with the consequences. It functions as a systematic check. Is it perfect? surely not.
I agree that greed is the problem. But greed isn't a bug in capitalism, it's a feature. Like the nuclear reaction, we're trying to harness the human motivation behind greed and seize it. But, we don't want it to get too far out of control. It's a balancing act.

Figure 7: Example Transcript, Part 4 of 4.