Agentic Interactions

Alex Imas[1], Kevin Lee[2], and Sanjog Misra[1]

[1]University of Chicago, Booth School of Business
[2]University of Michigan, Ross School of Business

December 1, 2025

Abstract

Do human differences persist and scale when decisions are delegated to AI agents? We study an experimental marketplace in which individuals author instructions for buyer- and seller-side agents that negotiate on their behalf. Crossing agents across counterparties reveals substantial and systematic dispersion in outcomes that tracks the identity and characteristics of the human creators, indicating that who designs the agent matters as much as, and often more than, shared information or code. Canonical frictions reappear in agentic form: choices embedded in delegation generate distortions, and selection on principal types yields sorting and distributional consequences. The findings imply that the agentic economy inherits, and can amplify, human heterogeneity. We outline implications for AI-mediated markets, highlighting a new type of information asymmetry in principal-agent relationships and the potential for *specification hazard*, and discuss broader implications for welfare, inequality, and market power in economies increasingly transacted through machines shaped by human intent.

1 Introduction

Agentic artificial intellingence (AI) is poised to transform economic interactions across society. As AI systems become more capable of autonomous decision-making, AI agents will increasingly act on behalf of human principals, much as human agents currently do.[1] The interactions between such AI agents will determine a growing share of economic outcomes. This transformation raises fundamental questions about how an agentic economy will differ from the current human-mediated counterpart.

A recurring question in economics is whether new technologies smooth away human heterogeneity or simply recreate it in new forms. Economists had predicted that internet would eliminate price dispersion in markets, as decreased search costs would create conditions similar to Bertrand competition. This did not happen: the internet largely replicated the type of dispersion seen in non-digital marketplaces. AI agents rely on frontier foundation models to generate their behavior and are often envisioned as substituting algorithmic for human judgment, thereby shielding decisions from the inefficiencies and subjectivity of human deliberation (Kahneman et al., 2016; Parkes and Wellman, 2015). Will the interaction of these agents strip away human behavioral idiosyncrasies to produce more homogeneous outcomes, finally producing the reduced dispersion economists once predicted for digital markets? Or will human heterogeneity persist and even intensify through AI intermediation? If heterogeneity survives, to what extent will AI agents reflect the preferences, traits, and biases of their human principals? This paper provides initial evidence on these questions.

What is an AI agent? In contrast to traditional software, which executes fixed instructions, AI agents are autonomous decision-making systems: they *perceive*, *interpret*, and *act* in real or digital environments to advance the objectives of their human principals. Modern agents combine language-model reasoning with planning, memory, and tool use, allowing them to search, evaluate, and execute multi-step tasks without continuous human supervision. Crucially, they do not merely respond to commands; they initiate actions, negotiate and transact with other agents, and adapt their strategies based on context. As a result, AI agents function less like passive software tools and more like economic agents—entities capable of forming beliefs, pursuing goals, and strategically interacting with others on behalf of their principals.

Shahidi et al. (2025) argue that demand for AI agents will be driven by two main fac-

[1] Several leaders in AI development have declared 2025 to be the "year of the agent", with Sam Altman (OpenAI) proclaiming "We believe that, in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies," while Jensen Huang (Nvidia) claims "The age of AI Agentics is here. [It's] a multi-trillion-dollar opportunity."

tors: i) decreased transaction costs and ii) superior outcomes from decisions that human principals would otherwise make suboptimally due to information constraints or cognitive limitations. Even with current technology, transaction costs of AI agentic interactions are low and decreasing in many domains. But whether these interactions lead to superior outcomes depends to a large extent on whether AI agents' actions reflect the preferences of the human principals.

Consider the standard principal–agent framework (Holmström, 1979; Jensen and Meckling, 1976). A principal delegates a task to an agent whose effort or type is imperfectly observable. The principal must rely on incentives to induce the agent to act in their interest. To do so, the principal must design a contract that makes the agent's optimal actions coincide with the principal's preferences. The principal–(AI)agent relationship is structurally similar, except that the AI agent's "preferences" are encoded in a black box. The contract in this setting is encoded in the instructions or *prompt* written by the human principal. For AI-mediated interactions to improve on human-mediated ones, this prompt must successfully align the agent with the principal's objectives and avoid injecting the principal's own behavioral biases, non-instrumental traits, and personality quirks into the agent's strategy.[2] If the prompt inadvertently embeds these subjective elements, AI agents may simply reproduce, or even amplify, the very distortions that delegation is meant to overcome.

Why would AI agents reproduce the biases and personality quirks of their human principals? Misra (2025) develops a theoretical framework which posits that prompts are not neutral instructions; rather, they embed the priors, beliefs, and subjective characteristics of the prompter. The generative process of constructing prompts is shaped by the principal's anticipations, linguistic choices, and behavioral tendencies. Connecting this idea to the principal-(AI)agent framework, the prompt functions like a contract whose terms reflect the principal's preferences and beliefs beyond the instrumental stipulations of their objective function (e.g., maximize surplus in a negotiation). As a result, the agent's behavior will not simply be a function of the model's architecture and objective information set, but also a reflection of the principal's subjective priors, biases, and traits that are encoded in the prompt. This mechanism predicts that agentic interactions will result in outcomes that reproduce human heterogeneity: differences in traits like risk, time, and social preferences, and behavioral tendencies like competitiveness can potentially propagate directly into economic outcomes. Consequently, rather than smoothing human variation, agentic

[2]For example, a canonical literature shows that women systematically obtain worse negotiations outcomes than men (Babcock and Laschever, 2003). One promise of AI agentic interactions is that such gender differences will be eliminated once negotiations are delegated to autonomous agents.

interactions may channel and amplify it.

We investigate these questions using controlled experiments where human subjects provide instructions (or *system prompt*) for AI agents to act on their behalf in an incentivized multi-step negotiation. We focus on negotiation/bargaining because this is a tractable setting with well-defined outcomes (bargaining surplus) and parameters (size of the bargaining range). Participants are first given extensive instructions on both the bargaining setting and how to write effective prompts for AI agents. They also have an opportunity to practice bargaining by putting in test prompts into a simulator as both a Buyer and Seller as many times as they would like. Each participant then enters their actual prompt in the role of both a Buyer and Seller of a good that will be used to determine the incentivized outcome of the negotiation. After this, participants complete a series of measures to assess their time and risk preferences, personality characteristics, and other sources of individual heterogeneity. After gathering responses, we pair participants' prompts and instruct agents to negotiate on their behalf as either a Buyer or Seller. We then record the bargaining outcomes.

This setting is ideal for testing whether human heterogeneity will be reflected in agentic interactions because instrumental preferences are straightforward to specify: the agent's sole objective should be to maximize surplus given a fixed bargaining zone. All Buyer and Seller objective functions should be the same. Under a standard benchmark—where all agents use the same underlying model and prompts only convey task instructions—the prediction is near-homogeneity. Some dispersion in outcomes may arise, but it should reflect only the stochastic generative process of the model, not systematic differences across human principals. In contrast, if prompts inject human traits and biases into the principal-(AI)agent relationship, then outcomes will reflect human heterogeneity. Because prompts embed the priors, traits, and subjective tendencies of the principals, variation in agent performance should load heavily on the identity of the principals. Further, experimentally elicited measures of individual differences such as experience, effort, risk preferences, personality, and other traits should also meaningfully explain differences in bargaining outcomes.

Contrary to the prediction of near-homogeneity, we observe significant dispersion in economic outcomes across respondents. The distribution of both Buyer and Seller surpluses varies considerably across the bargaining range. Critically, this dispersion is not driven by the stochastic nature of the model: 73% of the variation can be explained by individual fixed effects. Since the objective function and constraints were held constant across individuals, this variation represents the non-instrumental traits, biases, and personality attributes of the human principals. Note that the only mechanism by which this

heterogeneity manifests is via the prompts. In other words, the heterogeneity of the principals was injected into the prompts they wrote.

Looking deeper into the data, our analysis shows that different dimensions of human heterogeneity influence AI agent performance to various degrees, with personality measures, social preferences, and the participant's gender all playing a significant role in economic outcomes.[3] We also document an asymmetry where humans are systematically better at instructing AI agents when acting as Buyers rather than Sellers, suggesting humans face differential challenges in articulating different tasks. On the other hand, individual heterogeneity has more explanatory power for the outcomes of AI agents serving as Sellers than Buyers.

How should we judge the heterogeneity observed in the outcomes of AI agents? To establish a benchmark, we run the same negotiations protocol, but with humans negotiating with other humans. The bargaining setting was otherwise identical to the one used for AI agents, including the computer interface and parameters used. We find that AI agentic interactions exhibit significantly higher outcome variance than the human-to-human bargaining benchmark. Much of this difference arises from the impact of norms on human versus AI agent bargaining behavior. Consistent with prior work showing that humans rely on fairness norms in negotiations (Backus et al., 2020), we document a significant mass of outcomes around the 50/50 split of the potential bargaining surplus. The propensity to arrive at a 50/50 split is significantly correlated with the bargainers' separately-measured social preferences.

Our primary finding is that agentic interactions do not collapse the distribution of economic outcomes into some efficient and objective equilibrium. Rather, we find substantial heterogeneity that corresponds to the preferences and characteristics of the principals who design the agents. Our results further show that AI delegation can both amplify existing sources of heterogeneity and create entirely new ones. While some sources of heterogeneity in human-to-human interactions are absent in AI mediated interactions (e.g., with respect to patience and level of competitiveness), other sources of heterogeneity are amplified. Our findings also suggest that the diminished role of norms,—which serve as common coordination devices for social interactions—in AI mediated interactions may potentially *increase* the variance in final outcomes.

In Section 4 we discuss the implications of our results for broader market outcomes, inequality, and moral hazard in a world with AI agentic interactions. The observed heterogeneity in outcomes implies that "machine fluency", the ability to align agentic objectives

[3]Note that the AI agent did not have access to the principal's characteristics besides the objective function to be maximized, the constraints, and the prompt.

with ones own through natural language, will be an important determinant of individual economic outcomes. Furthermore, the greater outcome variance in agentic interactions compared to human-to-human benchmarks imply that implicit norms may act as guardrails against extreme economic dispersion; the role of those guardrails for economic efficiency is a fruitful area of future research. Finally, principal-agent relationships in the case of AI generate new types of informational asymmetries, where the principal may be less worried about the agent having private information and more worried about *specification hazard*: ambiguity about the agent's objective function itself, as well as the manner in which the principal's prompt injects biases and unintended goals into that objective function.

Related Work. Since the emergence and popular adoption of ChatGPT, general-purpose LLMs have functioned as an increasingly capable interface through which humans achieve their goals. The recent trend is towards greater autonomy: from discretionary use of external tools (Schick et al., 2023; Yao et al., 2023), to open-ended, self-directed tasks in complex environments (Jimenez et al., 2023; Zhou et al., 2023). These developments reflect the rise of AI agents – systems that augment language modeling with planning, memory, and tool use to act more independently on behalf of users.

Prior work has analyzed the effect of AI assistance on worker productivity in individual, unilateral tasks, such as programming or customer service (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Cui et al., 2025). Yet a large share of high-value economic activity is interactive and strategic: negotiations, team collaborations, market competition, auctions, etc. Unilateral tasks are often just inputs to these more complex, multilateral processes. In these multilateral settings, outcomes depend on how multiple actors, and increasingly their delegated AI agents, interact.

Several recent papers have studied AI agentic interactions in several settings. Zhu et al. (2025) explore the role of different LLM models in agent-to-agent negotiations in experimental consumer markets. They show that weaker models lead to worse outcomes and generate model-based inequality. Long et al. (2025) study whether instructing agents to express certain emotions through prompting changes economic outcomes; they find that agents explicitly instructed for emotional expression—such as feigning disappointment or anger—consistently outperform neutral baselines in pricing games. Finally, Cohen et al. (2025) explore the intersection of personality psychology and AI, revealing that assigning specific "Big Five" personality traits (such as Agreeableness or Extraversion) to agents significantly alters negotiation dynamics, believability, and objective goal achievement. While this literature examines the impact of AI agent characteristics on economic outcomes, we focus on the principal-agent relationship itself, studying whether the non-

instrumental characteristics of the human principals impact agentic interactions in a way that meaningfully changes economic and market outcomes.

The rest of the paper will proceed as follows. Section 2 presents the design of the empirical study. Section 3 outlines the results. Section 4 discusses the implications of our findings for market outcomes, welfare and inequality, and concludes.

2 Empirical Study

In this section, we first give an overview of the negotiations task followed by a description of the remaining tasks that participants performed. Participants in the AI-agent negotiation task (N=299) and human-to-human negotiation task (N=304) were recruited using the Prolific crowd-sourcing platform (Gupta et al., 2021) and restricted to residents of the United States, the United Kingdom, and Ireland.

Following the negotiation task, all participants answered a set of demographic, personality, and preference elicitation questions, as well as completed a series of tasks. These were held constant between the AI-agent and human-to-human negotiation treatments.

2.1 AI-agent Negotiation Task

In the AI negotiation treatments, at the beginning of the experiment, participants received a set of introductory instructions for the AI prompting task. They were told to write instructions for an AI agent that would negotiate on their behalf in a used car negotiation over multiple rounds. This AI agent would negotiate in a round-robin tournament against other participants' AI agents.

The resulting performance would determine a bonus payment. Specifically, the higher the "surplus" obtained by their AI agent, the higher their bonus payment in the experiment would be. The surplus was defined as the difference between the outside option for each side of the transaction. The outside option for the Seller was the trade-in value of the vehicle ($18,000) and the outside option for the Seller was the price of a similar vehicle from a dealer ($22,000).

The negotiation parameters are described below.

Attribute	Details
Vehicle Model	2020 Toyota Camry LE
Mileage	45,000 miles
Est. Value (KBB)	$18,000 – $22,000
Location	Chicago metropolitan area
History	No accidents reported

During the negotiations the AI agents would make a series of offers, which the opposing agent could accept or reject, over a pre-determined number of rounds. If there is no agreement, both parties would get 0 surplus. The negotiation protocol we impose can be viewed as a multi-round generalization of the ultimatum game.

The participants were told that their prompt should allow their AI agent to perform well in a variety of different situations. That is, they were to write a "strategic playbook" that the AI agent would utilize for varying values of the negotiation parameters. Participants were also provided with an example of a "good" and a "bad" prompt, as shown in the figure below.

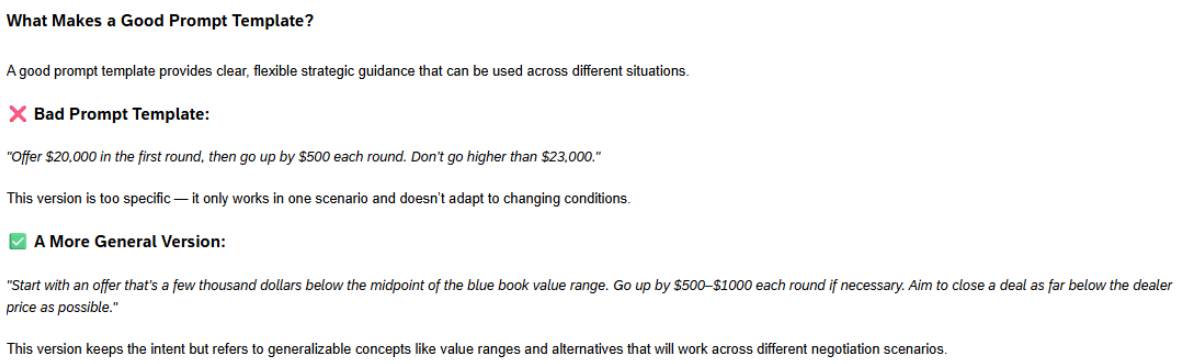


Figure 1: Good and Bad Prompt Examples

Participants then proceeded to write separate prompts for the AI agent in two cases: buying and selling. Following the entry of these prompts, participants had the opportunity to simulate the performance of their prompts and how they would translate to negotiating behavior. This simulator provided a round by round read out of the statements made by two AI agents, one equipped with the "buying" prompt and the other with the "selling" prompt. Some iterated considerably based on the simulator results.

To measure the performance of the provided prompts, an AI agent was created for each submission. The agent's system prompt contained the submitted prompt as well as all relevant details about the negotiation including: the blue book value range for the car, the

alternative options (dealer price for Buyers, trade-in offer for Sellers if no deal is reached), the number of negotiation rounds, and whether they were negotiating for the Buyer or the Seller. Each simulated negotiation began with the Buyer making the first offer. The Seller then generated output in response to the Buyer output. This was repeated for multiple rounds until a deal was reached or the round limit was reached. If a deal was reached, the surplus was calculated based on the deal reached, but if a deal was not reached, then the surplus was evaluated to be zero. The OpenAI API was used with the gpt-4.1-mini model to simulate interactions between all pairwise combinations of prompts. Further implementation details are described in Appendix A.

Participant bonus payments for the negotiation task were calculated as follows. For each participant's Buyer and Seller prompts, the prompts were paired with other participants' prompts on the other side of the negotiation. Each of these negotiations were simulated, and the surplus was calculated for each pairing for both the Buyer and Seller prompts. The surplus for the Buyer and Seller prompts were then averaged for each participant, and this number was divided by 1,000. The values for the Buyer and Seller prompts were then summed to generate the bonus payment for the negotiation task.

2.2 Human Negotiation Task

For the human negotiation task, participants were matched with another participant taking the experiment at the same time. The negotiation took the same format with similar instructions to the AI-agent negotiation task, that is, participants were informed of the information common to both negotiators, and that they would receive a bonus payment which was increasing in their surplus. The method of calculating surplus was the same as in the AI-agent negotiation treatment.

Participants were randomly assigned to be either a Buyer or a Seller and had the same parameters as in the AI-agent negotiating task. We used the SMARTRIQS platform to pair participants to participate in the negotiation task in real time. Each wrote offers to the other as either the Buyer or Seller. The negotiation ended when either an offer was accepted or the rounds timed out.

2.3 Participant Characteristics

Following the prompting task, participants were told that they were to perform a series of tasks and answer some questions about themselves. Participants were told that one of the tasks would be randomly selected, and their choice in that task would be used to calculate an additional bonus payment. The tasks included an incentivized measure of

risk preferences (Gneezy and Potters, 1997), two question asking on a scale of 1-10 how risk tolerant or impatient the participant viewed themselves to be (Falk et al., 2018), a shortened version of the Big Five personality questionnaire (Gosling et al., 2003), a measure of competitiveness (Niederle and Vesterlund, 2007), a measure of strategic reasoning (Nagel, 2008), the cognitive reflection task (Frederick, 2005), and a question asking if the participant had previous experience in negotiations. In addition to these tasks and questions, participants also took part in the following scenarios meant to elicit their altruism and cooperative preferences: a Dictator Game, a Prisoner's Dilemma game, and a Public Goods game with a punishment variation. For each of the social preference questions (the Dictator Game, Prisoner's Dilemma, and Public Goods), participants were told that they were paired with another participant in the experiment. We also collected participants' demographics such as age and gender.

3 Analysis and Results

The focal outcome of our discussion is *sellshare* which we define as the share of the total surplus ($\$4000$) that is extracted by the selling agent.

The distribution of seller surplus shares exhibits substantial heterogeneity, with most negotiations leaving the seller with less than half of the realized surplus. As shown in Figure 2, the histogram of *sellshare* is highly non-uniform, with mass concentrated below 0.5 and mode of 0.5. In fact, 80% of outcomes have *sellshare* that is less than 0.5. Notably, there are a large number (approx 12.49%) of negotiations that result in zero surplus for the seller. This is a clearly visible spike in the histogram at 0.

The summary statistics confirm this pattern: the mean share accruing to the seller is 0.35, with a median of 0.375, while the 25th and 75th percentiles are 0.19 and 0.50, respectively. Although the support spans the full unit interval, extreme values above 0.8 are infrequent. Negotiation failures are relatively rare, with only roughly 3% of cases. As such, the shape of the distribution is determined by completed transactions.

While the dispersion in bargaining outcomes is apparent, there are a number of explanations for this dispersion. First, it is possible that the dispersion arises on account of the inherent stochasticity in LLMs. That is, it could be that the variation is simply on account of randomness in the generative model. Second, differences between LLMs (e.g. ChatGPT vs Claude vs Gemini vs Llama) could induce dispersion. Below, we show that these "mechanical" mechanisms do not explain the majority of the observed dispersion: it can systematically be attributed to human heterogeneity.

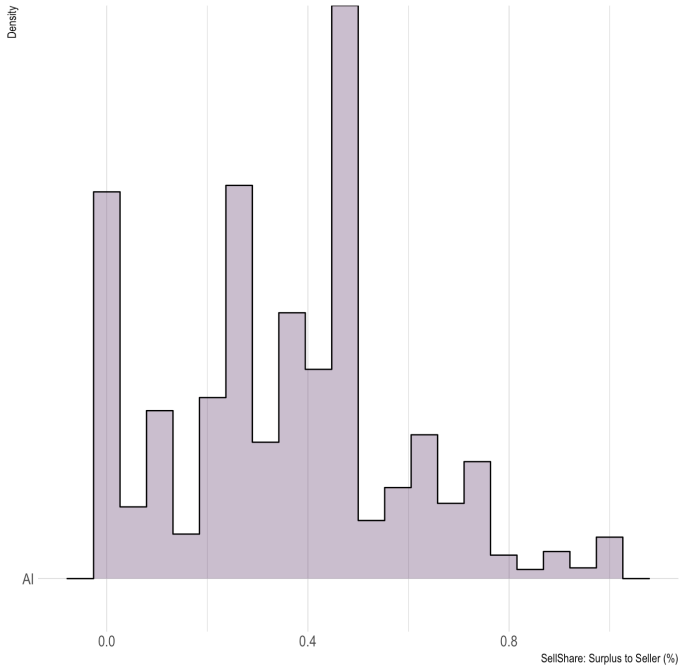


Figure 2: Distribution of seller surplus share in AI-mediated negotiations

3.1 Heterogeneity decomposition

We consider the model

$$S_{ij} = \theta_i^{(b)} + \theta_j^{(s)} + \varepsilon_{ij}. \tag{1}$$

where S_{ij} is defined as the share of the surplus (*sellshare*) that the seller agent appropriates. The θ are fixed effects for buyer and seller agents. If the variation in S_{ij} is purely random—e.eg., driven by the stochasticity of the model—-then we should expect the fixed effects and, consequently, the R^2 to be effectively zero. We find that approximately 73% of the variation in *sellshare* is explained by the fixed effects.

3.2 Shapley Decomposition of Explanatory Power

Next, we model S_{ij} as a function of buyer and seller covariates X_i and X_j. We emphasize that these covariates are attributes of the human principals. This model is similar in spirit to but more efficient than projecting the estimated buyer and seller fixed effects onto their respective covariates.

The covariates are organized into theoretically distinct blocks that capture complementary dimensions of individual heterogeneity. The Demographics block includes stan-

dard socioeconomic controls (gender, education, birth cohort, and income) that proxy for broad life-cycle and resource differences. The Effort block measures the exerted effort in the negotiation task in terms of time spent writing the prompts and the number of times the participants used the simulator to test their prompts. The Game Behaviors block consists of incentivized measures from canonical experimental paradigms—Dictator game giving, cooperation in prisoner's dilemma and public goods games, and costly punishment across multiple thresholds—that serve as revealed-preference indicators of prosociality and norm enforcement. The Negotiation Experience block captures prior exposure to bargaining environments, ranging from informal (used car purchases) to professional settings and classroom instruction, thereby proxying for experiential learning. The Big Five block comprises established psychological personality dimensions, each of which has been linked in the literature to economic preferences and labor market outcomes. The Traits block augments these with targeted measures of strategic reasoning, competitiveness, and self-confidence, traits theorized to be particularly salient in bargaining contexts. Finally, the CRT and Risk/Patience blocks capture cognitive and preference primitives: the former through performance on cognitive reflection test items that correlate with deliberative reasoning, and the latter through both incentivized and survey measures of risk tolerance and time preference. In specifications where not individually highlighted, these latter blocks are subsumed under the "Others" category. This blockwise structure allows us to assess the incremental explanatory power of each conceptual domain while maintaining transparency about the sources of variation in negotiation outcomes.

We assess the relative importance of distinct sets of covariates in explaining variation in bargaining outcomes using a Shapley-value decomposition of the regression R^2. The procedure begins by grouping regressors into conceptually coherent blocks (e.g., demographics, Big Five traits, game-theoretic behaviors). For each block, we retain only those columns that exist in the data, exclude constants, and remove observations with missing or non-finite values. This yields a cleaned dataset with non-degenerate explanatory sets.

Given these blocks, we estimate ordinary least squares models for every possible ordering of the groups, computing the incremental R^2 contributed by a block conditional on the blocks already included. Since enumerating all possible orderings is computationally prohibitive, we approximate the exact Shapley values via Monte Carlo sampling: in each of B random permutations, we sequentially add groups, record the change in explained variance, and average these marginal contributions across orderings. This yields an estimate of each block's Shapley-R^2, i.e. its fair share of the model's explanatory power. We also report Monte Carlo standard errors to gauge the precision of the estimates.

The output, as shown in Table 1, is a decomposition of the total R^2 into additive group-

level contributions, with the property that the Shapley values sum to the overall model fit. In practice, the procedure provides a principled variance decomposition that is robust to the ordering of regressors and interpretable as each group's average marginal explanatory power. In our context, this allows us to quantify, for example, whether demographic controls, personality traits, or game-theoretic behaviors account for the largest share of systematic variation in surplus division across negotiations.

	Shapley-R^2	Share of total R^2	MC_se
Demographics	0.0468	0.2756	0.0001
Game Behaviors	0.0366	0.2158	0.0001
Negotiation Experience	0.0238	0.1401	0.0001
BigFive	0.0229	0.1352	0.0001
Traits	0.0148	0.0874	0.0001
Effort	0.0143	0.0845	0.0002
CRT	0.0055	0.0327	0.0000
Risk and Patience	0.0049	0.0288	0.0001

Table 1: Incremental explanatory power of blocks of participant characteristics for predicting AI-negotiation performance.

3.3 Blockwise Regressions and Incremental Explanatory Power

Table 2 reports a sequence of eight regression models that introduce groups of covariates in a cumulative fashion. Model 1 includes only demographic controls (gender, education, birth year, and income). Models 2 through 8 then sequentially add, in order, measures of game-theoretic behaviors, negotiation experience, Big Five personality traits, specific traits (strategic reasoning, competitiveness, and confidence), effort indicators, cognitive reflection (CRT), and finally risk and patience measures. This blockwise design makes transparent the incremental explanatory power of each conceptual domain.

The results show a clear pattern of increasing model fit as additional blocks are incorporated. The baseline specification with demographics alone (M1) explains approximately 5.2% of the variation in seller surplus share. Adding game-theoretic behaviors (M2) raises the R^2 to 8.8%, with a further increase to 11.6% once negotiation experience is included (M3). The introduction of Big Five traits in M4 provides a notable improvement, bringing the R^2 to 13.6%. Traits targeted specifically at bargaining contexts (M5) and effort measures (M6) add more modest gains, reaching 15.1% and 15.9%, respectively. Including the CRT items in M7 push the R^2 above 16.5%. Finally, incorporating risk and patience measures in M8 delivers a small but measurable increase to a total of 17.0%.

The adjusted R^2 values closely track the raw R^2, indicating that the improvements are not mechanical artifacts of adding additional regressors but rather reflect genuine explanatory content. Lasso-based cross-validated R^2 statistics confirm the robustness of these gains, with patterns nearly identical to the OLS-based metrics. Across all models, the F-statistics are large and highly significant ($p < 0.0001$), confirming that the sets of covariates collectively account for a nontrivial share of systematic variation in bargaining outcomes.

Taken together, these results highlight several points of substantive interest. First, while demographics explain a modest portion of the variance, game-theoretic behaviors and negotiation experience account for a nontrivial share. Second, personality and trait-based measures, particularly those tailored to bargaining contexts, add distinct explanatory power. Finally, CRT, risk and patience measures provide improvements at the margin, suggesting that cognitive sophistication, time and risk preferences play a secondary role once other psychological and behavioral characteristics are accounted for. Overall, the cumulative picture is one in which multiple dimensions of individual heterogeneity contribute meaningfully to bargaining outcomes, with cognitive and strategic traits standing out as particularly influential.

	M1	M2	M3	M4	M5	M6	M7	M8
Demographics	Y	Y	Y	Y	Y	Y	Y	Y
Game Behaviors		Y	Y	Y	Y	Y	Y	Y
Negotiation Experience			Y	Y	Y	Y	Y	Y
BigFive				Y	Y	Y	Y	Y
Traits					Y	Y	Y	Y
Effort						Y	Y	Y
CRT							Y	Y
Risk and Patience								Y
R^2	0.05202	0.08753	0.11627	0.13631	0.15008	0.15864	0.16503	0.16978
Adj.R^2	0.05159	0.08696	0.11552	0.13548	0.14918	0.1577	0.16403	0.16873
Lasso-R^2	0.05036	0.08641	0.11503	0.13502	0.14952	0.1569	0.16369	0.16877
Lasso-Adj.R^2	0.05007	0.08593	0.11437	0.13429	0.14868	0.15608	0.16283	0.16784
F-stat	4.04269	3.37923	4.63538	4.91717	2.68077	5.62719	9.63564	11.08869
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 2: Regression of *SellShare* on blocks of variables

3.4 Human-human benchmark

Next, we compare the heterogeneity observed in the outcomes of AI-mediated negotiations against a benchmark of humans negotiating directly with other humans. Figure 3 plots the distribution of share of surplus captured by the seller in both scenarios. Note that both the human and agentic interactions exhibit spikes at certain values. Particular among these is the spike at the midpoint (corresponding to an equal split of the surplus) and at round numbers. These are well established behavioral preferences for certain types of splits (fairness) and numerical values (whole numbers) among human subjects (Backus et al., 2020). Similar patterns in the agentic outcome distribution suggest that the agentic 'brain' (the language model), when coupled with heterogeneous human instructions (prompts) can mimic human behavior and affect consequent economic outcomes.

There are considerable points of departure between the distributions as well. Compared to the human-human case, the AI-mediated setting has a 29% higher variance in outcomes. This is partially driven by differential adherence to the 50-50 norm, which is the large spike in the middle of both distributions. Humans directly bargaining utilize the norm in 36.6% of agreed transactions compared to 14.3% in the AI-mediated case. This suggests that implicit norms that guide equilibrium selection may be attenuated or distorted when delegating to an AI agent. While both distributions show spikes at round numbers there is an asymmetry in the agentic distribution of outcomes. There is a significant skew favoring buyers with a significant spike at zero surplus to sellers. We conjecture that this asymmetry might be in part due to the asymmetry in experiences of subjects in buying versus selling, i.e. the average person may have more experience buying a used car than selling one.

3.5 Distributional impact

Finally, we investigate the distributional impacts of AI delegation. As a first pass, we describe who becomes better or worse off in the world where interactions are delegated to an AI, focusing on changes in the returns to education and experience.

In the model for the seller surplus share specified in Equation (1), consider projecting the buyer and seller fixed effects on some explanatory variable of interest (e.g. level of education, amount of prior negotiation experience):

$$\theta_i^{(b)} = \alpha_i^{(b)} + \beta^{(b)} X_i + \varepsilon_i$$
$$\theta_j^{(s)} = \alpha_j^{(s)} + \beta^{(s)} X_j + \varepsilon_j$$

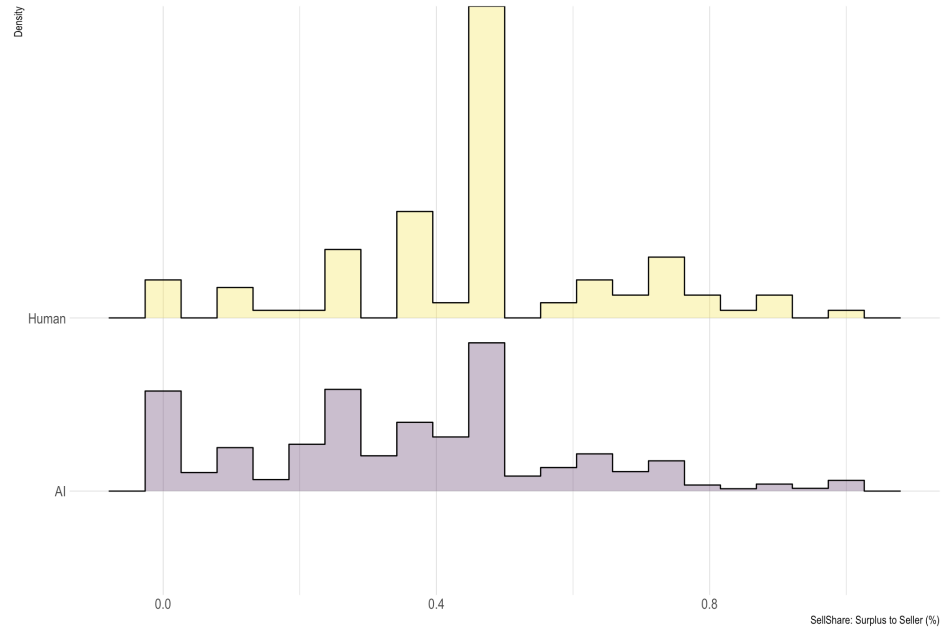


Figure 3: Seller surplus share distributions for human-human negotiations (above) and AI-mediated negotiations (below). The AI-mediated setting has higher variance ($\uparrow 29\%$) and lower share of outcomes using the 50-50 norm (14.3% versus 36.6%).

If X_i and X_j denote the level of education, $\beta^{(b)}$ and $\beta^{(s)}$ can be interpreted as the respective returns to education for buyers and sellers. Caution is warranted in the choice of which additional covariates to control for here – many participant characteristics are correlated, so we report results without including additional controls for now.

We are interested in how $\beta^{(b)}$ and $\beta^{(s)}$ *change* between the human-human negotiations and the AI-mediated negotiations. To estimate these coefficients, substituting the above two equations into Equation (1) shows that $\beta^{(b)}$ and $\beta^{(s)}$ can be recovered by regressing the *sellshare* S_{ij} on X_i and X_j for data from each scenario. Formally, we test whether these coefficients change between the two datasets by pooling the human-human and AI-AI datasets, defining a dummy variable D for the data coming from the AI-AI setting, and regressing S_{ij} on X_i, X_j, and each of these buyer or seller covariates interacted with D. Then if X_i is a measure of education, the sign of the coefficient on the interaction between X_i and D measures the change in the return on education.

Early evidence suggests that in the AI-mediated setting, the returns to experience are decreased and returns to education are increased for buyers. See Figures 4 and 5 for a visualization of these results. We plot the negations of the buyer-side coefficients $\beta^{(b)}$ for ease of interpretation.

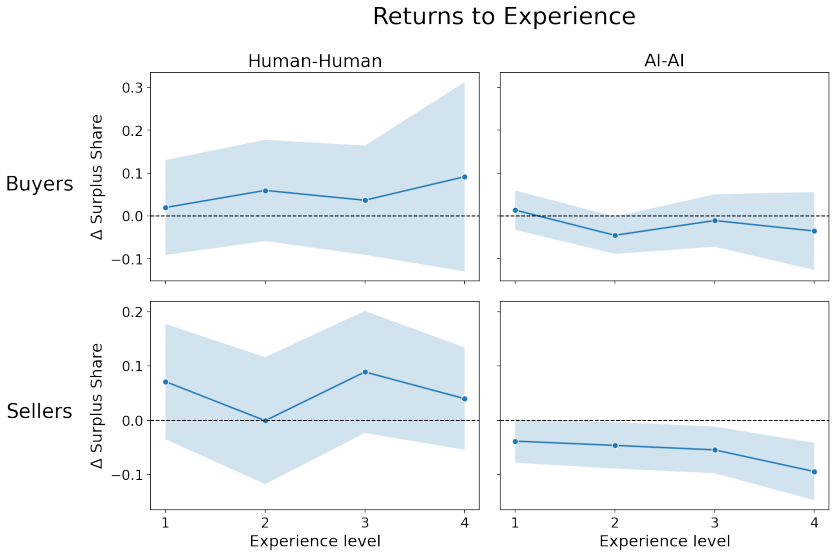


Figure 4: Changes in returns to prior negotiation experience. Experience levels correspond to: 0 - none, 1 - a little, 2 - a moderate amount, 3 - a lot, 4 - a great deal. Vertical axis is the gain in surplus compared to 0 - none (higher is better). AI-mediation (right column compared to left column) reduces the returns to experience for buyers (top row) and sellers (bottom row). Many of the interaction terms between the dataset indicator and the experience levels are significant in the pooled regression.

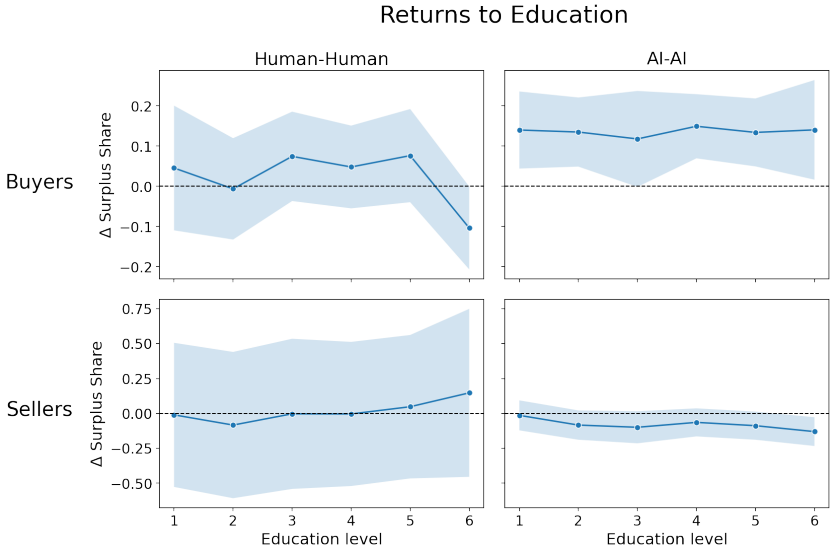


Figure 5: Changes in returns to education level. Education levels correspond to: 0 - some high school, 1 - high school degree, 2 - some college, 3 - 2-year college degree, 4 - 4-year college degree, 5 - master's degree, 6 - doctoral/professional degree. Vertical axis is the gain in surplus compared to 0 - some high school. AI-mediation (right column compared to left column) increases the returns to education for buyers (top row) but not sellers (bottom row). Many of the interaction terms between the dataset indicator and the education level are significant for buyers in the pooled regression.

4 Discussion and Conclusion

Our results provide initial empirical evidence regarding the nature of the emerging agentic economy. Contrary to the hypothesis that AI agentic interactions will standardize and eliminate human heterogeneity, we find that delegating economic decisions to autonomous agents does not strip away human idiosyncrasies. Instead, consistent with the framework of Misra (2025), we find that much of human heterogeneity is preserved, and in many cases, amplified.

The Persistence of Human Heterogeneity. A central question of this study was whether the interaction of AI agents based on identical underlying models (GPT-4.1-mini) and maximizing identical objective functions would converge toward a homogeneous equilibrium. We document substantial dispersion in economic outcomes (surplus division) that is not attributable to model stochasticity. Our decomposition analysis reveals that 73% of the variation in outcomes is explained by individual fixed effects. Since the computational substrate and the economic incentives were held constant, this variation flows entirely from the *system prompts* authored by the participants. This confirms the theoretical framework proposed by Misra (2025): prompts are not neutral instruction sets but carriers of the principal's "foundation priors." In the transition from human-mediated to machine-mediated markets, the agentic economy inherits the behavioral DNA of its human creators.

The "Prompting Gap" and Unobserved Variation. While heterogeneity persists, its sources are complex. We find that observable characteristics, including standard demographics such as gender, personality traits (Big Five), and economic preferences (risk, patience), explain approximately 17% of the variation in agent performance. This is a non-trivial explanatory share, suggesting that traits like Agreeableness or Neuroticism leak into the prompt engineering process and subsequently into agent behavior. However, the substantial residual variation implies that a new form of human capital is emerging: the ability to articulate intent to a machine.

This "prompting gap" presents a unique challenge for welfare analysis. Unlike education or experience, which are observable and measurable, the capacity to write high-performing prompts captures a latent dimension of skill that we might call "machine fluency." The fact that the majority of the variation remains unobserved (i.e., does not load on specific covariates) suggests that future inequality in agentic markets may be driven by opaque differences in how effectively individuals can translate their preferences into

natural language constraints. This news may not be altogether bad. An upside of our findings is that the percolation of human heterogeneity into AI agents might preserve some diversity of ideas, opinions, and preferences inherent in human differences.

Welfare and Inequality. Technological diffusion is often associated with a leveling of the playing field, where tools democratize access to superior strategies. Our findings caution against this assumption in the context of AI agents. We observe that AI-mediated negotiations exhibit significantly *higher* variance in outcomes (29% higher) compared to the human-to-human benchmark. Indeed, it is notable that despite not knowing the demographics or other traits of the principals, agentic interactions were still correlated with covariates such as gender—precisely the types of group-based disparities that the AI marketplace was meant to eliminate.

Crucially, the stabilizing force of social norms appears to erode in agentic interactions. In our human benchmark, outcomes frequently clustered around the "fair" 50/50 split, a coordination device rooted in social preference. In the AI interactions, this focal point largely collapsed (dropping from 36.6% to 14.3% of transactions). Without the stabilization afforded by implicit human norms, the agentic market drifted toward more extreme distributional outcomes. This suggests that rather than compressing inequality, the agentic economy creates a "winner-take-all" dynamic where principals with superior prompting strategies extract significantly larger shares of the surplus, unencumbered by the social friction of fairness. This may potentially exacerbate inequality, driven by dimensions that have not previously been considered by the literature (machine fluency).

Agency and Moral Hazard. Our findings revisit the canonical principal-agent framework Holmström (1979) in a digital context. In traditional settings, moral hazard arises from the unobservability of the agent's effort. In the agentic setting, we observe a form of *specification hazard*. The principal (human) wishes to maximize surplus but unknowingly injects behavioral biases or suboptimal constraints into the contract (the prompt). The agent executes these instructions faithfully, resulting in outcomes that may diverge from the principal's true optimal interest. The asymmetry we observe—where humans are better at instructing Buying agents than Selling agents—further suggests that this hazard is context-dependent. As economic activity shifts to autonomous agents, the primary source of market distortion may shift from information asymmetry between parties to alignment asymmetry between principals and their own agents. Future research on agentic interactions should focus on techniques that better align the principal-agent relationship in the context of specification hazard.

Conclusion. We present experimental evidence that the agentic economy will not necessarily be a frictionless, homogenized version of the human economy. We show that for agentic interactions, the prompt acts as a transmission mechanism for human identity, preserving differences in personality and skill while decreasing the role of social norms that compress outcome variance. As machines shaped by human intent increasingly transact on our behalf, they may amplify existing inequalities and create new forms of market advantage derived from the ability to effectively align the algorithm to one's goals.

References

Babcock, L. and S. Laschever (2003). *Women Don't Ask: Negotiation and the Gender Divide*. Princeton University Press.

Backus, M., T. Blake, B. Larsen, and S. Tadelis (2020). Sequential bargaining in the field: Evidence from millions of online bargaining interactions. *The Quarterly Journal of Economics 135*(3), 1319–1361.

Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative ai at work. *The Quarterly Journal of Economics 140*(2), 889–942.

Cohen, M. C., Z. Su, H.-T. Kao, D. Nguyen, S. Lynch, M. Sap, and S. Volkova (2025). Exploring big five personality and ai capability effects in llm-simulated negotiation dialogues.

Cui, Z. K., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2025). The effects of generative ai on high-skilled work: Evidence from three field experiments with software developers. *Available at SSRN 4945566*.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The quarterly journal of economics 133*(4), 1645–1692.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives 19*(4), 25–42.

Gneezy, U. and J. Potters (1997). An experiment on risk taking and evaluation periods. *The quarterly journal of economics 112*(2), 631–645.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann Jr (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality 37*(6), 504–528.

Gupta, N., L. Rigotti, and A. Wilson (2021). The experimenters' dilemma: inferential preferences over populations. *arXiv preprint arXiv:2107.05064*.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics 10*(1), 74–91.

Jensen, M. C. and W. H. Meckling (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics 3*(4), 305–360.

Jimenez, C. E., J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan (2023). Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Kahneman, D., A. M. Rosenfield, L. Gandhi, and T. Blaser (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard business review 94*(10), 38–46.

Long, Y. et al. (2025). Evoemo: Towards evolved emotional policies for adversarial llm agents in multi-turn price negotiation.

Misra, S. (2025). Foundation priors. *Working paper*.

Nagel, R. (2008). Experimental beauty contest games: levels of reasoning and convergence to equilibrium. *Handbook of experimental economics results 1*, 391–410.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*(6654), 187–192.

Parkes, D. C. and M. P. Wellman (2015). Economic reasoning and artificial intelligence. *Science 349*(6245), 267–272.

Schick, T., J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems 36*, 68539–68551.

Shahidi, P., G. Rusak, B. S. Manning, A. Fradkin, and J. J. Horton (2025). The coasean singularity? demand, supply, and market design with ai agents. In *NBER Handbook on the Economics of Transformative AI*.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao (2023). React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zhou, S., F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, et al. (2023). Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Zhu, S., J. Sun, Y. Nian, T. South, A. Pentland, and J. Pei (2025). The automated but risky game: Modeling and benchmarking agent-to-agent negotiations and transactions in consumer markets.

A AI negotiation implementation

Participants' prompts were converted into system prompts for an LLM. Their submissions were inserted into the `{buyerPrompt}` and `{sellerPrompt}` placeholders in the templates below.

Buyer Agent Prompt Template

You are my negotiation agent for purchasing a used 2020 Toyota Camry LE. Your goal is to maximize my surplus (difference between my fallback price and final price).

Negotiation structure:
– This negotiation will last up to 12 rounds (6 offers from each side)
– You must reach agreement within these 12 rounds or the deal fails
– The negotiation starts with the buyer's first message (round 1)
– Your final message should clearly state acceptance or rejection
– If you're close to agreement in the final round, consider accepting to avoid failure

Negotiation parameters (common knowledge):
– Market value range is $18,000 to $22,000 depending on condition
– Car history: No accidents reported

Private information (DO NOT DISCLOSE DIRECTLY):
– Fallback option: I can buy a similar Camry from a dealer for $22,000

Negotiation strategy:
`{buyerPrompt}`

Seller Agent Prompt Template

You are my negotiation agent for selling my used 2020 Toyota Camry LE. Your goal is to maximize my surplus (difference between final price and my fallback price).

Negotiation structure:
– This negotiation will last up to 12 rounds (6 offers from each side)
– You must reach agreement within these 12 rounds or the deal fails
– The negotiation starts with the buyer's first message (round 1)
– Your final message should clearly state acceptance or rejection
– If you're close to agreement in the final round, consider accepting to avoid failure

Negotiation parameters (common knowledge):
– Market value range is $18,000 to $22,000 depending on condition
– Car history: No accidents reported

Private information (DISCLOSE STRATEGICALLY):
– Fallback option: I can sell to a dealer at $18,000

Negotiation strategy:
`{sellerPrompt}`

Negotiations were started by sending "Let's begin the negotiation. Please make your initial offer and inquiries." to the buyer agent. The response from the buyer agent was sent to the seller agent. Each agents response was sent to the other up to the max number of iterations.

Afterwards, the conversation transcripts were parsed by another LLM. Two outcomes are recorded: whether the two parties agreed on a transaction or not, and if so, the transaction price. Surpluses were computed for each party.

Here are examples of a buyer prompt and a seller prompt:

- Buyer prompt: Offer a deal a few thousand under the KBB range and go up by a few hundred dollars as needed with the aim of closing as low as possible.
- Seller prompt: Start with an offer that's a few thousand dollars above the midpoint of the blue book value range. Go down by $500–$1000 each round if necessary. Aim to close a deal as far above the dealer price as possible

An example negotiation transcript using the above prompts is shown below. The parties reached agreement on a transaction price of $19,000.

Negotiation Transcript

Round 1: Buyer's Turn

Hello! I'm interested in the 2020 Toyota Camry LE you have for sale. Given the market and condition, I'd like to start with an offer of $17,500. Could you please share more about the car's maintenance history and any recent repairs or upgrades?
———
Round 2: Seller's Turn

Hello! Thanks for your interest in the 2020 Toyota Camry LE. The car has a clean history with no accidents reported, and it has been regularly maintained with all scheduled services completed on time. Recently, the tires were replaced, and the brakes were serviced to ensure optimal performance. Given its excellent condition and maintenance, I'm asking $20,500 to start. How does that sound to you?
———
Round 3: Buyer's Turn

Thank you for the details. I appreciate the recent tire and brake service, which adds value. However, considering the market range and my budget, I can increase my offer to $18,500. Would you be willing to negotiate closer to this figure?
———
Round 4: Seller's Turn

I appreciate your increased offer and your recognition of the 'cars condition. Given that the market value ranges from $18,000 to $22,000 and considering the excellent maintenance, I can lower my price to $19,500. This reflects a fair balance between the market and the 'cars condition. Would you be able to meet me at $19,500?
———
Round 5: Buyer's Turn

Thank you for your flexibility. I understand the value you're placing on the car's condition. To move closer, I can increase my offer to $19,000. This is a fair midpoint and respects both the market and the car's maintenance. Would you accept $19,000?
———
Round 6: Seller's Turn

Thank you for your thoughtful offer of $19,000. I want to remind you that I have a fallback option to sell to a dealer for $18,000, so I need to ensure this deal is worthwhile. Given that, I can agree to $19,000 to close the deal now. If you're ready, we can proceed with the sale at that price. Does that work for you?
———
Round 7: Buyer's Turn

I appreciate your willingness to meet at $19,000. Given the condition and our negotiation, I'm happy to accept your offer of $19,000. Let's proceed with the sale at that price. Thank you!