EVAL · BILATERAL PRICE NEGOTIATION

TERMS-Bench

A diagnostic benchmark for LLM negotiation agents.

Erica Zhang, Fangzhao Zhang, Aneesh Pappu, Batu El,
Jose Blanchet, Susan Athey, Jiashuo Liu, James Zou

Presentation for DeepMind Gemma Team
Open-model benchmark

Stanford University
Human-Centered
Artificial Intelligence

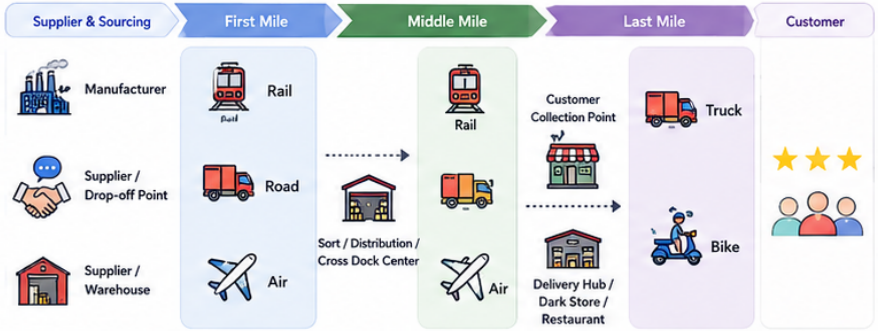
Stanford
ENGINEERING

Gemma | DeepMind

Why Negotiation Matters

Practitioner motivation and evaluation challenge for TERMS-Bench

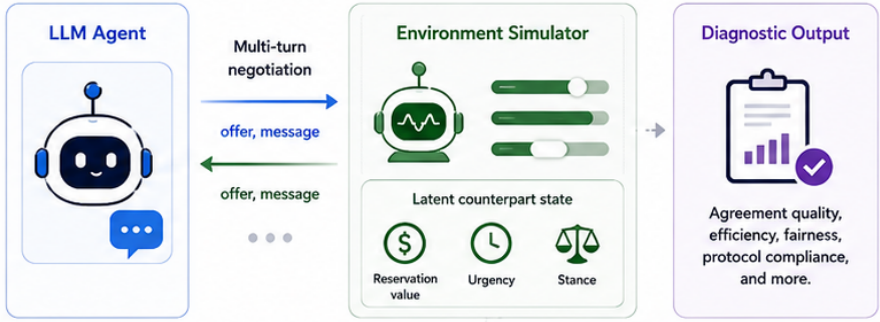
Negotiation is central to real-world economic exchange



Negotiation shapes procurement, contracting, pricing, logistics, labor agreements, and resource allocation. In commercial workflows such as spot freight and middle-mile operations, human negotiators still play an important role.

Procurement · Pricing · Logistics · Resource allocation

A compact testbed for agentic language models



Multi-turn interaction · Hidden preferences · Strategic communication · Binding constraints

Negotiation requires reasoning over latent counterpart state while coordinating offers, language, and constraints across multiple rounds.

Why evaluation is methodologically hard

Intrinsic verifiers

Math · Code

Automatic correctness signals

→

Negotiation

Semi-verifiable: outcomes are objective, but multi-turn reasoning and language strategy are not.

→

Open-ended generation

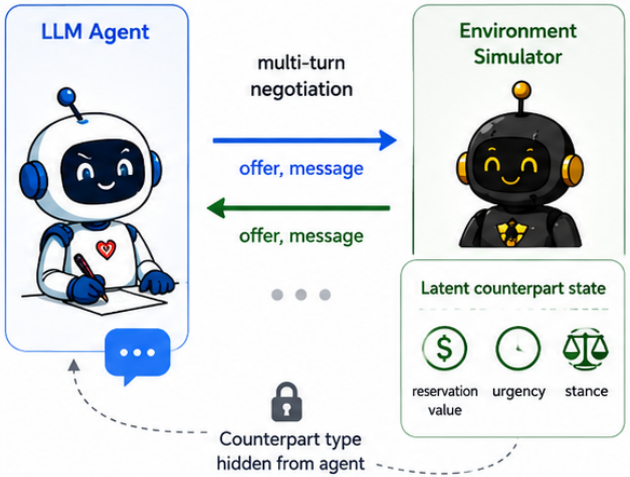
Human or LLM-as-a-judge proxies

Negotiation sits in the middle: objective outcomes exist, but the verifier must be constructed, and that determines what failures can be diagnosed.

TERMS-Bench: A Controlled Negotiation Testbed

Constructing the verifier through fixed regimes and counterpart families

1. Episode protocol



Agent negotiates over price and language, observing only public history while inferring the counterpart's latent state.

2 Three regimes

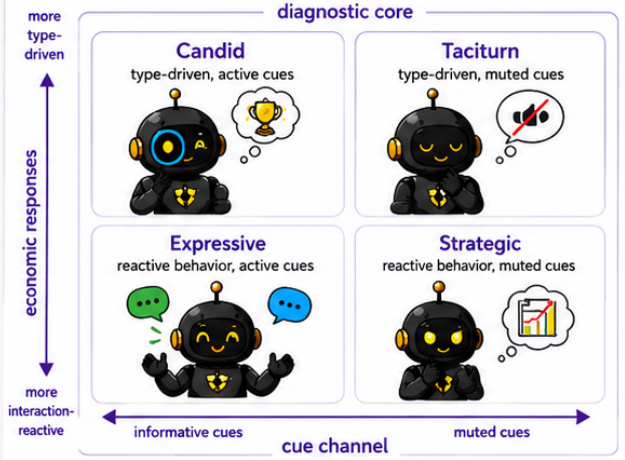
Overlap
Feasible bargaining
Tests agreement and surplus extraction when a deal is possible.

Urgency-shift
Hidden time pressure
Tests adaptation when urgency changes bargaining behavior.

No-deal
Infeasible episodes
Tests exit discipline when no individually rational agreement exists.

Regime draw sets reservation geometry and urgency conditions.

3 3. Six counterpart families



Fixed simulator + hidden types + controlled regimes = diagnostic evaluation

Because the regime generator and counterpart policy are held fixed across agents, performance differences can be attributed to the agent policy.

What We Measure: Six Diagnostic Metrics

*From logged traces and simulator ground truth, we compute **six orthogonal metrics**—no human or LLM judge.*

How diagnostics are computed

Logged trace
prices, messages, actions, timestamps

Environment ground truth
latent type (r, κ, η), utilities, regime

Metric computation
programmatic metrics from trace + ground truth

Orthogonal diagnostics
no composite score, no LLM-as-a-judge

1 Surplus Efficiency (SE↑)

How much value the agent extracts when a deal is actually possible.

Measures
- Normalized concluded surplus relative to optimal
- Captures value capture, not just agreement

0 — 1
higher is better

2 Feasible Agreement Rate (AGR↑)

How often the agent reaches agreement when a deal should be possible.

Measures
- Agreement rate in feasible (overlap / urgency-shift) cases
- Captures ability to close when value exists

0 — 1
higher is better

3 Conditional Deal Quality (CSE↑)

How good the deal is given that an agreement is reached.

Measures
- Normalized surplus conditional on agreement
- Rewards leaving more value on the table

0 — 1
higher is better

4 No-deal False Agreement Rate (FAGR↓)

How often the agent agrees when no deal should exist (irrational agreements).

Measures
- Rate of agreements in no-deal episodes
- Penalizes failure to walk away

0 — 1
lower is better

5 Opponent Modeling Error (BE_{type}↓)

How accurately the agent infers the opponent's hidden latent type.

Measures
- Distance between inferred and true latent type (r, κ, η)
- Reflects belief accuracy from behavior + cues

0 — 1
lower is better

6 Critical Protocol Violations (CritViol%↓)

How often the agent violates critical protocol rules.

Measures
- % of turns with hard rule violations
- Includes price bounds, reservation, time, etc.

0% — 100%
lower is better

These six metrics diagnose distinct capabilities—from value extraction to belief accuracy and rule compliance.
Together they provide **actionable insight** for improving negotiation agents.

Gemma 4 31B delivers frontier-level negotiation at commodity cost

On the new bankroll leaderboard run, Gemma finishes #2 overall, statistically neck-and-neck with Claude Opus 4.6, while the measured base-suite cost is about 71× lower.

Cost edge

71x cheaper per evaluated episode than Claude Opus 4.6

$0.023 Gemma vs $1.647 Claude Opus 4.6

New bankroll leaderboard

10 sessions, 10-period horizon, $1,000 start

1. **Claude Opus 4.6** — $1,187
   #1, terminal balance mean, SEM $12.75
2. **Gemma 4 31B** — $1,183
   #2, terminal balance mean, SEM $12.04
3. **Grok 4.20** — $1,181
   #3, Gemma also matches Claude on survival and agreements

Full base suite validates the signal

1,800 episodes across regimes and counterpart families

Overall SE+	Agreement rate	Critical violations	%Oracle
#4	99.8%	0.06%	69.9
behind only Claude 4.6, GLM 5.1, Claude 4.7	1198 / 1200 agreed in the core suite	only 1 violation in 1,800 episodes	Claude Opus 4.6 reaches 76.3 (upper bound); GPT-4o-mini provides a lower reference at 22.2.

Trace-level behavior is calm and economically coherent: in a curated success trace, Gemma captured 38.9 surplus over five rounds with zero bound, monotonicity, turn-budget, or IR violations.

Takeaway: Gemma combines near-frontier negotiation quality with dramatically lower evaluation cost.

The bankroll result is limited but directionally strong; the larger base suite provides durability evidence.

Demo / Leaderboard