



# Evaluating Rationality in Natural-Language Contracting

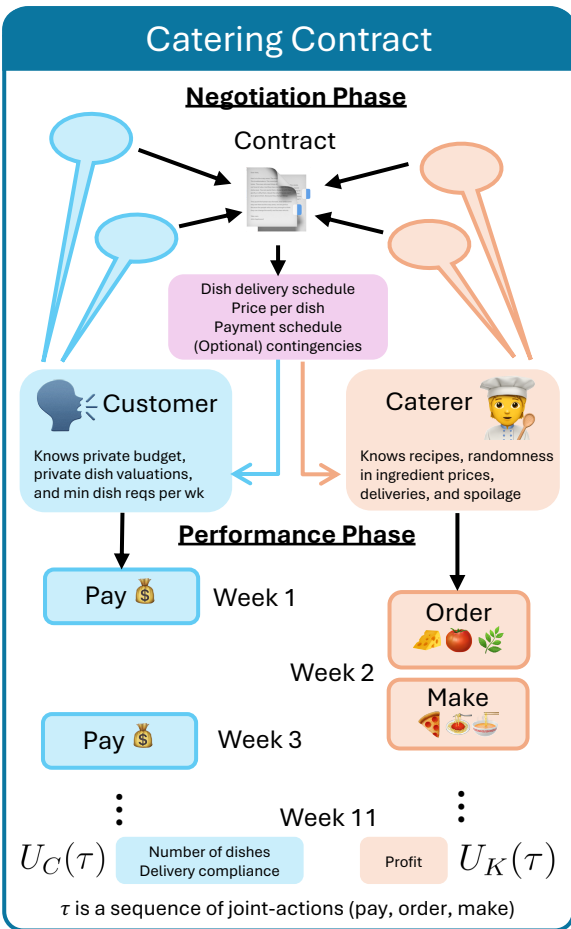
Bhavyesh Sajja (Mentors: Dr Max Kleiman-Weiner, Dr Tan Zhi-Xuan; Manager: Dr Sam Brown)



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”?



## Methodology

### Evaluation Metrics

What are agents maximizing?

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

Customer's Preference: Reliability first, then maximize value

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

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  $\alpha$  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

$$\pi_K^* = \arg \max_{\pi_K} \mathbb{E}[U_K(\tau) | \theta, \mathbf{m}, \pi_K]$$

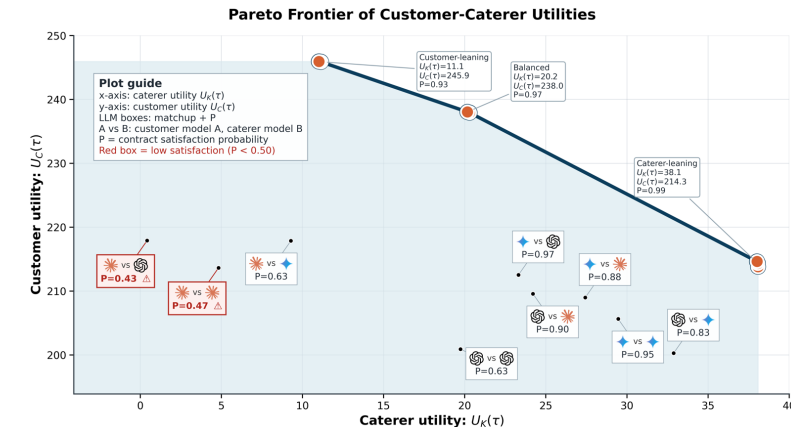
s.t.  $P(\text{meet demand} | \theta, \pi_K^*) \geq 1 - \epsilon$

### 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

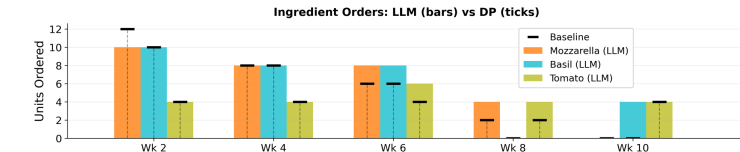
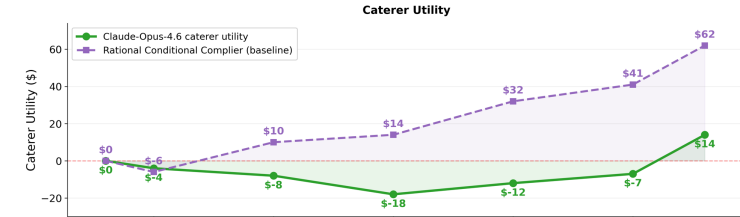
**Negotiation Phase:** Agents negotiate contracts of different qualities. Opus is a ruthless customer; note the low caterer utility and odds of contract satisfaction. Gemini as a customer seeks reliability



**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!

GPT-5.4 (customer) vs Claude-Opus-4.6 (caterer) – Performance Overview

Paid: +\$28	+\$24	+\$24	+\$22	+\$21	+\$21
Contract: \$28	\$24	\$24	\$22	\$21	\$21



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

Customer	Opus 4.6	238.3 RCC 258.3	234.0 RCC 208.7	204.7 RCC 200.3
	Gemini 3.1 Pro	211.7 RCC 211.7	206.7 RCC 177.3	190.7 RCC 139.3
	GPT-5.4	211.7 RCC 211.7	213.7 RCC 200.0	191.3 RCC 171.0
	Opus 4.6	Gemini 3.1 Pro	GPT-5.4	

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

Customer	Opus 4.6	1.0 RCC 0.0	1.3 RCC 4.0	8.0 RCC 7.3
	Gemini 3.1 Pro	0.0 RCC 0.0	0.3 RCC 3.0	3.7 RCC 8.0
	GPT-5.4	0.0 RCC 0.0	4.0 RCC 5.0	2.7 RCC 4.0
	Opus 4.6	Gemini 3.1 Pro	GPT-5.4	

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

Customer	Opus 4.6	21.0 RCC 49.7	18.7 RCC 53.0	45.3 RCC 40.0
	Gemini 3.1 Pro	32.3 RCC 72.0	41.0 RCC 81.3	49.0 RCC 85.7
	GPT-5.4	20.7 RCC 66.7	27.3 RCC 54.7	49.3 RCC 89.3
	Opus 4.6	Gemini 3.1 Pro	GPT-5.4	

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

Customer	Opus 4.6	\$29	\$34	-\$5
	Gemini 3.1 Pro	\$40	\$40	\$37
	GPT-5.4	\$46	\$27	\$40
	Opus 4.6	Gemini 3.1 Pro	GPT-5.4	

Performance Measures