Bridging Context Gaps: Enhancing Comprehension in Long-Form Social Conversations Through Contextualized Excerpts

Shrestha Mohanty[†], Sarah Xuan[†], Jacob Jobraeel[†], Anurag Kumar[‡], Deb Roy[†], Jad Kabbara[†]
[†] Massachusetts Institute of Technology, [‡] Meta
shresmoh@mit.edu

Abstract

We focus on enhancing comprehension in small-group recorded conversations, which serve as a medium to bring people together and provide a space for sharing personal stories and experiences on crucial social matters. One way to parse and convey information from these conversations is by sharing highlighted excerpts in subsequent conversations. This can help promote a collective understanding of relevant issues, by highlighting perspectives and experiences to other groups of people who might otherwise be unfamiliar with and thus unable to relate to these experiences. The primary challenge that arises then is that excerpts taken from one conversation and shared in another setting might be missing crucial context or key elements that were previously introduced in the original conversation. This problem is exacerbated when conversations become lengthier and richer in themes and shared experiences. To address this, we explore how Large Language Models (LLMs) can enrich these excerpts by providing socially relevant context. We present approaches for effective contextualization to improve comprehension, readability, and empathy. We show significant improvements in understanding, as assessed through subjective and objective evaluations. While LLMs can offer valuable context, they struggle with capturing key social aspects. We release the Human-annotated Salient Excerpts (HSE) dataset to support future work. Additionally, we show how context-enriched excerpts can provide more focused and comprehensive conversation summaries.

1 Introduction

In recent years, there has been a troubling rise in polarization, both on social media and even in public spaces in society, with these spaces being increasingly dominated by loud and extreme voices. Motivated by this reality, we focus on small-group conversations as a medium to bring people together, strengthen community building and understanding, and present people a space to share personal stories and experiences related to crucial matters that affect their daily lives. These conversations can be recorded (with consent) thus allowing the sharing of excerpts from one conversation in many other conversations. This can help in promoting a collective understanding of relevant issues, highlighting perspectives and experiences to other groups of people who might otherwise be unfamiliar with, and thus unable to relate to, these experiences.

A key challenge here is that excerpts taken from one conversation and shared in another setting might be missing key elements or crucial context that was previously introduced in the original conversation, e.g., background information about the speaker, their intent, or key details about the story or perspective being shared. This could lead to misunderstandings or misinterpretation and could elicit negative reactions from the receiving side as they may not fully grasp the background that shapes the speaker's words. This issue is increasingly common in today's digital landscape, where missing context often fuels misunderstandings (Mauranen, 2006) and perpetuates false and harmful stereotypes that contribute to dehumanization (Roy, 2023). Moreover, this problem is exacerbated when, as in our case, conversations become longer and richer in themes and shared experiences.

In this work, we examine the role of highlighted salient excerpts and the influence of social factors in understanding long-form group conversations, spanning tens of thousands of words. We assess how well these excerpts can convey socially relevant information and explore the effectiveness of contextualizing excerpts to improve readability and overall understanding, a process we term *effective contextualization* of the excerpt. More specifically, we examine how Large Language Models (LLMs) can retrieve and synthesize additional information
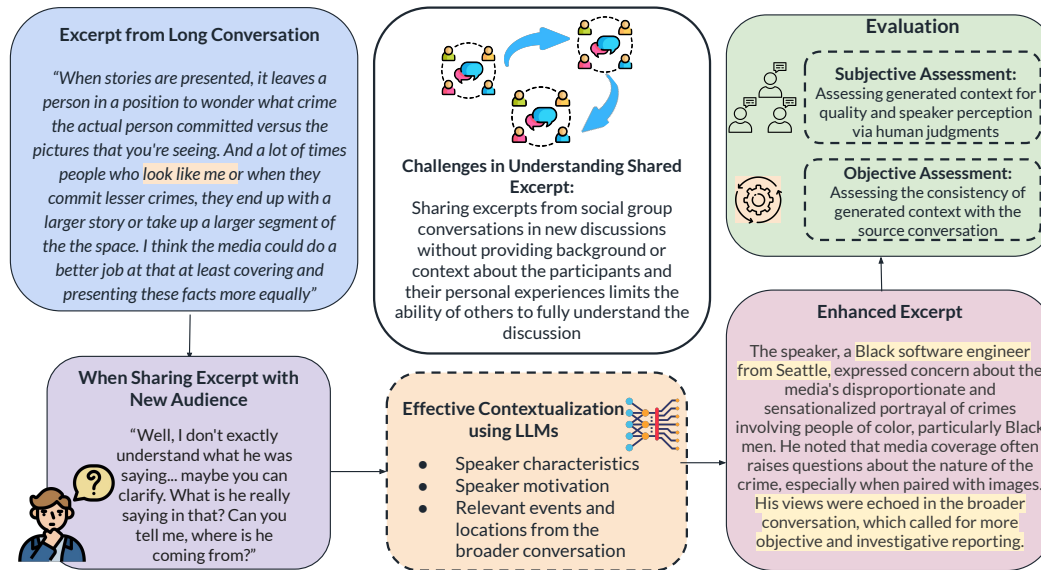
Figure 1: **Overview of Effective Contextualization of Salient Excerpts in Social Conversations:** The figure illustrates the challenges posed when excerpts from longer social group conversations are shared without sufficient context. To address this, a large language model (LLM) is employed to generate context that incorporates key social attributes, aiming to enhance the comprehension of the excerpt, a process we term effective contextualization. This enhanced context is then evaluated through human judgment to gauge its usefulness and through objective measures of faithfulness to assess its accuracy.

related to social attributes to augment the excerpts from lengthy social conversations.

The conversations (Schroeder et al., 2024) we focus on involve diverse speakers discussing and sharing personal stories on topics that include public health, affordable housing, and neighborhood safety among others. While LLMs have been effective in synthesizing information in structured settings like news or meetings (Goyal et al., 2022; Feigenblat et al., 2021; Laskar et al., 2023), their application to social conversations is less studied.

Figure 1 illustrates the novel problem we examine in this work. It presents a human-highlighted excerpt from a long group conversation about systemic bias in media representation. The excerpt, itself, without additional context, is difficult to interpret. Key elements such as the speaker's identity, their intent, and the broader conversational topic remain unclear. To address this, we employ a large language model (LLM) for effective contextualization with relevant social attributes, improving the excerpt's comprehensibility. We develop two approaches for effective contextualization of a given excerpt from a conversation. The first one relies on the zero-shot (Kojima et al., 2022) capabilities of LLMs whereas, for the second one, we instruct the LLM to explicitly focus on socially grounded

attributes to effectively contextualize the excerpt. The quality of the enhanced excerpt is then assessed through both human judgment for usefulness and objective measures of faithfulness for accuracy.

The key contributions of this work are:

1. We explore the use of LLMs for the analysis and understanding of long-form social group conversations through *highlighted salient excerpts*. To the best of our knowledge, this is the first work to explore the use of salient excerpts for understanding long multi-party social conversations. These conversations present unique challenges due to their distinctive characteristics, such as personal storytelling, social cues, topic shifts, informal language, and diverse perspectives. While our method is tailored for social conversations, it can also be adapted to other domains, such as structured meetings.

2. We introduce a new dataset, *Human-annotated Salient Excerpts* (HSE), built on the Fora dataset (Schroeder et al., 2024), to enable exploration of this novel problem.

3. We introduce the concept of *effective contextualization*, exploring two approaches for enriching salient excerpts with social context using

LLMs. Our analysis shows that while LLMs can synthesize useful context, they struggle to accurately extract key social information.

4. We demonstrate the utility of salient excerpts and their contextualized versions for summarizing lengthy social conversations. Summaries based on effectively contextualized excerpts are rated higher in subjective evaluations, underscoring the value of integrating social attributes when designing automated methods for understanding social group conversations.

2 Related Work

Social dialogue: Social group conversations, in the context of our study, refer to discussions within small groups where participants share experiences, perspectives, and viewpoints on topics of communal relevance. These conversations are rich in social dynamics, often featuring personal stories, emotional expressions, and diverse perspectives. They have been recognized as valuable across various fields, offering insights into collective understanding, civic governance, and nuanced perspectives in the social sciences (Roy, 2023; Schroeder et al., 2024). Past work highlighted the importance of such dialogues in fostering better decision-making and legitimizing democratic outcomes. Dialogue networks (Roy, 2023) rely on such conversations among a small group of people to capture their perspectives on local issues and enable participation in civic and democratic processes. Deliberative polling has demonstrated the role of discourse in improving decision outcomes (Fishkin, 1997), while storytelling has been shown to enhance productive deliberation in forums (Ryfe, 2006). Sustained dialogue, as noted by (Saunders, 1999), plays a critical role in transforming relationships amidst deeply rooted conflicts. This body of work underscores the significance of social group conversations as a medium for understanding collective reasoning, resolving conflicts, and studying broader societal dynamics. These conversations can be lengthy, on average over an hour long (Schroeder et al., 2024) necessitating approaches to distill key points.

Socially aware NLP models: Recent advances in NLP and LLMs have enabled the development of more socially aware models. Ziems et al. (2024) discuss the growing ability to create such models for computational social science problems. Several works (Flek, 2020; Hovy and Yang, 2021; Yang et al., 2024) emphasized the importance of understanding the relationship between language and social context, advocating for the development of socially aware NLP systems. Our work builds on these efforts by integrating attention mechanisms to better capture social factors within conversations.

Extracting social dynamics and speaker characteristics in NLP: LLMs have been applied to understand both emotional undertones and social meaning in conversations, as well as to extract speaker characteristics. Dutt et al. (2024) use machine-generated rationales to deduce the emotional and social meaning behind conversational statements, while (Chae and Davidson, 2023) focuses on extracting social stances from online dialogue. Additionally, (Jurafsky et al., 2009) and (Broniatowski, 2012) presented methods for identifying the speaker's personality and identity from the text. We are inspired by these methods to explore the use of LLM-generated enriched contexts to enhance the understanding of both the speaker's identity and the social dynamics surrounding the conversation excerpt.

Human-in-the-loop for contextualization and synthesis: Human-in-the-loop methods have proven effective in refining LLM-generated content. (Jiang et al., 2024) involves human experts to improve the quality of LLM-generated content used to help in the understanding of complex legal concepts. Additionally, (Chen et al., 2023) illustrates how human-highlighted information can improve LLM-generated summaries while (Yao et al., 2023) shows enhanced domain-specific summarization by incorporating human edits into machine-generated outputs. In our work, we leverage human-identified salient excerpts and use LLMs to generate contextual information, improving comprehension of social conversations. Additionally, we focus on much longer, multi-party conversations.

3 Human-annotated Salient Excerpts Dataset (HSE)

To enable the salient excerpt-driven understanding of social group conversations, we create the *Human-annotated Salient Excerpts Dataset (HSE)*[1]. This dataset is built on top of the Fora dataset (Schroeder et al., 2024), which includes social group conversations organized by the non-

---

[1]Accessible at `https://github.com/shresh02/bridging_context_gaps`

profit Cortico in partnership with various US organizations. These are small-group facilitated conversations that revolve around topics such as *affordable housing, education, public health, and neighborhood safety*. Participants often share their perspectives and personal stories during these discussions. The dataset includes 262 conversations having an average of 66 minutes in length, where a conversation has a median of 6 participants and an average of 152 speaker turns. Figure 3 (Appendix) shows the distribution of conversation length within the dataset, with most conversations ranging from *7,000 to 12,500 words*.

Humans were asked to highlight salient excerpts in each conversation. That led to 1,774 salient excerpts across 177 conversations in the Fora dataset. Each excerpt averages 111 words, with over 75% tagged by topics such as civic engagement, educational inequality, and financial literacy.

These excerpts, while salient and representing important points in the conversation, often lack the necessary context for full comprehension. Our study focuses on effective contextualization of the excerpts. To be able to assess the models' performance in generating enhanced contexts, we annotated excerpts with social and contextual attributes from the conversation, such as speaker demographics, motivations, locations, and relevant events. We crowd-sourced annotators through the Prolific platform, providing them with 35 full social group conversations from the Fora dataset and 100 excerpts across these conversations. Annotators responded to specific questions listed in Appendix F.2. After filtering out incomplete or inattentive responses based on task completion times, we retained approximately 90 excerpts, each annotated by 4-5 annotators. We conducted a secondary review of the annotations to ensure consistency, establishing a ground truth for evaluating the models' objective performance.

Given the need for developing socially aware models (Yang et al., 2024) and the limited availability of relevant datasets, the HSE dataset presents a valuable resource for advancing research in this area. Although the additional tags, such as topics, are not the primary focus of this paper, they can support other socially oriented tasks and benchmarks (e.g., topic detection). The excerpts can also be applied to summarization tasks, as demonstrated in a later section, to enhance overall comprehension of the full conversation.

4 Problem Setup: Effectively Contextualizing Excerpts

Understanding excerpts from long-form social group conversations could be non-trivial and time-consuming given their length. Our primary goal is to contextualize these excerpts to provide meaningful and useful information, which we term *effective contextualization*. We formalize the problem as follows:

The data consists of long-form group conversations $C = \{C_1, C_2, \ldots, C_n\}$, where each C_j is a conversation among K participants on some topic(s). For each C_j, a set of salient excerpts (in our case human-highlighted) is available, $E_j = \{e_{j1}, e_{j2}, \cdots, e_{jm}\}$, where e_{jl} is a salient excerpt. For a given excerpt from a conversation, we aim to generate additional context so that the overall comprehensibility and understanding of the conversation are improved.

Formally, $\tilde{e}_{jl} = f(C_j, e_{jl})$ represents the *context-enriched excerpt* for e_{jl}. Here, f is an LLM that is used for extracting the relevant information and generating $\tilde{e}_{jl}$. This process of generating the effective context for a given excerpt is referred to as *effective contextualization*.

The key questions here are: *(1)* What additional information should be sought for effective contextualization in a social group conversation? *(2)* How do we evaluate effective contextualization?

4.1 Factors in Effective Contextualization

Our work is inspired by (Hovy and Yang, 2021) which uses linguistic frameworks such as systemic functional linguistics (Halliday and Matthiessen, 2013) and the Cooperative Principle (Grice, 1975), proposing factors for effective communication in social conversational settings. We focus on the *Speaker Characteristics* and *Context* factors to improve the understanding of excerpts and specifically emphasize the following factors to improve the comprehensibility and usefulness of excerpts:

- **Speaker characteristics:** Demographic details (age, gender, ethnicity), educational, and occupational references, along with relevant personal or cultural experiences.

- **Speaker motivation:** Understanding the reasons behind a speaker's statements, including the context of the questions or comments they respond to.

10245

- **Relevant locations or events:** Identifying key locations or events in the conversation to aid comprehension.

4.2 Evaluating Effective Contextualization

Once an effectively contextualized excerpt $\tilde{e}_{jl}$ has been generated, we evaluate the effectiveness of this contextualization by evaluating how well $\tilde{e}_{jl}$ improves the comprehension and informativeness through the additional information. We also evaluate the relevance and consistency of the generated excerpt $\tilde{e}_{jl}$ with respect to the original excerpt e_{jl} and the conversation C_j. The overall evaluation is done through subjective evaluations involving humans as well as through objective faithfulness analyses. The subjective evaluations are done across dimensions such as understandability, readability, completeness, and cohesiveness.

4.3 Implicit and Explicit Effective Contextualization

We utilize Large Language Models (LLMs) for the effective contextualization of conversation excerpts. For a given excerpt e from a conversation C, we use an LLM to generate a *Context-Enriched Excerpt (CEE)*, using the surrounding contexts within C. We use two approaches:

- $CEE_i(e)$ represents the excerpt obtained through **implicit contextualization**, where zero-shot capabilities of the LLM (Kojima et al., 2022) are leveraged to generate context without explicit instructions on social attributes.

- $CEE_e(e)$ represents the excerpt obtained through **explicit contextualization**, where the LLM is guided by *In-Context Learning (ICL)* (Brown, 2020) and is explicitly instructed to incorporate specific social attributes from the conversation.

In both cases, the input consists of both the conversations and the excerpts. The LLM is tasked with gathering relevant information from the conversation, whether by focusing on explicit social attributes (in CEE_e) or by relying on inherent language patterns in the implicit case (in CEE_i). Prompts for both methods are provided in Appendix D.

More formally, $CEE_e(e_{jl}) = f_e(C_j, e_{jl})$ and $CEE_i(e_{jl}) = f_i(C_j, e_{jl})$ represent the explicit and implicit context-enriched excerpts for e_{jl} respectively. f_e corresponds to the LLM's generation with an explicit focus on social attributes, and f_i corresponds to the zero-shot generation without explicit instructions.

5 Experimental Setup

We now describe the overall experimental setup and details of the evaluation approach. We investigate the effective contextualization of excerpts in the following three directions.

CEE_e vs CEE_i: Our first empirical evaluation studies the difference in explicit vs implicit effective contextualization. More specifically, we conduct subjective evaluations to analyze and compare the excerpts and their context-enriched version (CEE_e and CEE_i) across various dimensions such as understandability, readability, and cohesiveness (Table 1). The contextualization function f here is *GPT-4 Omni (GPT4-o)* (Achiam et al., 2023). The details of the evaluation process are described in the subsequent section.

LLMs for contextualization: We next compare different LLMs for the task of effective contextualization. We compare 3 LLMs *GPT-4 Omni (GPT4-o)* (Achiam et al., 2023), *Llama 3.1-70b* (Dubey et al., 2024) and *Claude Opus* (Anthropic, 2024). Llama 3.1-70b is an open-source LLM whereas the other two are not. Due to cost considerations, we limit our experimentation on Claude and Llama to only explicit contextualization[2] (as opposed to both implicit and explicit). This comparative analysis helps us evaluate the relative effectiveness of different state-of-the-art LLMs in generating meaningful and faithful context for conversation excerpts. Complete prompts and further details are provided in Appendix D.

Clarification through extrinsic knowledge: To further probe the efficacy of LLMs in generating context-enriched excerpts for social conversation, we investigate their ability to clarify uncommon terms and phrases which can further improve the generated excerpt. As part of the explicit contextualization method (CEE_e), we prompt the LLMs (Appendix D.3) to clarify any specialized terms or phrases that might not be common knowledge by requiring the LLM to provide additional explanations

---

[2]As we see later in Section 6, this is the better performing effective contextualization approach. Given that we are restricted by cost considerations, we opted to limit the performance comparison of the 3 LLMs on this better performing case.

or references to extrinsic knowledge to enhance the comprehensibility of the context. We analyze how well the LLMs can identify such terms and provide useful clarifications, particularly when using models like GPT-4 Omni's web search capability. This capability was expected to enrich the context by offering definitions that could otherwise be unclear to the reader.

5.1 Evaluation

We conduct an evaluation of the enriched context for 90 annotated excerpts from the HSE dataset (Section 3) through a combination of subjective human assessments and objective faithfulness measures. We assessed the quality of the enriched excerpts produced by various models, focusing on dimensions such as *faithfulness, text quality*, and *speaker perception*.

Subjective human evaluation: We recruited 75 human evaluators from Prolific with a 99-100% approval rating to evaluate the enriched contexts. Participants assessed the excerpts on textual quality (*understandability, readability, redundancy, completeness, cohesiveness*) and speaker perception (*agreement with the speaker's point of view, perception of the speaker as honest or trustworthy, respect for the speaker, empathy for the speaker, the ability to see the speaker's point of view (POV)*). This evaluation determined how well the enriched contexts convey the speaker's original message and their social perception.

Each excerpt was rated by at least three evaluators. Quantitative data was collected using likert scale ratings [1-5], and qualitative insights were gathered through open-ended responses. Details of the survey and prompts are available in Appendix F.

Faithfulness and objective metrics: To measure the consistency of the enriched context with the source conversation, we assess the faithfulness (Li et al., 2022) of the information extracted from the conversations and compare it with the annotated social attributes from the HSE dataset. We categorize the extracted attributes into short-response factors and long-response factors, based on the type and depth of the information being conveyed:

- **Short-response factors** are brief, factual details like the speaker's name, gender, age, occupation, race, and economic status. These require minimal elaboration and can typically

be conveyed in a few words.

- **Long-response factors** involve more nuanced and context-rich information, such as the speaker's background, personal experiences, significant events, or locations. These require detailed explanations to capture complexity. For instance, locations could be specified at multiple levels (e.g., city, state, country), adding depth to the context, which makes them long-response factors.

We calculate the F1-score to balance precision and recall, where precision was measured by the cosine similarity between the SBERT embeddings (Reimers and Gurevych, 2019) of extracted attributes from the context and the annotations in the HSE dataset, and recall was determined based on whether essential information was missed. The extraction of attributes was done using GPT-4o and details are provided in Appendix D.4

6 Results and Discussion

Dimension	e	CEE_i	CEE_e
Understandability	3.43	4.11*	**4.13***
Readability	3.40	3.90*	**4.10***†
Low redundancy	3.24	3.53*	**3.75***†
Completeness	3.07	**3.96***	3.83*†
Cohesiveness	3.23	3.93*	**4.03***
Agreement with the speaker's p.o.v.	3.57	**4.01***	3.99*
Viewing speaker as honest/trustworthy	3.55	3.91*	**3.98***
Respect for speaker	3.58	3.89*	**3.97***
Empathy for speaker	3.50	3.59	**3.66***
Ability to see the speaker's p.o.v.	3.77	**4.08***	4.06*

Table 1: Average ratings from the human evaluation of the original excerpt e, and context enriched excerpts CEE_i (implicit) and CEE_e (explicit) generated using GPT-4o where * denotes a statistically significant difference for the ratings from the original excerpt and † shows a statistically significant difference between the ratings for CEE_i and CEE_e for Welch t-test for $p < 0.05$. The horizontal line divides dimensions grouped into two broader categories: textual quality (upper half) and speaker perception (lower half).

6.1 Textual quality and Speaker Perception

CEE_e vs CEE_i: A total of 90 annotated excerpts from the HSE dataset were used for this analysis, and each excerpt was rated by an average of 4-5 crowdsourced evaluators, resulting in 433 total responses. On average, each excerpt contained 128 words. The implicit contextualization (CEE_i) contexts have an average length of 267 words, while the enriched (CEE_e) contexts average around 55

Dimension	GPT	Claude	Llama
Understandability	3.70	**3.81**	3.58†
Readability	3.86	**3.90***	3.71*†
Low redundancy	3.66	**3.72***	3.51*†
Completeness	3.67	**3.72**	3.58
Cohesiveness	3.66	**3.81***	3.47*†
Agreement with the speaker's p.o.v.	4.07	**4.10**	4.00
Viewing speaker as honest/trustworthy	4.07	**4.17**	4.02†
Respect for speaker	4.19	**4.21**	4.18
Empathy for speaker	4.19	**4.22**	4.18†
Ability to see the speaker's p.o.v.	4.20	**4.29**	4.11†

Table 2: Average ratings from the human evaluation of explicitly enriched contexts generated by GPT, Claude, and Llama where * denotes a statistically significant difference between the ratings for Llama and Claude against GPT and † denotes a statistically significant difference between the mean ratings for Llama and Claude for the Welch t-test for $p < 0.1$. The horizontal line divides dimensions grouped into two broader categories: textual quality (upper half) and speaker perception (lower half).

words. This difference arises from the approaches: CEE_i includes all relevant details from the full conversation, whereas CEE_e focuses solely on extracting key social factors.

In Table 1, we observe that both methods of contextualization received significantly higher ratings than the original salient excerpts alone ($p < 0.05$, Welch t-test). Moreover, the explicit contextualization (CEE_e) method generally outperformed implicit contextualization (CEE_i) on most dimensions. The ratings for CEE_e were particularly high in comprehensibility, readability, low redundancy, and cohesiveness as well as in cultivating respect and empathy for the speaker. This was attributed to the inclusion of key social factors that improved the understanding of the excerpts. In contrast, participants found CEE_i more complete, as it included more details from the conversation. Notably, explicit contextualization results in shorter enriched contexts while improving performance across several dimensions.

LLMs for contextualization: We compare three LLMs, GPT-4 Omni (GPT-4o), Claude Opus, and Llama 3.1-70b, on their ability to generate enriched contexts using CEE_e. We recruited 30 evaluators to rate the contexts generated by each model. Each excerpt was rated by an average of three evaluators. Claude emerged as the most preferred model, achieving the highest ratings across nearly all dimensions, including understandability, readability, and speaker perception (Table 2).

The enriched contexts varied in length among the models. GPT-4o produced contexts averaging

296 words, while Claude Opus generated slightly shorter contexts, averaging 264 words. In contrast, Llama 3.1-70b produced notably longer contexts, averaging 538 words. The differences between Llama 3.1-70b and the other models were statistically significant ($p<0.05$), particularly in areas like understandability, readability, redundancy, empathy, and the ability to see the speaker's POV.

6.2 Qualitative Analysis of Responses

Beyond the quantitative evaluation, the comments from the evaluators provided nuanced insights into how the inclusion of social attributes and background information influenced their perceptions. We provide some qualitative insights through these comments.

Evaluators emphasized the value of background information such as the speaker's name and other contextual cues, which helped them connect more deeply with the speaker while preferring conciseness. This enhanced trustworthiness, empathy, and the overall utility of the contexts. One participant stated, *"Giving background information about the speaker gave it a personal touch and helped me empathize with the reader more."* Another participant echoed this sentiment, saying that the CEE_e context *"succinctly introduces [speaker name] as a high school student from Gardiner, Maine, sharing her thoughts on essential career skills such as collaboration, writing, and speaking."*

The preference for CEE_e was frequently attributed to its ability to offer a broader and more socially rich context without excessive detail. As noted earlier, CEE_e produces considerably shorter enriched contexts and is yet preferred on factors such as readability and cohesiveness. One participant commented, *"I preferred the CEE_e context because it offers a broader context for the excerpt, explaining the community's overall goals and how the speaker's concerns fit into those goals. This context makes the excerpt more meaningful and easier to understand."* This underscores how CEE_e successfully situates the excerpt within the larger conversation while maintaining conciseness, thus enhancing the overall understandability.

We also observe that disagreement with the speaker's POV could coexist with high ratings in empathy and the ability to understand the speaker's perspective. In several cases, evaluators rated their agreement with the speaker's perspective as low (1-2), while still giving high ratings (3-5) for empathy and related perception questions. The CEE_e con-

text had slightly more such cases, with 15 instances of high empathy and the ability to see the speaker's POV despite disagreement, compared to 11 cases of high empathy in the CEE_i context.
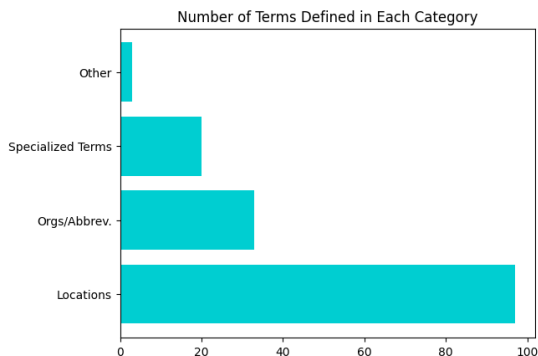

Figure 2: Number of defined terms in each category

6.3 Clarification through Extrinsic Knowledge

As part of the explicit contextualization (CEE_e) process, we prompted the LLMs to clarify uncommon or specialized terms to enhance comprehensibility by incorporating extrinsic knowledge not directly mentioned in the conversation. This included providing additional context about locations (e.g., Cornwallis, Durham), explaining abbreviations (e.g., a school district acronym or a church's name), or elaborating on specialized terms not defined in the conversation (e.g., "green roofs," "heat island"). Figure 2 illustrates the distribution of these categories. The "Other" category refers to terms that did not fall into the three prevalent categories. Terms in this category were largely names of other participants in the conversation. Of the 433 responses collected during the human evaluations, 128 responses included specifically flagged terms that required elaboration. The CEE_e process successfully addressed only **55** of these instances, highlighting that LLMs still struggle to consistently identify and explain uncommon terms critical for tailoring content to the reader. However, when the models correctly defined terms using extrinsic knowledge, the evaluators found the elaboration helpful, particularly for specialized terms that could otherwise cause confusion, which helped in improving their understanding of the excerpt.

6.4 Faithfulness Analysis

It is important that the enriched contexts remain consistent with the original conversation. We evaluate the consistency of the enriched contexts with the

Factor	CEE_i	CEE_e
Name of speaker	0.87	**0.93**
Gender	0.96	**0.98**
Race	**0.91**	**0.91**
Education	0.85	**0.90**
Occupation	0.71	**0.86**
Age	0.88	**0.89**
Economic status	0.87	**0.88**
Additional info on speaker background	0.43	**0.69**
Personal experiences	0.41	**0.48**
Question/response that prompted speaker	**0.74**	0.52
Significant events/locations mentioned	**0.77**	0.68
Location of conversation	**0.51**	0.34

Table 3: F1 score of CEE_i and CEE_e enriched contexts for Faithfulness Analysis. Factors are divided into two broader categories: short-response factors (upper half) and long-response factors (lower half).

original conversation through an objective faithfulness analysis, focusing on both short-response and long-response factors (Table 3). Both CEE_e and CEE_i performed well in extracting short-response information, such as the speaker's name, gender, and race, demonstrating high precision. The explicit contextualization (CEE_e) exhibited slightly higher recall rates, suggesting it was more adept at capturing a broader range of relevant details, aligning with its design to identify social attributes efficiently. However, the LLMs struggled with long-response factors, such as speaker background, shared personal experiences, and location context, likely due to the high variability and open-ended nature of these elements suggesting areas for future improvement. Despite this challenge, CEE_e showed a stronger ability to extract fine-grained information, as it focused on highlighting key attributes within the conversation. Comparisons of GPT-4o, Claude Opus, and Llama 3.1-70b for faithfulness are detailed in Appendix A. Claude Opus generated the most faithful responses.

7 Context-Enriched Excerpts for Enhanced Comprehension of Conversation

Intuitively, highlighted salient excerpts could be useful in highlighting key information or providing some form of a summary of the full conversation. To evaluate whether context-enriched excerpts lead to better summaries, we compare the effectiveness of these excerpts against direct summarization of the full conversations. Specifically, we generate and compare three types of summaries (using GPT-

4) with a word limit of 200 for each:

- **Full Conversation Summary**: Generated by providing the LLM with the entire conversation.

- **Excerpt-Only Summary**: Generated using only the salient excerpts identified for the conversation.

- **Context-Enriched Excerpt Summary**: Generated using all excerpts after they have been effectively contextualized.

This comparison aims to determine whether enriching excerpts with context can effectively capture the key points of the conversation and enhance overall comprehension within the set word limit.

A total of 50 conversations were analyzed, with 3-4 evaluators rating each summary, resulting in 156 responses from 34 evaluators. Table 4 shows the average ratings on various dimensions. The Context-Enriched Excerpt Summary received the highest ratings in coherence, informativeness, and level of detail, outperforming both the Full Conversation Summary and the Excerpt-Only Summary. While the Excerpt-Only Summary scored slightly higher in readability—likely due to its straightforward merging of salient excerpts—it was rated lower in informativeness and detail, as it lacked the additional context from the conversation. Participants noted that the Context-Enriched Excerpt Summary effectively captured the main points while providing essential background information, thereby enhancing overall comprehension.

A key strength cited in the qualitative responses for this approach was that the additional details including names and background information of speakers induced empathy and connection. As one participant noted, it *"gives light to names of those in school and their educators... It gives you a reason to get behind these people because you feel close to them and can see their struggle."* . This underscores the importance of effectively contextualizing salient excerpts with social aspects to convey the essential elements of the conversation, offering a focused yet comprehensive understanding.

8   Conclusion

Long-form social group conversations often convey rich information but are challenging to process due to their length. Human-highlighted salient excerpts provide anchor points for understanding but can

Dimension	Full Conv.	Excerpt-Only	CEE
Readability	4.21	**4.25**	4.11
Coherence & structure	4.04	3.98	**4.19**
Informativeness	4.16	3.86	**4.26**
Detail level	4.06	3.66	**4.16**

Table 4: Average ratings for different summarization approaches of the conversation.

lack sufficient context, potentially leading to misunderstandings, particularly given the informal and personal nature of these conversations. To address this, we introduced a method using LLMs to augment salient excerpts with effective contextualization. This approach significantly improved comprehension, readability, and empathy by enriching excerpts with meaningful context. While LLMs show promise, challenges remain in capturing speaker background, personal experiences, contextualizing locations, and addressing unique terminology. We also demonstrated the utility of contextualized excerpts for summarizing lengthy conversations, with subjective evaluations highlighting their improved effectiveness. To support further research, we release the Human-annotated Salient Excerpts (HSE) dataset as a resource for advancing understanding and contextualization in social conversations.

9   Limitations

Consistency of generated context: LLMs can produce different results with each iteration. To ensure consistency, we set the model temperature to 0 for deterministic outputs and collected a large number of responses to accurately reflect the representative average rating from the population.

Faithfulness evaluation: F1 scores are lower for open-ended questions due to their subjective nature, which complicates standardization. For instance, responses about a speaker's background varied among annotators, highlighting the challenge of standardizing such tasks. Further research is needed to develop better methods and metrics for measuring faithfulness, as noted by (Risch et al., 2021) and (Xu et al., 2024). Our final method for assessing faithfulness was chosen as other methods, such as Named Entity Recognition and entailment approaches (Goyal and Durrett, 2021; Manakul et al., 2023), were less suitable for our specific use case.

Number of excerpts: Due to cost constraints, we analyzed only 90 excerpts. Each required multiple ground truth annotators to get accurate ground truth data, making the process expensive.

10 Ethics Statement

Code of Conduct The informed consent of all survey participants and annotators was obtained before participation. This study received approval from the Institutional Review Board (IRB) at our university, and participants were compensated at a rate above the minimum wage for the state of Massachusetts.

Privacy of Models: We are aware of the privacy concerns associated with releasing information to LLM models for analysis. We used GPT and Claude via APIs, and both OpenAI and Anthropic have stated that data sent to the API would not be used for training (OpenAI; Anthropic). As a note, the Fora dataset (Schroeder et al., 2024) upon which we build our HSE dataset involves conversation data that has received the proper consent to be publicly released and underwent anonymization to further preserve the privacy of the conversation participants.

Acknowledgements

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. How do you use personal data in model training?

Anthropic. 2024. Introducing the next generation of claude.

David A Broniatowski. 2012. Extracting social values and group identities from social media text data. In *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pages 232–237. IEEE.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.

Jiaao Chen, Mohan Dodda, and Diyi Yang. 2023. Human-in-the-loop abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9176–9190, Toronto, Canada. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ritam Dutt, Zhen Wu, Kelly Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Penstein Rose. 2024. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. *Preprint*, arXiv:2406.19545.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm–a dialog summarization dataset for customer service. *arXiv preprint arXiv:2111.11894*.

James S Fishkin. 1997. *The voice of the people: Public opinion and democracy*. Yale university press.

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7828–7838.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219, Bangkok, Thailand. Association for Computational Linguistics.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. *arXiv preprint arXiv:2310.19233*.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Anna Mauranen. 2006. Signaling and preventing misunderstanding in english as lingua franca communication. *International Journal of the Sociology of Language*, 2006(177):123–150.

OpenAI. How we use your data.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*.

Deb Roy. 2023. The internet could be so good. really. *The Atlantic*.

David M Ryfe. 2006. Narrative and deliberation in small group forums. *Journal of Applied Communication Research*, 34(1):72–93.

Harold Saunders. 1999. *A public peace process: Sustained dialogue to transform racial and ethnic conflicts*. Springer.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2024. Fora: A corpus and framework for the study of facilitated dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13985–14001, Bangkok, Thailand. Association for Computational Linguistics.

Shaochen Xu, Zihao Wu, Huaqin Zhao, Peng Shu, Zhengliang Liu, Wenxiong Liao, Sheng Li, Andrea Sikora, Tianming Liu, and Xiang Li. 2024. Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv preprint arXiv:2402.11398*.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2024. The call for socially aware language technologies. *arXiv preprint arXiv:2405.02411*.

Zonghai Yao, Benjamin J Schloss, and Sai P. Selvaraj. 2023. Improving summarization with human edits. *Preprint*, arXiv:2310.05857.

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

Factor	GPT	Claude	Llama
Name of Speaker	0.94	0.95	**0.97**
Gender	**1.00**	**1.00**	**1.00**
Race	**0.98**	**0.98**	**0.98**
Education	**0.97**	**0.97**	0.96
Occupation	**0.90**	0.88	0.89
Age	0.97	**0.98**	0.83
Economic Status	0.73	0.70	**0.85**
Additional info on speaker background	0.80	**0.81**	0.78
Personal Experiences	**0.78**	0.77	0.76
Question/Response that prompted speaker	0.80	0.83	0.82
Significant Events/Locations mentioned	0.88	**0.90**	0.86
Location of Conversation	**0.77**	0.71	**0.77**

Table 5: **F1 score of GPT-4o, Claude Opus, and Llama 3.1 enriched contexts for Faithfulness Analysis**. The horizontal line divides the factors into two broader categories: direct, short-response factors (upper half) and open, long-response factors (lower half)

A   Faithfulness evaluation of LLM enriched contexts

When comparing the different LLMs using the CEE_e approach, we found that GPT-4o and Claude Opus demonstrated slightly better precision and recall for short-response factors than Llama 3.1-70b, as shown in Table 5. in the Appendix. However, Llama tended to capture more details related to personal experiences and the preceding conversation topics, possibly due to its generally longer responses. This verbosity, while occasionally helpful, sometimes led to a decrease in overall cohesiveness and readability, reflecting the need for a careful balance in context generation.

B   Examples of Explicit Effective Contextualization of Excerpts

Table 6: Examples of Effective Contextualization

Original Excerpt	Generated Context	Annotator Comments	Social Attributes Present
I just have one quick thing and it's the concern for our teachers. I have a lot of concern for teachers leaving, good teachers, leaving the industry because for financial reasons ... And I have teachers in my family, I have teachers that are clients, I have teachers that are friends, and they're exhausted and we're losing the good ones ...	*[Speaker name]*, a parent from Kennebec County, expressed her deep concern about the exodus of quality teachers from the education sector ... where participants shared their personal experiences and visions for the future of education in their community.	I found the second version of the excerpt and accompanying information to be the most informative and clear. It provided relevant details about *[speaker name]'s* concerns regarding teachers leaving due to financial reasons and inadequate retirement programs, without being redundant or vague ... offering a comprehensive understanding of the excerpt's context and significance.	• Additional information on speaker's background • Speaker's Name • Personal Experience

Original Excerpt	Generated Context	Annotator Comments	Social Attributes Present
... And then the greater question then becomes, what happens to community when the folks who are from this place are no longer of this place? It can go down, that's the avenue, and there are all these new really big houses, which is great ... They don't even have a backyard. It's just the house. There's no front yard, just a house. And so the purpose is in order to be as isolating and as isolated as humanly possible...	**[Speaker name]**, a pastor at **[church name]** in **Durham**, has been receiving increasing calls for rental assistance since the pandemic, highlighting the severe housing crisis in the area; *he moved to Durham in 2018 to pastor and has witnessed the city's demographic and socioeconomic shifts, which have exacerbated housing instability and poverty.*	I chose to focus on understanding the broader context and details of the housing and community issues discussed by *[speaker name]* because it provides a deeper insight into the underlying causes, impacts, and potential solutions to the problems he highlights, enriching the interpretation of the excerpt. This additional information helps contextualize his concerns and vision within the larger socioeconomic landscape of **Durham**.	• Personal experience • Location • Speaker's name
I want to help my community by making it so everyone in my community is able to be in affordable housing, whether that's giving fundraisers as well, I'm going to be an electrician when I grow up so I can help them give discounts and I think that's just going to help everyone continue on with their lives better.	*[Speaker name]*, a participant *from Kennebec County, Pittston*, shared his aspiration to support his community by ensuring access to affordable housing through fundraisers and offering discounts as an *electrician*; this was in response to [participant name]'s question about *how they want to contribute to their community and how their career fits into that*.	This version provided *context about the conversation*, *[speaker name]'s* thoughts and motivations, and clarified how his career as an *electrician* will help with affordable housing, making it easy to understand his perspective and goals without needing additional information.	• Speaker's name • Occupation • Question / Response that motivated the speaker

| Original Excerpt	Generated Context	Annotator Comments	Social Attributes Present
... and there may be special items needed to bring that about. I think that's something that we can work on. Just like, for instance, what's done with museums, sometimes people with vision can go and see and hear, and do everything. But for us, they sometimes have some kind of technology that makes it accessible for us.	The speaker, *[speaker name]*, is ***visually impaired*** and is advocating for the inclusion of facilities and activities that cater to the needs of visually impaired or deaf individuals; she emphasizes the importance of accessibility technology in public spaces, similar to what is available in museums.	The added context about the committee's goals and *[speaker name]'s* role provided a clear understanding of the issues being addressed and the significance of her request, helping to see the relevance of her comments in the larger discussion about accessibility.	• Speaker's name • Additional information on speaker's background • Personal experiences

C Datasets Statistics
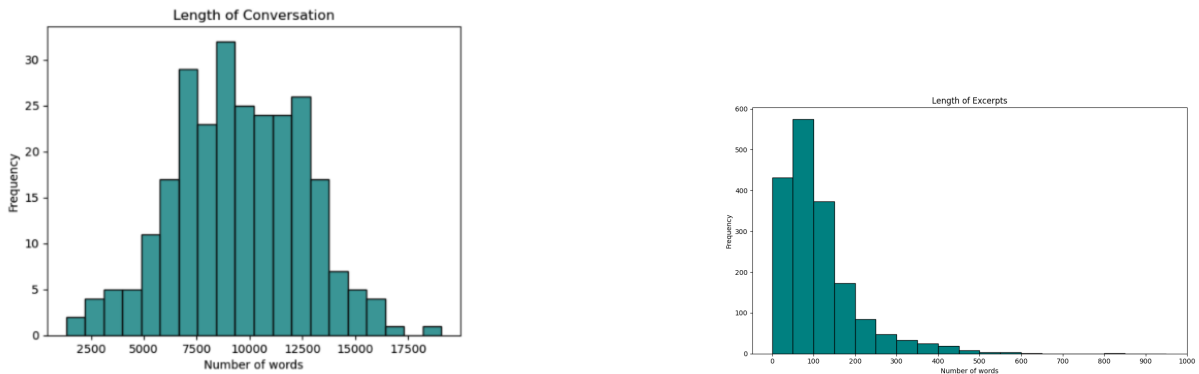
Figure 3: a) Length of conversation (in words) in Fora dataset and b) Length of excerpts (in words) in human-annotated excerpts dataset

D Prompts

D.1 Prompt for CEE_i Context Generation

System Prompt

```
Summarize and contextualize an excerpt using the full conversation transcript for
    enhanced understanding and significance.
Your goal is to build a foundation for the excerpt that fosters reader
    comprehension.
Additional information should only be retrieved from the full conversation
    transcript.
Do not make assumptions if the information is not explicitly mentioned in the
    conversation.
```

10255

Reflect on the main message or idea conveyed in the excerpt. Consider the full
    conversation and the broader context in which the excerpt occurred.
Add additional information to better understand the significance of this excerpt.
The transcript of the full conversation is:
{conversation}
The excerpt is:
{text}

D.2   Prompt for CEE_e Context Generation

System Prompt

Summarize and contextualize an excerpt using the full conversation transcript for
    enhanced understanding and significance.
Your goal is to build a foundation for the excerpt that fosters reader
    comprehension.
Additional information should only be retrieved from the full conversation
    transcript.
Do not make assumptions if the information is not explicitly mentioned in the
    conversation.
You should also add justification of the context added in the summary of the
    excerpt.
The objective is to enhance clarity by supplying both context and its rationale,
    aiding readers in fully comprehending the discussion's meaning and importance.

User Prompt

Reflect on the main message or idea conveyed in the excerpt. Consider the full
    conversation and the broader context in which the excerpt occurred.
Output the following in three separate lines without the headers.
It is important that there is ONLY ONE LINE for each of the following points
    SEPARATED BY A NEW LINE:
1.Add additional information from the conversation which would help better
    understand the excerpt and its significance in the broader context of the
    conversation. If the below is outputted in multiple lines, join them by a ;
Some of the items that could be added are:
- Background details of the speaker such as: name, demographic(such as age, gender)
    , cultural, ethnic (including race), educational, occupational and economic
    references made by or about the speaker in the conversation.\
Do not make assumptions about the speaker if not mentioned explicitly by the
    speaker or referred to directly about the speaker.
- Motivation or intent of the speaker as to why they said what they said. At which
    point in the conversation was the excerpt brought up. To what question or
    statement was the excerpt a response.
- Relevant locations or events mentioned in the conversation that could help ground
     the excerpt better and make it easier to understand such as the location of
    the conversation.
- Also, include background information about where the speaker is currently staying
     or where they grew up, if it would help in understanding the excerpt better.
- Share any specific personal experiences mentioned by the speaker that is related
    to the excerpt in the conversation.
2. Include exact sentences from the conversation that support all of the additional
     information added. Cite the statements, speakers, and line numbers that

support these additional anecdotes. Join them by a ;
3. Reasoning of added information: Explain why the added contextual information is
    crucial for a deeper understanding of the excerpt and its significance. Put
    this in one paragraph.
The transcript of the full conversation is:
{conversation)
The excerpt is:
{text}

D.3   Prompt for Terms Clarification using Extrinsic Knowledge Retrieval

System Prompt

You are a knowledgeable assistant that helps identify and define terms and
    locations that might not be common knowledge.
For each of these, look them up on the web and provide a concise definition or
    explanation (1-2 lines) that clarifies their meaning or significance within the
     context of the text.

User Prompt

I have a piece of text that may contain some terms and locations which might not be
     common knowledge.
Please identify those terms and locations that are less likely to be familiar to
    the general public.
For each of these, look them up on the web and provide a concise definition or
    explanation (1-2 lines) that clarifies their meaning or significance.
Multiple terms should be joined by a ; and the output should be in one line.

Example:
Text:
The speaker expressing concern about safety is Peter, a third-grade teacher from
    Cumberland County. The conversation is part of a larger discussion about the
    state of education, specifically in Maine, and even more specifically in MSAD
    11.

Output:
Term: The term "MSAD 11" refers to Maine School Administrative District 11, which
    is a school district in Maine that includes Gardiner, Pittston, Randolph, and
    West Gardiner.
Term: The term "Cumberland County" refers to a county in the state of Maine, which
    is located in the United States.

Here is the text that needs to be considered:

{text}

D.4   Prompts for Faithfulness Factor Extraction

System Prompt

You are an expert tasked with answering questions about a context paragraph.
    Follow these guidelines:
    1. Do not make assumptions beyond the provided text.
    2. Do not quote the question.

From the context below, answer the following questions:
Context: {context}

If any of the responses are not mentioned, please indicate that it is not mentioned
    . Do not quote the question in the answer.

Questions:

4.What is the name of the speaker? If the name is not directly mentioned but the
    number is, please specify the number only. If both are mentioned, mention the
    name only. If not mentioned, please answer "Not mentioned" exactly.
5.What is the gender of the speaker? If not mentioned, please answer "Not
    explicitly mentioned by the speaker" exactly.
6.What is the speaker's gender? If not, please answer "Cannot be inferred from the
    conversation" exactly.
7.What is the race mentioned by the speaker? If it is not mentioned, please answer
    "Not mentioned" exactly.
8.What is the speaker's educational background? If it is not mentioned, please
    answer "Not mentioned" exactly.
9.What is the speaker's occupation? If it is not mentioned, please answer "Not
    mentioned" exactly.
10.What is the speaker's age? If it is not mentioned, please answer "Not mentioned"
    exactly.
11.Indicate the speaker's economic status, such as low-income, middle-income, high-
    income, unemployed, retired, student, or self-employed, without making
    assumptions beyond the provided transcript. If it is not mentioned, please
    answer "Not mentioned" exactly.
12.Is there any additional information about the speaker's background, such as
    cultural details, ethnic background, or where they grew up? If yes, please
    specify.
14.Share any specific personal experiences mentioned by the speaker that are
    related, if applicable. If none, please answer "None".
15.What is the question or response from another person that prompted the speaker's
     statement? If the exact question or response is not given but the current
    conversation topic is, give the topic. Otherwise, please answer "Not mentioned".

16.Are there any significant events or locations that would provide additional
    context for understanding? If it is not mentioned, please answer "No".
17.Where did the group conversation take place? Please specify the county, city,
    state, or country if mentioned. If it took place over Zoom, you can specify
    that it is on Zoom, but also mention the city, state, or county. If not
    mentioned, please answer "Not mentioned".

E Demographics of Participants

E.1 Demographics of Participants of Effective Contextualized Excerpts - Human Evaluation
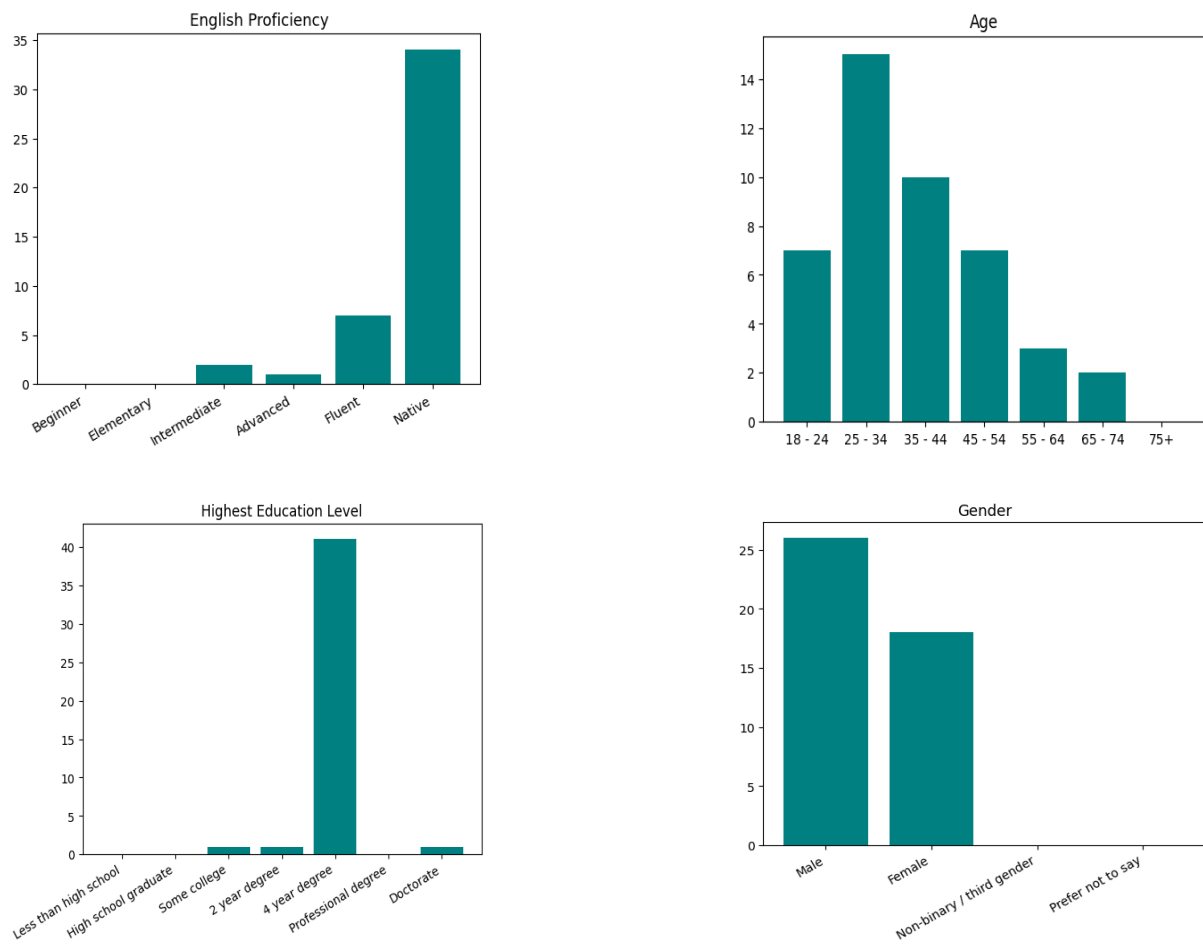
Figure 4: GPT Prompts Comparison Survey Demographics

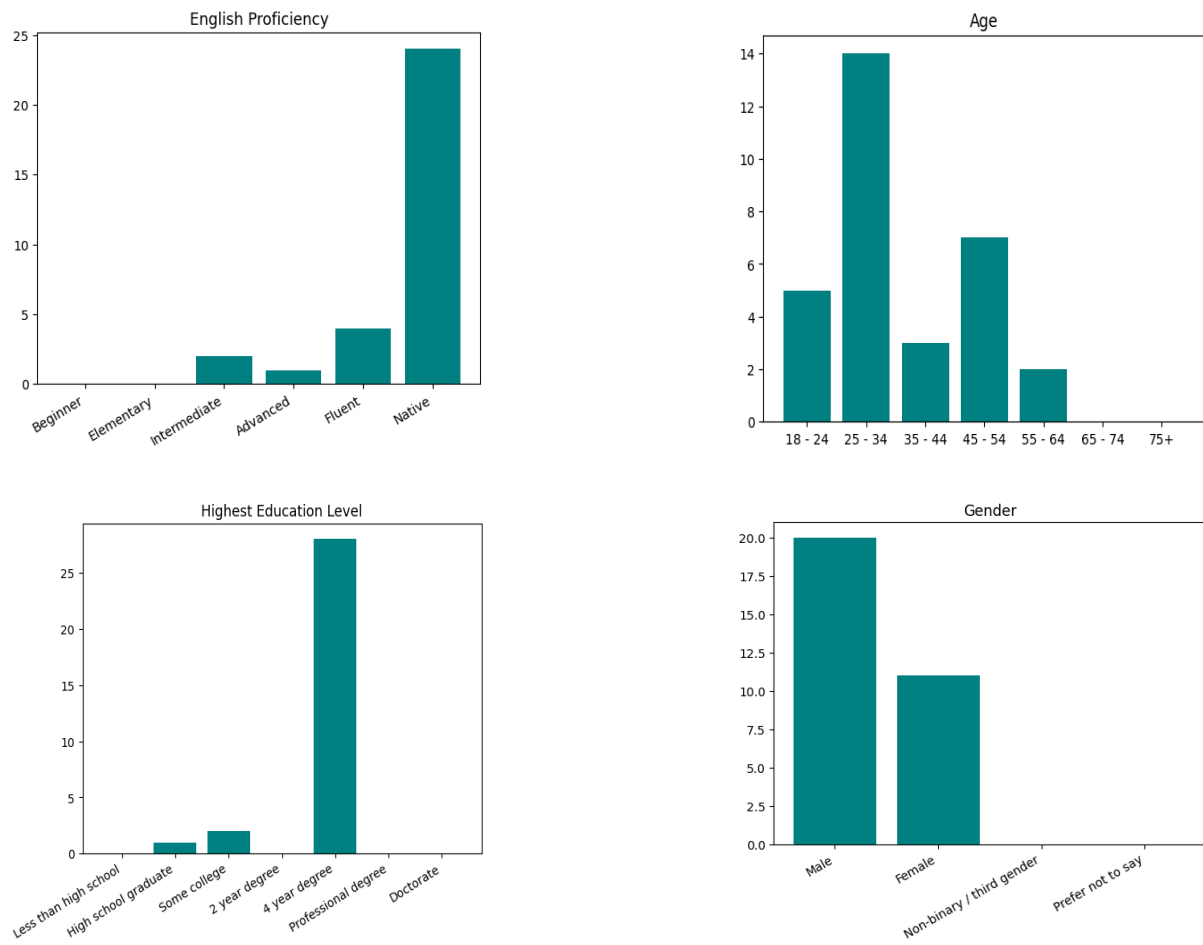## E.2 Demographics of Participants of LLM Contexts Comparison

Figure 5: LLM Contexts Comparison Survey Demographics

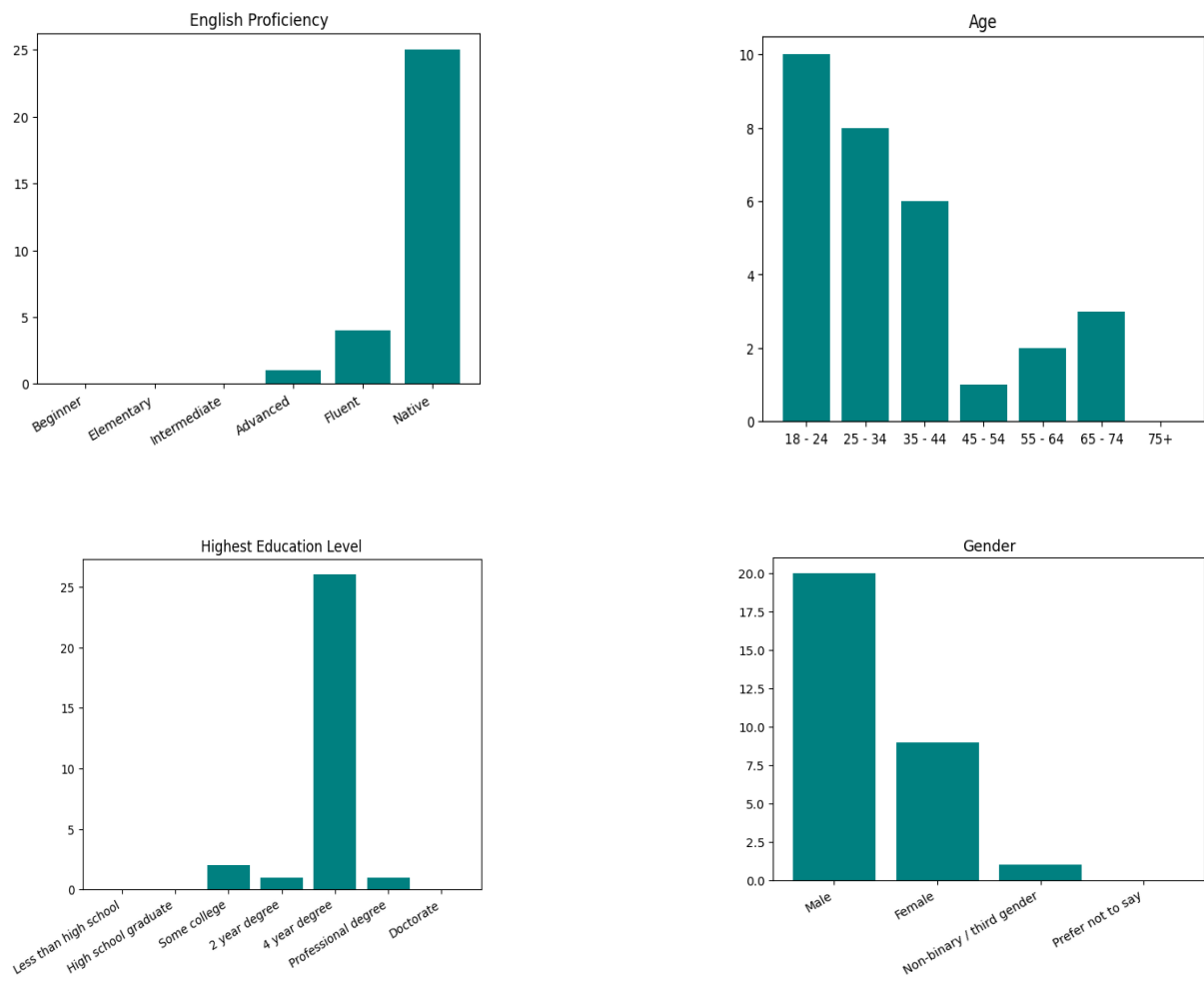## E.3 Demographics of Participants of Summary Task 2

Figure 6: Summary Survey 2 Demographics

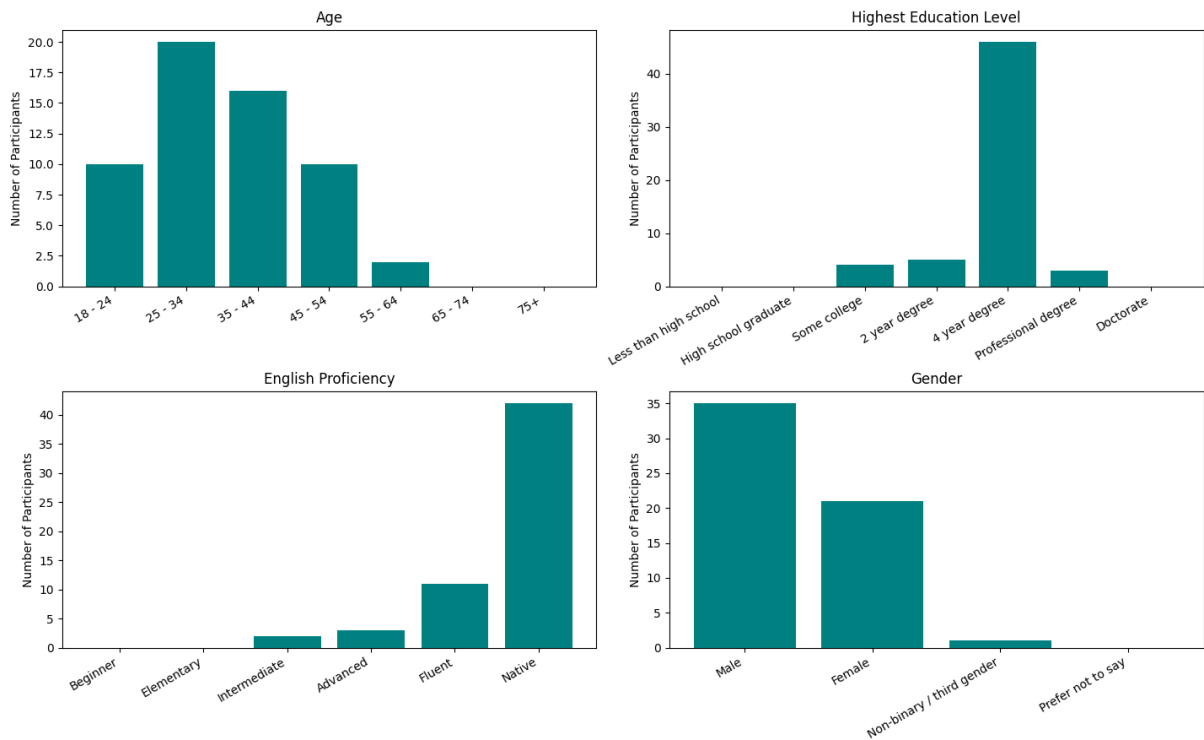### E.4 Demographics of HSE dataset Annotators

Figure 7: Crowd Source Survey Demographics

F Screenshots of surveys

F.1 Comparison of Explicit and Implicit Contextualisation Survey

Please rate how well you **understand** the meaning of what the speaker of the <u>excerpt</u> is talking about without needing additional context from the full conversation:

Understandability
1 (bad): The meaning of the excerpt is difficult to understand; it is confusing and lacks necessary context.
5 (good): The meaning of the excerpt is easy to understand fully; it is clear and comprehensive on its own without needing extra context.

| 1 | 2 | 3 | 4 | 5 |
| ○ | ○ | ○ | ○ | ○ |

Figure 8: Understandability Question

Please rate the **readability** of the <u>excerpt</u>:

Readability

1 (bad): The excerpt is highly difficult to read with rare words and complex structures.
5 (good): The excerpt is easy to read, well-structured, and flows naturally.

1	2	3	4	5
○	○	○	○	○

Figure 9: Readability Question

Please rate the **redundancy** of the <u>excerpt</u>:

Redundancy

1 (bad): The excerpt is excessively repetitive, containing unnecessary repetitions of the same information.
5 (good): The excerpt is concise and free from redundancy, explaining only essential information.

1	2	3	4	5
○	○	○	○	○

Figure 10: Redundancy Question

Please rate the **completeness** of the <u>excerpt</u>:

Completeness

1 (bad): The excerpt is incomplete missing key relevant information to understand the meaning.
5 (good): The excerpt is comprehensive, accurate, and includes all relevant information.

1	2	3	4	5
○	○	○	○	○

Figure 11: Completeness Question

Please rate the **cohesiveness** of the excerpt:

Cohesiveness
1 (bad): Sentences in the excerpt are highly incoherent as a whole. For instance, they are illogical, lack self-consistency, or contradict each other.
5 (good): Sentences in the excerpt fit together well. They are logically organized and coherent.

1	2	3	4	5
○	○	○	○	○

Figure 12: Cohesiveness Question

Do you agree with the speaker's **perspective** on the topic addressed in the excerpt?

*1: I **strongly disagree** with the perspective on the topic in the excerpt.*
*5: I **strongly agree** with the perspective on the topic in the excerpt.*

1	2	3	4	5
○	○	○	○	○

Figure 13: Perspective Question

Do you view the speaker of the excerpt as **honest and trustworthy**?

*1: I **strongly disagree** that the speaker is honest and trustworthy.*
*5: I **strongly agree** that the speaker is honest and trustworthy.*

1	2	3	4	5
○	○	○	○	○

Figure 14: Honest and Trustworthy Question

Do you have **respect** for the speaker of the <u>excerpt</u>?

1: *I strongly disagree* that I have respect for the speaker.
5: *I strongly agree* that I have respect for the speaker.

1	2	3	4	5
○	○	○	○	○

Figure 15: Respect Question

Can you feel the speaker's **emotions** in this <u>excerpt</u>?

1: *I strongly disagree* that I can feel the speaker's emotions.
5: *I strongly agree* that I can feel the speaker's emotions.

1	2	3	4	5
○	○	○	○	○

Figure 16: Empathy Question

Can you see the speaker's **point of view** in the <u>excerpt</u>?

1: *I strongly disagree* that I can see the speaker's point of view.
5: *I strongly agree* that I can see the speaker's point of view.

1	2	3	4	5
○	○	○	○	○

Figure 17: POV Question

What information according to you is missing and would have helped you **understand** this <u>excerpt</u> better?

Figure 18: Missing Info Question

Are there any **specific terms** in the excerpt that need to be explained further?

```
[                                                    ]
```

Figure 19: Clarify Terms Question

Do you have any other comments about this excerpt?

```
[                                                    ]
```

Figure 20: Comments Question

Do the explanations of the terms presented to you above help you understand the excerpt or any of the additional information provided for the excerpt?

1 (bad): The explanations **do not help me** understand the excerpt or any additional information at all.
5 (good): The explanations **are extremely helpful** and make the excerpt and any additional information completely clear and easy to understand.

1	2	3	4	5
○	○	○	○	○

Figure 21: ICL Explanation of Terms Question

Which version of the excerpt did you find clearer and prefer overall in understanding the intended meaning?

Figure 22: Rank Contexts Question

What made you choose the option of the **version** of the excerpt in the <u>previous question</u>?
Please provide your reasoning in a sentence or two.

Figure 23: Justify Rank Question

F.2 Crowd-Sourced Annotations for HSE

Please answer the following questions regarding its understandability:

Is the meaning of this snippet clear to you without further context?

○ No

○ Yes

Figure 24: Understandability YN Question

What specific aspects would enhance the understandability of this snippet if additional context is required?

Figure 25: Enhance Understandability Question

Are there any specific concepts or terms in the snippet that require further clarification? Please specify them.

Figure 26: Further Clarification Question

10267

Is the speaker's race mentioned in the **snippet**? For example, racial backgrounds may include but are not limited to: African American, Hispanic/Latino, Asian, Caucasian, Native American, Middle Eastern, etc.

○ No

○ Yes

Figure 27: Speaker Race YN Question

What is the race mentioned by the speaker in the **conversation/ snippet**?
If the snippet/conversation does not mention it, please indicate that it is not mentioned.

Figure 28: Speaker Race Text Entry Question

Is the speaker's occupation mentioned in the **snippet**?

○ No

○ Yes

Figure 29: Speaker Occupation YN Question

What is the speaker's occupation?
If the snippet/conversation does not mention it, please indicate that it is not mentioned.

Figure 30: Speaker Occupation Text Entry Question

Speaker Background

This section pertains to the background of the speaker of the snippet above. The task is to identify the speaker and gather information regarding their demographic, cultural, ethnic, educational, and economic references as explicitly mentioned in the **conversation in the file above**.

No assumptions or inferences should be drawn from sources other than the transcript of the conversation in the file above.

Is the name of the speaker mentioned in the snippet?

○ Yes

○ No

Figure 31: Speaker Name YN Question

What is the name of the speaker as mentioned in the conversation transcript or snippet directly?
If the name of the speaker is not directly mentioned but the number is. You can specify the number. If both are mentioned, mention the name.

[   ]

Figure 32: Speaker Name Text Entry Question

Is the gender of the speaker mentioned in the **snippet**?

○ Yes

○ No

Figure 33: Gender YN Question

10269

Did the speaker personally identify their gender during the **conversation**? Please answer only if the speaker themselves mentioned their gender, **NOT if others in the conversation referred to it.**

○ Male

○ Female

○ Non-binary / third gender

○ Not explicitly mentioned by the speaker

Figure 34: Speaker Gender Select Question

If the speaker **did not explicitly** mention their gender, did others in the conversation refer to the speaker in a manner that implied their gender? If so, please specify.

○ Male

○ Female

○ Non-binary / third gender

○ Cannot be inferred from the conversation

Figure 35: Speaker Gender Select 2 Question

Is the speaker's education background mentioned in the **snippet**?

○ No

○ Yes

Figure 36: Speaker Education YN Question

10270

What is the speaker's educational background?
If the snippet/conversation does not mention it, please indicate
that it is not mentioned.

Figure 37: Speaker Education Text Entry Question

Is the speaker's economic status referred to in in the **snippet**?

◯ No

◯ Yes

Figure 38: Speaker Economic Status YN Question

Indicate the speaker's economic status, such as low-income,
middle-income, high-income, unemployed, retired, student, or
self-employed, without making assumptions beyond the
provided transcript.

If the snippet/conversation does not mention it, please indicate
that it is not mentioned.

Figure 39: Speaker Economic Status Text Entry Question

Is the speaker's age mentioned in the **snippet**?

◯ No

◯ Yes

Figure 40: Speaker Age YN Question

10271

What is the speaker's age?
If the snippet/conversation does not mention it, please indicate that it is not mentioned.

[ ]

Figure 41: Speaker Age Text Entry Question

Broader Context

This section focuses on the broader context of the snippet and identifying relevant information from the conversation.

Share any specific personal experiences mentioned by the speaker that is related to the snippet, if applicable.

[ ]

Figure 42: Personal Experience Text Entry Question

Speaker Motivation

The following section aims to understand the speaker's motivation for their statement in the snippet.

What motivated the speaker to express their statement, and what was their underlying intent?

[ ]

Figure 43: Motivation Text Entry Question

Where did the group conversation take place? Please specify the county, city, state, or country if mentioned in the conversation. If the conversation is over Zoom, you can specify that it is on Zoom, but also mention the the city, state, or county.

[ ]

Figure 44: Location Text Entry Question

Are there any significant events or locations mentioned in the conversation that would provide additional context for understanding the snippet? For example, if the conversation is about the COVID-19 pandemic and the snippet is related to it, please specify.

[ ]

Figure 45: Events Text Entry Question

What is the question or response from another person that prompted the speaker's statement?

[ ]

Figure 46: Context Text Entry Question

Is there any additional information needed to better understand the context of the snippet, which has not been addressed by the questions above?

[ ]

Figure 47: Additional Info Text Entry Question

Consent Form

What follows is a brief overview of the research we are completing with your assistance.

In this survey, we'll present you with conversation snippets, transcripts of conversations, and related questions, and ask for demographic information.

This survey takes approximately **2 hours** to complete.

This survey is part of a scientific research project. Your decision to complete this survey is voluntary. If you give us permission by completing the survey, we plan to discuss/publish the results in an academic forum. In any publication, information will be provided in such a way that you cannot be identified. Only members of the research team will have access to the original data set. Before the data is shared outside the research team, any potentially identifying information will be removed. Once identifying data has been removed, the data may be used by the research team, or shared with other researchers, for both related and unrelated research purposes in the future.

Your anonymized data may also be made available in online data repositories such as the Open Science Framework, which allow other researchers and interested parties to use the data for further analysis.

Clicking on the arrow at the bottom of this page indicates that you are **at least 18 years of age** and agree to complete this survey voluntarily.

Figure 48: Consent Form