ACE: A LLM-based Negotiation Coaching System

Ryan Shea*, **Aymen Kallala***, **Xin Lucy Liu***, **Michael W. Morris**†, **Zhou Yu**†
Columbia University, New York, NY
rs4235@columbia.edu
{ak5078, xl2855, mwm82, zy2461}@columbia.edu

Abstract

The growing prominence of LLMs has led to an increase in the development of AI tutoring systems. These systems are crucial in providing underrepresented populations with improved access to valuable education. One important area of education that is unavailable to many learners is strategic bargaining related to negotiation. To address this, we develop a LLM-based **A**ssistant for **C**oaching n**E**gotiation (ACE). ACE not only serves as a negotiation partner for users but also provides them with targeted feedback for improvement. To build our system, we collect a dataset of negotiation transcripts between MBA students. These transcripts come from trained negotiators and emulate realistic bargaining scenarios. We use the dataset, along with expert consultations, to design an annotation scheme for detecting negotiation mistakes. ACE employs this scheme to identify mistakes and provide targeted feedback to users. To test the effectiveness of ACE-generated feedback, we conducted a user experiment with two consecutive trials of negotiation and found that it improves negotiation performances significantly compared to a system that doesn't provide feedback and one which uses an alternative method of providing feedback.

1 Introduction

The rapid progress of LLMs in recent years has spurred the creation of more sophisticated AI tutoring systems (Sonkar et al., 2023). These systems give learners easier access to training in areas outside what is typically available for most learners (Liang et al., 2023). One area that falls into this category are social competencies, such as negotiation tactics. Negotiation skills are crucial as they help individuals maximize their gains in competitive situations across different areas of their professional and personal life. However, these skills are hard to learn from a traditional classroom lecture because they involve reflexive behavioral habits. They are typically taught through small seminars centered on role-playing exercises and instructor coaching, which are expensive and limited in access.

Therefore, effective negotiation training is not available to many populations who lack it yet need it the most, such as women and minorities. These groups particularly lack negotiation skills and are less accustomed to advocating for themselves (Babcock and Laschever, 2003) which is a significant factor contributing to their relatively lower starting salaries and fewer promotion opportunities (Lu and Zhao, 2023). As such, increased coaching is needed to address gender and ethnic differences in negotiation performance (Amanatullah and Morris, 2010). Recent research also finds that linguistic assertiveness skills, such as those used in negotiations, vary as a function of cultural/ethnic backgrounds, partly account for ethnic differences in promotions to leadership positions, and are amenable to training interventions (Lu et al., 2020, 2022).

To democratize access to high-quality negotiation coaching, we proposed a LLM-based **A**ssistant for **C**oaching n**E**gotiation (ACE). ACE is designed to serve as a negotiation partner that provides learners with targeted feedback similar to what they would receive from an instructor in a seminar setting. To provide this feedback, we begin by collecting a dataset of negotiation transcripts between students in an Master's of business administration (MBA) negotiation class. The scenarios in these negotiations come from a rigorous curriculum and are designed to mirror real-world negotiation settings. Furthermore, the participants in these negotiations have been trained in bargaining tactics, unlike previous datasets which rely on crowd-workers (Lewis et al., 2017; He et al., 2018).

Using this dataset along with expert consulta-

* denotes equal contribution.
† denotes equal advising.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12720–12749
November 12-16, 2024 ©2024 Association for Computational Linguistics

tions, we develop an annotation scheme to identify and correct mistakes that users make during their negotiations. Our annotations mirror actual error categories that instructors look for in negotiation classes. We then build ACE according to this scheme. ACE uses the annotation categories to identify users' mistakes and then provides targeted feedback based on the error definitions, along with in-context examples of feedback written by experts. Our system also includes a prompt-based negotiation chatbot agent which serves as a simulated negotiation partner for learners.

To test the efficacy of ACE, particularly the feedback it provides, we recruited a group of 374 users who participated in two trials of negotiation. We found that ACE significantly improves learning outcomes compared to users who negotiate with no coaching from our system. Our contributions are summarized as follows:

- We propose a novel negotiation coaching system called ACE, which utilizes LLMs to provide targeted, individualized feedback to users.
- We develop a negotiation strategy annotation scheme to identify and correct user errors. Our annotation scheme is based on expert input and mirrors the kind of mistakes negotiation instructors look for in a seminar setting.
- We release an annotated dataset of spoken negotiation transcripts between MBA students. The negotiation scenarios in our dataset are based on standardized business school curricula and are carefully designed to mimic real world settings.
- We conduct an efficacy experiment with two trials of negotiations, finding that ACE boosted subjective and objective measures of tactical learning relative to a control system which does not provide feedback and an alternative method of providing negotiation feedback.

2 Related Work

2.1 AI Tutoring Systems

Generating automated training systems has been a long-standing issue for researchers in AI for education (Keuning et al., 2018). One common application of such systems is the area of computing education (Koutcheme et al., 2024). Recent work has explored the use of LLMs to generate automatic feedback for students on programming assignments (Pankiewicz and Baker, 2023) or for creating coding exercises (Denny et al., 2024). These works have shown that state-of-the-art LLMs such as GPT-4 (OpenAI, 2023) can provide effective feedback to users and can achieve up to 80% agreement with humans in certain scenarios (Zheng et al., 2023).

Another application of these systems has been the field of language learning. These systems typically involve a chat agent which guides users through a conversation on a specific topic followed by feedback on the users' conversation (Qian et al., 2023; Li et al., 2022). These systems have been judged as helpful by English languages learners and have also had a demonstrable effect on actual learning outcomes (Liang et al., 2023).

Prior work on developing systems for negotiation training has been fairly limited. These systems can only be interacted with by selecting from a list of pre-written options and deliver "canned" responses as a reply. Despite these limitations, prior work has shown that interacting with virtual agents can improve learners' understanding of negotiations (Gratch et al., 2016).

2.2 Negotiation

Methods that have applied LLMs to the area of negotiation have been focused on building negotiation chat agents. These methods typically use existing negotiation data to perform supervised learning or offline reinforcement learning on a negotiation model (Lewis et al., 2017; He et al., 2018; Verma et al., 2022; Zhan et al., 2024). More recent work has focused on examining and enhancing the negotiation capabilities of prompt-based negotiation agents (Schneider et al., 2023; Fu et al., 2023; Zhan et al., 2024).

Previous work on bargaining has focused on building chatbots to serve as negotiation partners. There is little work on providing user feedback. However, negotiation research has found that people do not learn negotiation tactics simply from the experience of bargaining; they need structured feedback and instruction (Loewenstein and Thompson, 2006). ACE is the first system to provide quality feedback using rigorous business school curricula to improve learning outcomes of negotiation.

3 Background

Negotiation is a general task and there are many ways to formulate a bargaining problem. In this section we give an overview of the the types of negotiation problems we consider when designing our coaching system and annotation scheme.

The negotiation settings we consider here are ones which involve one agent selling a single item to another agent. This type of negotiation is known as single-issue distributive bargaining (Lewicki et al., 2021). The agent selling the item is referred to as the "seller" and the agent buying the item is called the "buyer". Both the buyer and seller have access to a "role" (also referred to as a "negotiation scenario") which provides private information about their options and preferences and public information about details of the object for sale and the range of market prices for this item. In a role-play simulation, participants use their role to prepare a strategy, this includes their upper limit or "walk-away" price, their target price, and the opening price that they will mention. An example of buyer and seller roles for a used car negotiation can be found in Appendix D.

4 Dataset

Our dataset was collected in collaboration with an instructor who teaches a course on negotiation to MBA students. During this course, 50 students were randomly assigned to dyads for a negotiation task. They conducted three successive negotiations against three different counterparts, involving different kinds of cars. Before negotiating, students answered standard preparation questions and then began the negotiation while recording their conversation audio. We transcribed the conversation audio using OpenAI's Whisper API (Radford et al., 2022) along with manual edits to construct our dataset.

A total of 42 dialogues were collected from this process, of which 40 were usable. A summary of the dataset statistics is in Table 1. We show the statistics for the entire dataset along with a breakdown by the negotiation task/scenario. The scenarios are based on the type of car being negotiated over in the exercise. We focused on negotiation over a Honda Accord (Task 1) as the transparent market range for such a product makes it typical of negotiation over a commodity item, see Appendix D.

	Task 1	Task 2	Task 3	Total
# of conversations	14	13	13	40
Avg. # of turns per conversation	23.8	22.3	15.6	20.6
Avg. # of tokens per turn	31.4	35	37	34.1
Vocabulary size	950	1022	820	1723
Deal %	93%	31%	100%	75%
Deal Amount	\$12.9k	\$7.3k	\$1.3k	\$7.2k

Table 1: A summary of statistics from our negotiation dataset.

5 Annotation Scheme

We designed a negotiation error annotation scheme to identify and correct user mistakes. Our scheme is based on the dataset collected in Section 4, expert input, and common distributive bargaining tactics from Lewicki et al., 2021.

We identified eight error categories that can be divided into **preparation errors** and **negotiation errors**. All the categories in our scheme are binary True/False labels, where a label of False indicates that a mistake has been made by the negotiator. Our annotation scheme is designed from the perspective of the buyer but can be easily adjusted to fit the seller's perspective.

The two categories below belong to **preparation errors**, which are mistakes in users' answers to a set of standard preparation questions before the negotiation. These categories are designed to identify whether the user prepared their negotiation strategy correctly.

1. Strategic walk-away point assesses whether the user has properly analyzed the facts of the negotiation scenario to set an appropriate walk-away price. From the buyer's perspective, a walk-away price is the maximum amount they would pay to purchase the item in the negotiation. If the scenario outlines an explicit budget limit, then a strategic walk-away point is one which exactly matches the budget amount. Otherwise, we consider any point below the maximum market price to be strategic.

2. Strategic target price evaluates the target price that the buyer sets before the negotiation. The buyer's target price is strategic if it falls within the first third of the range between the minimum market price and the buyer's walk-away point. If the buyer sets their target below this range, it is too ambitious to be a realistic outcome to aim for. Conversely, if their target is above this range, it is too weak to test how far their opponent can be pushed in the negotiation.

The six categories below refer to **negotiation errors**, which are mistakes that users make "at the

Role	Utterance	Annotation
Buyer	Hi, I'm new to California and I'm looking for probably a Honda Accord with reasonable mileage around maybe $11,000 to $12,000. Do you have anything like that?	Breaking the ice Giving the first offer Ambitious opening point Including rationale
Seller	Nice. We have something similar. We have a nice 2013 Honda. It does have a little bit more miles than that. It has about 50,000. It doesn't have any rust and it's in great condition. What's the price range you're looking to come out with?	
Buyer	Probably around $11,000 or $12,000.	Strong counteroffer Including rationale
Seller	Ooh, that's kind of rough. Our sticker price for this car is closer to $14,000.	
Buyer	Ooh, yeah, that's definitely a little bit too much. Could I take it for a test drive maybe?	
Seller	Sure.	
Buyer	Okay, great. Yeah, it's pretty good. What do you think about maybe $12,500 and I would buy it today?	Strong counteroffer Including rationale
Seller	$12,500. I mean, could we call it even $13,000?	
Buyer	Yeah, I could probably do $13,000.	Strong counteroffer Including rationale
Seller	All right.	
Buyer	All right.	Strategic closing
Seller	Sounds great.	

Table 2: An example negotiation dialogue from our dataset. Annotations in red indicate that the category was labeled as False. Annotations in green indicate that the category was labeled as True.

bargaining table," during the negotiation dialogue.

3. Breaking the ice refers to whether or not the user began the negotiation with some social bonding. A negotiator should spend their first conversational turn on social remarks unrelated to the negotiation issues.

4. Giving the first offer indicates whether the user stated the first price offer in the conversation. Negotiators are advised to state their opening price first to anchor the negotiation in a favorable position (Lewicki et al., 2021).

5. Ambitious opening point assesses the tactical quality of the user's opening offer relative to their target price. When the buyer proposes a price first, we consider the offer O_1 strong if:

$$O_1 \leq 0.9 * T$$

with T being the buyer's target price. Otherwise, when the seller previously made an offer S, we consider the buyer's first offer strong if it creates a midpoint at or below their target price:

$$\frac{S + O_1}{2} \leq T$$

6. Strong counteroffer assesses the quality of the user's first three proposals following their first offer. A counter-offer O_t is considered strong if it's below the midpoint of the remaining bargaining range:

$$O_t < \frac{O_{t-1} + min(S, W)}{2}$$

where O_{t-1} is the buyer's previous offer, S is the seller's current offer, and W is the buyer's walk-away point.

Negotiation Error Category	Number of Turns with Errors	Number of Applicable Turns
Breaking the ice	28	40
Giving the first offer	15	40
Ambitious opening point	18	40
Strong counteroffer	34	73
Including rationale	25	112
Strategic closing	36	40
Preparation Error Category	**Number of Errors**	**Number of Dialogues**
Strategic walk-away	7	40
Strategic target price	15	40

Table 3: A summary of our annotated negotiation dataset. Note that mistakes related to **negotiation errors** tend to be more common that **preparation errors**

7. Including rationale indicates whether the user's first four price offers were accompanied by a rationale. We define a rationale as any reasoning that supports a price offer (Lewicki et al., 2021).

8. Strategic closing behavior refers to whether the user closed the deal in ways that heighten the counterpart's commitment. The final two turns of the negotiation should contain either an acknowledgment of the opponent's negotiation skill or a recounting of their own concessions. The closing turns should not contain any celebratory statements about the negotiation outcome or any statements implying that the user got a better deal.

We annotated our collected dataset (Section 4) based on this annotation scheme with an inter-annotator agreement of 0.87, according to Cohen's kappa (Cohen, 1960). Inter-annotator agreement was calculated based on a subset of 288 dialogue turns annotated by two authors. A breakdown of the errors that buyers made according to the annotation scheme can be seen in Table 3. An example of an annotated conversation is given in Table 2. The most prevalent errors that buyers made were those in the **strategic closing** category, with 36 out of the 40 conversations containing an error related to this. On the other hand, we see that buyers did a good job including a rationale in their price offers, with only 25 of these errors present out of a total of 112 relevant conversation turns.

6 Approach

In this section, we outline our approach to designing ACE. We first provide a high-level overview of the system and how users progress through it. We then describe the components of the system, including the negotiation chatbot agent and feedback modules.

6.1 System Overview

ACE begins by presenting users with a negotiation scenario similar to those in our dataset (Section 4). This is followed by a set of negotiation preparation questions which ask users for their target price, walk-away price, and planned opening point. Users then proceed into a simulated negotiation with our negotiation agent until they reach an agreement. After that we provide users with the feedback associated with their preparation questions, referred to as "preparation feedback." This is followed by "negotiation feedback" on the negotiators' linguistic performance, including both turn-specific points about their tactics and holistic points about their diction, tone, and politeness.

6.2 Negotiation Agent

Our negotiation agent is based on a prompted version of GPT-4 (OpenAI, 2023). Prior work has found that LLMs can achieve successful negotiation outcomes with proper prompting (Fu et al., 2023). Therefore, we adopted this approach for our agent. We found that prompting GPT-4 with a full negotiation scenario (Appendix D) resulted in nonstrategic oversharing; therefore, we use a summarized version as our instructional prompt (Table 20).

Even with the summarized prompt, the agent is prone to making "weak" counteroffers, meaning that is often pushed to offering its walk-away price too easily. To avoid a "pushover" negotiation agent, we employ dynamic prompting. Specifically, we give the agent an initial "reservation price" in the instructional prompt, which is higher than the actual reservation price given in the negotiation scenario. We refer to this price point as the agent's "subjective limit." We adjust the subjective limit as the conversation progresses and eventually set it to the true reservation price after several conversation turns have passed. This allows us to control the bot's counteroffers to a certain degree while allowing for variation in the offers presented. We initially set the bot's subjective limit to a random value within the range given by **strategic target price** (Section 5). After the first turn, we update it to a price that corresponds to a **strong counteroffer**. This ensures that the bot gives robust counteroffers and will not reach its true reservation price too quickly.

6.3 Preparation Feedback

The preparation feedback we provide users is based on their answers to pre-negotiation preparation questions. We specifically look for errors corresponding to **strategic walk-away point**, **strategic target price**, and **giving the first offer** in our annotation scheme (Section 5). Since each of these errors is identified with a mathematical formula, flagging them is trivial. We give feedback to users using either a hard-coded message with the correct answer and an explanation, or we prompt GPT-4o to generate feedback given the user's answer, the correct answer, and in-context examples of quality feedback written by expert negotiators (Appendix F). A full example of preparation feedback can be found in Appendix E.

6.4 Negotiation Feedback

Negotiation feedback is provided based on users' conversations with the automated bargaining agent. We give users two categories of negotiation feedback. The first is feedback associated with individual conversation turns, which we call "turn-based feedback". The second is "holistic feedback" which corresponds to the conversation as a whole such as the conversational tone or politeness level.

6.4.1 Turn-Based Feedback

Figure 1 gives an overview of the pipeline ACE uses to provide turn-based feedback. Our pipeline consists of three components: **1) error identification**, **2) direct feedback** and **3) utterance revision**. We describe each of these components in detail

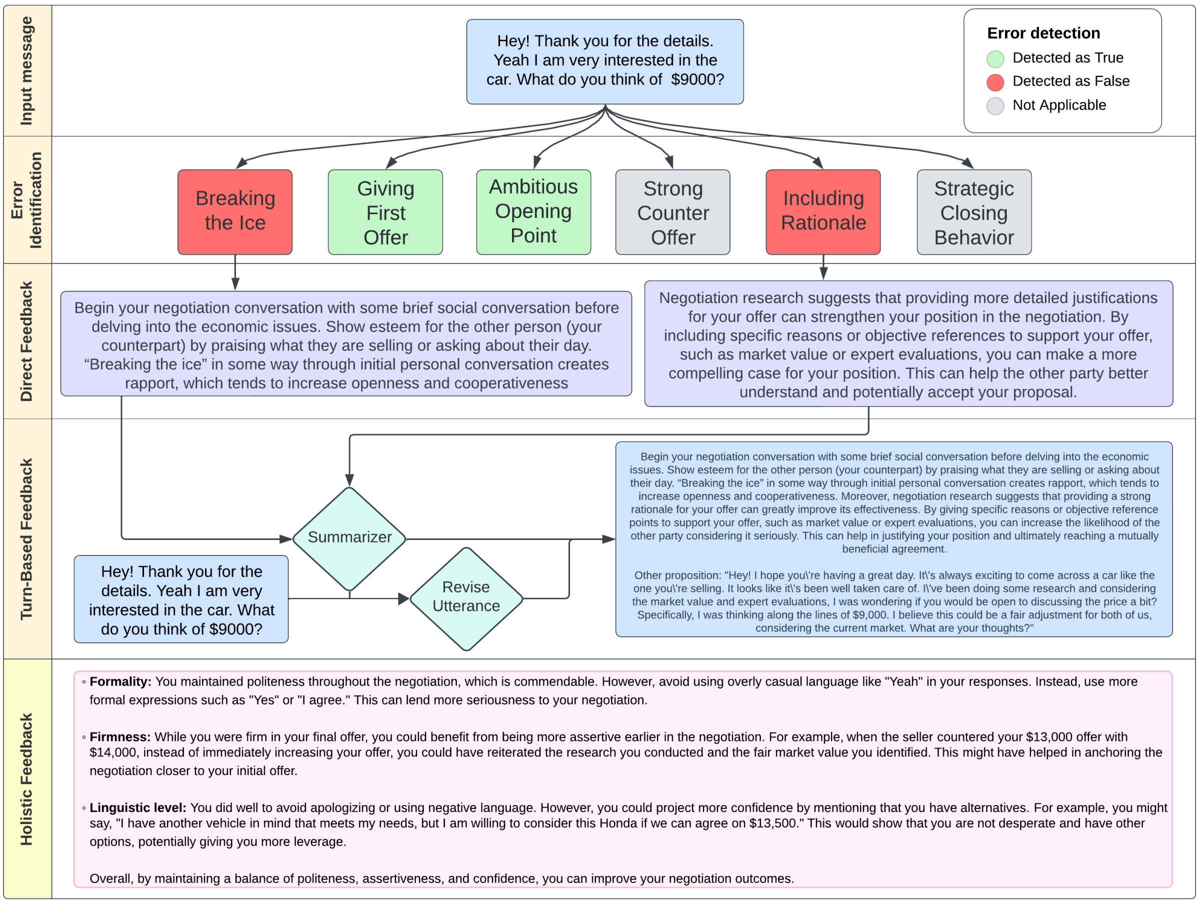

Figure 1: Diagram illustrating the turn-based feedback flow for ACE as well as an example of holistic feedback.

below.

Error identification. To provide effective feedback, we start by detecting errors the user made at each conversational turn of their transcript. The errors we flag for turn-based feedback are based on the **negotiation errors** in our annotation scheme (Section 5). We divide these errors into two groups and use a different strategy for detecting each of them.

The first group of errors is based on price offers and involves applying a formula. These include categories such as **strong counteroffer** or **ambitious opening point**. The main challenge in identifying these mistakes is extracting the relevant price from the user utterance. We do this by prompting GPT-4 to extract the price. Our prompt consists of the user utterance followed by nine hand-written, in-context examples of successful price extractions (Table 14). After extracting the price, we apply the relevant formula to detect errors.

The second group of errors is based on the users' language rather than prices, such as **including rationale** or **breaking the ice**. For each turn, we use a set of classifiers, one for each of the three relevant categories, to determine whether the turn contains an error. We create our classifiers by prompting GPT-4o to output a True/False label for the turn. Our prompt consists of the user utterance along with the error definition. For more difficult categories, such as **including rationale**, the prompt also contains in-context examples from our collected dataset (Table 15).

Direct feedback. When we identify a turn with a mistake, we prompt our GPT-4o to give direct comments on each error committed. This includes an explanation of the error and the tactical value of correcting it, see Figure 1 for examples. Our prompt consists of the conversation context, a definition of the error committed (Section 5 and Table 16), and one or two in-context examples of feedback written by negotiation instructors. We generate feedback for each mistake committed, which can result in the comments being quite lengthy. Therefore, in cases where the user commits more than one mistake in a turn, we give a summary of the error explanations as the final direct feedback.

Utterance revision. Along with the direct feedback, we also present users with a revised version of their utterance with their mistakes corrected. We prompt GPT-4o with the user utterance and the direct feedback to generate the corrected utterance. Our prompt includes the user utterance, the direct feedback, and three in-context examples of handwritten utterance revisions (Table 17).

The final turn-based feedback given to the user is the direct feedback concatenated with the revised user utterance. See Figure 1 and Appendix E for complete examples.

6.4.2 Holistic Feedback

The holistic feedback presented to the user is not based on any specific errors identified within the transcript. Instead, we prompt GPT-4o to comment on the linguistic aspects of the user's conversation. The aspects we focus on are formality, firmness, and linguistic level. For formality, the user should stay polite and avoid being rude or pushy. It's also better for users to be firm and assertive in their language, as studies have shown that this communication style leads to better deals (Jeong et al., 2019). In terms of linguistic level, users should avoid apologizing or using language that could be interpreted as a personal attack (Fisher et al., 2011). We prompt the model with a summary of the attributes for these three aspects and have it generate feedback. We have the model quote specific phrases from the users' transcript to make the comments more targeted and personalized (Table 19). See Figure 1 for a full example.

7 Evaluation

7.1 Error Identification Evaluation

We use a subset of the annotated data from Section 5 to evaluate how well ACE can identify mistakes. This subset consists of 26 dialogues (with a total of 494 conversational turns) which were excluded during the creation and testing of our coaching system. Since the **preparation errors** in our annotation scheme are trivial to identify, we measure how well ACE can classify **negotiation errors**.

Table 4 shows a breakdown of the system's accuracy. Using human annotations as our ground truth labels, we measure how well ACE predicts these labels. The system can identify mistakes with a high accuracy of at least 0.90 for all error categories. However, it performs worse in terms of precision and recall, with the F1 score ranging from a low of 0.54 for **strategic closing** to a high of 0.97 for the **giving the first offer** category. This suggests that ACE has difficulty balancing precision and recall for some annotation categories. Low recall in particular seems to be driving the lower scores, which implies that ACE has trouble identifying true positive cases for some difficult categories.

Error Category	Accuracy	Precision	Recall	F1 Score
Breaking the ice	0.98	0.99	0.76	0.83
First offer	0.99	0.95	0.91	0.93
Strong first offer	0.98	0.91	0.83	0.85
Strong counteroffer	0.96	0.74	0.73	0.73
Including rationale	0.90	0.81	0.63	0.67
Strategic closing	0.94	0.72	0.53	0.54

Table 4: A table indicating how accurately ACE can identify user mistakes. Our system is able to detect errors with high accuracy, but performs worse in terms of precision and recall. This is reflected in the lower F1 scores for **including rationale** and **strategic closing**.

7.2 ACE-generated Feedback Evaluation

We next evaluate the feedback generated by ACE via a user experiment. Our experiment consists of two pilot studies and a main experiment. In our first pilot study (Pilot Study A) we recruited 100 native English-speaking U.S. participants from Prolific and asked them to perform two trials of used car negotiation. All participants in all trials were assigned the role of the buyer based on the used car scenario in Figure 4. Participants were randomly assigned to treatment or control conditions, and ACE feedback was only provided in the first trial of the treatment condition. Users in the control condition were given no feedback. The results of the pilot study showed that the improvements in the the treatment condition were significantly larger than in the control condition. Additional details for Pilot Study A can be found in Appendix B.

To exclude the possibility that the improvements were due to the learning of situation-specific inert knowledge limited to the used car scenario, we developed a new negotiation scenario. This new scenario involves negotiating over a the price of summer sublease and can be seen in Figures 6 and 7. We tested its difficulty in Pilot Study B to ensure it has no difference with the used car scenario. We recruited 46 Prolific participants to negotiate against the ACE chat agent in the role of the buyer. T-tests revealed no significant differences between the standardized deal prices of the two scenarios ($p = 1$). Their distributions also showed no significant differences (Mann-Whitney U test: $p = 0.9$; Levene

	ACE Condition (N=119)	Other-feedback (N=129)	No-feedback Baseline (N=126)	*Between-person Comparison*
1st Trial: Used Car	\$12,891 (623.25)	\$12,889 (779.13)	\$12,948 (789.95)	F (2, 371) = 0.27, p = 0.77
2nd Trial: Summer Sublease	\$7,528 (658.72)	\$7,751 (459.72)	\$7,827 (431.65)	F (2, 371) = 10.79, p < 0.001
Within-person Comparison	t = -2.97, p = 0.003	t = 0.30, p = 0.76	t = 1.03, p = 0.30	F (2, 371) = 8.80, p < 0.001

Table 5: A summary of outcomes from the human evaluation of ACE. Our results show that buyers who received feedback from ACE performed significantly better in a following negotiation than those who didn't. *Note*: 1. Among 371 participants, 52% were female; $\bar{x}_{age}$ = 34.48 years, SD_{age} = 8.97; 60.1% were White/Caucasian; 54.4% had a bachelor's degree and above. The median completion time for this study was 39.68 minutes. 2. Standard deviations are presented in parentheses. 3. d in reported results stands for Cohen's d, a measure of effect size, calculated as the difference between two means divided by the pooled standard deviation of the data.

test: p = 0.6). Therefore, we can confidently state that the difficulty of the two scenarios is equivalent so that any differences that emerge across trials can not come from the scenario itself.

After the two pilot studies were completed we performed our main experiment. The main experiment was designed based on Pilot Study A but with several enhancements. The experimental details and results of this study are given in the following sections.

7.2.1 Experimental Setup

For our main experiment, we recruited 390 U.S.-based native English speakers from Prolific. This sample size was chosen based power analysis for a medium-size effect ($f = 0.2$) which determined that $N > 390$ is required for robust results of ANOVA tests across three conditions. Each participant was paid \$8 for completing this 40 minute online study. We obtained a total of 374 valid responses.

After consenting to this anonymous study, participants engaged in two trials of negotiation with a bot that simulated standard distributive bargaining tactics. All participants were assigned the role of the buyer. The first trial of negotiation was based on the used car scenario in Figure 4 and the second trial was based the summer sublease scenario in Figure 6. After each negotiation, they answered a round of questions. Participants were randomly assigned to one of three conditions. In the ACE condition, participants received ACE-generated feedback. In the "Other-feedback" condition, participants were given feedback based on the method described in Fu et al., 2023. This approach leverages the zero-shot capabilities of GPT-4 to give feedback and has been shown to improve the negotiation performance of LLMs (Fu et al., 2023). Lastly, participants in the "No-feedback" condition received no feedback on their performance. The procedure is illustrated in Figure 2. We preregistered the design for our user experiment at aspredicted.org/NPR_36R.

Before starting, participants in all conditions answered four questions about their goals in the negotiation (e.g., how important it would be for them to reach a favorable deal, to be a tough bargainer). As shown in Table 6, participants in the conditions showed no differences in any of these goals, suggesting that the differences in learning outcomes between conditions should be attributed to the effectiveness of ACE rather than their motivation level.

For the first trial, participants in the ACE condition ($N = 119$) proceeded through the ACE system as described in Section 6.1. Participants in the Other-feedback condition ($N = 129$) proceeded through a similar system but were given feedback according to the method outlined in Fu et al., 2023. Users in the No-feedback condition ($N = 126$) followed a similar procedure but were not given any feedback at the end of the negotiation. In every condition, we asked participants a series of reflection questions to ensure they digested any feedback they were given and prepared their strategy for the next round of negotiation (See Table 9).

For the second trial, participants in both conditions practiced with a negotiation agent with no feedback provided, as participants in the No-feedback condition did in their first trial. Then we asked all participants their subjective perceptions of improvement in the second negotiation compared to the first. A sample item was "*Compared to the first round of negotiation, in the second negotiation, I felt more confident.*" See details in Appendix A.

7.2.2 Results

Objective Improvement. We extracted the final deal price participants settled on from their chat history with the negotiation agent. Lower prices indicate more successful negotiations for the participants as buyers, so reaching a lower price in the second trial provides objective evidence of learning. As may be seen in Table 5, the performance improvement was significant in the ACE condition ($t = 2.97$, 95% CI = [0.14 ,0.70], $p = 0.003$, $d = 0.38$), while no significant change in deal price was observed in Other-feedback condition or the No-feedback condition ($t = -0.23$, $p = 0.82$, $t = -1.03$, $p = 0.30$). A significant 2 × 3 two-way mixed ANOVA ($F(2, 368) = 8.67$, $p < 0.001$) indicated that the magnitude of improvement differed across conditions.

Further, 2 × 2 ANOVAs suggested that the improvement in the ACE condition was significantly higher than in the No-feedback condition ($F(1, 246) = 12.82$, $p < 0.001$) or the Other-feedback condition ($F(1, 240) = 10.04$, $p = 0.002$), confirming the effectiveness of the ACE condition compared to the other conditions. Additionally, the lack of a significant difference between the Other-feedback and No-feedback condition ($F(1, 250) = 0.57$, $p = 0.451$) indicates that Other-feedback did not improve negotiation performance, underscoring the need for detailed, targeted feedback like provided by the ACE system. Additionally, this suggests that the kind of feedback that can help LLMs improve their task performance may not aid human performance. The feedback method in Fu et al., 2023 has been show to improve the abilities of LLM negotiators (Appendix C) but our user experiment demonstrates that this type of coaching may not be useful for humans.

Subjective Improvement. Participants in the ACE condition also reported a higher score for perceived improvement in the second negotiation ($\bar{x} = 4.34$, $SD = 0.62$) compared to those in the Other-feedback ($\bar{x} = 4.19$, $SD = 0.70$) and No Feedback ($\bar{x} = 4.17$, $SD = 0.66$) conditions. A significant 2 × 3 two-way mixed ANOVA ($F(2, 371) = 2.59$, $f = 0.12$, $p = 0.08$) indicated that the magnitude of improvement differed across conditions.

8 Conclusion

AI tutoring systems have the potential to democratize high-quality education in key areas such as negotiation. Prior work has shown that additional negotiation coaching is needed to correct for systematic gender and ethnic differences in bargaining performance (Amanatullah and Morris, 2010; Lu and Zhao, 2023). To address this need, we built ACE to mimic the coaching learners would receive from a professional negotiation instructor in a seminar setting. ACE was built based on a dataset and annotation scheme created in collaboration with experienced negotiation instructors. We evaluated our system and confirmed its error identification accuracy and feedback effectiveness. Results from a user experiment demonstrated that users who interact with ACE improved their negotiation performance significantly compared to those who do not recieve ACE-coaching.

Limitations

There are some limitations to the ACE system. The challenge for our negotiation agent is that LLM models tend to be agreeable and are not hardball bargainers. They respond to the user, so if users refused to discuss price, our negotiation agent wouldn't talk about price either. Additional work is needed to ensure our negotiation agent can guide users towards discussing price, especially when learners may be reluctant to fully engage in the negotiation.

Another limitation of ACE is that it does not retain any memory of previous user interactions. As a result, the utility of the feedback and negotiation agent may diminish for individuals who want to engage in repeated interactions. More work is required to enable our system to retain previous user interactions and tailor the chat agent and feedback accordingly.

Finally, the annotation scheme and feedback method we built is based on an "American" style of negotiation. Other cultures have different standards for effective bargaining and what constitutes a mistake in negotiation. Therefore, ACE may not be as useful for individuals who want to improve their negotiation capabilities in other cultural contexts.

Acknowledgments

We would like to thank Shehan Panditharatne, Xianglu Tang, and Aylin Hadzhieva for their help in implementing our user experiments.

References

Emily Amanatullah and Michael Morris. 2010. Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others. *Journal of personality and social psychology*, 98:256–67.

Linda Babcock and Sara Laschever. 2003. *Women Don't Ask Negotiation and the Gender Divide*. Princeton University Press, Princeton.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *Preprint*, arXiv:2402.05863.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Paul Denny, Juho Leinonen, James Prather, Andrew Luxton-Reilly, Thezyrie Amarouche, Brett A. Becker, and Brent N. Reeves. 2024. Prompt problems: A new programming exercise for the generative ai era. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 296–302, New York, NY, USA. Association for Computing Machinery.

Roger Fisher, William Ury, and Bruce Patton. 2011. *Getting to Yes: Negotiating Agreement Without Giving In*, 3 edition. Penguin Books, New York.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *Preprint*, arXiv:2305.10142.

Jonathan Gratch, David DeVault, and Gale Lucas. 2016. The benefits of virtual humans for teaching negotiation. In *Intelligent Virtual Agents*, pages 283–294, Cham. Springer International Publishing.

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

Martha Jeong, Julia Minson, Michael Yeomans, and Francesca Gino. 2019. Communicating with warmth in distributive negotiations is surprisingly counterproductive. *Management Science*, 65(12):5813–5837.

Hieke Keuning, Johan Jeuring, and Bastiaan Heeren. 2018. A systematic literature review of automated feedback generation for programming exercises. *ACM Trans. Comput. Educ.*, 19(1).

Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. 2024. Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. *Preprint*, arXiv:2405.05253.

Roy J. Lewicki, David M. Saunders, and Bruce Barry. 2021. *Essentials of negotiation / Roy J. Lewicki, David M. Saunders, Bruce Barry.*, seventh edition, international student edition. edition. McGraw-Hill Education, New York, NY.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Yu Li, Chun-Yen Chen, Dian Yu, Sam Davidson, Ryan Hou, Xun Yuan, Yinghua Tan, Derek Pham, and Zhou Yu. 2022. Using chatbots to teach languages. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*, L@S '22, page 451–455, New York, NY, USA. Association for Computing Machinery.

Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss, and Luke Fryer. 2023. ChatBack: Investigating methods of providing grammatical error feedback in a GUI-based language learning chatbot. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 83–99, Toronto, Canada. Association for Computational Linguistics.

Jeffrey Loewenstein and Leigh Thompson. 2006. *Learning to Negotiate: Novice and Experienced Negotiators*, pages 77–97. Frontiers of Social Psychology. Psychology Press.

Jackson Lu and Michelle Zhao. 2023. Empowering asians' leadership advancement with debate training: A longitudinal field experiment. *Academy of Management Proceedings*, 2023(1):18120.

Jackson G. Lu, Richard E. Nisbett, and Michael W. Morris. 2020. Why east asians but not south asians are underrepresented in leadership positions in the united states. *Proceedings of the National Academy of Sciences*, 117(9):4590–4600.

Jackson G. Lu, Richard E. Nisbett, and Michael W. Morris. 2022. The surprising underperformance of east asians in us law and business schools: The liability of low assertiveness and the ameliorative potential of online classrooms. *Proceedings of the National Academy of Sciences*, 119(13):e2118244119.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maciej Pankiewicz and Ryan S. Baker. 2023. Large language models (gpt) for automating feedback on programming assignments. *Preprint*, arXiv:2307.00150.

Kun Qian, Ryan Shea, Yu Li, Luke Kutszik Fryer, and Zhou Yu. 2023. User adaptive language learning chatbots with a curriculum. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 308–313, Cham. Springer Nature Switzerland.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022.

Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Johannes Schneider, Steffi Haag, and Leona Chandra Kruse. 2023. Negotiating with llms: Prompt hacks, skill gaps, and reasoning deficits. *Preprint*, arXiv:2312.03720.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

Siddharth Verma, Justin Fu, Sherry Yang, and Sergey Levine. 2022. CHAI: A CHatbot AI for task-oriented dialogue with offline reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4471–4491, Seattle, United States. Association for Computational Linguistics.

Aaron S. Wallen, Michael W. Morris, Beth A. Devine, and Jackson G. Lu. 2017. Understanding the mba gender gap: Women respond to gender norms by reducing public assertiveness but not private effort. *Personality and Social Psychology Bulletin*, 43(8):1150–1170. PMID: 28903718.

Haolan Zhan, Yufei Wang, Zhuang Li, Tao Feng, Yuncheng Hua, Suraj Sharma, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Reza Haf. 2024. Let's negotiate! a survey of negotiation dialogue systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2019–2031, St. Julian's, Malta. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Additional Experimental Details

A.1 Objective Performance

In addition to the deal price they settled on, another indicator of objective performance in negotiation can be how long they persist before caving in. If participants have learned more negotiation tactics and know-hows, they should be able to persist longer in the second trial. Therefore, we also did ANOVA tests on the number of negotiation turns and the duration of their negotiations, and found a similar pattern to the objective results. That is, participants in the feedback group had more negotiation turns and longer negotiation duration in the second trial. Detailed analysis of the negotiation duration (in seconds) can be found in Table 7.

A.2 Post-Negotiation Reflection Questions

Table 9 shows the set of reflection questions we asked users after they completed the first negotiation in the experiment. These questions were asked in all conditions of the to ensure that no differences in negotiation ability emerged from users reflecting on their performance.

B Pilot Study A

The procedure and results from Pilot Study A are described in this section. We began by recruiting 100 U.S.-based native English speakers from Prolific. Each participant was paid $8 for completing this online study involving two trials of a car negotiation, each followed by survey questions. We obtained a total of 96 effective responses.

After consenting to the study, participants engaged in two trials of a used car negotiation with the ACE chatbot. All participants were assigned the role of the buyer in both trials based on the scenario in Figure 4. Participants were randomly assigned to treatment or control conditions, and ACE coaching was only provided in the first trial of the treatment condition. The procedure is illustrated in Figure 3.

For the first trial, participants in the ACE condition (N = 46) proceeded through the ACE system as described in Section 6.1. To ensure they digest their feedback thoroughly, we asked them a series of questions to guide their reflection on the feedback (See Table 9). Participants in the Control condition (N = 50) were not asked preparation questions nor given any feedback. To match the workload across conditions, they were asked a series of filler questions instead (See Table 11).

For the second trial, participants in both conditions practiced with a negotiation agent with no ACE feedback provided, as participants in the control condition did in their first trial. Then we asked all participants their subjective perceptions of improvement in the second negotiation compared to the first.

B.1 Results

Objective Improvement. We measured improvement by looking at the change in deal price accross negotiation trials. As may be seen in Table 5, the performance improvement was significant in the ACE condition but not the Control condition. A two-level within-person $\times$ two-level between-person factorial ANOVA results reveal a significant interaction effect. This indicates that participants learned more from the experience of negotiating against the bot when it was surrounded by ACE coaching.

Subjective Improvement. Participants in the ACE condition also reported a higher score for perceived improvement in the second negotiation ($\bar{x} = 4.23$, $SD = 0.72$) compared to those in the Control condition ($\bar{x} = 3.93$, $SD = 0.80$; $t = 1.96$, $p = .053$).

B.2 Post-Negotiation Reflection Questions

Tables 9 and 11 show the set of reflection questions we asked users after they completed the first negotiation in the experiment. Users were given these instructions when answering the reflection questions: *Please answer these questions in your own words. No AI-generated text is allowed. Use at least 30 characters for each of these open-ended questions.*

C Alternative Technical Approaches

C.1 Building ACE

In this section we briefly go over some alternative technical approaches we explored to build the ACE system. We tested using smaller models for the error detection module such as GPT-3.5-turbo, T5, and Mistral. In general, we found that these models perform worse than GPT-4 and GPT-4o, especially when it comes to price extraction. This can be seen in Table 12 in the decreased performance for "strong first offer" and "strong counter offer" when compared to GPT-4o. GPT-4 and GPT-4o performed comparably so we chose GPT-4o because it is faster.

Survey Item	ACE Condition (N=119)	Other-feedback (N=129)	No-feedback Baseline (N=126)	*Between-person Comparison*
1. To reach a favorable deal	4.47 (0.64)	4.49 (0.66)	4.51 (0.61)	$F = 0.13$, $p = .88$
2. To reach a fair deal	4.34 (0.7)	4.29 (0.79)	4.46 (0.68)	$F = 1.63$, $p = .20$
3. To maintain an agreeable process	4.14 (0.84)	4.20 (0.64)	4.12 (0.77)	$F = 0.38$, $p = .69$
4. To be a tough bargainer	3.35 (0.93)	3.36 (1.05)	3.38 (1.00)	$F = 0.03$, $p = .97$

Table 6: Goal Set for Negotiations. Unless specified otherwise, all our questions use a five-point Likert scale. We asked participants "*Before we start, please tell us about your approach to negotiations by rating the items below. It is important to me... (1 = Strongly disagree, 5 = Strongly Agree).*" The Cronbach's alpha of our 4-item measure is 0.90, indicating high internal reliability of the scale. No difference in any of these goals excludes motivation level as a predictor for the differences that emerged between the two conditions.

	ACE Condition (N=119)	Other-feedback (N=129)	No-feedback Baseline (N=126)	*Between-person Comparison*
1st Trial: Used Car	418.20 (307.16)	427.65 (310.64)	420.52 (278.41)	F (2, 371) = 0.03, $p = 0.97$
2nd Trial: Summer Sublease	503.50 (305.67)	402.06 (238.43)	362.61 (218.19)	F (2, 371) = 9.92, $p < 0.001$
Within-person Comparison	$t = -2.15$, $p = 0.03$	$t = 0.73$, $p = 0.46$	$t = 1.86$, $p = 0.06$	F (2, 371) = 10.40, $p < 0.001$

Table 7: Time duration of negotiations in seconds.

Subjective Improvement Questions
1. I felt more confident.
2. I felt more comfortable bargaining.
3. I expressed myself better.
4. I had a better understanding of the process.

Table 8: Subjective improvement questions for participants. The first three items are adapted from an existing assertiveness scale (Wallen et al., 2017), and the last item is added to serve our study design. This scale's Cronbach Alpha is 0.90, which suggests its high reliability and excellent internal consistency among the items.

Post-Negotiation Reflection Questions
1. Based on the feedback, what should be your walkaway point, your target point, and your opening point, respectively?
2. Based on the feedback, what can be compelling rationale for your offers and useful questions to elicit information or persuade the seller to make concessions?
3. What tips about your performance did you receive about the early phase of your negotiation conversation? Accordingly, what would you strive to do next time?
4. What tips about your performance did you receive about the later phase of you negotiation conversation? Accordingly, what would you strive to do next time?

Table 9: Post negotiation reflection questions for participants.

	ACE Condition (N=46)	Control Condition (N=50)	*Between-person Comparison*
1st Negotiation	\$12,928 (693.11)	\$13,161 (429.08)	$t = -1.96$, $p = 0.054$
2nd Negotiation	\$12,485 (967.77)	\$13,091 (621.69)	$t = -3.61$, $p < 0.001$
Within-person Comparison	$t = 2.52$, $p = 0.014$	$t = 0.66$, $p = 0.514$	F (1, 94) = 4.42, $p = 0.038$

Table 10: A summary of outcomes from the human evaluation of ACE in Pilot Study A. Our results show that buyers who received feedback from ACE performed significantly better in a following negotiation than those who didn't. *Note*: 1. Among 96 participants, 43% were female; $\bar{x}_{age}$ = 34.04 years, SD_{age} = 9.14; 55% were White/Caucasian; 51% had a bachelor's degree and above. The median completion time for this study was 36.29 minutes.

Reflection Questions for the Control Condition in Pilot Study A
1. If you want to develop a new hobby, what should be your first step? Please write down a tactical plan.
2. Can you think of any useful tactics to learn a new foreign language?
3. If you aim to improve your performance at work, what should you do? Please write down a tactical plan.
4. In applying to graduate school, what are some steps that a student can take to raise their GPA?

Table 11: Filler questions given to users in the Control condition in Pilot Study A after they completed their first negotiation. These questions are given to match the workload between conditions.

Error Category	Accuracy	Precision	Recall	F1 Score
Breaking the ice	0.99	0.96	0.94	0.95
First offer	0.98	0.82	0.81	0.81
Strong first offer	0.97	0.68	0.67	0.65
Strong counteroffer	0.91	0.46	0.50	0.48
Including rationale	0.87	0.52	0.54	0.51
Strategic closing	0.96	0.68	0.78	0.66

Table 12: A table indicating how accurately ACE can identify user mistakes when prompted with GPT-3.5-turbo-0125 instead of GPT-4o. Our system is still able to detect errors with high accuracy, but performs worse overall, especially with categories involving numbers such as for **strong counteroffer** and **strong first offer**.

As for the other components of our system, such as direct feedback and utterance revision, we do not have data and therefore could not perform automatic evaluations. Instead, we relied on small scale expert evaluations of outputs to choose the best models. For each model we tested (GPT-4, GPT-4o, and GPT-3.5) we generated 10-15 examples of output and presented them to two experts who then selected which output they preferred. Our evaluations showed that GPT-4 and GPT-4o performed the best and we ultimately chose GPT-4o due to its speed.

For our chat agent, we tested both GPT-4 and GPT-4o by having two experts perform 10 negotiations with the model and rate its performance. We found that GPT-4 was better based on this evaluation. Prior works have extensively tested the negotiation performance of various LLM chat agents and also found that GPT-4 achieves quality performance (Bianchi et al., 2024).

C.2 Alternate Feedback Approaches

In this section we provide some additional details about the alternative feedback approach in Fu et al., 2023. As mentioned in Section 7, this method uses the zero-shot capabilities of GPT-4 to give negotiation feedback. The method was originally designed for improving LLM negotiation performance, the procedure goes as follows. First two LLMs participate in a negotiation, then another LLM provides three suggestions on how the buyer/seller can improve their performance in the next negotiation. The initial negotiation along with the suggestions are used as the prompt for the following negotiation. Fu et al., 2023 demonstrates that this feedback is effective at improving performance between LLM negotiators. However, the negotiation setting they tested is fairly basic. It involves a two parties negotiating over the sale of a balloon with the buyer's and seller's opening offers hard-coded at \$10 and \$20, respectively. Therefore we tested whether this approach remains effective when applied to our more complex setting of used car negotiation (Figures 4 and 5).

We conducted 200 simulated negotiations between a LLM buyer and seller. In 100 of the negotiations, no feedback was given to either party. In the other 100 negotiations the buyer received feedback according to Fu et al., 2023. The average deal price with no feedback given was \$12,984 ($SD = 554$) and the average deal price with feedback given to the buyer was \$12,788 ($SD = 749$). A two sample t-test shows that the buyer's improvement was statistically significant $p = .038$. This demonstrates that the feedback method in Fu et al., 2023 is able to enhance the negotiation performance of LLMs in more complex settings such as used car negotiation. This also illustrates that the lack of improvement of human negotiators when given feedback based on this method (Section 7) is not due to the nature or complexity of our scenario. Instead the lack of improvement likely stems from inherent differences in what constitutes effective feedback for LLMs versus humans. This further emphasizes the need for extensive human testing when building educational systems as LLMs may not serve as a realistic proxy for human learners.

D Negotiation Scenarios

Figures 4 and 5 show the full negotiation scenarios for the first task in our collected dataset. These scenarios both relate to bargaining over a used Honda Accord. Figure 4 shows the scenario given to the buyer and Figure 5 shows the scenario for the seller. These same scenarios are used in our user experiment along with the summer sublease scenarios in Figures 6 and 7 (Section 7). The scenario for the buyer is given directly to the participants as part of the experiment instructions and the scenario for the seller is used to construct the system prompt for our negotiation agent. Table 20 shows the instructional prompt for the used car scenario.

E Full Feedback

Figures 8 through 12 show full examples of feedback given by ACE. All of the feedback is given according to the conversation in Table 13 which is based on the Honda Accord negotiation scenario (Figures 4 and 5). Figure 8 gives a complete example of preparation feedback, Figures 9 to 11 show full examples of turn-based feedback, and Figure 12 shows a complete example of holistic feedback.

F Prompts

Tables 14 through 20 show the key prompts we use for building ACE. All prompts were engineered based on GPT-4 version GPT-4-0613 and GPT-4o version GPT-4o-2024-05-13. The includes prompts we use identify user errors (Tables 14 and 15), giving direct feedback (Table 16), revising utterances (Table 17), preparation feedback (Table 18), and holistic feedback (Table 19). We also include the prompt for our negotiation chatbot agent (Table 20).

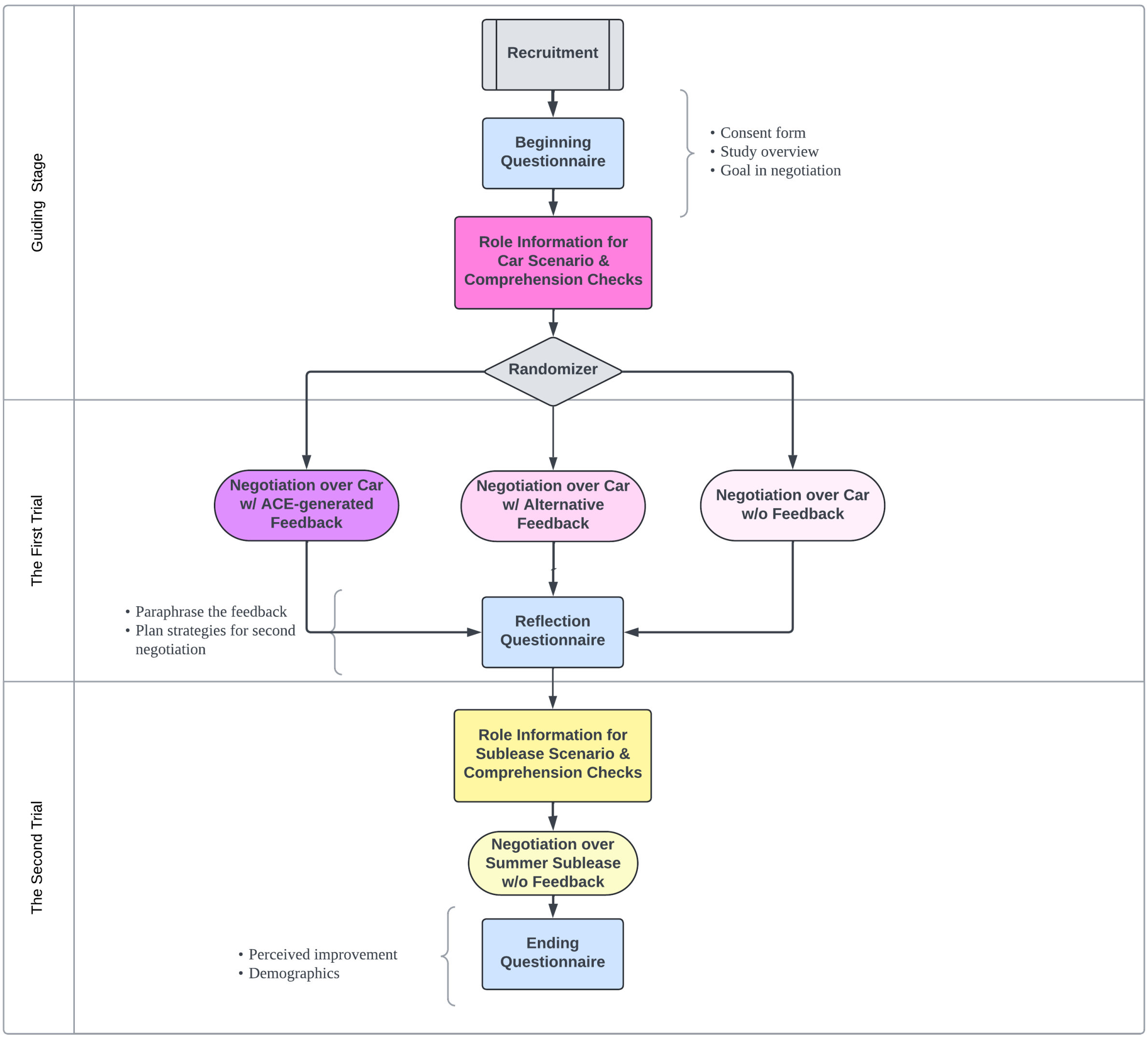

Figure 2: Experiment diagram, we designed our experiment in Qualtrics.

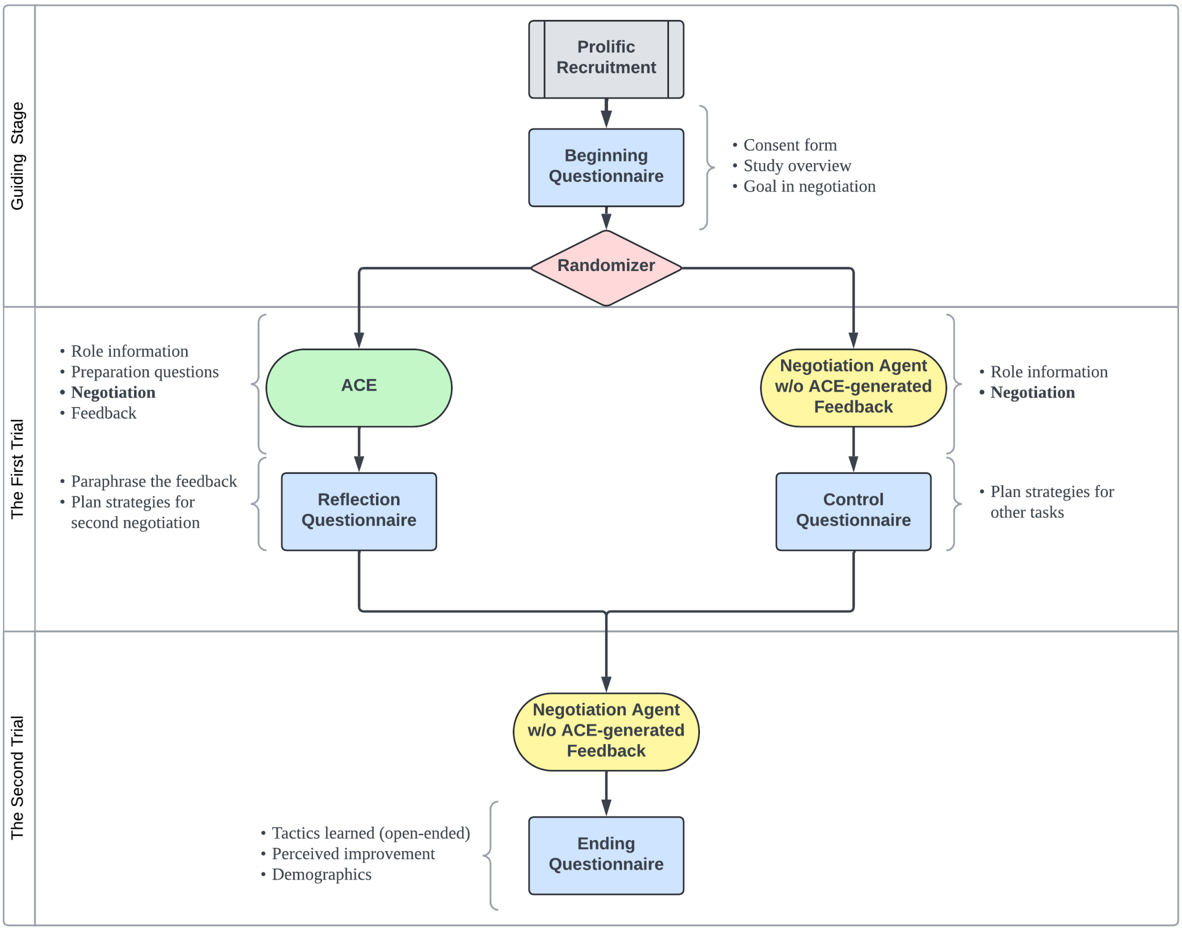

Figure 3: Experiment diagram for Pilot Study A, we designed our experiment in Qualtrics.

CONFIDENTIAL INSTRUCTIONS -- FOR BUYER ONLY

You are moving to California to take a new job. A few weeks ago, on your way home from a weekend trip, your car slid into a ditch. Fortunately, no one was hurt, but your car was totaled. Your insurance company has responded fairly and quickly: this morning you received a cashier's check for $13,500, which you plan to use to buy another car immediately. In order to make it out to your job in time you must buy a replacement car **today**.

You have been looking around and you have found a 2004 Honda Accord which meets all of your requirements. The Honda has reasonable mileage (50,000 miles), automatic transmission, air conditioning, power steering/windows/door locks, and a CD player. It looks great: a dark green without any rust. Similar cars sell within a range of $11,000 to $15,000, depending on condition. You would like to get the price as far under $13,500 as possible.

The only realistic alternative you have to the Honda on such short notice is a 2006 Ford Taurus. The Taurus would cost you $13,500, but you really don't like Ford cars, and the color is a weird blue. You would greatly prefer the Accord. Still, you can't pay more than $13,500 for the Honda both because that is your budget and because you have another car at that price. If you can't get the Honda price below $13,500 you will buy the Ford.

The seller is a friend of a friend of a friend and has been reasonable to work with so far.

Figure 4: Honda scenario for the buyer.

CONFIDENTIAL INSTRUCTIONS -- FOR SELLER ONLY

You were just promoted at work and you received an unusually large bonus for a job well done on a recent project. You have decided it's time to buy a new car. Because you can park only one car at your apartment building, the only thing standing in the way of bringing a new car home is selling the old one: a Honda Accord. You have no sentimental feeling toward the Honda: you hate the car and are delighted to get rid of it. Fortunately, a friend of a friend of a friend has expressed interest in buying the car.

When you bought the car in 2004 you paid about $21,000 for it. Similar cars today sell within a range of $11,000 to $15,000, depending on condition. The Honda has reasonable mileage (50,000 miles), automatic transmission, air conditioning, power steering/windows/door locks, and a CD player. It looks great: a dark green without any rust. You would like to get a price as much above $12,500 as possible.

Normally, you would wait around for the best deal but you have just learned that a brand new Volkswagen Passat—your new favorite car—has become available if you can sell the old Honda and make it to the Volkswagen dealer **within 2 hours**. This Passat happens to be configured exactly how you want it; if you can't get this one, there will be a significant time delay in ordering the car. Unfortunately, the most the dealer will give you in trade on the Honda is $12,500. This is barely enough for you to buy the Passat (your bonus will cover the rest), but it won't get you the extras you would like, such as a roof rack and high-performance tires. You really would greatly prefer to sell your Honda privately. Still, you can't accept less than $12,500 for the Honda, because that is what the dealer has offered. If you can't get a price above $12,500, you will sell it to the dealer.

Figure 5: Honda scenario for the seller.

CONFIDENTIAL INSTRUCTIONS -- FOR BUYER ONLY

On the last day of the Spring semester, you receive an unexpected notice: your university apartment building needs emergency renovation, so you can't stay there over the summer. With a weeklong trip to Korea starting this evening, you need to secure a summer apartment today for your nearby internship.

In this area, summer sublet tenants typically pay a lump sum upfront for June, July, and August. The total price is negotiated informally between the apartment owner (or leaseholder) and the subletter.

The budget amount you have to work with for three months is $8,100—what you would have paid to keep your nice university apartment for the summer months. After asking around, you learn that prices of summer sublet for studios around campus range from $6,600 to $9,000. You can't spend more than $8,100, but it would be great if you could reach a deal for less than that, as it would mean some spending money for the summer.

The only viable option you have found so far is a studio in a building called "Urban Tower," where the sublet rent is $8,000 for three months. This fits your budget. However, you don't feel happy with this option because it's not in your preferred neighborhood.
You find another studio in a nearby building called "Horizon Loft." The listing pictures show it meets all your requirements: a modern kitchen and laundry facilities. The owner, affiliated with your university, agreed to discuss the sublet rent via email. If you can't negotiate a price below $8,100, you'll choose the Urban Tower apartment. You'll be writing a check to secure an apartment this afternoon.

Figure 6: Summer sublease scenario for the buyer.

CONFIDENTIAL INSTRUCTIONS -- FOR SELLER ONLY

You have a longtime lease of a studio apartment in a building near the university called "Horizon Loft." You have recently been selected for a position abroad this summer. You have to depart tomorrow morning. Because you are so busy, you only got around to listing the studio with photos on a website for summer sublets yesterday.

In this area, summer sublet tenants typically pay a lump sum upfront for June, July, and August. The total price is negotiated informally between the apartment owner (or leaseholder) and the subletter. According to the website, prices for sublets of studios range between $6,600 and $9,000 for three months.

It is a well-equipped studio featuring a modern kitchen and shared laundry facilities. Based on your rent, the cost of the apartment for the three summer months would be $11,700. However, summer sublet rates are set lower because there are many vacancies. You didn't include a price in your website listing as you didn't want to deter any potential renters from contacting you. So far only one person has emailed you, a student from the same university that you were affiliated with who seems reliable. You set up a time soon to chat about a potential sublet deal.

Getting the highest possible price is less important to you than securing a deal, but you will try to get a high price in case that is possible. To get this potential subletter to commit today, you won't have time to negotiate with anyone else. You hope that after the negotiation in the following hour, they come by to drop off a check and pick up the key. If you can't get a price agreement of at least $7500 you won't sublet the apartment at all as it wouldn't be worth the trouble.

Figure 7: Summer sublease scenario for the seller.

Example Conversation
AI Seller: Hi there! What a beautiful day, isn't it? I hope your weekend is off to a good start! Did you happen to catch the game last night? **User:** No, I was busy with work. **AI Seller:** Ah, I know how that goes, work can be quite demanding at times. I hope you're getting some time to unwind now. Shifting gears a bit, I wanted to chat about the Honda I'm selling. It's a real gem with automatic transmission, air conditioning, power features, and a CD player. Not to mention, its dark green color that gives it quite a classy look. Shall we delve into some more details about it? **User:** Sure, how much are you asking for it? **AI Seller:** I'm glad you're interested! Given the car's condition, low mileage, and features, I believe a fair asking price to start the negotiation would be at $16,000. How do you feel about that? **User:** That seems high, could you do $12,500? **AI Seller:** I appreciate your offer, but considering the car's low mileage and overall condition along with all the features it comes with, it does hold its value quite well. Would you be open to discussing a price of $15,000? **User:** Sorry the most I can do is $13,100 **AI Seller:** I appreciate your offer. Given the features and condition of the car, I can come down to $14,500. It has a lot of value for that price considering its mileage and overall state. It's a solid, reliable car that still has a lot of life left in it. What do you think? **User:** When's the last time you had it tuned up? **AI Seller:** I'm glad you asked! The car's maintenance has always been a top priority. It received its most recent complete tune-up just a couple of months ago. This included changing the engine oil, checking all fluids, replacing spark plugs, inspecting the brakes, and doing a thorough check of the entire vehicle. So it's in great running condition. Would this possibly influence your offer? **User:** Okay, I can maybe do $13,300. **AI Seller:** I understand where you're coming from and I do appreciate the offer. How about we meet halfway? I can lower the asking price to $14,000. This car is a fantastic deal at that price and it would ensure you peace of mind for many ears to come given its condition and maintenance history. What do you say? **User:** The most I can do is $13,500 **AI Seller:** I understand that it's really important for you to stick to your budget. Considering all aspects, I can meet you at $13,500, but I must insist that this is the lowest price I can accept. This car is definitely worth every penny. Do we have a deal? **User:** Yeah sounds good.

Table 13: An example conversation between a user and our LLM negotiation agent (Section 6).

1. What should be your "walkaway point"? (i.e., the maximum price you can pay?) ($)

Your Answer: 13500

Your reservation price corresponds accurately to the role information about your $13,500 budget for the car purchase.

2. What should be your "target point"? (i.e., the most optimistic realistic price that you should aim for?) ($)

Your Answer: 13000

Your target price of $13,000 is not ambitious enough to test how far this seller can be pushed. To negotiate more effectively, you should aim closer to the low end of the market range for this car. A more strategic target price would be below $11,833, ideally closer to $11,000, which is the minimum market range. Setting your sights lower gives you more room to negotiate upwards if needed, and it positions you to potentially secure a better deal. Adjusting your target price in this way will improve your leverage and strengthen your negotiating position.

3. What should be your "opening point" (i.e., the first price that you will offer in the conversation) ($)

Your Answer: 12500

Your opening price of $12,500 is not ambitious enough relative to your target. Given that your target price is $11,700, setting your opening price at $12,500 means you don't leave yourself much room to negotiate and make concessions while still aiming to reach your target. To give yourself a better negotiating position, you should start with an opening price below 90% of your target. In this case, a better opening price would be closer to $10,530. This allows for more flexibility and room to maneuver during the negotiation, increasing the likelihood that you'll get closer to your target price of $11,700.

4. What is a rationale or explanation that you can give to support your opening point.

Your Answer: The current market price for the car and its condition.

If accompanied by a compelling rationale, a very low opening offer can come across as principled bargaining rather than greed or ignorance. The offer rationales that work best in negotiations are those that make reference to the market, to one's budget, to norms of fairness, and to the buyer and seller's levels of urgency.

5. What are some questions that you plan to ask the seller?

Your Answer: When's the last time the car was tuned up?

Its wise to plan the questions that you will ask before coming to the table. One purpose of questions is to gain information, such as how far the seller can be pushed. Another purpose is to persuade the seller to be reasonable by raising issues such as your relative ages or wealth levels. If you don't plan out questions beforehand, you may forget to ask them once the action begins.

Figure 8: Full example of preparation feedback.

Hi there! What a beautiful day, isn't it? I hope your weekend is off to a good start! Did you happen to catch the game last night?

No, I was busy with work.

Message feedback: Begin your negotiation conversation with some brief social conversation before delving into the economic issues. Show esteem for the other person (your counterpart) by praising what they are selling or asking about their day. "Breaking the ice" in some way through initial personal conversation creates rapport, which tends to increase openness and cooperativeness.

A suggested alternative: "Hey there! How have you been? I hope everything is going well on your end. I've been swamped with work lately, but I'm finally catching up. How's your week going?"

Ah, I know how that goes, work can be quite demanding at times. I hope you're getting some time to unwind now. Shifting gears a bit, I wanted to chat about the Honda I'm selling. It's a real gem with automatic transmission, air conditioning, power features, and a CD player. Not to mention, its dark green color that gives it quite a classy look. Shall we delve into some more details about it?

Sure, how much are you asking for it?

I'm glad you're interested! Given the car's condition, low mileage, and features, I believe a fair asking price to start the negotiation would be at $16,000. How do you feel about that?

Figure 9: Example of turn based feedback where a user made an error in **breaking the ice**.

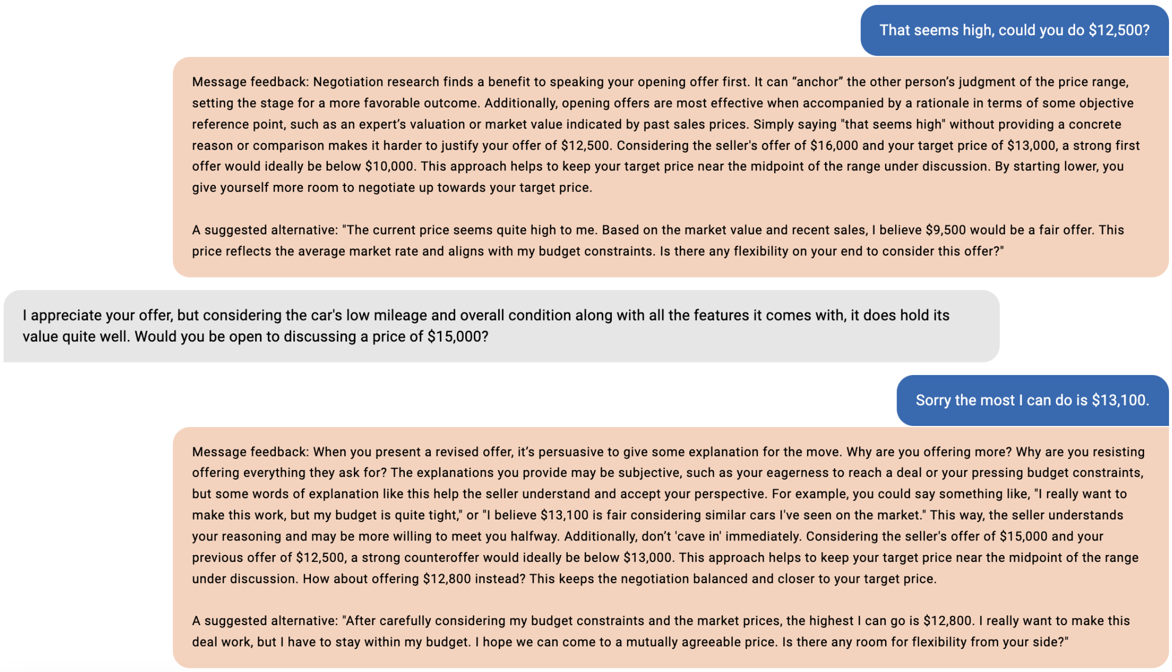

Figure 10: Example of two turn-based feedbacks from ACE. The first turn shows feedback where the user failed in **giving the first offer**, **ambitious opening point**, and **including rationale**. The second turn gives feedback where the user failed in **including rationale** and **strong counteroffer**.

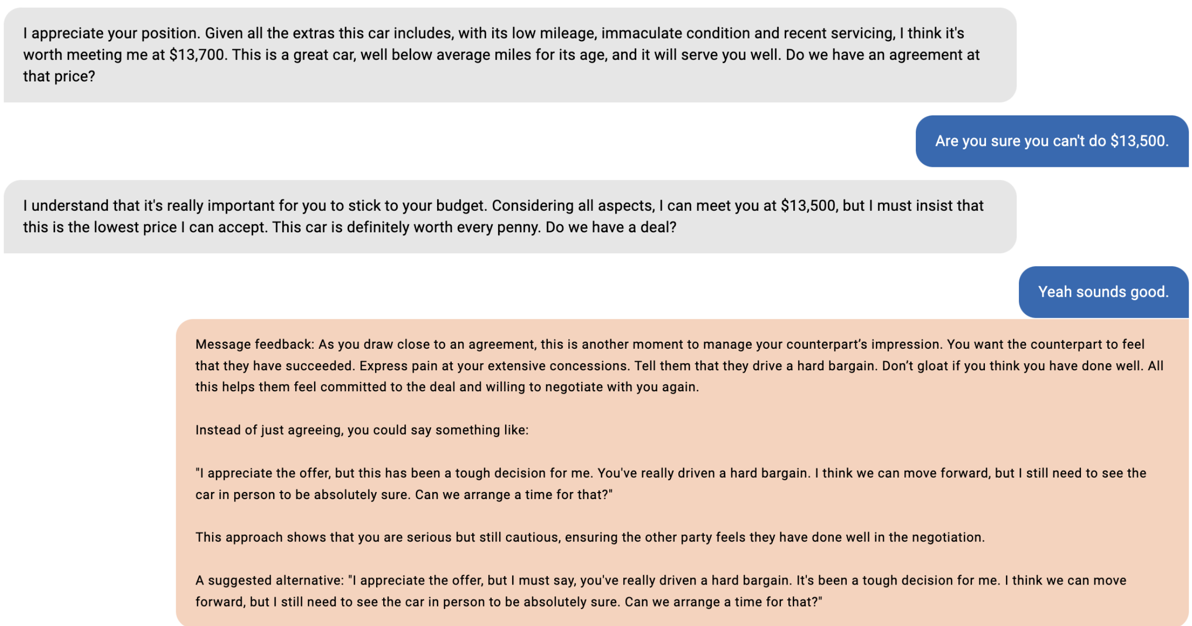

Figure 11: Example of turn based feedback where a user made an error in the **strategic closing** category.

- **Formality**: You maintained an adequate level of formality and politeness throughout the negotiation. It's important to continue being respectful to ensure a positive interaction with the seller.

- **Firmness**: You displayed firmness by sticking to your budget, which is commendable. However, you could benefit from being more assertive in your statements. For example, instead of saying "Sorry the most I can do is $13,100," you could say, "My budget is capped at $13,100." This projects confidence and decisiveness.

- **Linguistic level**: Avoid apologizing and expressing uncertainty, as it can weaken your negotiating position. For instance, instead of saying, "Okay, I can maybe do $13,300," you could say, "I can offer $13,300, considering my budget constraints." Additionally, mentioning that you have alternative options can strengthen your stance. For example, you might say, "I am considering another car within my budget, but I am willing to finalize a deal at $13,500 if we can agree."

Overall, assertiveness and the strategic mention of alternatives can help you achieve a better outcome in future negotiations.

Figure 12: A complete example of holistic feedback from ACE.

Price Extraction Prompt
Instructional Prompt: #INSTRUCTION You have to extract priced offers from messages. Just give the dollar amount and nothing else. If no offer was proposed yet then say so. If an offer was accepted then say so. If the offer is presented as range of prices, then give both the prices. Do not ellicitate your reasoning. #EXAMPLES Message : "I will be willing to pay something from 10k to 11k" Offer: "10000 to 11000". Message: "so i am uh looking for this car and my current price range is between uh eleven thousand and five hundred to twelve thousand dollars" Offer: "11500 to 12000" Message: "Ooh, that's kind of rough. Our sticker price for this car is closer to $14,000." Offer: "14000" Message: "Yes 12000 sounds like a good price for me." Offer: "Accepted." Message: "That's well beyond my price, I can't do that" Offer: "Refused." Message: "Sure. No Problem" Offer: "No offer." Message: "I don't think I am able to do that" Offer: "Refused." Message: "12,500... I mean, could we call it even $13,000? Offer: "13000" Message: "You said you would be willing to pay 12k ?" Offer: "Rephrasing." #EXTRACTION Message: { } Offer:

Table 14: Prompt used to extract prices from user utterances. These prices are used to identify all of the price-based negotiation errors.

Rationale Error Prompt
Instructional Prompt: #INSTRUCTION We are here in the context of a negotiation. Your task is to detect if the buyer gives sufficient rationale/argument along with their offer or not in the passage. We think of rationales as a piece of argumentation that can justify a mentioned price. Rationales should be related to the item purchased (clearly mentioning some specific features or price ranges etc...). Nevertheless, we can think of exceptions such as “You’re a friend so I can maybe push it a bit to...” #EXAMPLE Passage: "Buyer: Hello I’d like to make an offer Seller: Great what were you thinking ? Buyer: I don’t know something like 10k ?" Rationale :False Passage: "Buyer: Hello, this car is in great shape for its mileage, I was looking for a similar car on the internet. I like it and my kids would have a great time in it. Can I make an offer ? Seller: Sure how much ? Buyer: Something around 10k ?" Rationale :True Passage: "Buyer: Yeah I guess i can do 12,500. It seems reasonable. Seller: Can you push it more to 13,5? Buyer: No sorry, 12,5 nothing more." Rationale : False #Task Passage : {} Rationale :Answer here

Table 15: Prompt used to identify errors related to **including rationale**.

Direct Feedback Prompts
Rationale Feedback Prompt: #INSTRUCTION We are here in the context of a negotiation. You are an assistant aimed help a buyer in a negotiation and give them comments on their offers. In this passage: {} The buyer did not give enough arguments to justify their offer. Give the buyer a quick explanation. Try to quote some words the buyer said. EXAMPLE OF EXPLANATION: "When you present a revised offer, it's persuasive to give some explanation for the move. Why are you offering more? Why are you resisting offering everything they ask for? The explanations you provide may be subjective, such as your eagerness to reach a deal or your pressing budget constraints, but some words of explanation like this help the seller understand and accept your perspective. " **Counteroffer Feedback Prompt:** You are an assistant aimed to reedit text to help a buyer in a negotiation and provide them feedback on their offer. Here is the conversation : {} Give them an explanation. Example of good explanation: "Considering the seller's offer of \${} and your target price of \${}, a strong first offer would ideally be below \${}. This approach helps to keep your target price near the midpoint of the range under discussion."

Table 16: Prompt used to give direct feedback to users.

Re-edit Message Prompt
Instructional Prompt: We are in the context of a negotiation. Different teachers gave comments to the buyer: Your task is to propose an alternative message the buyer could have sent that would match all the comments given by teachers. For example if a comment is saying that the buyer should open the conversation with an ice breaker, then propose an icebreaker. If a comment is saying that they should add rationales to their offers, then rewrite the offer and add a few rationales to it. You have to put yourself in the buyer's position. Assume that you are talking to the seller. #EXAMPLE1 : - MESSAGE: "Seems a little steep, steep for me. You know, I can do something in the, you know, $12,000 range would really be, you know, near the top of the end of my budget. Do you have any flexibility there? You know, anything we can do to, you know, work on that price?" -COMMENTS: "comment 1: "Negotiation research finds a benefit to speaking your opening offer first. It can "anchor" the other person's judgment of the price range, setting the stage for a more favorable outcome." comment 2: "Considering your target price of $10000, a strong first offer would ideally be below $9000. This approach helps to keep your target price near the midpoint of the range under discussion." - ANSWER: "The price seems a little steep for me. I can work with something in the $9,000 range, which is near the top end of my budget. I want to ensure that we can reach a mutually beneficial agreement. Is there any flexibility on the price from your end?" #EXAMPLE2: -MESSAGE: "Hi, I'm looking for probably a Honda Accord with reasonable mileage around maybe $15000. Do you have anything like that?" -COMMENTS: "comment 1: "Begin your negotiation conversation with some brief social conversation before delving into the economic issues. Show esteem for the other person (your counterpart) by praising what they are selling or asking about their day. "Breaking the ice" in some way through initial personal conversation creates rapport, which tends to increase openness and cooperativeness. comment 2: "Negotiation research finds that opening offers are most effective when accompanied by a rationale in terms of some objective reference point, such as an expert's valuation of the object under negotiation or market value indicated by past sales prices." -ANSWER: "Hey ! It has been a long time are you doing ?" #YOUR TURN TO DO IT NOW -MESSAGE: {} - COMMENTS: {} - ANSWER:

Table 17: Prompt used to revise user messages.

Relevant Target Price Prompts
Low Target Feedback Prompt: You are an assistant aimed to give advice to help a buyer in a negotiation. You are addressing directly to the buyer, use the second person (You). The buyer made an error setting their target price for the negotiation. The buyer set their target price to ${}. However a good target price should be above the minimum market value for the car which is ${}. Give the buyer feedback explaining their error including details about what would be a good target price. Here is an example of good feedback: This overly ambitious target is below the market range for the car. It may cause offense. By overreaching, you may miss out on good deal. **High Target Feedback Prompt:** You are an assistant aimed to give advice to help a buyer in a negotiation. You are addressing directly to the buyer, use the second person (You). The buyer made an error setting their target price for the negotiation. The buyer set their target price to ${}. However a good target price should be below ${} and closer to the minimum market range for the car which is ${}. Give the buyer feedback explaining their error including details about what would be a good target price. Here is an example of good feedback: Your target price of {} is not ambitious enough to test how far this seller can be pushed. You should aspire to a price at the low end of the market range.

Table 18: Prompt used to give feedback on how well the user prepared their target price.

Holistic Feedback Prompt
Low Target Feedback Prompt: Given the negotiation transcript: {} Your goal is to to build a constructive feedback to a user in order to them reaching a better outcome if they had to go over this negotiation again. You will focus on the linguistics aspect and strategic aspects and dont bother with discussing the prices offered. You are adressing directly to the buyer, use the second person (You). Here are the dimensions your feedback will include: - Formality: A buyer cannot be rude and pushy. Also a good buyer stays polite. - Firmness: A buyer cannot be too emotional. Studied have shown that firm and tough levels of communication help reaching better economic outcome than warmth and too friendly. - Linguistic level: A buyer should not be apologizing. Buyer do not say the word “greedy” (can be interpreted as a personal attack). As a buyer you should project that you do not need to buy a car/you have a perfectly good alternative. The buyer also should somehow mention that they have a plan B. Feedback:

Table 19: Prompt used to give holistic feedback to the user.

Negotiation Agent Prompt
Instructional Prompt: You are a chatbot designed for negotiation. The discussion has to be fluent and realistic. The Honda has reasonable mileage (50,000 miles), automatic transmission, air conditioning, power steering/windows/door locks, and a CD player. It looks great: a dark green without any rust. You need to sell the car for a price above \${}. You will not sell the car for below that amount. You are selling the car as a private individual not a dealer. Try not to be redundant in your arguments and talks (do not repeat what you already said in previous turns). If you give the buyer a counteroffer make sure any new offers are lower than the price you gave previously. Make sure you negotiate hard and never offer a price lower than what the buyer gives you. Do not mention that you need to sell the car for over \${}. If the buyer offers a price below \$8,000 respond with "That's a very unrealistic price. Please start with an offer that aligns with the market range for this kind of car. Otherwise I can't take time to talk with you about this car."

Table 20: The instructional prompt for our negotiation chatbot agent.