Evaluating Rationality in Natural-Language Contracting

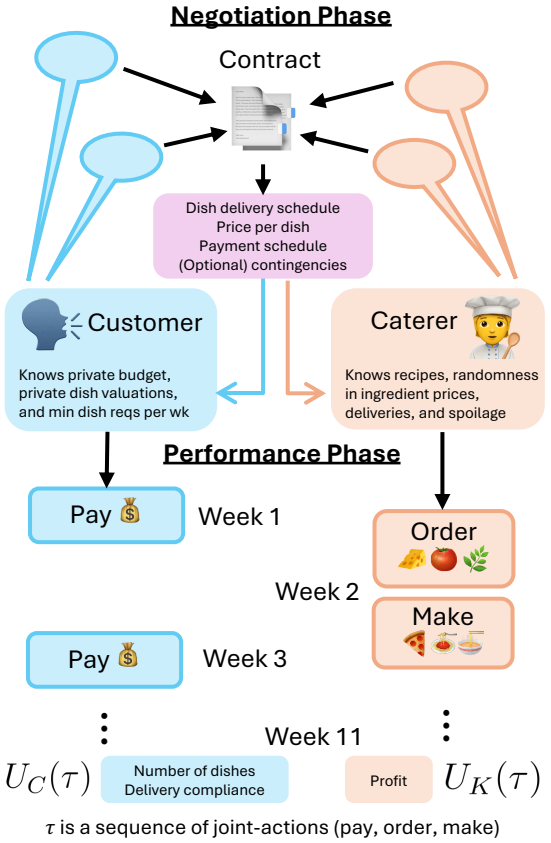
Bhavyesh Sajja (Mentors: Dr Max Kleiman-Weiner, Dr Tan Zhi-Xuan; Manager: Dr Sam Brown)

Who am I?

AI Agents are entering the economy, but their ability to negotiate contracts and perform them (**"rational contracting"**) is unclear.

What is the theory of rational contracting, and are modern AI agents "rational"?

Catering Contract



τ is a sequence of joint-actions (pay, order, make)

Methodology

Evaluation Metrics

What are agents maximizing?

Contract $\theta = (q, p, m)$; delivery schedule, prices, payments

Customer's Preference: Reliability first, then maximize value

$$U_C(\tau) = \mathbf{1}[\underbrace{\text{del} \geq q}_{\text{demand met}}](\underbrace{M}_{\text{budget}} + \underbrace{\sum_{w,d} v_d \, \text{del}_{w,d}}_{\text{private value}} - \underbrace{\sum_w m_w}_{\text{total paid}})$$

Caterer's Preference: Maximize profit

$$U_K(\tau) = \underbrace{\sum_w m_w}_{\text{revenue}} - \underbrace{\sum_{w,i} b_{w,i} \cdot p_{w,i}}_{\text{ingredient cost}}$$

Negotiation Phase

Are agents negotiating the optimal contract?

Optimal contract – maximizing social welfare

$$W = \alpha \cdot U_C(\tau) + (1-\alpha) \cdot U_K(\tau), \alpha \in [0,1]$$

Hill-climbing search over a contract space with varying α creates a convex hull over the Pareto frontier (upper-bound)

Performance Phase

How are agents performing the contract in the environment?

Baseline: **Rational Conditional Complier - RCC** (best-play); adheres to contract, presses grim-trigger upon violation

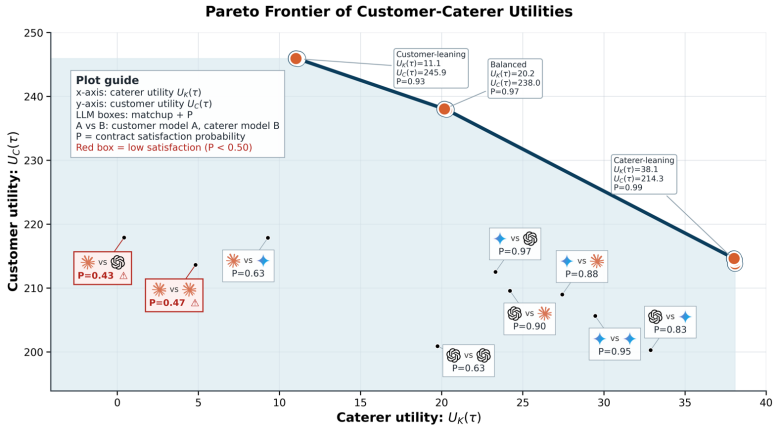
$$\pi_K^* = \arg\max_{\pi_K} \mathbb{E}\left[U_K(\tau) \mid \theta, \mathbf{m}, \pi_K\right]$$
$$\text{s.t.} \quad P(\text{meet demand} | \theta, \pi_K^*) \geq 1 - \epsilon$$

Measures

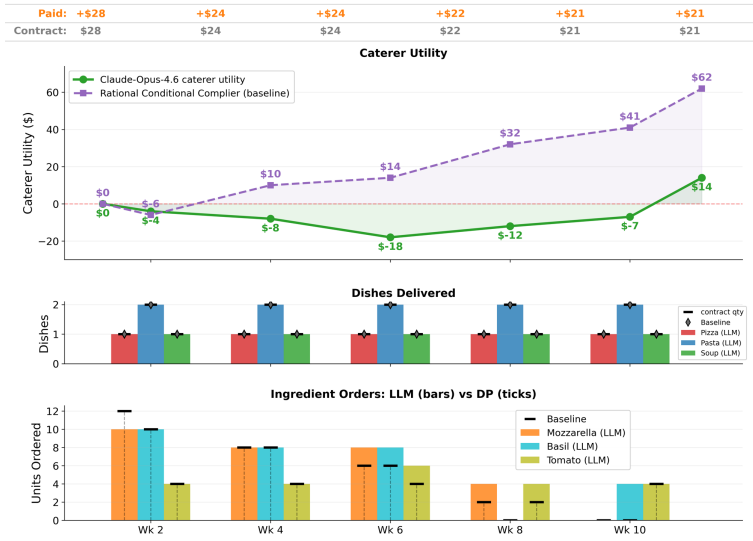
- **Customer Utility:** Cumulative private value (M = 200)
- **Customer Delivery Shortfall:** Number of failed deliveries
- **Caterer Utility:** Profit
- **Caterer Spend Regret:** How much did they overspend on ingredients?

Results and Conclusions

Negotiation Phase: Agents negotiate contracts of different qualities. Opus is a ruthless customer; note the low caterer utility and odds of contract satisfaction. Gemini as a customer seeks reliability

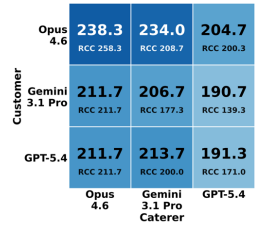


Single-run analysis: Opus as a caterer satisfies delivery requirements but overspends on ingredients. Compared to the baseline, lots of value left at the table. Now, imagine an agentic economy with agents like so!

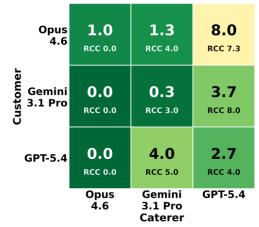


Performance Measures

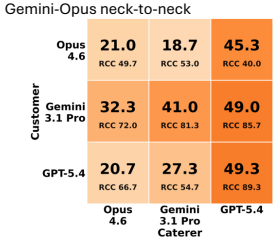
Customer Utility: Opus gets most value; GPT as a caterer makes customer lose money



Customer Delivery Shortfalls: GPT as a caterer is unreliable; Opus is steady



Caterer Utility: GPT is the richest (deceptive!); Gemini-Opus neck-to-neck



Caterer Spend Regret (vs RCC): All LLMs overbuy ingredients; Gemini is economical