Can LLMs Simulate Economic Agents? Testing Production Theory with GPT-Based Firm Behavior

Yijiashun QI *, Hanzhe Guo , Yijiazhen Qi

Posted Date: 26 February 2026

doi: 10.20944/preprints202602.1648.v1

Keywords: large language models; agent-based simulation; economic agents; production theory; technology adoption; computational economics

 doi:10.20944/preprints202602.1648.v1

Article

Can LLMs Simulate Economic Agents? Testing Production Theory with GPT-Based Firm Behavior

Yijiashun Qi [1,*], Hanzhe Guo [1] and Yijiazhen Qi [2]

[1] University of Michigan, USA
[2] The University of Hong Kong, Hong Kong
* Correspondence: elijahqi@umich.edu

Abstract

Agent-based computational economics (ACE) relies on hand-coded heuristics or stylized utility maximization to model firm behavior. We propose using large language models (LLMs) as economic agents—firms that reason in natural language about investment decisions under realistic market conditions. We design a simulation framework in which GPT-based agents, representing heterogeneous manufacturing firms, decide whether to adopt CNN-based visual inspection technology across five rounds with declining costs. We test whether emergent aggregate behavior aligns with propositions from production theory regarding scale economies in adoption, declining adoption thresholds, and capital-labor substitution. Across 150 firm-round decisions (30 firms $\times$ 5 rounds), decision-level logistic regressions show that time strongly predicts adoption ($\hat{\beta}_{\text{round}} = 0.88$, $p < 0.001$; marginal effect: $+13$ percentage points per round), while firm size has no statistically significant effect ($\hat{\beta}_{\log q} = -0.11$, $p = 0.55$). Among adopters, larger firms exhibit greater proportional QA labor displacement ($\hat{\beta}_{\log q} = -0.05$, $p = 0.007$), consistent with scale-dependent substitution elasticity. Robustness experiments across three prompt personas and six temperature settings reveal dramatic persona sensitivity—rational optimizers reach 100% adoption by Round 4 while risk-averse owners never adopt—interpretable as distinct behavioral types (risk neutrality, moderate risk aversion, high loss aversion). We identify behavioral phenomena not predicted by standard theory—status quo bias, production disruption anxiety, and reasoning heterogeneity (97 distinct decision factors)—suggesting LLM agents capture bounded rationality absent from stylized models. Our open-source framework provides a tool for hypothesis generation and behavioral prior elicitation for agent-based computational economics.

Keywords: large language models; agent-based simulation; economic agents; production theory; technology adoption; computational economics

1. Introduction

Agent-based computational economics (ACE) has been a productive methodology for studying emergent phenomena in markets, industries, and macroeconomies [1,2]. Traditional ACE agents follow hand-coded decision rules—zero-intelligence traders, genetic algorithm learners, or stylized utility maximizers—that capture some aspects of economic behavior but miss the nuanced, context-dependent reasoning that characterizes real firm decision-making.

The recent emergence of large language models (LLMs) with human-like reasoning capabilities opens a new frontier. Horton [3] demonstrated that GPT can replicate classic experimental economics results, while Park et al. [4] showed that LLM agents can simulate believable human social behavior. These findings raise a fundamental question: *Can LLMs serve as realistic economic agents for testing and validating economic theory?*

We address this question by constructing a simulation framework in which GPT-based agents represent heterogeneous manufacturing firms deciding whether to adopt CNN-based visual inspection technology. The economic environment features declining technology costs, rising wages, and

 doi:10.20944/preprints202602.1648.v1

heterogeneous firm characteristics—conditions that generate theoretical predictions about adoption patterns.

Our testing ground is a formal production-theoretic model of AI capital adoption that yields four propositions:

P1. Diminishing returns to inspection capital (convex defect cost reduction).
P2. Scale economies: optimal investment is increasing and concave in output.
P3. Size-dependent adoption threshold that declines as technology costs fall.
P4. Capital-labor substitution elasticity increases with firm size.

Of these, P1 requires observing continuous investment variation, which our binary adoption design cannot provide (see Section 3.5). We therefore focus empirical testing on P2–P4, using decision-level logistic regressions with clustered standard errors rather than descriptive metrics alone.

Our main findings are:

- **Strong temporal effect (P3)**: Adoption increases significantly over rounds ($\hat{\beta}_{\text{round}} = 0.88$, $p < 0.001$), with a marginal effect of +13 percentage points per round.
- **No significant scale effect (P2)**: Firm size does not significantly predict adoption ($\hat{\beta}_{\log q} = -0.11$, $p = 0.55$), with large firms exhibiting disruption-averse behavior that counteracts theoretical scale advantages.
- **Scale-dependent labor substitution (P4)**: Among adopters, larger firms reduce proportionally more QA workers ($\hat{\beta}_{\log q} = -0.05$, $p = 0.007$, $R^2 = 0.21$).
- Agents display **behavioral richness** absent from standard theory: production disruption anxiety dominates cost-benefit reasoning, with 97 distinct decision factors cited across 150 firm-round decisions.
- Results are **highly sensitive to prompt persona**: rational optimizers reach 100% adoption while risk-averse owners never adopt, interpretable as a spectrum of behavioral types.

We contribute a simulation framework, empirical methodology, and robustness protocol for LLM-based economic simulation. We note important limitations: our monotone environment design (all variables move pro-adoption simultaneously) inflates temporal alignment metrics, prompt personas effectively define agent behavior rather than merely perturbing it, and agents reason from structured prompts rather than first principles. Our code is open-source.

2. Related Work

2.1. LLMs as Simulated Agents

The idea of using LLMs to simulate human behavior has gained rapid traction. Horton [3] coined the term "Homo Silicus" and showed that GPT-3 reproduces results from ultimatum games, dictator games, and labor market experiments. Aher et al. [5] replicated multiple human subject studies using LLMs, demonstrating that language models can serve as proxies for human experimental subjects. Argyle et al. [6] showed that LLMs can simulate demographic subgroups with distinct political attitudes. Chen et al. [7] demonstrated that GPT exhibits emergent economic rationality in trading tasks.

Our work extends this literature in two ways. First, we test *theoretical predictions* from formal economic models using decision-level regressions with proper statistical inference, rather than relying solely on descriptive alignment metrics. Second, we focus on *firm behavior* (production decisions, capital investment, technology adoption) rather than individual consumer or experimental subject behavior. We do not claim to be the first to use LLMs as economic agents—Horton [3] established this research direction—but we contribute a structured methodology for mapping theoretical propositions to testable empirical proxies in an LLM-agent setting.

 doi:10.20944/preprints202602.1648.v1

2.2. Generative Agent Architectures

Park et al. [4] developed a multi-agent simulation with memory, reflection, and planning capabilities, demonstrating emergent social behaviors. We adopt key architectural ideas—state persistence across rounds, structured decision output, role-based prompting—but apply them to economic decision-making rather than social interaction. Our agents make structured investment decisions (JSON output) rather than free-form conversational interactions.

2.3. Agent-Based Computational Economics

ACE has a long tradition of bottom-up economic modeling [1]. Axtell [8] demonstrated that simple agent rules can produce emergent macroeconomic patterns (Zipf distribution of firm sizes). Dawid and Delli Gatti [2] surveyed macro ACE models using heterogeneous interacting agents. LLM agents represent a qualitatively different agent type: they reason in natural language, adapt to context, and exhibit behavioral heterogeneity without explicit programming of decision rules. We do not compare against heuristic ACE agents (logit rules, bounded-rational rules, structural adoption models) in this paper; a proper benchmark would pit LLM agents against these baselines on the same environment. We leave this for future work.

2.4. Economic Theory of Technology Adoption

Our testing ground draws on the technology adoption literature, particularly the task-based automation framework of Acemoglu and Restrepo [9,10], which models automation as expanding the set of tasks performed by capital. The specific model we test treats CNN inspection systems as specialized capital with exponential defect-reduction properties, yielding predictions about scale-dependent adoption thresholds.

3. Framework

3.1. Overview

Our simulation framework consists of three components: (1) a set of heterogeneous *firm agents*, each represented by an LLM with a firm-specific prompt; (2) an *environment* that broadcasts market conditions and technology costs to all agents each round; and (3) an *aggregation layer* that collects individual decisions and computes both descriptive and regression-based metrics.

3.2. Firm Agent Design

Each firm agent is defined by:

- **Identity**: Industry, firm size (monthly output), number of employees, current QA workforce.
- **State**: CNN adoption status, cumulative CNN investment, QA worker count. State persists across rounds.
- **Behavioral anchoring**: A system prompt establishing the agent as a "practical operations manager" who reasons about costs, payback periods, and risk.

The agent receives a structured prompt each round containing its firm state, current market conditions (CNN costs, wages, accuracy, competitor adoption), friction costs (integration time, production disruption, false positive costs), and a "Considerations" block highlighting qualitative factors (institutional knowledge, severance costs, technology obsolescence risk, vendor reliability). It responds with a JSON object containing:

- `reasoning`: 2–4 sentence step-by-step explanation.
- `adopt_cnn`: Boolean adoption decision.
- `invest_amount`: Dollar investment if adopting.
- `qa_workers_after`: QA workforce after decision.
- `confidence`: 1–10 self-reported confidence.
- `key_factor`: Single most important decision factor.

 doi:10.20944/preprints202602.1648.v1

Design rationale. We use structured JSON output (via OpenAI's JSON mode) to enable automated analysis while preserving natural-language reasoning. The system prompt explicitly states "You are NOT an economist or theorist" to prevent agents from reasoning about theoretical predictions rather than making practical decisions. State persistence allows agents to consider sunk costs and prior decisions, introducing path dependence.

Information provided to agents. Each decision prompt includes: firm identity (industry, output, employees), QA labor costs, defect rates, CNN hardware costs, maintenance costs, integration time, expected production loss during integration, false positive rates and costs, QA inspector wages, defect costs per unit, and industry-wide CNN adoption rates. Agents also receive a "Considerations" block listing qualitative factors (institutional knowledge, severance costs, technology risk, vendor reliability, fallback planning). This structured information introduces experimenter degrees of freedom analogous to questionnaire framing in surveys—see Section 6.5 for discussion.

3.3. Environment and Scenarios

We simulate five rounds spanning 2020–2028, with market conditions calibrated from manufacturing and computer vision literatures (Table 1).

Table 1. Market conditions by round. All variables move monotonically in a pro-adoption direction (costs decrease, wages increase, accuracy increases, competitor adoption increases). This design choice simplifies interpretation but limits the strength of alignment tests—see Section 6.2.

Round	Year	F_{CNN}	Wage	Accuracy	Adoption
1	2020	\$8,000	\$18/hr	92%	5%
2	2022	\$5,000	\$20/hr	95%	12%
3	2024	\$2,500	\$22/hr	97%	25%
4	2026	\$1,000	\$25/hr	98%	40%
5	2028	\$500	\$28/hr	99%	55%

3.4. Firm Heterogeneity

We sample firms from 8 industries (visual-defect share: 50–90%) and 4 size classes (Table 2).

Table 2. Firm size classes.

Size Class	Output (units/mo)	Employees
Micro	200–800	5–15
Small	800–3,000	15–50
Medium	3,000–15,000	50–200
Large-SME	15,000–50,000	200–500

3.5. Proposition-to-Test Mapping

Table 3 maps each theoretical proposition to its empirical proxy, statistical test, and falsification criterion. We note that P1 (diminishing returns to inspection capital) cannot be tested with our binary adoption design, which elicits only adopt/not-adopt decisions rather than continuous investment amounts. Testing P1 would require a design that allows agents to choose investment levels on a continuous scale. We therefore focus on P2–P4.

 doi:10.20944/preprints202602.1648.v1

Table 3. Proposition-to-test mapping. Each proposition from the theoretical model is mapped to an empirical proxy testable with our simulation data. P1 cannot be tested with binary adoption decisions.

	Formal Statement	Empirical Proxy	Statistical Test	Falsification
P1	MC_{QA} decreasing convex in K_{CNN}	Not directly testable (binary adopt decision)	N/A	N/A
P2	K^* increasing in q	Adoption rate by firm size	Logit: adopt $\sim \log(q)$, clustered SEs	$\hat{\beta}_{\log q} \leq 0$
P3	Threshold $\bar{q}$ declining over time	Adoption rate by round	Logit: adopt $\sim$ round, clustered SEs	$\hat{\beta}_{\text{round}} \leq 0$
P4	σ_{KL} increasing in q	QA worker reduction by size	OLS: Δworkers $\sim \log(q)$, among adopters	Interaction ≤ 0

4. Experimental Design

4.1. Main Experiment

We simulate $N = 30$ firms (randomly sampled from 8 industries and 4 size classes) across $R = 5$ rounds, yielding 150 firm-round decisions. We use GPT-4o-mini (`gpt-4o-mini`, accessed via OpenAI API in February 2026) at temperature $\tau = 0.8$ with the practical manager prompt. Firm generation uses a fixed random seed (`seed=42`) for reproducibility.

Sample composition. The 30 firms comprise 7 micro, 6 small, 12 medium, and 5 large-SME firms, randomly drawn from 8 industries. The unequal allocation across size classes reflects random sampling rather than stratification; we address this by using $\log(\text{output})$ as a continuous predictor in our regressions rather than relying on size-class comparisons.

4.2. Robustness Experiments

R1: Prompt sensitivity.

We test three prompt variants with $N = 20$ firms each (100 firm-round decisions per variant):

- **Practical manager** (baseline): Cautious, cost-focused, considers competitors.
- **Rational optimizer**: Risk-neutral, NPV-calculating, no behavioral biases.
- **Risk-averse owner**: Family business, skeptical, prefers proven technology.

R2: Temperature ablation.

We test $\tau \in \{0.0, 0.3, 0.5, 0.8, 1.0, 1.5\}$ with $N = 20$ firms each (100 firm-round decisions per setting). At $\tau = 0$, decisions should be near-deterministic. At high τ, increased stochasticity should add individual-level noise while potentially preserving aggregate patterns.

4.3. Metrics

Regression-based tests (primary).

We use decision-level logistic regressions with standard errors clustered by firm to test P2–P4 (Table 3). These provide proper statistical inference (coefficients, confidence intervals, p-values) and control for the panel structure of the data.

Descriptive alignment metrics (supplementary).

- **Scale monotonicity** (SM): Fraction of adjacent size-class pairs where adoption rate weakly increases. SM = 1.0 indicates perfect ordinal alignment with P2. This metric is coarse—it treats a 51%/50% difference the same as 90%/10%—and is reported as a supplement to the regression results.
- **Temporal monotonicity** (TM): Fraction of adjacent round pairs where aggregate adoption weakly increases. TM = 1.0 indicates perfect ordinal alignment with P3. As discussed in Section 6.2, the

 doi:10.20944/preprints202602.1648.v1

monotone environment design means TM primarily measures agent consistency with commonsense monotonicity rather than alignment with a specific economic theory.

Behavioral realism.

- **Reasoning diversity**: Number of distinct `key_factor` categories across all firm-round decisions.

4.4. Error Handling Protocol

When the LLM returns a response that fails JSON parsing, the decision is recorded as a non-adoption with default values: `adopt_cnn=False`, `confidence=0`, `key_factor="error"`. This conservative default biases against adoption, making any positive adoption signal more credible. In the main experiment ($\tau = 0.8$, $N = 30$), zero JSON parse errors occurred across all 150 queries. In the temperature ablation, errors occurred only at $\tau = 1.5$: 19 out of 100 firm-round queries (19%) failed to parse, producing garbled text fragments. All other temperature settings ($\tau \leq 1.0$) produced zero errors.

5. Results

5.1. Main Experiment

We simulate 30 firms across 5 rounds using GPT-4o-mini at $\tau = 0.8$ with the practical manager prompt. Table 4 reports adoption rates by firm size in the final round, and Table 5 reports the adoption trajectory over time.

Table 4. Adoption rate by firm size (final round, $N = 30$).

	Micro ($n = 7$)	Small ($n = 6$)	Medium ($n = 12$)	Large-SME ($n = 5$)
Adoption rate	57%	83%	58%	40%

Table 5. Adoption rate by round ($N = 30$, practical manager prompt).

	R1 (2020)	R2 (2022)	R3 (2024)	R4 (2026)	R5 (2028)
Adoption	0%	10%	20%	43%	50%

The descriptive alignment metrics are: scale monotonicity SM = 0.33 (weak), temporal monotonicity TM = 1.00 (perfect). Agent reasoning produced 97 distinct decision factors, with "production disruption during integration" as the most cited factor (14 mentions), followed by "risk of production disruption during integration" (10) and "disruption to existing workflows" (9). Notably, cost—the primary driver in the formal model—ranked below operational disruption concerns, suggesting LLM agents prioritize qualitative risk over quantitative cost-benefit analysis.

Labor displacement averaged 11.2% across all firms: agents retained QA workers for functional (non-visual) defects, consistent with the task-heterogeneity prediction that CNN capital substitutes for visual inspection labor but complements complex inspection labor.

5.2. Decision-Level Regression Analysis

Table 6 reports logistic regression results for the 150 firm-round decisions in the main experiment, with standard errors clustered by firm.

 doi:10.20944/preprints202602.1648.v1

Table 6. Decision-level regression results (150 firm-round observations, clustered SEs by firm). Dependent variable: $\text{adopt}_{it} \in \{0, 1\}$.

	(1) P2	(2) P3	(3) Full
$\log(q)$	-0.11		-0.92^{*}
	(0.19)		(0.47)
Round		0.88^{***}	-0.24
		(0.15)	(1.12)
Round $\times \log(q)$			0.18
			(0.13)
Industry FE	No	No	Yes
Pseudo-R^2	0.004	0.189	0.395
N	150	150	150

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Clustered SEs in parentheses.

P2 (Scale economies).

Column (1) shows that firm size (log output) does not significantly predict adoption: $\hat{\beta}_{\log q} = -0.11$ (SE $= 0.19$, $p = 0.55$, 95% CI: $[-0.48, 0.26]$). We fail to reject the null of no scale effect. The point estimate is slightly negative, consistent with the descriptive finding that large-SME firms (40% adoption) adopted at lower rates than small firms (83%), driven by disruption aversion.

P3 (Declining threshold).

Column (2) shows that round strongly predicts adoption: $\hat{\beta}_{\text{round}} = 0.88$ (SE $= 0.15$, $p < 0.001$, 95% CI: $[0.58, 1.18]$). The marginal effect is $+0.13$ (an additional 13 percentage points per round). This confirms the descriptive TM $= 1.00$ finding with proper statistical inference. However, as discussed in Section 6.2, this result must be interpreted cautiously given the monotone environment design.

Interaction model.

Column (3) includes the full specification with round $\times \log(q)$ interaction and industry fixed effects. The interaction term is not statistically significant ($\hat{\beta} = 0.18$, $p = 0.17$), indicating that the effect of time on adoption does not vary significantly with firm size. The pseudo-R^2 increases to 0.40 with industry fixed effects, suggesting substantial cross-industry heterogeneity in adoption propensity.

P4 (Labor substitution).

Among the 37 adopting firm-round observations, we estimate: $\Delta\text{workers\_pct} = 0.27 - 0.05 \times \log(q)$ ($R^2 = 0.21$, $p = 0.007$). Larger adopting firms reduce proportionally more QA workers, consistent with P4's prediction that capital-labor substitution elasticity increases with scale. This is the strongest theory-aligned finding in our data.

5.3. Scale Economies (P2)

The failure to find a positive size effect (P2) merits discussion. Contrary to theoretical predictions, large-SME firms (40% final-round adoption) adopted at *lower* rates than small firms (83%). Qualitative analysis of agent reasoning reveals why: large firms disproportionately cited "production disruption during integration" as their primary concern, reasoning that their larger production volumes amplify the absolute cost of integration downtime. This finding is consistent with the behavioral economics literature on loss aversion—large firms have more to lose from disruption, even though their per-unit savings are greater.

Importantly, the regression analysis (Table 6, Column 1) shows that this pattern is not statistically significant. With only 30 firms, we lack power to distinguish a true negative size effect from sampling noise. The descriptive SM $= 0.33$ overstates the evidence against P2 by treating each size-class comparison as equally informative regardless of sample size within each class.

 doi:10.20944/preprints202602.1648.v1

5.4. Adoption Threshold (P3)

Adoption increases monotonically from 0% (Round 1) to 50% (Round 5), following a realistic S-curve trajectory: 0%, 10%, 20%, 43%, 50%. The first adopters appear in Round 2 when CNN costs drop to $5,000. The regression confirms this pattern is statistically significant ($p < 0.001$), with each additional round increasing adoption probability by approximately 13 percentage points at the mean.

5.5. Behavioral Phenomena

Qualitative analysis of agent reasoning reveals phenomena not predicted by standard production theory:

- **Status quo bias**: Firms that declined adoption in early rounds express reluctance to adopt even when conditions improve ("we've managed fine without it").
- **Herding**: Many agents cite competitor adoption rate as a key factor, even though it is not decision-relevant in a price-taking framework. This behavior may be partly an artifact of including industry adoption rates in the decision prompt.
- **Cash flow framing**: Agents disproportionately focus on upfront costs relative to ongoing savings, consistent with behavioral economics findings on present bias.
- **Reasoning heterogeneity**: Agents in the same size class cite different key factors (payback period, competitor pressure, defect severity, disruption risk), producing a richer behavioral landscape than any single-heuristic ACE model.

5.6. Robustness

R1: Prompt sensitivity as behavioral type specification.

We test three prompt variants with 20 firms each (Table 7). Results reveal dramatic persona sensitivity: the rational optimizer reaches 100% adoption by Round 4, the practical manager follows an S-curve to 50%, and the risk-averse owner *never adopts* across all 5 rounds (0% throughout).

Table 7. Prompt sensitivity results ($N = 20$ per variant, 100 firm-round decisions each).

Prompt	SM	TM	Final Adopt.	Labor Disp.	Diversity
Practical manager	0.67	1.00	50%	10.2%	83 factors
Rational optimizer	1.00	1.00	100%	40.1%	73 factors
Risk-averse owner	1.00	1.00	0%	0.0%	62 factors

Rather than interpreting these results as evidence for or against robustness, we propose reframing prompt sensitivity as *behavioral type specification via prompt engineering*. Each prompt maps to an economic primitive:

- **Rational optimizer** $\leftrightarrow$ risk neutrality, zero adjustment costs.
- **Practical manager** $\leftrightarrow$ moderate risk aversion, positive adjustment costs.
- **Risk-averse owner** $\leftrightarrow$ high risk aversion, large adjustment costs, loss aversion.

The finding is not "results are fragile" but rather "LLMs can capture a spectrum of behavioral types whose adoption rates bracket real-world adoption curves (0% to 100%)." The rational optimizer's key decision factor is overwhelmingly "cost savings from reduced defects" (16 of 100 firm-round mentions), while the risk-averse owner focuses on "cash flow stability" (27 of 100 firm-round mentions)—a clean separation of decision heuristics by persona. Temporal monotonicity is perfect (TM = 1.00) across all three variants, indicating that the temporal ordering prediction is robust across behavioral types.

R2: Temperature ablation.

We test six temperature settings with 20 firms each (Table 8).

 doi:10.20944/preprints202602.1648.v1

Table 8. Temperature ablation results ($N = 20$ per setting, practical manager prompt).

τ	SM	TM	Final Adopt.	Labor Disp.	Diversity	Errors
0.0	0.67	0.75	35%	19.6%	39	0
0.3	0.33	1.00	45%	9.4%	52	0
0.5	0.67	0.50	35%	10.5%	55	0
0.8	0.67	0.75	35%	14.9%	75	0
1.0	0.67	1.00	50%	9.4%	86	0
1.5	0.67	0.75	30%	16.4%	81	19

Scale monotonicity is stable at SM $\approx$ 0.67 across most temperatures. Temporal monotonicity varies between 0.50 and 1.00 without a clear trend. Reasoning diversity increases monotonically with temperature (39 distinct factors at $\tau = 0.0$ vs. 86 at $\tau = 1.0$), consistent with the interpretation that higher temperature induces more varied reasoning strategies. At $\tau = 1.5$, the model produces 19 JSON parse errors out of 100 firm-round queries (19%), with garbled text fragments, establishing an upper bound on usable temperature. These 19 error decisions are treated as non-adoption per our error handling protocol (Section 4.4), which biases the $\tau = 1.5$ results toward lower adoption rates.

6. Discussion

6.1. When LLM Agents Align with Theory

Our results suggest that LLM agents reliably reproduce "intuitive" economic predictions—those that align with common-sense reasoning about costs and timing. The temporal effect (cheaper technology gets adopted by more firms) is statistically significant and robust across conditions. The scale effect (larger firms benefit more from fixed-cost technologies) is not statistically significant in our data, though the small sample size limits power.

More subtle theoretical predictions (precise convexity properties, elasticity variation) are harder to test behaviorally. The significant P4 result (larger adopters reduce more workers) is encouraging but based on only 37 adopter observations. We hypothesize that LLMs encode a *folk economics*—a reasonable but imprecise version of economic reasoning absorbed from training data—rather than formal utility maximization.

6.2. Environment Design Limitations

A fundamental limitation of our environment design is that *all five environmental variables move in a pro-adoption direction simultaneously*: CNN costs decrease, wages increase, accuracy increases, defect costs increase, and competitor adoption increases across rounds (Table 1). This means that temporal monotonicity (TM) primarily measures whether agents respond consistently to monotone incentives, not whether they align with any specific economic theory. Any agent with monotone preferences over these variables—including a simple threshold rule—would produce TM $\approx$ 1.0.

A stronger test of P3 would: (i) vary environmental variables independently (e.g., hold costs constant while varying only wages), (ii) include counter-directional shocks (e.g., a cost reversal in Round 3), or (iii) randomize the order of scenarios. We leave these designs for future work.

6.3. LLMs as Bounded Rationality Agents

The behavioral phenomena observed in our simulation—status quo bias, herding, present bias in cash flow assessment—are well-documented in the behavioral economics literature but absent from standard production theory. LLM agents naturally exhibit bounded rationality without requiring it to be explicitly programmed. This suggests a use case: LLM agents as *behavioral enrichment* for ACE models that currently rely on either full rationality or ad hoc behavioral rules.

6.4. Training Data Contamination

LLM agents may reproduce theory-consistent patterns because economics texts are in their training data, not because they reason economically from first principles. The "folk economics" hypothesis

 doi:10.20944/preprints202602.1648.v1

implies that GPT-4o-mini has absorbed common economic intuitions ("cheaper is better," "bigger firms can afford more") from its training corpus. This is not necessarily a limitation for ACE applications—real economic agents also absorb economic intuitions from their environment—but it means our results should not be interpreted as evidence that LLMs "understand" production theory.

6.5. Limitations

Several limitations merit discussion:

- **Not real agents**: LLM agents encode training data priors, not genuine utility maximization. Their "reasoning" is pattern completion, not decision theory. There is no evidence that LLM adoption decisions map to real firm behavior.
- **Prompt determines behavior**: Prompt personas effectively define agent behavior rather than merely perturbing it. The rational optimizer and risk-averse owner produce adoption rates of 100% and 0% respectively—these are not perturbations around a baseline but fundamentally different agents. This means the researcher's prompt choices are a primary degree of freedom.
- **Information scaffolding**: Agents receive structured financial parameters (costs, wages, revenue estimates, competitor adoption rates) rather than reasoning from raw observations. The "Considerations" block in the prompt introduces framing effects analogous to questionnaire design in surveys. Future work should include: (i) a minimal-information condition (only CNN cost and firm size) as a baseline, (ii) placebo variables to test whether agents react to irrelevant information, and (iii) adversarial prompt variations.
- **Monotone environment**: All environmental variables move pro-adoption, inflating temporal alignment metrics (Section 6.2).
- **No genuine learning**: Agents do not update beliefs from experience within a session; they react to provided context. State persistence provides a crude form of memory but not true Bayesian updating.
- **Model specificity**: All results are specific to GPT-4o-mini (`gpt-4o-mini`) accessed in February 2026. OpenAI periodically updates models, potentially changing results. Model version IDs are not guaranteed to be stable.
- **Sample size**: With 30 firms and 150 firm-round decisions, our main experiment has limited statistical power, particularly for detecting size effects across 4 size classes with unequal allocation.
- **No ACE baseline**: We do not compare LLM agents against standard ACE agent types (logit adoption rules, bounded-rational rules, structural models). Such a comparison would be necessary to evaluate whether LLM agents offer meaningful advantages over simpler alternatives.
- **Cost**: Large-scale simulations with frontier models (GPT-4o) are expensive. Our 150-decision main experiment costs approximately $0.50–1.00 with GPT-4o-mini.

6.6. When to Use LLM-Agent Simulations

Based on our experience, we recommend LLM-agent simulations for:

- **Hypothesis generation**: Discovering behavioral phenomena (herding, status quo bias) that formal models miss, and generating testable predictions for empirical work.
- **Behavioral prior elicitation**: Using prompt personas to map out the space of plausible agent behaviors before committing to a specific behavioral model in ACE simulations.
- **Communication**: Making economic models tangible by showing how "realistic" agents respond to the model's environment.

We caution against using LLM agents for:

- Precise quantitative predictions (adoption rates, welfare calculations).
- Mechanism design where incentive compatibility is critical.
- Claims about "validating" economic theory—our results show alignment with intuitive predictions but cannot distinguish theory-driven reasoning from training data pattern matching.

 doi:10.20944/preprints202602.1648.v1

7. Conclusions

We have investigated whether LLM agents can serve as a useful tool for exploring economic theory. In our simulation of CNN adoption in manufacturing, GPT-based firm agents exhibit statistically significant temporal adoption patterns consistent with declining cost thresholds ($p < 0.001$), but no significant scale-dependent adoption as predicted by theory ($p = 0.55$). Among adopters, larger firms show greater labor displacement, consistent with scale-dependent substitution elasticity ($p = 0.007$). The agents also exhibit behavioral richness (bounded rationality, herding, status quo bias) absent from standard models, suggesting their potential value as behavioral enrichment for computational economics.

Important caveats apply: our monotone environment design inflates temporal alignment, prompt personas effectively define rather than perturb agent behavior, agents reason from structured prompts rather than first principles, and we provide no comparison against heuristic ACE baselines. These limitations mean our results should be interpreted as a methodological contribution—demonstrating a framework for mapping theoretical propositions to LLM-testable proxies—rather than as evidence that LLMs can validate economic theory.

Our open-source framework is general: it can be applied to any economic decision environment with clear structure and testable predictions. We envision it as complementary to—not a substitute for—traditional tools in the economist's toolkit: formal theory, laboratory experiments, and empirical estimation.

Appendix A. Reproducibility

Appendix A.1. Model and API Configuration

- **Model**: `gpt-4o-mini` (OpenAI API, accessed February 22–23, 2026).
- **API parameters**: `temperature` varies by experiment (see Tables 7 and 8); `max_tokens=400`; `response_format={"type": "json_object"}`.
- **Firm generation seed**: `seed=42` for reproducible firm sampling. Note that this seeds the Python random number generator for firm attribute generation only; the LLM API does not support deterministic seeding of model outputs.
- **Error handling**: JSON parse failures are recorded as non-adoption (`adopt_cnn=False`, `confidence=0`, `key_factor="error"`). See Section 4.4.

Appendix A.2. Error Rates by Experiment

- Main experiment ($\tau = 0.8$, $N = 30$): 0 errors / 150 queries (0%).
- Prompt sensitivity ($\tau = 0.8$, $N = 20$ per variant): 0 errors / 300 queries (0%).
- Temperature ablation ($N = 20$ per setting): 0 errors at $\tau \leq 1.0$; 19 errors / 100 queries (19%) at $\tau = 1.5$.

Appendix A.3. Prompt Versioning

Full system prompts for all three persona variants are reproduced in Appendix B. The decision prompt template is reproduced in Appendix C. All prompts are also available in the code repository.

Appendix A.4. Data and Code Availability

All simulation code, raw JSON results, analysis scripts, and prompt templates are available in our public repository. The analysis script (`scripts/analyze_results.py`) reproduces all regression results reported in Section 5.2.

 doi:10.20944/preprints202602.1648.v1

Appendix B. Prompt Templates

Appendix B.1. System Prompt: Practical Manager (Baseline)

```
You are the operations manager of a small manufacturing
firm. You make practical business decisions about quality
assurance investments. You think in terms of costs,
benefits, payback periods, and risk.

You are NOT an economist or theorist. You are a practical
manager who:
- Cares about the bottom line and cash flow
- Is cautious about new technology
- Knows your production line intimately
- Weighs upfront costs against long-term savings
- Considers what competitors are doing
- Worries about disruption to existing workflows

IMPORTANT: Respond with valid JSON only.
```

Appendix B.2. System Prompt: Rational Optimizer

```
You are a perfectly rational, profit-maximizing firm
manager. You make optimal investment decisions by
comparing expected costs and benefits.

You:
- Calculate NPV of all investments
- Are risk-neutral
- Make decisions purely on expected value
- Ignore sunk costs
- Have no behavioral biases

IMPORTANT: Respond with valid JSON only.
```

Appendix B.3. System Prompt: Risk-Averse Owner

```
You are the owner of a small family manufacturing
business. You are naturally risk-averse and cautious
about big technology investments.

You:
- Prefer proven technologies over cutting-edge ones
- Worry about cash flow and liquidity
- Value stability and predictability
- Are skeptical of vendor promises
- Only adopt when you've seen others succeed
- Think about your employees' jobs

IMPORTANT: Respond with valid JSON only.
```

 doi:10.20944/preprints202602.1648.v1

Appendix C. Firm Decision Prompt Template

Each round, the agent receives a structured prompt containing:

1. **Firm identity**: Industry, monthly output, employees, QA workers, monthly QA cost, defect rate, visual defect share, CNN status, annual revenue, cash reserves, and prior-round decision summary.
2. **Market conditions**: CNN cost, monthly maintenance, integration time, expected production loss, detection accuracy, false positive rate and cost, QA wage, defect cost per unit, industry adoption rate.
3. **Considerations block**: Qualitative factors including institutional knowledge, severance/morale costs, technology management requirements, obsolescence risk, vendor reliability concerns, and fallback planning.
4. **Decision task**: JSON schema specifying the required output fields.

The complete template is included in our code repository.

References

1. Tesfatsion, L. Agent-Based Computational Economics: A Constructive Approach to Economic Theory. *Handbook of Computational Economics* **2006**, *2*, 831–880.
2. Dawid, H.; Delli Gatti, D. Agent-Based Macroeconomics. *Handbook of Computational Economics* **2018**, *4*, 63–156.
3. Horton, J.J. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper* **2023**.
4. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023, pp. 1–22.
5. Aher, G.V.; Arriaga, R.I.; Kalai, A.T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 337–371.
6. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.R.; Rytting, C.; Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* **2023**, *31*, 337–351.
7. Chen, Y.; Jensen, T.X.; Zheng, A. The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2316205120.
8. Axtell, R.L. Zipf Distribution of US Firm Sizes. *Science* **2001**, *293*, 1818–1820.
9. Acemoglu, D.; Restrepo, P. The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review* **2018**, *108*, 1488–1542.
10. Acemoglu, D.; Restrepo, P. Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives* **2019**, *33*, 3–30.