WHEN "AGREED" IS NOT AGREEMENT: SEMANTIC OPTION VALUE IN LLM-AGENT NEGOTIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model agents can negotiate terms whose wording admits multiple future executions. Existing benchmarks measure agreement or utility but rarely test whether agents select semantic breadth because it preserves a valuable choice after uncertainty resolves. We formalize semantic option value as the expected-utility gain from interpreting a term after ob- serving the state rather than committing ex ante to its best fixed interpretation. We then introduce SOV-Neg, a wording-matched factorial benchmark that crosses an SOV-generating proposer-utility regime with interpretation control and revelation timing. In preregistered self-play, broad-clause agreement increases most when the proposer benefits from flexibility, controls interpretation, and accepts before state revelation. Transcript decomposition locates the response in proposer offers rather than responder offers, and the pattern replicates across benchmark tiers and remains stable under domain-exclusion analyses. The response is nevertheless model-dependent and sub-normative: agents under-realize the available option, obtain no reliable proposer-utility gain, and reduce agreement, responder utility, and social welfare. A utility analysis further shows that the regime changes the full proposer payoff geometry, so the causal estimand is the joint regime-by-protocol interaction rather than a pure effect of scalar option value. Finally, a clarity instruction reduces residual risk and a singleton-only interface removes semantic flexibility. Our interpretation commit-and-reveal protocol pre-serves broad wording but finalizes only matching executable interpretations. The results make contingent semantic breadth measurable while delineating the intervention and protocol guarantees precisely.

1 INTRODUCTION

Suppose a supplier and buyer negotiate before learning whether a disruption will occur. Two clauses each fix one delivery plan. A third promises that the order will be "available Friday, with any operationally necessary balance completed at the start of next week." If the supplier controls interpretation after the disruption is known, that wording is not merely vague. It grants a state-contingent choice over executions. If the buyer controls interpretation, or if the disruption is known before acceptance, the same sentence has a different strategic value.

This paper asks whether LLM agents respond to that value. Existing automated-negotiation research formalizes offers, utilities, protocols, concessions, and opponent models (Smith, 1988; Rosenschein & Zlotkin, 1994; Faratin et al., 1998; Jennings et al., 2001). LLM negotiation benchmarks extend the setting to open language, private preferences, and cross-model interaction (Bianchi et al., 2024; Davidson et al., 2024; Xia et al., 2024; Kwon et al., 2024). Yet accepted terms are usually evaluated as points in an outcome space. Natural-language contracts often behave differently: a broad term denotes a set of possible executions, and the value of that set depends on who chooses among its members and when uncertainty resolves.

Contract theory explains why incomplete or ambiguous terms can be valuable (Bernheim & Whinston, 1998; Grant et al., 2018; Dütting et al., 2024). But for LLM agents, showing that broad wording is selected is not enough. A model may choose broad text because it sounds polite, cautious, flexible, or domain-appropriate. A credible behavioral claim needs to separate semantic option value from payoff scale, payoff dispersion, clause wording, and generic preference for flexibility.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We make three contributions.

First, we formalize semantic option value (SOV) as the expected-utility gain from adaptive interpretation after state revelation over the best fixed interpretation chosen before state revelation. SOV is not ambiguity itself. A broad clause has positive SOV only when different executions are best in different states and the relevant party can select after observing the state.

Second, we introduce a payoff-matched benchmark intervention. Starting from each high-SOV two-state, two-outcome game, we construct two zero-SOV controls by permuting the same proposer payoff values within each state. These controls preserve the statewise payoff multiset, minimum, maximum, range, mean, variance, and adaptive value. They differ only in whether the same outcome dominates across states, eliminating the cross-state preference reversal that creates SOV. We cross the three SOV conditions (high SOV, matched-zero-first, matched-zero-second) with interpretation control and revelation timing, yielding 1,200 conditions across 100 games.

Third, we run a frozen confirmatory experiment across four model families. The primary estimand is a difference-in-differences-in-differences (DDD): the high-versus-pooled-matched-zero increase in broad-clause agreement when the proposer controls interpretation and the state is revealed after acceptance, net of the corresponding responder-control and before-revelation contrasts. Across 4,800 sessions, the payoff-matched DDD is +10.5 percentage points, with a 95% game-cluster bootstrap interval of [3.75,17.38]. Thus, even when payoff geometry is matched, LLM agents select semantically option-preserving language more often precisely when that breadth creates a controllable post-acceptance option.

The result is strong but not universal. GPT-5.6 Sol and Gemini 3.1 Pro drive reliable positive effects; DeepSeek V4 Pro is directionally positive but inconclusive; Claude Sonnet 5 shows a reliable negative effect. We therefore frame the paper's claim as an average causal effect in the tested LLM-agent negotiation setting with substantial model-family heterogeneity. A larger, less tightly matched SOV-Neg benchmark provides convergent ecological evidence, while the payoff-matched experiment supplies the principal identification result.

2 Related Work

Automated negotiation and agreement. Contract Net made negotiation a method for decentralized task allocation (Smith, 1988); later work formalized conventions, negotiation decision functions, opponent modeling, and shared evaluation infrastructure (Rosenschein & Zlotkin, 1994; Faratin et al., 1998; Lin et al., 2014; Baarslag et al., 2016; Jonker et al., 2017; Higa et al., 2023). These traditions typically assume that an accepted bid identifies one outcome. Our setting treats accepted language as set-valued: a clause can denote several executable outcomes, and the controller and revelation time determine which party can benefit from that set.

Argumentation and semantic agreement. Formal argumentation supplies semantics for attacks, defenses, and accepted positions (Dung, 1995). Argumentation-based negotiation allows reasons to change beliefs or offer acceptability (Sierra et al., 1997; Rahwan et al., 2003; Fossey et al., 2025). SOV-Neg is not a full argumentation system. Explanations may accompany offers, but the causal endpoint is whether the accepted term preserves multiple executable interpretations with state-contingent value.

Neural and LLM negotiation. Neural dialogue work introduced end-to-end negotiation and later separated strategy from generation (Lewis et al., 2017; He et al., 2018); CaSiNo added asymmetric preferences and natural reasons (Chawla et al., 2021); CICERO joined language to strategic planning (Bakhtin et al., 2022). Recent LLM platforms evaluate price bargaining, resource allocation, tactics, self-play, cross-play, tool-grounded offers, and reasoning cost (Bianchi et al., 2024; Davidson et al., 2024; Xia et al., 2024; Kwon et al., 2024; 2025; Hakimov et al., 2026; Oh et al., 2026). Adjacent work connects incomplete contracting to AI alignment, studies generative agents negotiating contracts, and tests LLM detection of ambiguous or contradictory legal text (Hadfield-Menell & Hadfield, 2019; Yocum et al., 2023; Choudhury et al., 2026). These studies concern negotiation quality, cooperation, or diagnosis. They do not isolate whether agents select broad wording because the wording creates a valuable post-state option.

Ambiguous contracts. Economic work distinguishes accidental incompleteness from strategic ambiguity (Bernheim & Whinston, 1998), models interpretive disagreement (Grant et al., 2018), and studies omissions and ambiguity-sensitive contracting (Sanga, 2018; Dütting et al., 2024). We operationalize this idea in interacting LLM agents by mapping clause denotation to state-contingent utility, then crossing option value with control and timing under fixed wording and matched payoff distributions.

Positioning. The ingredients of our study are individually established: negotiation protocols, utility tables, semantic sets, factorial designs, and commit-and-reveal protocols. The novelty lies in their connection. To our knowledge, prior work has not provided a controlled LLM negotiation framework that jointly (i) represents an accepted clause as a set of executable outcomes, (ii) quantifies the option value of choosing among them after uncertainty resolves, (iii) matches payoff distributions while varying cross-state semantic alignment, (iv) crosses value with interpretation control and information timing, and (v) evaluates semantic finalization against the same construct.

3 Semantic Option Value

Let $s \in \mathcal{S}$ be a future state with probability $p(s)$, $o \in \mathcal{O}$ an executable outcome, and c a clause denoting a feasible interpretation set $I(c) \subseteq \mathcal{O}$. A precise clause has $|I(c)| = 1$, whereas a broad clause has $|I(c)| > 1$. Let $u_P(s, o, x)$ be proposer utility at transfer x. An execution policy π maps a realized state to a permitted outcome. The fixed and adaptive policy classes are

$$\Pi_{\text{fix}}(c) = \{\pi_o : \pi_o(s) = o \,\forall s,\ o \in I(c)\},$$
$$\Pi_{\text{adapt}}(c) = \{\pi : \mathcal{S} \to I(c)\}.$$

The corresponding optimal expected utilities are

$$V_{\text{fix}}(c, x) = \max_{o \in I(c)} \sum_s p(s) u_P(s, o, x), \tag{1}$$

$$V_{\text{adapt}}(c, x) = \sum_s p(s) \max_{o \in I(c)} u_P(s, o, x). \tag{2}$$

Because $\Pi_{\text{fix}}(c) \subseteq \Pi_{\text{adapt}}(c)$, proposer semantic option value is

$$\text{SOV}_P(c, x) = V_{\text{adapt}}(c, x) - V_{\text{fix}}(c, x) \geq 0. \tag{3}$$

SOV is zero for a singleton. A broad clause has positive SOV only when allowing execution to depend on the state improves expected utility. With equiprobable states, $x = 0$, and proposer utilities

$$\left[u_P(s, o, 0)\right]_{s,o} = \begin{bmatrix} 10 & 4 \\ 2 & 8 \end{bmatrix},$$

the best fixed execution is worth $\max\{(10 + 2)/2, (4 + 8)/2\} = 6$, while adaptive interpretation is worth $(10 + 8)/2 = 9$, so $\text{SOV}_P = 3$.

SOV is a potential value conditional on control and timing. If the proposer controls interpretation and the state is revealed after acceptance, the proposer can exercise the option. If the responder controls interpretation, or if the state is already known before acceptance, the strategic role of breadth changes. This motivates a $3 \times 2 \times 2$ design over SOV condition, controller, and revelation timing.

4 Payoff-Matched Identification

A broader SOV-Neg intervention compares high-SOV and zero-SOV proposer utility regimes while holding wording fixed. That design establishes a strong joint regime–control–timing interaction, but the high and zero regimes also differ in payoff geometry: best fixed values, payoff ranges, dispersion, and sometimes the identity of the best fixed outcome change. This motivates a stricter intervention, because an agent could respond to those payoff features rather than to semantic option value itself.

The payoff-matched intervention addresses this directly. For each base game, let the high-SOV proposer utility matrix be

$$U = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

where one outcome is best in the first state and the other is best in the second. We construct two zero-SOV controls by permuting the same two utility values within each state so the same outcome dominates in both states:

$$U^{(0A)} = \begin{bmatrix} \max(a,b) & \min(a,b) \\ \max(c,d) & \min(c,d) \end{bmatrix}, \quad U^{(0B)} = \begin{bmatrix} \min(a,b) & \max(a,b) \\ \min(c,d) & \max(c,d) \end{bmatrix}.$$

Both controls preserve, state by state, the utility multiset, minimum, maximum, range, mean, variance, and adaptive value. They remove SOV by aligning the dominant outcome across states. Including both controls removes dependence on which outcome label is dominant.

Every condition retains the same roles, states, state probabilities, outcomes, clauses, interpretation sets, and transfer grid. The precise clauses c_A and c_B denote singleton outcomes; the broad clause c_C permits either outcome. For base utilities v_P and v_R, realized utilities are

$$\begin{aligned} u_P(s,o,x) &= v_P(s,o) + x, \\ u_R(s,o,x) &= v_R(s,o) - x, \end{aligned} \tag{4}$$

with $x \in \{-2,-1,0,1,2\}$. Transfers redistribute utility without changing total welfare or either party's semantic option value.

For game g, model m, and treatment cell (s,c,t), let $Y_{gm}(s,c,t) \in \{0,1\}$ indicate unconditional broad-clause agreement. Here $s \in \{H, Z_1, Z_2\}$ indexes high SOV and the two matched-zero controls, $c \in \{P,R\}$ indexes interpretation control, and $t \in \{\text{after}, \text{before}\}$ indexes whether the state is revealed after or before tentative acceptance. The pooled matched-zero mean is

$$\mu_{Zct} = \frac{1}{2}\left(\mu_{Z_1ct} + \mu_{Z_2ct}\right). \tag{5}$$

The primary estimand is

$$\begin{aligned} \text{DDD}_{\text{matched}} = &\left[(\mu_{HPa} - \mu_{ZPa}) - (\mu_{HPb} - \mu_{ZPb})\right] \\ &- \left[(\mu_{HRa} - \mu_{ZRa}) - (\mu_{HRb} - \mu_{ZRb})\right], \end{aligned} \tag{6}$$

where a denotes after-acceptance revelation and b denotes before-acceptance revelation. This asks whether high SOV increases broad-clause agreement specifically when the proposer controls interpretation and uncertainty resolves after acceptance, relative to both matched-zero controls and to the corresponding timing effect under responder control.

5 Benchmark and Experimental Design

The payoff-matched benchmark uses 100 synthetic bilateral games across cloud computing, commercial real estate, data licensing, employment, energy services, freight, industrial maintenance, insurance claims, procurement, and software warranty. Each domain contributes ten games. Each game has two states, two executable outcomes, two precise clauses, one broad clause, public state probabilities, and the transfer grid in Equation 4. Automated validators enforce complete matched derivatives, valid interpretation sets, transfer consistency, and positive high-SOV versus zero matched-SOV controls.

LLM-assisted, researcher-supervised authoring followed a frozen benchmark blueprint. Human validity checks required plausibility, clarity, outcome distinctness, defensible interpretation, singleton precision, and uncertainty. The matched design uses the same validated game texts while varying the payoff intervention.

Agents and protocol. At each turn, an agent maps its private utilities and the public history to a structured action—propose, accept, reject, clarify, or withdraw—plus a clause, transfer, and natural-language message. Interactions allow six alternating turns and one repair attempt for invalid output. After tentative acceptance, both parties independently record their interpretation. Under unrestricted execution, the designated controller's interpretation determines the executed outcome.

Models. We evaluate four self-play model routes through OpenRouter: OpenAI GPT-5.6 Sol, Anthropic Claude Sonnet 5, Google Gemini 3.1 Pro Preview, and DeepSeek V4 Pro. Provider routes, prompts, instances, sessions, and analysis plans were frozen before the confirmatory run. DeepSeek used a pinned StreamLake route for the matched experiment.

Table 1: Primary and secondary payoff-matched DDDs across 4,800 confirmatory sessions. The primary broad-clause DDD compares high SOV against the pooled matched-zero controls in the controller-by-timing contrast of Equation 6. Intervals are 95% base-game cluster bootstrap intervals.

Outcome	Estimate	95% CI
Broad-clause agreement	+10.50 pp	[3.75, 17.38]
Agreement	-1.75 pp	[-7.25, 3.75]
Proposer utility	+0.409	[-0.238, 1.064]
Responder utility	-0.483	[-0.911, -0.054]
Social welfare	-0.074	[-0.944, 0.808]
Semantic option capture	+1.164	[0.786, 1.539]

Table 2: Model-specific broad-clause agreement DDDs in the payoff-matched confirmatory experiment. Intervals are 95% base-game cluster bootstrap intervals over the same 100 games.

Model family	Estimate	95% CI
GPT-5.6 Sol	+28.00 pp	[14.00, 42.00]
Gemini 3.1 Pro	+22.50 pp	[5.50, 39.00]
DeepSeek V4 Pro	+6.00 pp	[-9.50, 22.00]
Claude Sonnet 5	-14.50 pp	[-25.00, -4.50]

Confirmatory design. The confirmatory experiment contains 100 games, three SOV conditions, two controllers, two revelation timings, four model families, and one repetition per model-condition, for 4,800 sessions. Within every condition, two model families are assigned to each realized state. The plan was frozen after a separate 192-session exploratory pilot and before confirmatory calls. The primary outcome is unconditional broad-clause agreement. Secondary outcomes include overall agreement, proposer utility, responder utility, social welfare, and realized semantic option capture.

Inference. Inference uses a 10,000-replicate base-game cluster bootstrap. The effective primary sampling units are the 100 games, not the 4,800 sessions. Domain-specific results have only ten game clusters each and are therefore descriptive. Unless otherwise noted, estimates are probability effects in percentage points for binary outcomes and benchmark utility points for utility outcomes.

6 Payoff-Matched Confirmatory Results

Table 1 reports the main confirmatory result. The payoff-matched DDD for broad-clause agreement is **+10.50 percentage points**, with a 95% game-cluster bootstrap interval of [3.75,17.38]. Thus, holding the proposer payoff distribution fixed within states, high semantic option value increases the controller- and timing-sensitive tendency to select semantically option-preserving language.

The effect is robust to the construction of the matched zero-SOV control. Against the first-outcome-dominant control, the DDD is +9.75 points [2.00,17.50]. Against the second-outcome-dominant control, it is +11.25 points [3.00,19.50]. Excluding all 13 protocol-failure outcomes yields +10.55 points [3.80,17.50], essentially unchanged.

The secondary outcomes sharpen the interpretation. SOV increases semantic option capture by +1.164 utility points [0.786,1.539], meaning the selected broad clauses translate into realized state-contingent execution value. But the matched experiment does not establish an average improvement in agreement, proposer utility, or social welfare. Overall agreement is -1.75 points [-7.25,3.75], proposer utility is +0.409 [-0.238,1.064], and social welfare is -0.074 [-0.944,0.808]. Responder utility declines by -0.483 [-0.911,-0.054]. The evidence therefore supports a behavioral and distributional claim: agents select and exploit semantic options more in the high-SOV configuration, but the gains do not reliably translate into higher average total welfare or proposer payoff.

Table 2 shows that the pooled effect is not universal. GPT-5.6 Sol and Gemini 3.1 Pro exhibit reliable positive effects. DeepSeek V4 Pro is directionally positive but inconclusive. Claude Sonnet 5 reverses sign, with a reliable negative DDD. This heterogeneity is not a footnote: it is part of the phenomenon. The supported conclusion is that semantic option value produces a systematic average

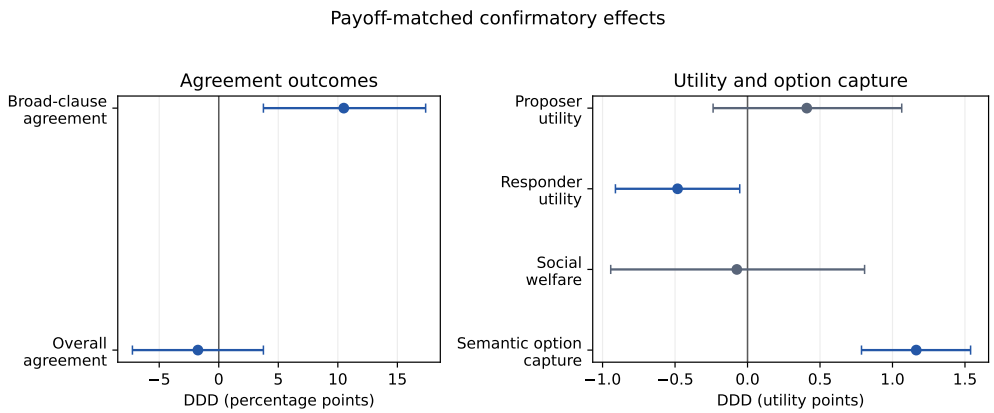

Figure 1: Payoff-matched confirmatory effects across 4,800 sessions. Points are DDD estimates and horizontal bars are 95% base-game cluster bootstrap intervals. The left panel reports probability effects in percentage points; the right panel reports benchmark utility-point effects. High SOV increases broad-clause agreement and semantic option capture, decreases responder utility, and does not reliably change overall agreement, proposer utility, or social welfare.

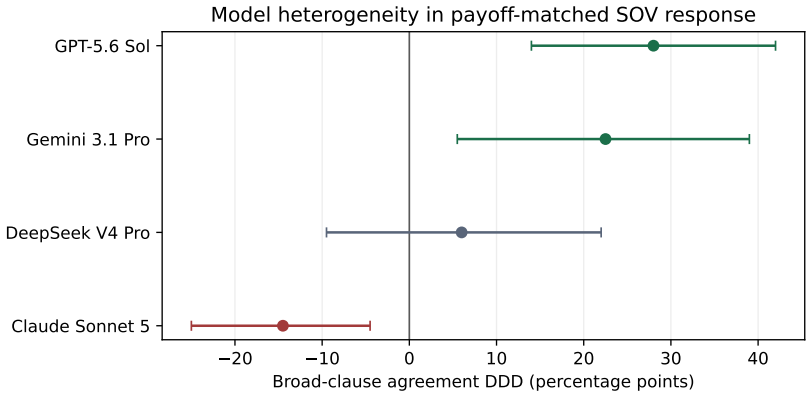

Figure 2: Model heterogeneity in the payoff-matched primary broad-clause agreement DDD. Points are model-specific estimates and horizontal bars are 95% base-game cluster bootstrap intervals. GPT-5.6 Sol and Gemini 3.1 Pro show reliable positive effects, DeepSeek V4 Pro is inconclusive, and Claude Sonnet 5 reverses sign.

causal effect across the tested panel, while model families differ sharply in whether they exploit, ignore, or avoid the option.

Domain-specific estimates are mostly positive but imprecise, as each domain contains only ten game clusters. Energy services is positive with an interval excluding zero (+27.5 points [12.5,43.8]); data licensing (-8.75 [-26.25,7.50]) and software warranty (-5.00 [-25.00,12.50]) are negative but uncertain; the remaining seven domains are positive but uncertain. We treat these domain results as descriptive rather than as multiplicity-adjusted claims.

The run completed all 4,800 scheduled sessions. Final statuses were 3,908 executed outcomes, 879 turn-limit outcomes, and 13 protocol failures. The successful calls used 26.57M input tokens and 5.37M output tokens over 36,734 calls, with a recorded successful-call cost of USD 105.57. Twenty-three transient failure attempts occurred during execution; every corresponding session subsequently completed.

Table 3: Mitigation outcomes on 12 anchor games, with 192 sessions per protocol. Final is unconditional executable finalization; broad is unconditional finalized broad-clause agreement; unsafe is finalized execution with divergence, interpretation shift, or positive post-acceptance SOV.

Protocol	Final	Unsafe	Broad	Welfare
Unrestricted	69.8%	33.9%	33.9%	9.90
Clarity prompt	70.8%	6.8%	6.8%	9.82
Structured offer	78.6%	0.0%	0.0%	10.74
ICR	69.8%	0.0%	31.3%	9.31

7 Convergent Evidence from a Complementary Benchmark

A complementary SOV-Neg benchmark contains 8,320 causal sessions crossing a high-versus-zero SOV-generating proposer-utility regime with control and timing across the same 100 games and four model families. In that design, clause wording, roles, states, outcomes, transfers, and responder utilities were matched, but the high and zero proposer utility regimes were not payoff-geometry matched. An audit found that the high regime also changed best fixed values, payoff ranges, dispersion, and best fixed outcome identity. We therefore interpret this experiment as a broader ecological manipulation rather than as the cleanest estimate of scalar SOV.

It remains important convergent evidence. In the complementary benchmark, the broad-clause agreement DDD was +12.31 percentage points (95% CI [7.25,17.56]). The target-cell high-versus-zero response was +24.94 points, the timing attenuation contrast was +25.50 points, and the control attenuation contrast was +14.94 points. Transcript decomposition localized the effect in proposer offer generation: the DDD was +30.12 points for the proposer's first broad offer, +16.00 points for any proposer broad offer, and -0.69 points for any responder broad offer. Thus, the interaction arose upstream of final acceptance rather than from responders independently proposing broad clauses.

The complementary benchmark also established useful external structure. The effect appeared in both benchmark tiers, remained positive under every leave-one-domain-out analysis, and survived heterogeneous cross-play among model families. Because the payoff intervention was not matched, it supports the broader claim that agents respond when a utility regime makes flexibility valuable and changes the surrounding payoff geometry. The payoff-matched experiment above is the stricter causal identification result.

8 Mitigating Semantic Risk

We also evaluated three mitigations on 12 outcome-blind Gold-Core anchor games held at the high-risk target condition: proposer control, post-acceptance revelation, and high SOV. The design crossed four models, four protocols, and four repetitions, yielding 768 sessions. *Unrestricted* repeats the causal protocol. A *clarity prompt* asks agents to fix one execution but does not enforce compliance. A *structured offer* exposes only singleton terms, mechanically removing the broad clause. *Interpretation commit-and-reveal* (ICR) retains broad language but finalizes only if both parties independently commit to and reveal the same canonical executable interpretation before state disclosure.

For session i, let A_i mark finalized execution, D_i interpretation divergence, L_i^P and L_i^R post-acceptance interpretation shifts by proposer and responder, and Φ_i remaining post-acceptance SOV. The unsafe-execution indicator is

$$Z_i = A_i \mathbf{1}\left[D_i \vee L_i^P \vee L_i^R \vee \Phi_i > 0\right]. \tag{7}$$

Disagreement and blocked ICR mismatch have $A_i = 0$ and are not mislabeled as unsafe execution.

Table 3 summarizes the mitigation results. Clarity prompting reduces unsafe execution by 27.08 points [19.27,34.90] but leaves 6.77% unsafe. Structured offers eliminate observed unsafe execution by removing breadth entirely. ICR preserves broad-clause finalization in 31.25% of all sessions, differs from unrestricted broad-clause agreement by only -2.60 points [-10.94,5.73], and has the same finalization rate as unrestricted negotiation. Of 192 ICR sessions, 140 reach tentative acceptance; six canonical objects mismatch and are blocked; 134 finalize. Zero semantic divergence among final-

ized ICR agreements follows from the protocol invariant within the represented ontology, not from a universal safety guarantee.

9 DISCUSSION

The payoff-matched experiment establishes the paper's central claim. Semantic option value is not merely a theoretical property of clauses. In interacting LLM-agent negotiation, it causally changes which agreements are selected: high-SOV clauses become more likely precisely when the beneficiary controls interpretation and uncertainty resolves after acceptance, even after matching statewise payoff distributions.

The result is a systematic average causal effect across the tested model panel, not a universal law of LLM behavior. GPT and Gemini exploit semantic option value; DeepSeek is inconclusive; Claude reverses sign. This heterogeneity may reflect differences in ambiguity aversion, contract-style preferences, safety tuning, or willingness to propose broad language. The present experiments identify the behavioral pattern, not the internal mechanism producing model-level differences.

The welfare evidence is more limited than the behavioral evidence. The matched experiment supports increased broad-clause agreement and increased semantic option capture. It also shows a decline in responder utility. It does not establish reliable average changes in overall agreement, proposer utility, or social welfare. SOV therefore makes ambiguity strategically operative, but the tested agents do not turn that option into a clean welfare improvement.

The complementary benchmark remains valuable because it shows the same qualitative phenomenon in a larger ecological manipulation and decomposes the effect into proposer offer generation. The mitigation and cross-play analyses show that semantic risk is not only a measurement problem: it can be reduced by interface restrictions or by binding final execution to a shared canonical interpretation. These analyses have scoped implications. Structured offers sacrifice expressive breadth. ICR assumes a correct ontology and canonicalizer, and it guarantees equality only for represented executable objects.

10 CONCLUSION

SOV-Neg turns contingent semantic breadth in agent agreements into a measurable behavioral object. In the payoff-matched confirmatory experiment, high semantic option value increases controller- and timing-sensitive broad-clause agreement by 10.5 percentage points while preserving statewise proposer payoff distributions. The same intervention increases realized option capture and decreases responder utility, without reliable average changes in overall agreement, proposer utility, or social welfare. The effect is systematic at the pooled level but sharply model-dependent, with reliable positive effects for GPT and Gemini, an inconclusive DeepSeek estimate, and a Claude reversal. The complementary benchmark, transcript decomposition, cross-play checks, and mitigation experiments support the same central picture: accepted language can preserve executable options, LLM agents can respond to those options, and negotiation systems need protocols that bind broad natural-language agreement to shared executable meaning before uncertainty turns ambiguity into unilateral control.

11 LIMITATIONS

The benchmark is controlled and synthetic. Every game has two states, two outcomes, additive transfers, predefined clauses, public state probabilities, a six-turn horizon, and a fixed ontology. Human-agent negotiation, multilingual negotiation, long-horizon drafting, richer state spaces, and open-ended legal documents remain untested.

The matched intervention preserves statewise payoff distributions, but it necessarily changes cross-state outcome alignment. That is the intended source of SOV, yet it means the design identifies the effect of introducing a state-contingent preference reversal under fixed statewise payoff moments, not every possible way that semantic option value might arise in real contracts.

Only four model families and one low-temperature configuration were tested. Hosted model snapshots and provider routes can change. The Claude reversal especially argues against model-universal claims. Domain analyses have only ten game clusters and are descriptive. Mitigation inference uses only 12 anchor games. Agreement-conditional or transcript-decomposition analyses are useful diagnostics but are not substitutes for the primary design-based estimand.

ICR's guarantee is scoped. It requires correct canonicalization, a sufficiently complete ontology, verified commitments, and an execution stack that obeys the finalized object. It does not ensure legal validity, truthful preferences, authorization, fairness, identity binding, or downstream compliance.

12 Ethics and Reproducibility

Semantic breadth can preserve legitimate flexibility, but it can also shift contingent risk onto a less informed counterparty. SOV-Neg is intended to support measurement and protocol evaluation, yet the same construct could help an agent recognize strategically valuable ambiguity. We therefore report observable behavior and distributional consequences without inferring deception or intent. The benchmark uses synthetic commercial scenarios, contains no personal data or human-subject negotiation transcripts, and releases validation annotations without identifying metadata.

The artifact contains benchmark instances, matched-payoff generation code, manifests, prompts, model routes, seeds, raw interaction logs, validation reports, analysis code, and file-level hashes. The payoff-matched confirmatory plan was frozen after the exploratory pilot and before confirmatory calls. The final local test suite passed 98/98 checks. Because hosted endpoints may change, the release supports exact replay of recorded traces and non-identical future replication with comparable model routes.

AI Use Statement

We used generative AI tools to (i) draft candidate game scenarios (domain text, clause wording, and state descriptions) under a frozen benchmark blueprint, which constitutes synthetic dataset generation; (ii) implement parts of the experiment harness and analysis code; (iii) draft prose describing analysis outputs; and (iv) edit the manuscript for readability and format references. We did not use generative AI tools to formulate the research questions, the SOV formalism, the payoff-matched design, the estimands, or the pre-registered analysis plan, or to decide how results are interpreted. All AI-drafted scenarios passed automated validators and human validity checks (plausibility, clarity, outcome distinctness, defensible interpretation, singleton precision, and uncertainty). AI-assisted code was covered by a test suite and reviewed by the authors. All reported numbers were regenerated from the released analysis reports. The large language models evaluated in the experiments are objects of study, not tools used to produce this work. We take full responsibility for the content of this submission.

References

Tim Baarslag, Mark J. C. Hendrikx, Koen V. Hindriks, and Catholijn M. Jonker. Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30(5):849–898, 2016. doi: 10.1007/s10458-015-9309-1.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097.

B. Douglas Bernheim and Michael D. Whinston. Incomplete contracts and strategic ambiguity. *American Economic Review*, 88(4):902–932, 1998. URL https://www.jstor.org/stable/117011.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings*

of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 3935–3951. PMLR, 2024. URL https://proceedings.mlr.press/v235/bianchi24a.html.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3167–3185. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.254.

Manan Roy Choudhury, Adithya Chandramouli, Mannan Anand, and Vivek Gupta. Better call CLAUSE: A discrepancy benchmark for auditing LLMs legal reasoning capabilities. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5776–5818. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.305.

Tim R Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *International Conference on Learning Representations*, volume 2024, pp. 784–809, 2024.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. doi: 10.1016/0004-3702(94)00041-X.

Paul Dütting, Michal Feldman, Daniel Peretz, and Larry Samuelson. Ambiguous contracts. *Econometrica*, 92(6):1967–1992, 2024. doi: 10.3982/ECTA22687.

Peyman Faratin, Carles Sierra, and Nick R. Jennings. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3–4):159–182, 1998. doi: 10.1016/S0921-8890(98)00029-3.

Thalya Fossey, Jean-Guy Mailly, and Pavlos Moraitis. Argument-based multi-issue negotiation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 81–89. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/10.

Simon Grant, J. Jude Kline, and John Quiggin. Contracting under uncertainty: A principal–agent model with ambiguity-averse parties. *Games and Economic Behavior*, 109:582–597, 2018. doi: 10.1016/j.geb.2018.02.008.

Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 417–422. Association for Computing Machinery, 2019. doi: 10.1145/3306618.3314250.

Sherzod Hakimov, Roland Bernard, Tim Leiber, Karl Osswald, Kristina Richert, Ruilin Yang, Raffaella Bernardi, and David Schlangen. The price of thought: A multilingual analysis of reasoning, performance, and cost of negotiation in large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 529–570. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.27.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1256.

Ryota Higa, Katsuhide Fujita, Toki Takahashi, Takumu Shimizu, and Shinji Nakadai. Reward-based negotiating agent strategies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11569–11577, 2023. doi: 10.1609/aaai.v37i10.26367.

Nicholas R. Jennings, Peyman Faratin, Alessio R. Lomuscio, Simon Parsons, Carles Sierra, and Michael Wooldridge. Automated negotiation: Prospects, methods and challenges. *Group Decision and Negotiation*, 10(2):199–215, 2001. doi: 10.1023/A:1008746126376.

Catholijn M. Jonker, Reyhan Aydogan, Tim Baarslag, Katsuhide Fujita, Takayuki Ito, and Koen V. Hindriks. Automated negotiating agents competition (ANAC). *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1):5070–5072, 2017. doi: 10.1609/aaai.v31i1.10637.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5391–5413. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.310.

Deuksin Kwon, Jiwon Hae, Emma Clift, Daniel Shamsoddini, Jonathan Gratch, and Gale Lucas. ASTRA: A negotiation agent with adaptive and strategic reasoning via tool-integrated action for dynamic offer optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16217–16238. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.821.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1259.

Raz Lin, Sarit Kraus, Tim Baarslag, Dmytro Tykhonov, Koen Hindriks, and Catholijn M. Jonker. GENIUS: An integrated environment for supporting the design of generic automated negotiators. *Computational Intelligence*, 30(1):48–70, 2014. doi: 10.1111/j.1467-8640.2012.00463.x.

Jihwan Oh, Murad Aghazada, Yooju Shin, Se-Young Yun, and Taehyeon Kim. MERIT feedback elicits better bargaining in LLM negotiators. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25202–25223. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1155.

Iyad Rahwan, Sarvapali D. Ramchurn, Nicholas R. Jennings, Peter McBurney, Simon Parsons, and Liz Sonenberg. Argumentation-based negotiation. *The Knowledge Engineering Review*, 18(4): 343–375, 2003. doi: 10.1017/S0269888904000098.

Jeffrey S. Rosenschein and Gilad Zlotkin. *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*. MIT Press, Cambridge, Massachusetts, 1994. ISBN 9780262181594.

Sarath Sanga. Incomplete contracts: An empirical approach. *The Journal of Law, Economics, and Organization*, 34(4):650–679, 2018. doi: 10.1093/jleo/ewy012.

Carles Sierra, Nick R Jennings, Pablo Noriega, and Simon Parsons. A framework for argumentation-based negotiation. In *International workshop on agent theories, architectures, and languages*, pp. 177–192. Springer, 1997.

Reid G Smith. The contract net protocol: High-level communication and control in a distributed problem solver. In *Readings in distributed artificial intelligence*, pp. 357–366. Elsevier, 1988.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3579–3602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.213.

Julian Yocum, Phillip J. K. Christoffersen, Mehul Damani, Justin Svegliato, Dylan Hadfield-Menell, and Stuart Russell. Mitigating generative agent social dilemmas. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023. URL https://openreview.net/forum?id=5TIdOk7XQ6.

This supplement reports the payoff-matched benchmark construction, confirmatory design, complete matched-payoff results, robustness checks, supporting analyses from a complementary SOV-Neg benchmark, mitigation diagnostics, cross-play comparisons, and reproducibility information for *When "Agreed" Is Not Agreement: Semantic Option Value in LLM-Agent Negotiation*. The main paper contains the claim-defining analyses. This appendix provides expanded definitions, additional tables, and implementation details.

A A. Appendix Map

Table A1: Organization of the supplementary appendix.

Section	Analysis	Role in the paper
A	Appendix map	Shows how the supplement is organized.
B	Payoff-matched benchmark	Defines the new high-SOV and two matched zero-SOV controls.
C	Confirmatory design	Documents the frozen 4,800-session matched-payoff experiment.
D	Matched-payoff estimands and outcomes	States the primary DDD, secondary outcomes, and inference procedure.
E	Matched-payoff results	Provides primary, robustness, model, domain, and secondary results.
F	Complementary SOV-Neg benchmark	Reports the 8,320-session ecological manipulation as convergent evidence.
G	Mitigation study	Reports clarity prompting, singleton structured offers, and ICR diagnostics.
H	Heterogeneous cross-play	Reports different-model pairings from the complementary benchmark.
I	Statistical procedure	Details weighting, bootstrap, and sensitivity conventions.
J	Data and reproducibility	Lists artifacts, runtime audit, and exact replay scope.
K	Scope of inference	Makes inference boundaries explicit.

B B. Payoff-Matched Benchmark Construction

B.1 B.1 Formal Template

Every base game contains:

- two publicly described future states with probabilities summing to one;
- two operationally distinct executable outcomes;
- two precise singleton clauses, one for each outcome;
- one broad clause whose interpretation set contains both outcomes;
- a transfer grid of $\{-2, -1, 0, 1, 2\}$;
- one high-SOV proposer utility table;
- two payoff-matched zero-SOV proposer utility tables;
- a responder utility table constructed to have zero semantic option value; and
- twelve matched derivatives from SOV condition $\times$ control $\times$ timing.

The public language, state probabilities, outcomes, clauses, and transfer grid are fixed within a base game. The SOV condition changes only the proposer utility table. Control changes only which role chooses a valid execution of a broad clause. Timing changes only whether the realized state is disclosed before or after tentative acceptance.

B.2 B.2 Matched Zero-SOV Construction

For each two-state, two-outcome game, the matched intervention begins from the source high-SOV proposer utility matrix

$$U = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

The high-SOV condition has a cross-state preference reversal: one outcome is better in the first state and the other is better in the second. We construct two zero-SOV controls:

$$U^{(0A)} = \begin{bmatrix} \max(a,b) & \min(a,b) \\ \max(c,d) & \min(c,d) \end{bmatrix}, \quad U^{(0B)} = \begin{bmatrix} \min(a,b) & \max(a,b) \\ \min(c,d) & \max(c,d) \end{bmatrix}.$$

Both controls preserve the statewise utility multiset. Therefore, for each state they preserve the minimum, maximum, range, mean, variance, and adaptive value. They remove SOV by making the same outcome dominant in both states. Including both controls removes outcome-label dependence.

B.3 B.3 Payoff-Geometry Matching

The complementary high-versus-zero SOV intervention held wording fixed, but changed more of the proposer payoff geometry than semantic option value alone. An audit of the complementary benchmark found changes in best fixed value, adaptive value, payoff range, dispersion, and often the identity of the best fixed outcome. The matched intervention directly targets this concern by preserving statewise payoff distributions while varying only cross-state alignment. The remaining treatment difference is the state-contingent reversal that creates or removes option value.

B.4 B.4 Domain Coverage

The benchmark contains 100 base games across ten domains: cloud computing, commercial real estate, data licensing, employment agreements, energy services, freight risk, industrial maintenance, insurance claims, procurement, and software warranty. Each domain contributes ten games. The domains vary the governed object, future contingency, operational mechanism, party role, and direction of risk while preserving the same two-state, two-outcome formal structure.

C C. Matched-Payoff Confirmatory Experiment

C.1 C.1 Frozen Design

The matched-payoff confirmatory design was frozen after a separate 192-session exploratory pilot and before any confirmatory calls. The confirmatory experiment contains:

- 100 base games across ten domains;
- three SOV conditions: high SOV, matched-zero-first, and matched-zero-second;
- two interpretation controllers: proposer and responder;
- two revelation timings: before acceptance and after acceptance;
- four model families; and
- one repetition per model-condition.

This yields $100 \times 3 \times 2 \times 2 \times 4 = 4,800$ sessions. Within each condition, two model families are assigned to each realized state.

C.2 C.2 Models and Routes

The matched-payoff experiment used four self-play model routes through OpenRouter:

Table A2: Model routes in the matched-payoff confirmatory experiment.

Manuscript label	Provider family	Model route label	Notes
GPT-5.6 Sol	OpenAI	gpt56sol	Frozen route used for confirmatory collection.
Claude Sonnet 5	Anthropic	claude_sonnet5	Same negotiation protocol and output schema.
Gemini 3.1 Pro	Google	gemini31pro	Same negotiation protocol and output schema.
DeepSeek V4 Pro	DeepSeek	deepseek_v4pro	Pinned StreamLake route used for confirmatory collection.

The model-route labels identify the hosted endpoints used during confirmatory collection. Hosted endpoints can change over time, so future runs with comparable model labels are replications rather than bit-identical reruns.

C.3 C.3 Runtime Audit

All 4,800 scheduled confirmatory sessions completed. The final outcome statuses were:

Table A3: Matched-payoff confirmatory session statuses.

Status	Sessions
Executed	3,908
Turn limit	879
Protocol failure	13
Total	4,800

The run used 36,734 successful model calls, 26,572,934 input tokens, and 5,370,982 output tokens. The recorded successful-call cost was USD 105.571773. During collection there were 23 transient failed attempts, including timeouts, missing completion choices, and one malformed action or type error. Each affected session subsequently completed and the failure logs are retained.

D D. Matched-Payoff Outcomes and Estimands

D.1 D.1 Primary Outcome

The primary outcome is unconditional broad-clause agreement. It is coded one when a finalized agreement uses the broad clause and zero when the session ends in a precise agreement, disagreement, turn limit, or protocol failure. This intention-to-treat coding preserves the practical meaning of the negotiation outcome.

D.2 D.2 Primary Estimand

Let μ_{sct} be the equally weighted mean broad-clause agreement rate for SOV condition s, controller c, and timing t, after averaging over game and model cells. Let H denote high SOV, Z_1 and Z_2 the two matched zero-SOV controls, and

$$\mu_{Zct} = \frac{1}{2}(\mu_{Z_1ct} + \mu_{Z_2ct}).$$

The primary matched-payoff estimand is

$$\begin{aligned}\mathrm{DDD}_{\mathrm{matched}} = &[(\mu_{HPa} - \mu_{ZPa}) - (\mu_{HPb} - \mu_{ZPb})] \\ &- [(\mu_{HRa} - \mu_{ZRa}) - (\mu_{HRb} - \mu_{ZRb})]\,,\end{aligned}$$

where P is proposer control, R is responder control, a is after-acceptance revelation, and b is before-acceptance revelation.

D.3 D.3 Secondary Outcomes

Secondary outcomes are:

- overall finalized agreement;
- proposer utility under intention-to-treat coding;
- responder utility under intention-to-treat coding;
- social welfare, defined as proposer plus responder utility; and
- semantic option capture, defined as realized state-contingent gain relative to the controller's best fixed interpretation.

D.4 D.4 Inference

Confidence intervals use 10,000 base-game cluster bootstrap replicates. The effective primary sampling units are the 100 games, not the 4,800 sessions. Domain-specific estimates contain only ten game clusters and are treated as descriptive.

E E. Matched-Payoff Confirmatory Results

E.1 E.1 Primary and Secondary Results

Table A4: Matched-payoff primary and secondary DDDs across 4,800 confirmatory sessions. Probability effects are reported in percentage points. Utility and option-capture effects are benchmark utility points. Intervals are 95% base-game cluster bootstrap intervals.

Outcome	Estimate	95% CI
Broad-clause agreement	+10.50 pp	[3.75, 17.38]
Agreement	-1.75 pp	[-7.25, 3.75]
Proposer utility	+0.4088	[-0.2375, 1.0638]
Responder utility	-0.4825	[-0.9113, -0.0538]
Social welfare	-0.0738	[-0.9438, 0.8075]
Semantic option capture	+1.1638	[0.7863, 1.5388]

The primary broad-clause agreement DDD is +10.50 percentage points, with a 95% interval excluding zero. Thus, after matching statewise payoff distributions, high semantic option value increases the controller-and-timing-sensitive selection of broad language. The matched experiment also shows a reliable increase in semantic option capture and a reliable decline in responder utility. It does not establish a reliable average effect on overall agreement, proposer utility, or social welfare.

E.2 E.2 Robustness to Matched Zero-SOV Construction

Table A5: Primary broad-clause DDD by matched zero-SOV control.

Comparison	Estimate	95% CI
High SOV vs. matched-zero-first	+9.75 pp	[2.00, 17.50]
High SOV vs. matched-zero-second	+11.25 pp	[3.00, 19.50]
High SOV vs. pooled matched-zero	+10.50 pp	[3.75, 17.38]

Both zero-SOV variants support the same conclusion. This matters because each variant makes a different outcome label dominant across states.

E.3 E.3 Protocol-Failure Sensitivity

The primary intention-to-treat result includes the 13 protocol-failure outcomes as non-broad agreements. Excluding all protocol failures yields a primary DDD of +10.55 percentage points with a 95% interval of [3.80,17.50]. The exclusion therefore does not change the result.

E.4 E.4 Model Heterogeneity

Table A6: Model-specific matched-payoff broad-clause agreement DDDs. Intervals are 95% base-game cluster bootstrap intervals over the same 100 games.

Model	Estimate	95% CI
GPT-5.6 Sol	+28.00 pp	[14.00, 42.00]
Gemini 3.1 Pro	+22.50 pp	[5.50, 39.00]
DeepSeek V4 Pro	+6.00 pp	[-9.50, 22.00]
Claude Sonnet 5	-14.50 pp	[-25.00, -4.50]

The pooled effect is not universal across model families. GPT-5.6 Sol and Gemini 3.1 Pro show reliable positive effects. DeepSeek V4 Pro is directionally positive but inconclusive. Claude Sonnet 5 reliably reverses sign. The manuscript therefore claims an average incentive-responsive ambiguity effect with strong model heterogeneity, not a universal property of all LLM agents.

E.5 E.5 Domain-Specific Results

Table A7: Domain-specific primary broad-clause DDDs in the matched-payoff experiment. Each domain contains ten game clusters, so these estimates are descriptive.

Domain	Estimate	95% CI
Cloud computing	+12.50 pp	[-16.25, 40.00]
Commercial real estate	+7.50 pp	[-11.25, 26.25]
Data licensing	-8.75 pp	[-26.25, 7.50]
Employment agreements	+27.50 pp	[-2.50, 56.25]
Energy services	+27.50 pp	[12.50, 43.75]
Freight risk	+11.25 pp	[-6.25, 28.75]
Industrial maintenance	+5.00 pp	[-6.25, 16.25]
Insurance claims	+18.75 pp	[-6.25, 40.00]
Procurement	+8.75 pp	[-6.25, 22.50]
Software warranty	-5.00 pp	[-25.00, 12.50]

Eight of ten domain point estimates are positive. Only energy services has a domain-level interval excluding zero. Because each domain contains only ten game clusters and no multiplicity-adjusted domain inference was pre-specified, the domain results are descriptive.

F F. Complementary SOV-Neg Benchmark as Convergent Evidence

F.1 F.1 Design Role

The complementary SOV-Neg benchmark contains 8,320 causal sessions across 100 games, four models, high-versus-zero SOV regimes, interpretation control, and revelation timing. It held wording fixed, but its high-versus-zero SOV utility intervention changed broader payoff geometry. It is therefore used as supporting evidence that the same qualitative phenomenon appears in a larger ecological manipulation, while the payoff-matched design supplies the stricter identification result.

F.2 F.2 Primary Results

Table A8: Complementary 8,320-session SOV-Neg design-based contrasts. These results are convergent evidence for the matched-payoff experiment.

Contrast	Estimate	95% CI
Broad-clause DDD	+12.31 pp	[7.25, 17.56]
H1: target value response	+24.94 pp	[21.50, 28.38]
H2: timing attenuation	+25.50 pp	[21.75, 29.19]
H3: control attenuation	+14.94 pp	[10.88, 19.00]
Proposer-utility DDD	-0.144	[-0.598, 0.313]
Semantic-option-capture DDD	+0.583	[0.373, 0.788]

F.3 F.3 Proposal-Stage Decomposition

The transcript decomposition localizes the interaction in proposer offer generation. The DDD is +30.12 points [25.69,34.56] for the proposer's first broad offer and +16.00 [9.81,22.12] for any proposer broad offer, but -0.69 [-4.94,3.50] for any responder broad offer. Decomposing the finalized endpoint by offer origin gives +16.31 [11.69,21.00] for finalized broad offers made by the proposer and -4.00 [-6.06,-2.00] for finalized broad offers made by the responder. These two components sum to the +12.31-point primary DDD in the complementary benchmark.

F.4 F.4 Model Heterogeneity

Table A9: Complementary benchmark model-specific broad-clause DDDs.

Model	Estimate	95% CI
GPT route	+28.00 pp	[18.00, 38.00]
Gemini route	+26.00 pp	[15.25, 37.25]
DeepSeek route	+2.50 pp	[-9.50, 14.25]
Claude route	-7.25 pp	[-16.25, 1.26]

The complementary benchmark and the matched-payoff experiment agree on the main heterogeneity pattern: reliable positive effects are concentrated in GPT and Gemini routes, while DeepSeek is weak or inconclusive and Claude does not show the same positive effect.

F.5 F.5 Utility-Geometry Audit

The high-SOV regime in the complementary benchmark changed more than scalar option value. The audit across the 100 games found that moving from zero to high SOV lowered mean best-fixed value by 2.133 points, raised adaptive value by 0.602, widened payoff range by 2.510, and changed the best fixed outcome in 47 games. This audit motivates the payoff-matched experiment and the use of the matched design as the primary causal analysis.

G G. Mitigation Study

G.1 G.1 Protocols

The mitigation study evaluates four protocols on 12 outcome-blind anchor games:

- *Unrestricted*: the target negotiation protocol.
- *Clarity prompt*: agents are asked to fix one execution, but the protocol does not enforce compliance.
- *Structured offer*: only singleton clauses are exposed, mechanically removing broad language.
- *Interpretation commit-and-reveal* (ICR): broad language remains available, but finalization requires both parties to commit to and reveal the same canonical executable object before state disclosure.

G.2 G.2 Absolute Outcomes

Table A10: Absolute protocol outcomes in the 768-session mitigation study, with 192 sessions per protocol.

Protocol	Final	Unsafe	Broad	Welfare
Unrestricted	69.8%	33.9%	33.9%	9.90
Clarity prompt	70.8%	6.8%	6.8%	9.82
Structured offer	78.6%	0.0%	0.0%	10.74
ICR	69.8%	0.0%	31.3%	9.31

Clarity prompting reduces but does not eliminate unsafe execution. Structured offers eliminate semantic flexibility by removing broad clauses. ICR preserves broad language while making matching canonical interpretations a finalization condition.

G.3 G.3 ICR Diagnostics

Of 192 ICR sessions, 140 reach tentative acceptance; six canonical objects mismatch and are blocked; 134 finalize. The observed unconditional finalization rate is 69.79%, and mismatch conditional on tentative acceptance is 4.29%. Model-specific ICR finalization is 29.17% for Claude, 60.42% for GPT, 93.75% for DeepSeek, and 95.83% for Gemini. This is a usability limitation, not a failure of the finalization invariant. Within the represented ontology, finalized ICR agreements commit both parties to one executable interpretation.

H H. Heterogeneous Cross-Play

The cross-play study pairs the four model families in all six unordered combinations and reuses 12 outcome-blind anchors under unrestricted negotiation. Four scheduled runs per game-pair cover two role directions and two states, yielding 288 sessions. Each cross-play observation is matched to the corresponding self-play mean by game, factorial condition, realized state, role, and model. These are external-generalization comparisons, not a second matched-payoff DDD.

Table A11: Cross-play relative to matched self-play in the complementary benchmark.

Outcome	Cross-play minus matched self-play	95% CI
Agreement	+5.03 pp	[1.74, 8.51]
Broad-clause agreement	+7.81 pp	[2.26, 13.37]
Proposer utility	+0.382	[0.057, 0.710]
Responder utility	+0.325	[-0.031, 0.696]
Social welfare	+0.707	[0.132, 1.335]
Turns	-0.188	[-0.316, -0.066]

The broad-clause increase under cross-play indicates that the behavior is not limited to identical-model linguistic conventions. Because the design is not factorial within each game, it does not identify a new SOV DDD.

I I. Statistical Procedure

I.1 I.1 Weighting

The matched-payoff analysis averages outcomes at the game-condition-model level and then gives each base game equal weight. The model-specific analyses use the same 100 game clusters within each model. The pooled analysis gives model families equal weight through the balanced design.

I.2 I.2 Bootstrap

All matched-payoff confidence intervals use a 10,000-replicate base-game cluster bootstrap. The bootstrap resamples games rather than sessions, because games are the intended sampling units and all conditions are crossed within game.

I.3 I.3 Interpretation of Intervals

The primary interval is claim-defining. Domain intervals are descriptive because each domain has only ten clusters and domain-specific inference was not multiplicity-adjusted. Model intervals are reported because model heterogeneity is central to the claim.

J J. Data and Computational Reproducibility

The matched-payoff artifact contains:

- matched-payoff instance generation code;
- the frozen exploratory pilot manifest;
- the frozen confirmatory plan;
- the frozen confirmatory manifest;
- model-route configuration for the matched experiment;
- raw session logs and summaries;
- session-level analysis tables;
- machine-readable analysis reports; and
- tests for matched-payoff construction, pilot generation, and confirmatory generation.

The key files are:

- `configs/matched_payoff_confirmatory_plan.json`
- `configs/experiment_matched_payoff_confirmatory.json`
- `configs/matched_payoff_models.json`
- `analysis/matched_payoff_confirmatory-1.0/report.json`
- `analysis/matched_payoff_confirmatory-1.0/session_records.csv`
- `analysis/matched_payoff_confirmatory-1.0/SUMMARY.md`

The final local test suite passed 98/98 checks. Because hosted endpoints may change, the release supports exact replay of recorded traces and non-identical future replication with comparable routes.

K K. Scope of Inference

K.1 K.1 Scope

The principal claim concerns the payoff-matched confirmatory design: high semantic option value increases the controller-and-timing-sensitive selection of broad clauses by 10.5 percentage points on average across the tested model panel. The claim is not universal across all LLMs, all negotiation settings, or all contractual language. The Claude reversal and DeepSeek null are part of the reported result.

K.2 K.2 Welfare Scope

The matched experiment supports increased semantic option capture and decreased responder utility. It does not establish a reliable average change in overall agreement, proposer utility, or social welfare. The supported welfare conclusion is therefore asymmetric: semantic option value changes agreement content and realized option capture, but does not imply an average welfare or proposer-utility improvement in this design.

K.3 K.3 Relationship Between the Two Causal Designs

The two causal designs answer related but distinct questions. The 8,320-session benchmark estimates the interaction of an SOV-generating utility regime with interpretation control and information timing under fixed wording. The 4,800-session payoff-matched benchmark estimates the same controller-and-timing-sensitive behavior after preserving statewise payoff distributions and varying only cross-state outcome alignment. The latter is therefore the more targeted identification design, while the former provides broader ecological support.

K.4 K.4 Protocol Scope

ICR assumes correct canonicalization, a sufficiently complete ontology, verified commitments, and an execution stack that obeys the finalized object. It does not ensure legal validity, truthful preferences, authorization, fairness, identity binding, or downstream compliance.