E-COMMERCE BENCH: EVALUATING LLM AGENTS ON LONG-HORIZON AUTONOMOUS BUSINESS OPERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon agentic tasks go beyond chaining short tasks over more interaction turns. Their evolving dynamic environments and long-range dependencies require Large Language Models (LLMs) to continually explore, learn from experience, and adapt their policies over thousands of steps. We introduce **E-Commerce Bench**, the first open-source benchmark that integrates multi-round counterpart negotiation and dynamic events into a year-long business operation. Over a 365-day year, an LLM agent concurrently runs multiple online stores, researching the market, negotiating with suppliers to source inventory, optimizing sales strategies, fulfilling orders, handling returns, and managing cash flow to maximize its end-of-year total assets. To construct a realistic merchant-side operating environment, the product and supplier data are derived from a real e-commerce platform, while a year-long calendar of promotions, natural disasters, and supply-chain shocks continually reshapes demand. For reproducibility, both sides of the market are deterministic: customer purchases and returns follow a fixed demand model, while a negotiation kernel determines supplier pricing, concessions, and decisions, with an LLM used only to verbalize them. We evaluate 18 frontier models across seven dimensions, including year-end assets, and find that no single model dominates. GPT-5.6 Sol earns the most, growing the ¥100,000 opening stake into ¥1,431,425, yet it ranks 16th of 18 on fraud avoidance and trails Fable5 in operational efficiency. Among open-weight models, Qwen3.8-Max-Preview leads with ¥416,252, 38% above GLM 5.2 (high), and achieves the strongest learning over the horizon, progressively bargaining down prices across repeated orders.

1 INTRODUCTION

Evaluating large language model (LLM) agents over long horizons requires more than counting completed actions (Yao et al., 2023). An agent must maintain coherent goals, respond to changing conditions, and use earlier experience to improve later decisions. The distinction between *episodic* and *continuing* tasks (Sutton & Barto, 2018) clarifies this challenge. Episodic benchmarks evaluate a specified goal, such as resolving a GitHub issue (Jimenez et al., 2024), operating a computer (Xie et al., 2024), or producing expert-graded knowledge work (Patwardhan et al., 2025). Continuing tasks evaluate accumulated outcomes in an evolving environment (Qian et al., 2025; Chen et al., 2026). Business operations make this distinction concrete: the agent must keep a system functioning while its decisions change the conditions of future work.

A retail business exposes the dependencies that make this difficult. Buying more stock can support later sales while leaving too little cash for current expenses. A low supplier quote can conceal defective inventory, and a price that worked during a promotion may perform differently afterward. The horizon also creates opportunities to learn: a purchase price obtained in January can guide later replenishment if the agent retains or retrieves it, while deliveries and returns provide evidence for revising supplier assessments. A useful evaluation must therefore connect immediate actions to delayed consequences and measure whether experience improves decisions as the business continues to operate. Existing business benchmarks address parts of this problem. The Vending-Bench series (Backlund & Petersson, 2025; Andon Labs, 2025) supports supplier negotiation, but uses mock economic data and sampled LLM counterparts. YC-Bench (He et al., 2026) simulates a deterministic startup year without supplier sourcing. MerchantBench (Shi et al., 2026) and RetailBench (Zhang et al., 2026c) use real product data but omit negotiation and adversarial suppliers. These settings

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.

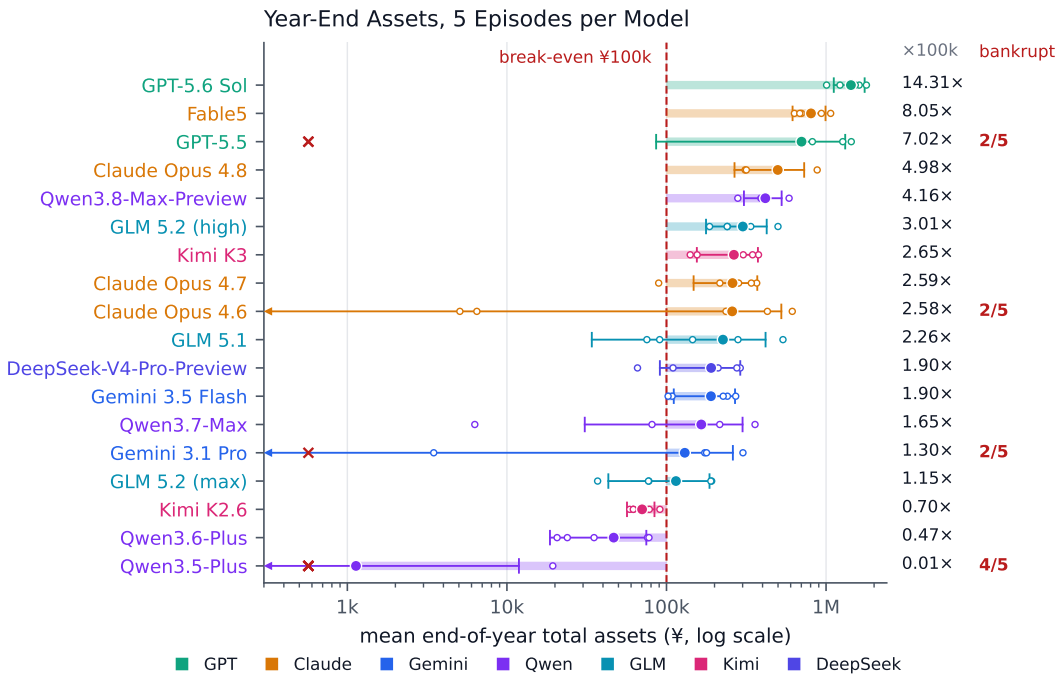


Figure 1: Mean end-of-year total assets, eighteen models, five 365-day episodes each. Bars start at the ¥100,000 stake. Whiskers span one 5-run standard deviation, hollow dots are single episodes, and red fractions count bankrupt episodes, which stay pooled into the mean. The right column gives mean assets as a multiple of the stake.

leave a gap in evaluating how agents coordinate data-grounded operations with repeated bargaining and supplier risk under reproducible economic rules.

We introduce **E-Commerce Bench**, the first open-source benchmark to combine deterministic supplier negotiation, dynamic events, and concurrent multi-store management over a year-long business operation (Figure 2). An agent starts with ¥100,000 and manages up to four stores over 365 simulated days, sourcing from 576 suppliers across 6,886 products derived from a real e-commerce platform. The central design connects *business consequences* to *opportunities for learning*. A purchase commits cash before revenue settles, determines what can be sold, and leaves delivery and return evidence for future sourcing. Promotions and supply shocks change these tradeoffs while product costs and demand parameters remain hidden. To make the resulting decisions measurable, a fixed demand model governs customer purchases and returns, while a seeded kernel governs supplier prices, concessions, and acceptance. An LLM renders the kernel's decisions as dialogue, retaining a natural-language interface to bargaining; Appendix I.4.2 describes the remaining variation in wording and advertised delivery times. Repeated purchases preserve the supplier's reservation price, giving past deals value as evidence for later negotiations. Persistent notes let the agent retain selected evidence beyond the context window (Packer et al., 2024; Park et al., 2023). Together, these choices test whether an agent can use experience while managing the financial commitments that experience creates. Total assets measure the accumulated outcome, and six operational diagnostics distinguish the capabilities behind it.

Our evaluation of 18 models reveals substantial financial gains alongside persistent operational weaknesses (Figure 1). GPT-5.6 Sol grows the initial stake roughly 14-fold, yet ranks 16th of 18 on fraud avoidance and earns less per tool call than Fable5. Claude Opus 4.7 leads negotiation and fraud avoidance while finishing mid-pack on assets. Qwen3.8-Max-Preview leads the ten open-weight entries at 4.2 times the stake and shows the strongest learning over the horizon. Across 8,647 repeat purchases from honest suppliers, 16 models show no clear improvement in purchase prices. Four models go bankrupt in some episodes. These findings motivate evaluating financial success together with the quality and consistency of the decisions that produce it.

Table 1: Comparison with continuing business-operation benchmarks. ✔ present, ✗ absent. *Steps/run* is the per-episode agent-action count, taken or inferred from each paper. *Op. units* counts business units (e.g., stores) run concurrently. *Reproducible* marks counterparts that follow fixed rules or seeded draws rather than a sampled LLM. [‡]In Vending-Bench 1/2, product and supplier names are real while prices, demand, and elasticity are LLM-generated. [§]Supplier economics are deterministic, while the renderer voicing them varies wording between runs (§I.4.2).

Benchmark	Domain	Steps/ run	Op. units	Real data	Nego- tiation	Multi- agent	Adver- sarial	Repro- ducible	Open source
Vending-Bench (Backlund & Petersson, 2025)	Vending	$\leq$2000	1	✗[‡]	✔	✔	✗	✗	✗
Vending-Bench 2 (Andon Labs, 2025)	Vending	3000–6000	1	✗[‡]	✔	✔	✔	✗	✗
YC-Bench (He et al., 2026)	Startup	$\sim$500	1	✗	✗	✔	✔	✔	✔
MerchantBench (Shi et al., 2026)	E-commerce	$\sim$700	1	✔	✗	✗	✗	✔	✔
RetailBench (Zhang et al., 2026c)	Supermarket	$\sim$1000	1	✔	✗	✗	✗	✔	✔
E-Commerce Bench (Ours)	E-commerce	1000–4000	$\leq$**4**	✔	✔	✔	✔	✔[§]	✔

Our contributions are: (1) a data-grounded business environment that couples year-long operations, dynamic events, and concurrent stores under shared cash and time constraints; (2) a seeded supplier kernel that extends TERMS-Bench (Zhang et al., 2026a) to repeated, cross-supplier bargaining, separating economic decisions from their presentation in language; and (3) an evaluation of 18 models across total assets and six capabilities, connecting the financial ranking to price quality, supplier risk, and the use of accumulated experience.

2 Related Work

2.1 Episodic Long-Horizon Agentic Tasks

Episodic benchmarks assess progress toward a terminal goal (Sutton & Barto, 2018). GDPval (Patwardhan et al., 2025) judges work products against expert assessments, while TheAgentCompany (Xu et al., 2025) models tasks in a software firm. SWE-bench (Jimenez et al., 2024) tests issue resolution, and OSWorld (Xie et al., 2024) tests computer use. Longer interaction can also require exploration: UltraHorizon (Luo et al., 2025) uses partially observable worlds, and EdgeBench (Zhu et al., 2026) studies learning over 12–72 hours. These settings test how agents use feedback to reach a goal. A running business adds a further challenge: each decision can create obligations that persist after its local task is done. Buying stock, for example, ties up cash that later sales must recover. Evaluating such agents requires tracking the business state across decisions as well as success on each task.

2.2 Continuing Long-Horizon Agentic Tasks

Continuing tasks assess accumulated outcomes, as in portfolio trading (Qian et al., 2025; Chen et al., 2026), skill-library growth (Wang et al., 2023), and agent self-modification (Zhang et al., 2026b). Business benchmarks bring this view to a running company. Vending-Bench (Backlund & Petersson, 2025) runs a vending machine for roughly two thousand messages; Vending-Bench 2 (Andon Labs, 2025) adds a year-long horizon and adversarial suppliers. Both use sampled LLM counterparts. YC-Bench (He et al., 2026) runs a deterministic startup year without supplier negotiation. MerchantBench (Shi et al., 2026) tests 365 days of seller operations; RetailBench (Zhang et al., 2026c) tests 180 supermarket days against an oracle. Neither includes negotiation or adversarial suppliers. Table 1 compares their scope, data, and counterpart design. TERMS-Bench (Zhang et al., 2026a) supplies a complementary piece: a rule-based bilateral counterpart without a business economy. E-Commerce Bench joins these strands by embedding a seeded supplier policy in data-grounded retail operations. Repeated purchases and competing suppliers make earlier deals useful evidence for later choices. For honest suppliers, fixed, data-grounded reservation prices provide a stable reference for measuring how agents use this history. The kernel sets economic decisions; a separate LLM renders dialogue. Bargaining outcomes determine what the agent pays for stock, how much cash each order commits, and which purchases remain affordable in subsequent operations.

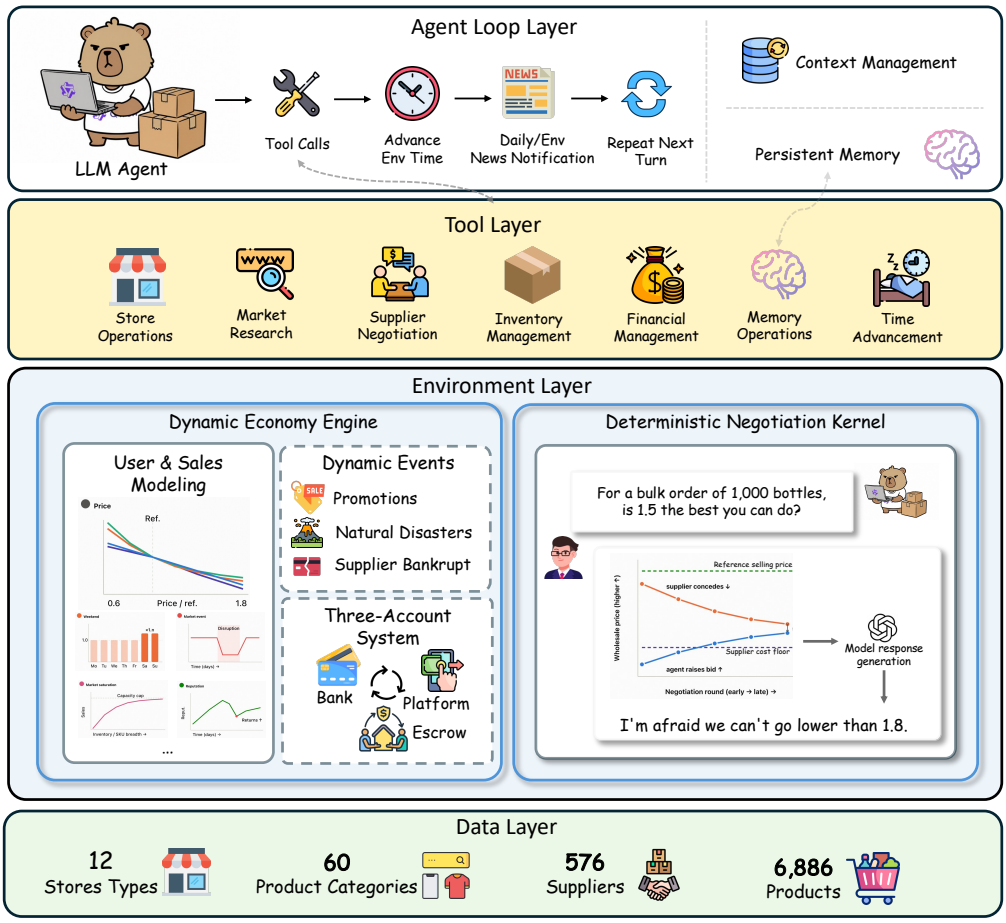


Figure 2: Overview of E-Commerce Bench. The agent loop manages turns, context, and persistent memory. The tool layer provides 18 e-commerce tools. The environment layer combines the sales-and-economy engine with the two-layer negotiation engine, using a data layer derived from a real e-commerce platform.

3 E-Commerce Bench

3.1 Overview

E-Commerce Bench gives the agent a merchant account and ¥100,000 to operate up to four stores during the simulated 2026 calendar year. The objective is to maximize end-of-year total assets. Figure 2 organizes the environment into four layers: the agent loop manages interaction and memory; tools expose the merchant's actions; the environment determines customer demand and supplier responses; and the data layer sets the catalog and economic parameters. Six diagnostics assess the decisions behind the financial outcome.

3.2 Agent Loop Layer

The agent operates under two budgets: model turns and simulated business time. Each model call produces a turn containing zero or more tool calls, executed in order. An episode allows up to 4,000 turns and ends early after three consecutive turns without a tool call. Tool use consumes time within a 600-minute working day; crossing the next day boundary triggers settlement. Since one turn can contain several actions, turn count and elapsed business time measure different constraints. All models share a 128,000-token context budget. At 120,000 tokens, the harness removes the oldest complete tool-call groups, releasing at least 60,000 tokens when enough groups are available. It preserves the system message, initial user turn, and two newest groups. Removing each group with

its tool responses preserves a valid interaction history, but can remove evidence that remains relevant to the business. To preserve selected evidence, the agent can maintain up to 20 persistent notes outside the transcript (Shinn et al., 2023; Packer et al., 2024; Zhong et al., 2024; Sumers et al., 2024). It chooses what to write, update, and retrieve; reading a note returns its contents to the context. Negotiated prices, supplier assessments, and pending commitments can therefore survive the eviction of their original observations, provided the agent records them. Appendix A specifies the interface and eviction algorithm; Appendix I.1 illustrates the loop and memory.

3.3 Tool Layer

Eighteen tools cover market research, stores, supplier negotiation, inventory, finances, memory, and time advancement (Table 3). Every tool except the day jump consumes simulated minutes: a balance check costs 10 minutes and a supplier message costs 30. Gathering information therefore competes with acting on it. A single chatbox message can address multiple suppliers at the same time cost, and a pricing call can update several products, allowing the agent to batch related decisions. The tools connect information gathering to the merchant workflow. Research reveals coarse market information; catalog and inventory queries support sourcing; publication, pricing, and shipping turn stock into orders and revenue. Financial actions remain explicit: a balance query reports funds, while a withdrawal makes settled revenue spendable. Only `chatbox` creates purchase orders, through free-form messages containing structured `negotiate` blocks (Appendix H.3).

3.4 Sales and Economy Engine

Customer demand combines a category-level sales baseline with six factors: price response, weekday, promotions, seasonality, events, and reputation. Category and store capacity limit demand; seeded rounding converts it to units, and stock limits sales. These constraints couple assortment and pricing: similar products compete for category demand, and promotions require inventory. Four elasticity families describe category-specific price responses. The quadratic family peaks at the reference price, so discounting does not uniformly increase demand. Appendix C.2 gives the equations and parameters. Fulfillment affects both current margins and future demand. Orders must ship within two days; missed deadlines cancel the order and reduce store reputation. Returns depend on the product, supplier defects, pricing, and shipping speed. Faster shipping costs more but lowers return probability, while higher prices can increase both unit revenue and returns. Shipped orders earn reputation credit; returns and cancellations reduce it, feeding service outcomes back into subsequent demand. The agent therefore has to evaluate sourcing, pricing, and shipping jointly over the life of an order. Calendar events change demand and delivery times, with different effects across store types. Joining a promotion trades a discount for a demand boost, making inventory timing part of the pricing decision.

Cash flow imposes a separate constraint through three accounts: the bank, escrow, and platform wallet. Purchases and operating costs debit the bank immediately; a customer order initially creates only a shipment obligation. Shipped sales enter escrow after commission, mature into the wallet after nine days, and become spendable after an explicit withdrawal. Refunds reduce the originating escrow batch. Ten consecutive days with a negative bank balance trigger bankruptcy. Positive total assets therefore do not guarantee solvency: the agent must coordinate purchasing, settlement, and withdrawals to keep funds available when costs are charged. Store choice commits working capital beyond the initial purchase. Opening a store incurs a setup fee, and each active store incurs daily operating costs. Closing stops that charge, but unsold inventory incurs storage costs until liquidated. Expansion must therefore cover both the stock and the continuing cost of carrying it. At each daily boundary, a fixed sequence charges costs, computes the previous day's sales, processes orders and payments, and updates state. Observed outcomes can thus reflect decisions made before the latest action. Appendix C specifies this settlement sequence, the costs, and the return and reputation rules.

3.5 Deterministic Negotiation Engine

Each purchase requires bargaining with a supplier whose reservation price is hidden. The public reference price and the supplier's private cost floor define the bargaining range used to measure procurement surplus. An agent must infer how much room remains from successive offers, while deciding whether to continue bargaining or try another supplier. The engine extends the counterpart

policy of TERMS-Bench (Zhang et al., 2026a) through two layers: a seeded kernel sets prices, concessions, acceptance, and walk-away decisions; an LLM renders them as dialogue. Separating these roles gives the counterpart an explicit economic policy while retaining a natural-language interface. The kernel responds to the negotiation history rather than selecting a fixed reply for each round. Its state tracks the agent's offers and concession behavior, with time pressure affecting acceptance and the possibility of walking away. Six honest behavior templates vary counterpart responsiveness and the cues exposed in dialogue. A deterministic parser reads the agent's structured offers, and an agent-side accept action is valid only when its price matches the standing kernel quote within ± 0.005. The kernel can also accept the agent's offer. This check ties completed purchases to the economic decision even if the renderer misstates a price. The supplier's membership program is handled separately through a prose intent classifier. The renderer's wording and advertised delivery times can vary between runs (Appendix I.4.2). Repeated purchases start a new negotiation cycle while retaining the supplier's type and reservation price. For an honest supplier, the best earlier deal bounds its hidden floor from above, providing evidence for restocking and comparing offers. Using that evidence requires carrying it into later negotiations, potentially after the original exchange leaves the context window. This separates single-session bargaining skill from the use of accumulated experience. Each exchange consumes working time, and a completed purchase immediately debits the bank, connecting price discovery to the operational and liquidity constraints above.

Of the 576 suppliers, 152 are fraudulent. Pre-deal scams inflate prices or charge fees; post-deal scams deliver short quantities or defective goods. Supplier types remain hidden, so agents must use dialogue and observed outcomes to assess risk. Post-deal evidence is especially indirect when stock from different suppliers is pooled: elevated returns need not identify which delivery caused them. The evaluation consequently distinguishes the price obtained from honest suppliers from the share of procurement spending sent to fraudulent ones. Appendix D provides the kernel functions, calibration, behavior templates, and transaction protocol. Appendix I.4 retains the bargaining formulation and examples.

3.6 Data Layer

The environment contains 6,886 products in 60 categories, 576 suppliers (424 honest and 152 fraudulent), 12 store types, 10 market events, and 8 promotions. Products, categories, return-rate tiers, monthly-sales bands, and the holiday calendar are desensitized from a real e-commerce platform. Other parameters, including elasticity, seasonality, reputation, and supplier attributes, are synthesized from analyses of platform logs. Promotion names are fictional to limit reliance on memorized events (Zhu et al., 2025). The catalog and supplier roster remain fixed across episodes. Agents observe coarse market summaries and their own outcomes, while demand parameters and supplier cost floors remain hidden. Appendix B lists the parameters and information available to the agent.

3.7 Evaluation and Metrics

The primary score is **end-of-year total assets**: bank balance, platform wallet, and unsettled escrow, also expressed as a multiple of the ¥100,000 stake. Bankrupt episodes remain in the averages. Total assets capture the cumulative effect of interdependent decisions. A profitable business can still pay fraudulent suppliers or waste tool calls. We therefore report six dimensions separately to expose the strengths and weaknesses behind that outcome (Appendix E). **Negotiation quality** uses CSE^{+}, the fraction of an honest supplier's bargaining range captured by a concluded deal, averaged without quantity weighting. **Fraud avoidance** uses $\mathrm{BadSpend}\%$, the share of procurement spending paid to fraudulent suppliers. **Cash flow and solvency** uses peak drawdown divided by peak total assets. **Operational efficiency** is mean profit divided by mean tool calls. **Operations execution** uses the controllable return rate, the pricing-induced change in expected returns (negative under discounts), excluding the defect channel. **Learning over the horizon** uses $\mathrm{AnchorRatio}$, overpayment on repeat orders from the same honest supplier for the same item, relative to the best earlier price and normalized by a null that shuffles each pair's price order. Shuffling keeps the achieved prices but removes their order, distinguishing low prices from sustained price discipline across repeat purchases. Higher is better for negotiation quality and efficiency; lower is better for the other four dimensions.

Table 2: Year-end assets, six diagnostic scores, and episode activity. Models are ordered by mean assets within each group; best scores are shaded. All five episodes, including bankruptcies, contribute to the financial means.

Model	Assets (¥k)		Diagnostic scores						Episode activity		
	Mean ↑	SD	CSE$^+$ ↑	Fraud % ↓	DD/peak ↓	¥/call ↑	Return pp ↓	Anchor ↓	Calls	Turns	Bankrupt runs
Proprietary											
GPT-5.6 Sol (max)	**1,431**	314	0.672	18.48	0.278	363	4.46	1.217	3,668	1,367	0/5
Fable5 (max)	805	188	0.772	3.46	0.141	**479**	0.22	1.573	1,469	945	0/5
GPT-5.5	702	616†	0.700	16.59	0.591	192	0.90	1.043‡	3,143	1,306	2/5
Claude Opus 4.8 (max)	498	231	0.662	5.41	0.130	266	0.20	1.309	1,497	812	0/5
Claude Opus 4.7 (max)	259	111	**0.811**	**0.12**	0.189	156	1.75	1.421	1,023	432	0/5
Claude Opus 4.6 (max)	258	266	0.649	14.98	0.587	129	0.76	1.268	1,221	788	2/5
Gemini 3.5 Flash	190	79	0.777	2.13	0.450	36	0.38	0.918	2,508	2,628	0/5
Gemini 3.1 Pro	130	130	0.660	12.58	0.688	19	2.03	1.376	1,605	1,141	2/5
Open-weight											
Qwen3.8-Max-Preview	416	111	0.713	6.13	0.242	173	0.38	**0.834**	1,826	962	0/5
GLM 5.2 (high)	301	124	0.693	1.99	**0.127**	137	0.59	1.434	1,467	672	0/5
Kimi K3	265	110	0.632	5.87	0.247	123	2.01	1.507	1,334	878	0/5
GLM 5.1	226	192	0.662	6.49	0.231	94	0.09	1.557	1,333	852	0/5
DeepSeek-V4-Pro-Preview (max)	190	100	0.654	8.99	0.166	68	1.54	1.235	1,337	678	0/5
Qwen3.7-Max	165	134	0.616	9.85	0.433	59	**−0.05**	1.433	1,108	782	0/5
GLM 5.2 (max)	115	71	0.632	19.68	0.360	10	2.24	1.362	1,430	931	0/5
Kimi K2.6	70	14	0.596	17.22	0.377	−27	0.42	1.548	1,087	1,017	0/5
Qwen3.6-Plus	47	28	0.625	10.68	0.583	−40	3.03	1.333	1,325	1,023	0/5
Qwen3.5-Plus	1.1	11	0.625	20.11	0.979	−92	5.88	1.860	1,076	800	4/5

Notes. Fraud is $\mathrm{BadSpend}\%$; DD/peak is normalized drawdown; ¥/call is profit per tool call; Return is the controllable return rate; Anchor is $\mathrm{AnchorRatio}$. Arrows indicate better performance. Calls, turns, and bankrupt runs are descriptive. Definitions: Section 3.7 and Appendix E. †GPT-5.5 reports population SD (sample estimate: ¥689k). ‡Its $\mathrm{AnchorRatio}$ uses three episodes; two bankrupt episodes have no repeat supplier–item order.

4 Experimental Results

4.1 Experimental Setup

We evaluate 18 models, grouped as eight proprietary and ten open-weight systems (Table 10). Each runs five independent 365-day episodes in the same fixed world, giving 90 episodes in total. Requests use provider defaults without an explicit temperature or `top_p`, and enable thinking where supported. All models share the tools, context policy, persistent memory, and supplier renderer. The longest episode takes 3,599 of the allowed 4,000 turns; all episodes end at the calendar horizon or in bankruptcy. Table 2 reports mean assets and their standard deviations, with bankrupt episodes included. Five episodes provide only a small sample (Miller, 2024; Kapoor et al., 2025); Appendix F reports the supporting per-dimension results.

4.2 Overall Performance

Financial performance varies widely both across models and across episodes. Mean final assets differ by a factor of 1,264 between GPT-5.6 Sol and Qwen3.5-Plus (Table 2, Figure 1). Proprietary models occupy the top four places; Qwen3.8-Max-Preview leads the open-weight group and ranks fifth overall. Within one model, Qwen3.7-Max finishes between ¥6,290 and ¥358,790 across five episodes. This spread makes the distribution of outcomes as relevant as the average when assessing consistent business operation. Ten of the 90 episodes end in bankruptcy, all from GPT-5.5, Claude Opus 4.6, Gemini 3.1 Pro, or Qwen3.5-Plus. GPT-5.5 still ranks third on mean assets despite two bankrupt episodes, separating strong average returns from consistent survival. These outcomes aggregate decisions across sourcing, pricing, fulfillment, and reinvestment. The six diagnostics below show why strong performance in one component need not yield the highest assets, and why high assets can coexist with costly mistakes (Appendix Figure 21).

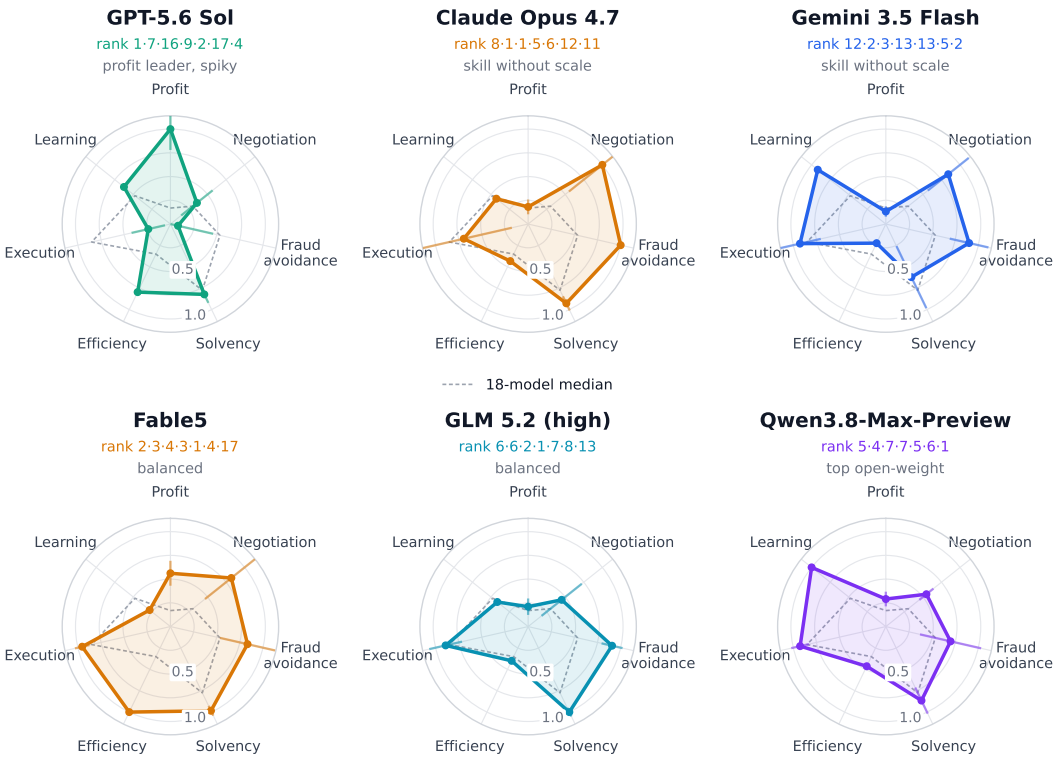


Figure 3: Capability profiles for six models. "Profit" denotes mean total assets; the other axes use the six metrics in Section 3.7. Each axis is min–max normalized over all 18 model means. Fraud, solvency, execution, and learning are sign-flipped, so higher is better throughout. Dashed polygons mark the field median; whiskers show episode variation. Efficiency is a ratio of means and has no whiskers; learning also omits whiskers because its archived spread uses a population standard deviation over varying episode counts.

4.3 Multi-Dimensional Analysis

The six diagnostics reveal distinct strengths behind the asset ranking (Figure 3).

Negotiation Quality. Agreement rates conceal differences in price quality. Claude Opus 4.7 leads on CSE^{+} (0.811), while Kimi K2.6 scores 0.596, versus roughly 0.50 for accepting the opening quote. Every model reaches agreement in at least 96% of concluded honest sessions. These rates exclude abandoned sessions, which account for 10.9% of sessions opened mid-year. Surplus distinguishes bargaining quality where agreement alone does not (Appendix F.2).

Fraud Avoidance. Every model pays fraudulent suppliers, with spending shares spanning more than 160-fold. GPT-5.6 Sol ranks 16th despite leading on assets; its absolute fraudulent spending is fifty times Qwen3.5-Plus's, despite a smaller share. The separation emerges between contact and purchase: 4.0% of fraudulent contacts become orders for Claude Opus 4.7, versus 31.7% for GPT-5.6 Sol (Appendix F.3). Asset growth thus coexists with costly supplier selection.

Cash Flow and Solvency. Revenue delays expose early cash-management failures. No episode receives platform-wallet revenue before day 15; two GPT-5.5 episodes fail on day 17 before settlement. Seventeen episodes have overdrawn mornings and ten go bankrupt. On 66 of 177 overdrawn days, the wallet already holds enough to cover the bank deficit. Solvency requires withdrawing settled revenue as well as controlling purchases: total assets can conceal cash shortfalls (Appendix F.4).

Operational Efficiency. Fable5 earns ¥479 per tool call versus GPT-5.6 Sol's ¥363 with 59.9% fewer calls. Routine operations consume 71.8% of calls, price updates 0.66%, and persistent memory 2.1% (Appendix F.5). Profit per call measures return on interaction, but tools differ in simulated time cost and can batch actions; API cost and latency are excluded (Kapoor et al., 2025; Patil et al., 2025).

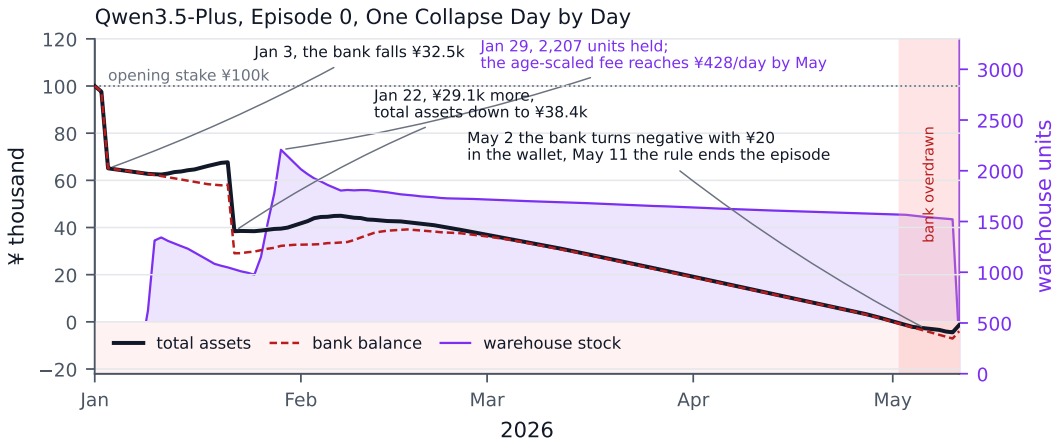


Figure 4: One bankruptcy episode day by day, Qwen3.5-Plus episode 0. Total assets and bank balance are on the left axis, warehouse stock on the right, and the shaded band marks the days the bank balance stayed negative.

Operations Execution. High assets can coexist with costly returns. GPT-5.6 Sol leads financially despite adding 4.46 percentage points of pricing-driven returns and incurring the largest refund bill, ¥1.62M. Shipping policies are stable within episodes: 72 of 88 episodes with sales use one tier for at least 98% of shipments. Yet nine models vary by over 30 points in fast-shipping share across episodes (Appendix F.6).

Learning over the Horizon. Repeated experience rarely improves price discipline. Only Qwen3.8-Max-Preview (0.834) and Gemini 3.5 Flash (0.918) achieve $\mathrm{AnchorRatio} < 1$; the median is 1.369. This baseline shuffles observed prices within supplier–item pairs. Fifteen models fare more than two standard deviations worse; only Qwen3.8-Max-Preview beats it by that margin. GPT-5.5 contributes three episodes because its two bankrupt runs have no repeat purchase. Sixteen models also capture less surplus in the second half of an episode. These outcomes do not isolate the effects of context eviction, memory use, or in-context adaptation (Appendix F.7).

4.4 Error Analysis

Ten detection rules connect aggregate scores to recurring behavior (Appendix G). Every model triggers at least five rules, and Gemini 3.1 Pro triggers all ten. Unsold inventory contributes to bankruptcy (Figure 4); 943 of 1,141 fraudulent agreements occur in supplier–item pairs traded at least twice. Agents repeat 746 of 2,180 catalog and supplier queries, while cancellations cluster near the beginning and end of the year. Fewer detected failures do not imply higher assets: the strongest bargainers have the cleanest profiles, while the asset leaders retain substantial weaknesses. These cases complement the ranking by identifying behaviors for further study (Appendix J).

5 Conclusion

E-Commerce Bench evaluates whether LLM agents can sustain a retail business as financial commitments, supplier evidence, and market conditions accumulate over a simulated year. Its data-grounded economy and seeded negotiation kernel connect individual decisions to their later consequences. Across 18 models, no model leads all seven evaluation axes: substantial asset growth coexists with weak fraud avoidance, inefficient tool use, and limited improvement in repeat purchasing. Evaluating the financial outcome together with operational diagnostics makes these weaknesses visible.

The results motivate studying how agents convert accumulated evidence into better decisions while meeting obligations created by earlier ones. Useful purchase history, timely withdrawals, and revised supplier assessments must support the same functioning business. E-Commerce Bench provides a setting for testing such policies; the present evaluation establishes their relevance without isolating the causal contribution of memory, planning, or supplier screening to final performance.

AI USE STATEMENT

We used AI tools to improve writing clarity and translate product names into English.

ETHICS STATEMENT

E-Commerce Bench evaluates agents in a simulated economy. Orders, payments, shipments, returns, and supplier dialogue do not execute real commercial transactions. The environment uses desensitized e-commerce data and synthetic economic and supplier parameters; the manuscript omits personal contact details and the identity of the source platform. Fraud scenarios represent dishonest counterparties to study agent robustness and financial risk. Reported transcripts are interactions between evaluated agents and simulated suppliers. Performance in this setting does not establish that an agent can safely operate a real business: real deployment requires separate evaluation of privacy, legal obligations, and safeguards for people affected by its decisions.

REPRODUCIBILITY STATEMENT

Appendices A–D describe the agent interface, data, economy, and negotiation kernel. Appendix E defines the metrics, Appendix F reports per-model results, and Appendix H provides the prompts. Additional design and experimental details appear in Appendices I and J.

REFERENCES

Andon Labs. Vending-Bench 2. https://andonlabs.com/evals/vending-bench-2, 2025. Andon Labs evaluation report.

Tim Baarslag, Katsuhide Fujita, Enrico H. Gerding, Koen V. Hindriks, Takayuki Ito, Nicholas R. Jennings, Catholijn M. Jonker, Sarit Kraus, Raz Lin, Valentin Robu, and Colin R. Williams. Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence*, 198:73–103, 2013. doi: 10.1016/j.artint.2012.09.004.

Tim Baarslag, Koen V. Hindriks, and Catholijn M. Jonker. Effective acceptance conditions in real-time automated negotiation. *Decision Support Systems*, 60:68–77, 2014. doi: 10.1016/j.dss.2013.05.021.

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.

Axel Backlund and Lukas Petersson. Vending-Bench: A benchmark for long-term coherence of autonomous agents, 2025. URL https://arxiv.org/abs/2502.15840.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.172. URL https://doi.org/10.18653/v1/2024.acl-long.172.

Kalyan Chatterjee and William Samuelson. Bargaining under incomplete information. *Operations Research*, 31(5):835–851, 1983. URL https://www.jstor.org/stable/170889.

Yanxu Chen, Zijun Yao, Yantao Liu, Amy Xin, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. StockBench: Can LLM agents trade stocks profitably in real-world markets?, 2026. URL https://arxiv.org/abs/2510.02209.

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.

Guillermo Gallego and Garrett van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, August 1994. ISSN 1526-5501. doi: 10.1287/mnsc.40.8.999. URL http://dx.doi.org/10.1287/mnsc.40.8.999.

John C. Harsanyi. Games with incomplete information played by "Bayesian" players, part I: The basic model. *Management Science*, 14(3):159–182, 1967.

Muyu He, Adit Jain, Anand Kumar, Vincent Tu, Soumyadeep Bakshi, Sachin Patro, and Nazneen Rajani. YC-Bench: Benchmarking AI agents for long-term planning and consistent execution, 2026. URL https://arxiv.org/abs/2604.01212.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=Zy4uFzMviZ.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://doi.org/10.1162/tacl_a_00638.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: an experimental study. *J. Comput. Soc. Sci.*, 9(1):11, 2026. doi: 10.1007/S42001-025-00435-2. URL https://doi.org/10.1007/s42001-025-00435-2.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, Zixuan Hu, Hongze Mi, Yibo Wang, Naiqiang Tan, Hong Chen, Yi R. Fung, Chun Yuan, and Li Shen. UltraHorizon: Benchmarking agent capabilities in ultra long-horizon scenarios, 2025. URL https://arxiv.org/abs/2509.21766.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): from tool use to agentic evaluation of large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374.

J. Andrew Petersen and V. Kumar. Are product returns a necessary evil? antecedents and consequences. *Journal of Marketing*, 73(3):35–51, May 2009. ISSN 1547-7185. doi: 10.1509/jmkg.73.3.035. URL http://dx.doi.org/10.1509/jmkg.73.3.035.

Hannes M. Petrowsky, Lea Boecker, Yannik A. Escher, Marie-Lena Frech, Malte Friese, Adam D. Galinsky, Brian Gunia, Alice J. Lee, Michael Schaerer, Martin Schweinsberg, Meikel Soliman, Roderick Swaab, Eve S. Troll, Marcel Weber, and David D. Loschelder. The power and peril of first offers in negotiations: A conceptual, meta-analytic, and experimental synthesis. *Organizational Behavior and Human Decision Processes*, 191:104448, 2025. doi: 10.1016/j.obhdp.2025.104448.

Lingfei Qian, Xueqing Peng, Yan Wang, Vincent Jim Zhang, Huan He, Hanley Smith, Yi Han, Yueru He, Haohang Li, Yupeng Cao, Yangyang Yu, Alejandro Lopez-Lira, Peng Lu, Jian-Yun Nie, Guojun Xiong, Jimin Huang, and Sophia Ananiadou. When agents trade: Live multi-market trading benchmark for LLM agents, 2025. URL https://arxiv.org/abs/2510.11695.

Howard Raiffa. *The Art and Science of Negotiation*. Harvard University Press, 1982.

Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000. doi: 10.1145/355112.355122. URL https://doi.org/10.1145/355112.355122.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982. doi: 10.2307/1912531.

Qiming Shi, Yulong Tao, Linbo Jin, Zhaolu Kang, Yibo Dou, Jiawen Zhu, Tianjun Pan, Shaokang Fu, Chengyu Wang, Siyue Li, Yaping Cheng, Di Weng, and Chengfu Huo. MerchantBench: Benchmarking LLM agents for long-term coherence in e-commerce operations, 2026. URL https://arxiv.org/abs/2607.28956.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=1i6ZCvflQJ.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

Yoshiki Takenami, Yin Jou Huang, Yugo Murawaki, and Chenhui Chu. How does cognitive bias affect large language models? a case study on the anchoring effect in price negotiation simulations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4481–4498, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.240. URL https://aclanthology.org/2025.findings-emnlp.240/.

Kian Siong Tey, Michael Schaerer, Nikhil Madan, and Roderick I. Swaab. The impact of concession patterns on negotiations: When and why decreasing concessions lead to a distributive disadvantage. *Organizational Behavior and Human Decision Processes*, 165:153–166, 2021. doi: 10.1016/j.obhdp.2021.05.003.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL https://arxiv.org/abs/2305.16291.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks, 2025. URL https://arxiv.org/abs/2412.14161.

Shu Yang, Shenzhe Zhu, Zeyu Wu, Keyu Wang, Junchi Yao, Junchao Wu, Lijie Hu, Mengdi Li, Derek F. Wong, and Di Wang. Fraud-R1 : A multi-round benchmark for assessing the robustness of LLM against augmented fraud and phishing inducements. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4374–4420, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.226. URL https://aclanthology.org/2025.findings-acl.226/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.773. URL https://aclanthology.org/2024.acl-long.773/.

Erica Zhang, Fangzhao Zhang, Aneesh Pappu, Batu El, Jose Blanchet, Susan Athey, Jiashuo Liu, and James Zou. TERMS-Bench: Diagnosing LLM negotiation agents beyond deal rate, 2026a. URL https://arxiv.org/abs/2605.13909.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Godel machine: Open-ended evolution of self-improving agents, 2026b. URL https://arxiv.org/abs/2505.22954.

Linghua Zhang, Jun Wang, Jingtong Wu, and Zhisong Zhang. RetailBench: Evaluating long-horizon autonomous decision-making and strategy stability of LLM agents in realistic retail environments, 2026c. URL https://arxiv.org/abs/2603.16453.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 19724–19731. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

Deyao Zhu, Xin Zhou, Shengling Qin, Xuekai Zhu, Hangliang Ding, Shu Zhong, Zixin Wen, Zhonglin Xie, Chenhui Gou, Linxuan Ren, Yueyang Wang, Junfeng Zhong, Rui Liu, Tian Gao, Yangguang Lin, Jingyuan Zhang, Maojia Song, Xuan Qi, Jinhong Wu, Chenyang Zhang, Yinzhu Piao, Ziru Niu, Hongbin Lin, Lingxiang Meng, Peng Tang, Chengyao Tang, Shanyu Wu, Huanyu Zheng, Yu Liu, Liya Zhu, He Wang, Ming Ding, Ziyu Wan, Hao Liu, Sibo Wang, Haotian Zhu, Xintian Zhang, Nan Chai, Yipeng Liu, Panhao Lai, Sihang Yuan, Zixin Su, Ge Zhang, Wangchunshu Zhou, Yantao Du, Wenhao Huang, and Guang Shi. EdgeBench: Unveiling scaling laws of learning from real-world environments, 2026. URL https://arxiv.org/abs/2607.05155.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet S. Sekhon, Jacob Steinhardt, Sarah Schwettmann, Arvind Narayanan, Matei A. Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices in building rigorous agentic benchmarks. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr

Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/f316275b44ee2de533102913828a8107-Abstract-Datasets_and_Benchmarks_Track.html.

A AGENT HARNESS AND INTERFACE

The harness defines what the agent can observe and how long it can retain those observations. Models interact through tool schemas and replies, with environment notifications embedded in the replies. This section specifies response formats and eviction within the 128,000-token context window. Sections 3.2 and 3.3 summarize the interface; Table 3 lists tool costs, Appendix H gives the prompts, and Section I.1.2 illustrates eviction.

Table 3: The 18-tool action space, grouped as in Figure 2, with the simulated minutes each call charges against the 600-minute working day.

Tool	min	Effect on environment state
Market Research		
`market_search`	30	Three disclosure levels over store types, their categories, and one category's sales band with qualitative return and margin labels
`supplier_search`	10	Name, contact, and categories served, no price, rating, or honesty signal
Store Operations		
`trace_return_sources`	10	Attributes pooled warehouse stock to the suppliers that delivered it
`open_store`	60	Opens one of the 12 store types for the ¥500 setup fee, at most 4 at once
`close_store`	30	Shuts a store for good, moving shelf stock to the warehouse or salvaging it at 10% of purchase cost
`check_store_status`	10	Yesterday's revenue, returns, shipping cost, and per-SKU velocity for one store, or a roll-up over all
`set_prices`	10	Rewrites retail prices on a shelf, effective at the next daily trigger
`join_promotion`	10	Enrolls a store in a calendar promotion at a discount of 0.05 to 0.50
Inventory Management		
`list_products`	10	Sellable SKUs with reference price and size, by store type or category
`check_warehouse`	10	Lists warehouse quantities with the purchase price paid for them
`publish_to_store`	20	Lists warehouse SKUs on a store page, storage fees still accruing
`return_to_warehouse`	15	Delists shelf units back into the warehouse allocation pool
`ship_orders`	20	Charges shipping, books net revenue into a 9-day escrow batch, and seeds the order's returns from the chosen speed
Financial Management		
`check_balance`	10	Reads bank, wallet, escrow, unshipped sales value, and settlement dates
`withdraw`	10	The only way cash leaves the platform wallet for the bank account
Supplier Negotiation		
`chatbox`	30	One free-form message to a supplier, or to many at the same charge, routing fenced `negotiate` actions to the Negotiation Kernel (§3.5)
Memory Operations		
`operate_memory`	10	Adds, reads, updates, or deletes one of 20 memory entries held outside the context window
Time Advancement		
`wait_for_next_day`	0	Jumps to 08:00 tomorrow, forfeiting the minutes left today, and fires the daily pipeline

A.1 THE INTERFACE CONTRACT

Every tool call returns a single string: structured results are serialized as JSON, while plain text is returned unchanged. Environment notices are included as JSON fields so the structured result remains parseable. Daily notices attach to the day-advance result if one exists in the batch, or to the last response otherwise. This associates the notice with the corresponding day boundary.

The morning update computes 17 fields but exposes at most three to the agent: the date, day index, and news feed. Across the 90 archived episodes, this channel carried 8,301 items, including 6,722 delivery notices aggregated into one line per day.

Listing 1 shows two real tool messages, one at a day boundary and one carrying the solvency warning that preceded a bankruptcy.

```
{"current_time": "2026-01-09T08:00:00", "day": 8, "system_notifications": [{"date":
    "2026-01-09", "day": 8, "news": [{"type": "delivery", "content": "Deliveries arrived:\n
    - 20x DreamMist LED Ceiling Light for Living Room and Bedroom, Mod from Crystal Trading
    "}]}]}
<system_warning>Token usage: 108593/128000 tokens (84%); 19407 remaining</system_warning>

{"pending_shipments": [], "count": 0, "note": "Ship with ship_orders (speed: fast/standard/
    slow). Faster ship costs more but reduces returns; slow is cheap but raises returns.
    Unshipped orders cancel after the deadline (lost sale + reputation hit).", "current_time
    ": "2026-01-03 16:40", "balance_reminder": "Your current balance (¥10.00) is below the
    daily operating cost (¥60.00). If your bank account stays negative for 10 consecutive
    days, you will go bankrupt."}
<system_warning>Token usage: 93856/128000 tokens (73%); 34144 remaining</system_warning>
```

Listing 1: Two tool messages exactly as they entered the request, from GPT-5.6 Sol run 0 on day 8 and from GPT-5.5 run 1 on 2026-01-03. The first carries the day notice, the second a low-balance reminder, and both close with the token gauge that sits outside the JSON. The second run went bankrupt.

A.2 Loop Control and Context Editing

Token counts come from a local tokenizer rather than the provider's own report, and cover the text, the tool calls and the reasoning trace of each message. Opaque artifacts such as replayed encrypted reasoning items are excluded, as are the tool schemas held outside the budget (§I.1.2). The system prompt specifies the 120,000-token threshold, the minimum release of 60,000 tokens, and the two protected groups. One pass proceeds as follows.

1. Sum the token counts of all messages not already cleared. Below the threshold, report usage without removing any messages.
2. Scan for groups from the third message onward, leaving the system message and the first user turn permanently protected. A group starts with an assistant message carrying tool calls and ends after its consecutive tool responses.
3. Set aside the two most recent groups not yet cleared. With two or fewer active groups the pass clears nothing at all, so a history made of a few very large groups can sit above the ceiling indefinitely.
4. Set the release target to the larger of the minimum release and the amount by which the current token count exceeds the threshold.
5. Clear eligible groups in chronological order until their combined token count reaches the release target.
6. Rebuild the request from the surviving messages, skipping cleared entries entirely (§I.1.2).

Eviction frequency depends on the model's context usage. Figure 5 reports the counts for all 90 episodes.

Plain-text assistant turns and idle warnings are never evicted: they neither start a group nor belong to one, so they accumulate throughout the episode. Evicting whole groups keeps each tool call together with its response, as required by the provider protocols used here.

B Data Layer

The data layer fixes the catalog and market conditions shared by all models and runs: 6,886 SKUs, 60 category parameter rows, 576 suppliers, 12 store types, 10 market events, and 8 promotions. The following sections describe these inputs and what tools expose to the agent. Appendix D specifies how supplier attributes enter bargaining.

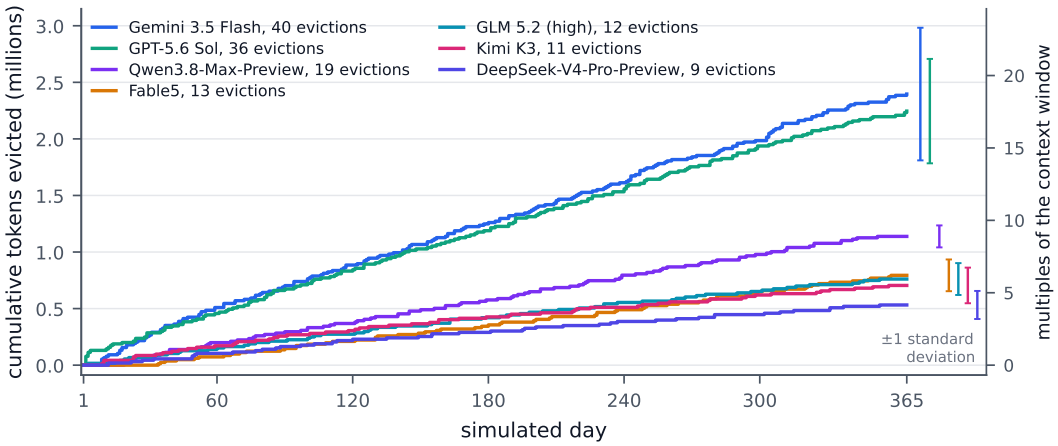


Figure 5: How often the context editor fired, over the 90 evaluated episodes. Each event is one eviction pass. A model that never reaches the ceiling evicts nothing.

Table 4: The 60 released categories, grouped by the store type that may list them, each heading giving that type's difficulty tier and daily operating cost. Reference-price bands are platform-wide and per SKU, and the agent never sees the numeric return band its categories carry. † marks the 20 categories whose agent-facing return sentence understates that band, which Table 5 gives per store type.

Category	Reference price (¥)
Auto & Hardware, T1, ¥130/day	
Auto Accessories	10–400
Hardware & Tools	5–200
Electrical Supplies	3–80
Electronic Components	1–50
E-Vehicles & Parts	100–1,200
Agricultural Supplies	10–400
Industrial & Lab Supplies	10–200
Daily & Office, T1, ¥130/day	
Cleaning & Hygiene†	5–60
Gifts & Party Supplies†	5–80
Lifestyle Accessories†	20–400
Stationery & Office Supplies†	3–80
Office Equipment†	20–600
Medical Devices†	10–800
Food & Beverage, T1, ¥130/day	
Snacks & Nuts†	8–60
Groceries & Staples†	10–80
Coffee & Beverages†	10–60
Fresh Food & Meat†	15–120
Alcoholic Beverages†	20–500
Health Supplements†	30–200

Category	Reference price (¥)
Home & Living, T1, ¥130/day	
Furniture	100–800
Bedding†	30–400
Household Essentials†	3–50
Storage & Organization†	10–100
Lighting†	15–500
Home Building Materials†	20–400
Basic Construction Materials†	10–200
Cleaning Tools†	5–50
Kitchenware†	8–200
Appliance & Digital, T2, ¥100/day	
Major Appliances	500–2,500
Small Appliances	50–600
Audio & Video Electronics	100–1,500
Smartphones	300–3,000
Laptops	800–5,000
Computer Hardware & Peripherals	30–800
3C Digital Accessories	5–150
Storage Devices	20–400
Networking Equipment	30–500
Fashion, T2, ¥100/day	
Women's Fashion	30–300
Men's Clothing	30–250
Underwear & Loungewear	15–120
Fashion Accessories	10–150
Sportswear & Casual	30–250

Category	Reference price (¥)
Mother & Baby, T2, ¥100/day	
Children's Clothing	15–100
Children's Shoes	20–120
Baby Products	15–400
Maternity Products	20–200
Toys & Educational	15–600
Shoes & Bags, T2, ¥100/day	
Women's Shoes	40–400
Men's Shoes	50–400
Sneakers	80–600
Bags & Luggage	50–1,000
Beauty, T3, ¥60/day	
Skincare & Beauty	15–400
Beauty Devices	80–1,000
Personal Care & Massage	30–600
Pet Supplies, T3, ¥60/day	
Pet Supplies	8–150
Sports & Outdoor, T3, ¥60/day	
Sports & Fitness	15–300
Sports Bags & Accessories	30–400
Outdoor & Camping	20–600
Toys & Entertainment, T3, ¥60/day	
Collectibles & Figures	15–500
Gaming & Accessories	30–1,000

B.1 Category Parameters

Several parameters that determine margin are hidden from the agent: the three per-category price ratios, the elasticity family and its parameter, and the return-rate bounds. A SKU's natural return rate becomes available only through `trace_return_sources` once stock has moved (§I.5.2). Each category belongs to exactly one store type. Table 4 groups all 60 categories by type and gives each type's difficulty tier and daily operating cost; Table 5 gives its return band and cost-floor ratio. Setup fee and commission are identical across the 12, leaving the seasonal curve and the demand terms of §3.4 as the only differences the two tables do not show. Reference price and natural return rate vary by SKU within each category's bounds, while base demand and the elasticity family are shared within a category. The wholesale ratio is pinned to the cost-floor ratio c_j/v_j in all 60 rows (§D.2).

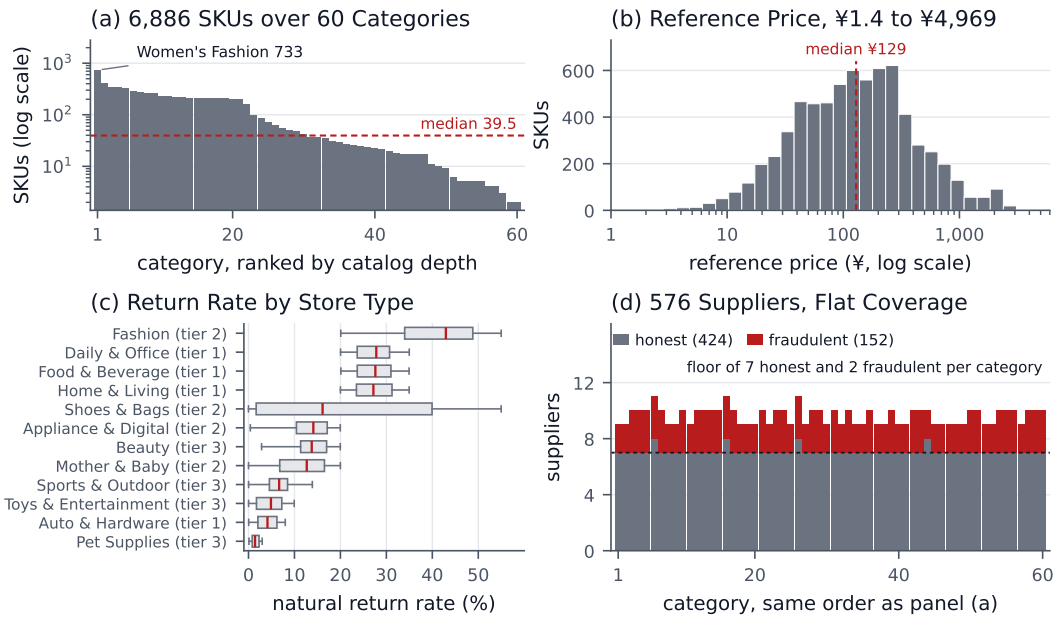


Figure 6: The released data. (a) Products per category. (b) Reference prices. (c) Natural return rate by store type, T the difficulty tier. (d) Suppliers per category against catalog depth.

B.2 Store Types, Seasonality, and Physical Cost

The merchant chooses a store type at setup; changing it requires another setup fee. Table 5 lists all 12 types with their difficulty tiers, operating costs, and supported categories.

Table 5: The 12 store types, by difficulty tier. Reference-price band, category count, seasonal range, tier and operating cost are visible to the agent, while the wholesale ratios are not, so the achievable cost must be discovered by negotiating. Opening ratio is a supplier's first quote over the reference price and floor ratio the lowest an honest supplier accepts, averaged over the store's categories, so their gap is the room negotiation can win. Return rate is natural, before pricing or shipping move it.

Store type	Cats	Ref. price (¥)	Return rate (%)	Opening / floor	Seasonal range	Ops (¥/day)
difficulty tier 1 hard						
Auto & Hardware	7	1 to 1,200	0 to 8	0.93 / 0.86	0.90 to 1.10	130
Daily & Office	6	3 to 800	20 to 35	0.77 / 0.54	1.00 to 1.20	130
Food & Beverage	6	8 to 500	20 to 35	0.78 / 0.55	0.60 to 1.60	130
Home & Living	9	3 to 800	20 to 35	0.86 / 0.72	0.90 to 1.30	130
difficulty tier 2 medium						
Appliance & Digital	9	5 to 5,000	0 to 20	0.79 / 0.58	0.90 to 1.40	100
Fashion	5	10 to 300	20 to 55	0.70 / 0.39	0.70 to 1.55	100
Mother & Baby	5	15 to 600	0 to 20	0.70 / 0.40	1.00 to 1.20	100
Shoes & Bags	4	40 to 1,000	0 to 55	0.80 / 0.60	0.80 to 1.40	100
difficulty tier 3 easy						
Beauty	3	15 to 1,000	0 to 20	0.69 / 0.38	1.00 to 1.30	60
Pet Supplies	1	8 to 150	0 to 3	0.61 / 0.22	1.00 to 1.10	60
Sports & Outdoor	3	15 to 600	0 to 20	0.58 / 0.18	0.80 to 1.30	60
Toys & Entertainment	2	15 to 1,000	0 to 10	0.65 / 0.29	1.00 to 1.70	60

Figure 7 gives the seasonal grid in full, one multiplier per store type and calendar month. Amplitude, not level, separates the types. Nine of the 12 attain their annual maximum in November, inside the same weeks as the calendar's longest promotion window (Table 6), while Food & Beverage inverts the pattern and peaks in February, so these store types favor different inventory schedules. Four types stay within ± 0.2 of neutral all year, indicating relatively mild seasonal variation. The index returned by `market_search` is exactly $100\,\sigma$. Freight and daily storage costs scale with parcel size. The

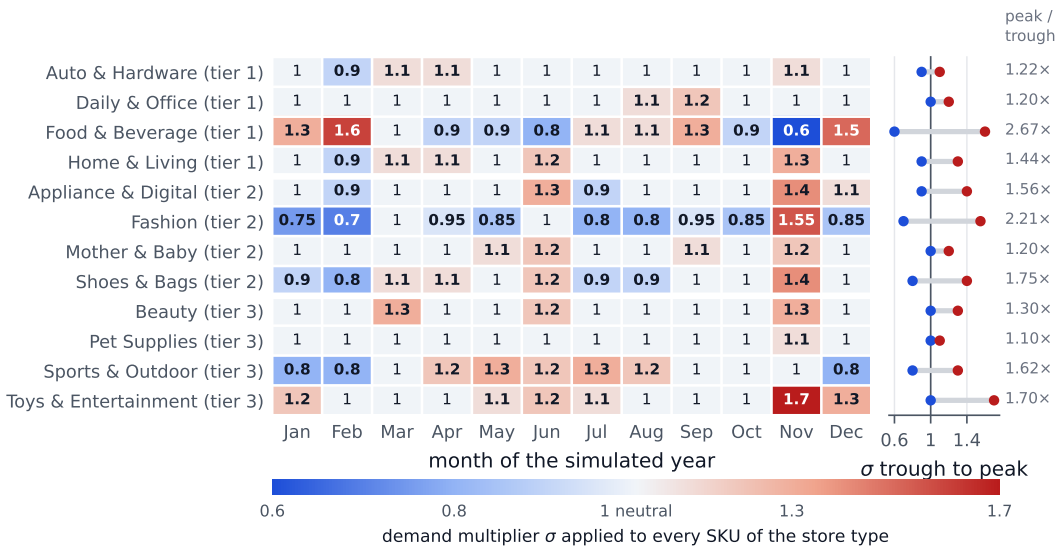


Figure 7: The monthly seasonality multiplier for all 12 store types, rows grouped by difficulty tier. The right panel gives each type's trough, peak and peak-to-trough ratio.

agent's choice of shipping speed also changes both freight cost and return probability. Appendix C gives all four schedules with the rest of the cost model.

B.3 Event and Promotion Calendar

Events are announced on their first day. Promotions are announced up to seven days in advance, and `join_promotion` accepts a target 30 days ahead, allowing advance inventory planning. The calendars overlap only when the Guangdong factory fire of March 1 to 7 coincides with the March 4 to 8 Spring Blossom Sale. This event affects supply rather than demand, so it does not directly change the promotion's demand multiplier.

An event with no demand term affects only purchase orders placed during its active window. Lead time is fixed when an order is placed, so goods already in transit keep their arrival dates. Table 6 lists all dates, phases, per-store-type demand multipliers, and delivery-delay factors.

B.4 The Supplier List

The fixed supplier list gives every model the same mix of counterpart difficulties. Each row records an identity, an honest-or-fraudulent label, a behavior template, a scam type if applicable, one served category, and a bankruptcy threshold measured in orders. There are 424 honest and 152 fraudulent suppliers across 60 categories, with 7 or 8 honest and 2 or 3 fraudulent suppliers per category. Supplier names do not reveal these labels, but email addresses do (§B.5). Appendices D.3 and D.5 specify the behavior templates and scams.

B.5 Properties a Reader Should Check

We checked four data inconsistencies against the code that reads these files. Two affect what an agent can infer from the available information.

1. **Return-rate wording understates the numeric band.** Twenty of the 60 categories carry a sentence reading low or very low against a numeric band reaching 20% to 35%, marked in Table 4. The sentence is what `market_search` shows, so a merchant who prices off the wording underestimates returns across categories holding 25.5% of all SKUs, and sees the gap only through `trace_return_sources` on stock already owned.

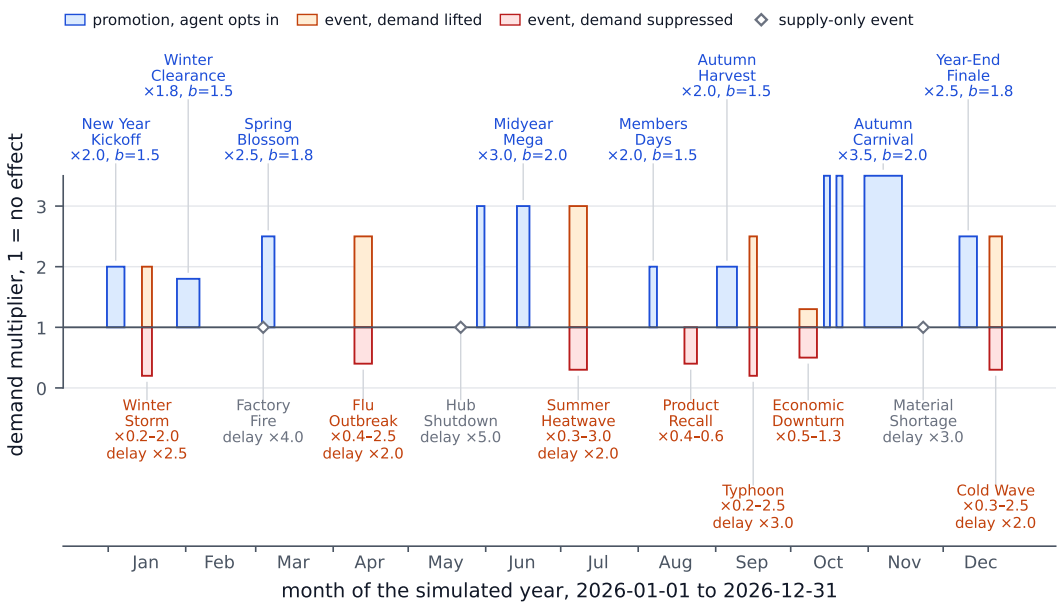


Figure 8: The 2026 calendar. Eight promotions the agent can join by choosing a discount, and ten events it cannot opt out of. Promotion bars show the multiplier a 30% discount buys, event bars their extreme per-store-type multiplier.

Table 6: The fixed calendar of the simulated year, 10 events above and 8 promotions below, store types named by their first word. Event demand multipliers are keyed by store type rather than by category, so an event misses a store whose type it does not name, and only the largest active delay applies. A promotion moves demand only if the agent joins it, at the ceiling shown for a 30% discount.

Market events, unavoidable Event	Window (2026)	Demand multiplier by store type	Supply effect
Winter Storm	Jan 15 to Jan 18	Food ×2, Home ×1.5, Sports ×0.2, Fashion ×0.3	lead time ×2.5, all store types
Factory Fire in Guangdong	Mar 1 to Mar 7	none	lead time ×4, Appliance, Toys
Flu Outbreak	Apr 10 to Apr 16	Daily ×2.5, Beauty ×1.5, Sports ×0.4, Toys ×0.5	lead time ×2, Daily
Logistics Hub Shutdown	May 20 to May 24	none	lead time ×5, all store types
Summer Heatwave	Jul 5 to Jul 11	Appliance ×3, Food ×2, Sports ×0.3, Home ×0.5	lead time ×2, Appliance
Celebrity Product Recall	Aug 20 to Aug 24	Beauty ×0.4, Fashion ×0.6	none
Typhoon Warning	Sep 15 to Sep 17	Food ×2.5, Daily ×2, Fashion ×0.2, Shoes ×0.2	lead time ×3, all store types
Economic Downturn News	Oct 5 to Oct 11	Food ×1.3, Daily ×1.2, Appliance ×0.5, Shoes ×0.5, Fashion ×0.6	none
Raw Material Shortage	Nov 20 to Nov 26	none	lead time ×3, Appliance, Home, Auto; cost term inert
Cold Wave	Dec 20 to Dec 24	Home ×2.5, Fashion ×2, Sports ×0.3, Food ×0.4	lead time ×2, all store types

2. **The supplier email leaks the fraud label.** `supplier_search` returns an email whose local part ends in the supplier identifier, and inside every one of the 60 categories the honest suppliers hold lower identifiers than the fraudulent ones, so an agent sorting on that numeric suffix could screen fraudulent counterparts without sending a single message. Section I.5.2 calls the supplier

Promotions, which the agent opts into with a discount			
Promotion	Windows (2026), every phase listed	Max demand	Elasticity boost
New Year Kickoff Sale	Jan 1 to Jan 7	×2	×1.5
Winter Clearance Festival	Jan 29 to Feb 6	×1.8	×1.5
Spring Blossom Sale	Mar 4 to Mar 8	×2.5	×1.8
Midyear Mega Sale	May 29 to May 31, Jun 14 to Jun 18	×3	×2
Summer Members Days	Aug 6 to Aug 8	×2	×1.5
Autumn Harvest Sale	Sep 2 to Sep 9	×2	×1.5
Grand Autumn Carnival	Oct 15 to Oct 16, Oct 20 to Oct 21, Oct 31 to Nov 14	×3.5	×2
Year-End Finale Sale	Dec 8 to Dec 14	×2.5	×1.8

list free of any ranking that sorts honest from fraudulent, which holds for the returned fields and their order while the suffix carries the signal anyway. Whether any evaluated model exploited it is read in §F.3, whose contact stage would show the effect.

The other two inconsistencies concern unused data fields. The released rent column lists ¥45 for the four easy-tier rows, while the environment charges ¥60 from the difficulty tier; using the column would understate those costs by a quarter. The cost-increase field for Raw Material Shortage is also unused. The November shortage therefore extends lead times without changing procurement prices, despite its news description.

Table 7 summarizes which environment quantities tools expose. The agent must infer the hidden quantities from outcomes, making sourcing and pricing challenging (Section 4).

C Economy Engine

This section formalizes the economy in §3.4, following the simulation's execution order. We use i for a store, j for a SKU, t for a simulated date, $\tau = \tau(i)$ for store type, and $c = c(j)$ for category. The prices are the public reference v_j, the paid purchase price c_j^{buy}, and supplier s's private floor c_j^s. Unit profit Π is distinct from the normalized price position π in §I.4.5. Appendix C.7 collects the remaining constants; Appendix B gives the seasonal grid.

C.1 The Daily Settlement Order

Settlement occurs whenever the clock crosses 08:00. Thirteen steps run in a fixed order at each boundary. Figure 9 lists the steps and the state each reads and writes; these internal updates are not shown to the agent (§A.1).

Costs are charged before any revenue is recognized, so solvency is tested each morning at the day's lowest balance. Sales for the *previous* calendar date use the inventory and prices the agent left in place, creating a one-day delay between decisions and sales reports.

An episode ends at the calendar horizon or when the unpaid streak reaches its limit (§C.7). Finalization cancels unshipped orders as lost sales, processes all seeded returns against their escrow batches, and transfers the remaining escrow to the platform wallet. Late sales therefore remain subject to returns even when those returns fall beyond the episode horizon.

C.2 Demand Model

Base demand starts from the category and is scaled twice, with $M_c^{\min}$ and $M_c^{\max}$ the platform-wide monthly sales range of category c.

$$\beta_{c\tau} \;=\; \frac{1}{30} \cdot \frac{M_c^{\min} + M_c^{\max}}{2} \;\cdot\; \lambda \;\cdot\; \psi_\tau, \qquad \lambda = 0.1, \tag{1}$$

The factor λ converts platform volume to store-level demand, and ψ_τ calibrates each store type, ranging from 0.2 for Shoes & Bags to 9.0 for Food & Beverage. All SKUs in a category share $\beta_{c\tau}$.

Table 7: Systematic information asymmetry. Tools disclose catalog facts and the agent's own realized outcomes, while the demand and cost-floor parameters that determine whether a decision pays stay latent, each priced in simulated minutes, in ¥, or in an irreversible loss. Minute costs are the clock charge of one call to the named tool, against a 600-minute working day.

Observable, and the tool that shows it	**Hidden**	**Discovery path, and what it costs**
Per-SKU reference price and per-category price band, plus a low/moderate/high margin label (`list_products` 10 min, `market_search` 30 min)	Supplier cost floor c_j^s, for an honest supplier between 0.17 and 0.90 of the reference price, and the nominal wholesale ratio	Bargain for it. The opening quote is calibrated to the nominal wholesale price, and every closed deal yields an upper bound on the floor while the best of them tightens it (`chatbox` 30 min per call, $K = 10$ rounds, and one call broadcast to any number of suppliers still costs 30 min)
A one-line qualitative return-rate note per category (`market_search` 30 min)	Per-SKU natural return rate, spanning 0.000 to 0.550	`trace_return_sources` prints the baseline rate, but only for a SKU already bought or sold (10 min), so the number costs one purchase
Own retail price and yesterday's units sold per SKU (`check_store_status` 10 min)	Elasticity family, one of four, and its curvature parameter	Price experiments (`set_prices` 10 min). One SKU-day yields one observation, confounded by the weekend uplift, the event multiplier, and two saturation terms, none of which any tool prints
Difficulty tier, a qualitative profit-potential grade, daily operating cost, the 12-month sales index, and a platform-wide sales range (`market_search` 30 min)	Store-type demand capacity, the per-category sub-cap, and the per-store share of platform volume, which jointly place the profit ceiling	Open the store (¥500 and 60 min) and run it. One year affords a handful of such trials, each returning one noisy trajectory
Supplier name, email, and categories served, listed in seed-shuffled order (`supplier_search` 10 min)	Behavior template of the supplier's kernel, one of six, with its urgency and stance	Read the concession sequence across rounds and across reopened cycles (`chatbox` 30 min per call)
No type label anywhere in the supplier list, and quotes from the 152 fraudulent suppliers sit inside the honest range by construction	Honest or fraudulent, and which of the five scams	Bargaining is where the scams first act, three of the five defending a floor inflated by 1.5×. A short delivery surfaces in `check_warehouse` (10 min) and defective stock in an elevated realized return rate, both after the cash has left
Own order count with each supplier, which no tool reports and which the agent must carry in its own transcript	That supplier's exit threshold, between 10 and 20 completed orders	Reaching it, which retires a calibrated price anchor for the rest of the year
Dated news prose for each of the 10 market events, which quotes the delivery-delay factor outright in 3 cases, and a name-only promotion announcement roughly a week ahead, both injected into tool results at no time cost	Per-store-type event demand multipliers (0.2 to 3.0), the delay factors the news leaves unquantified (2.0 to 3.0 for those five, against 2.0 to 5.0 over all eight), and each promotion's demand ceiling (1.8 to 3.5) and elasticity boost	`join_promotion` returns the demand ceiling once a store has joined (10 min). Demand multipliers are never printed and must be read off realized sales

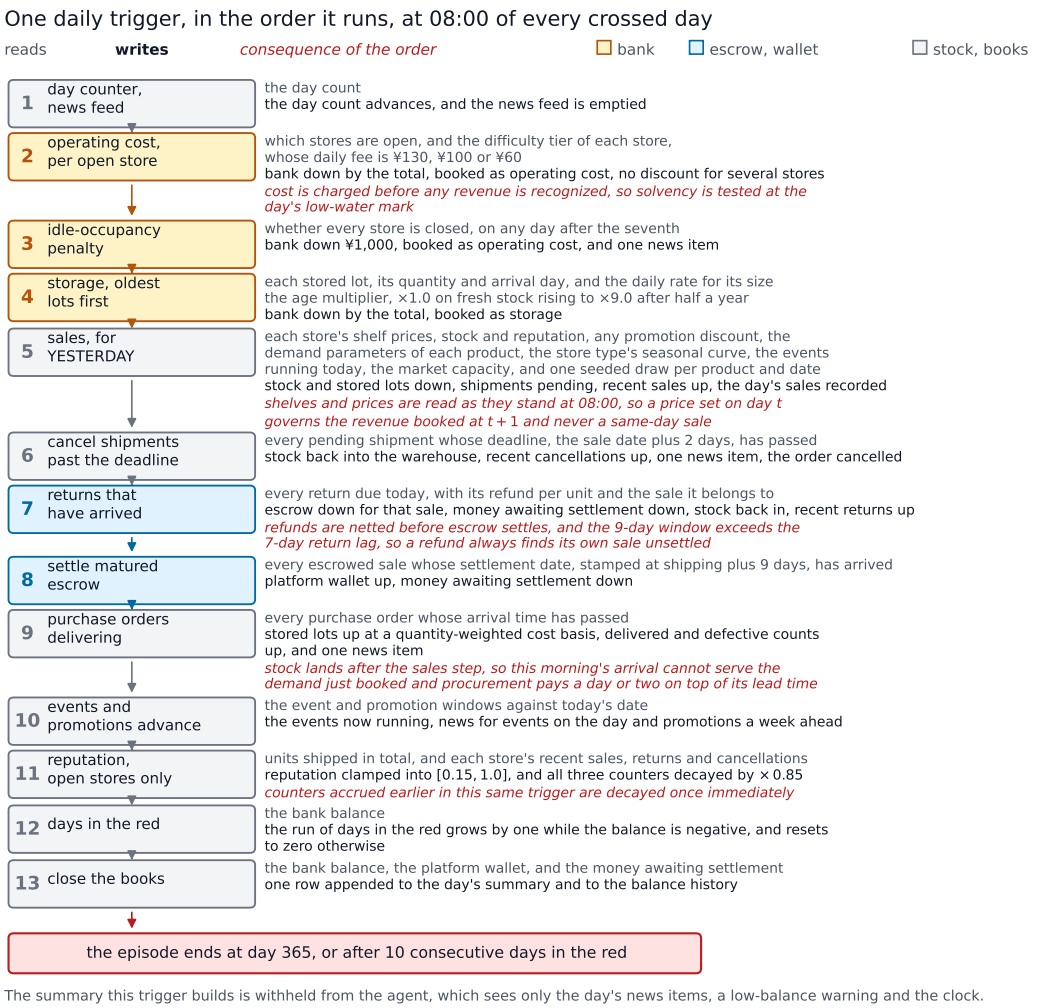


Figure 9: The thirteen settlement steps in the order the simulation runs them at each 08:00 crossing, with the state each step reads and writes. Fill colour marks which money bucket a step moves.

The price factor takes one of four closed forms, keyed by category and evaluated at the discounted price ratio $r_{ijt} = (1 - \delta_{it})\, p_{ijt}/v_j$.

$$\phi_f(r;\eta) \;=\; \begin{cases} \max\bigl(0,\; 1 - \eta\,(r-1)\bigr) & f = \text{LINEAR},\ \eta \in [2.185,\ 3.949],\ 11 \text{ categories}, \\ \exp\bigl(-\eta\,(r-1)\bigr) & f = \text{EXPONENTIAL},\ \eta \in [2.583,\ 4.944],\ 16, \\ r^{-\eta} & f = \text{CONSTANT-ELASTICITY},\ \eta \in [1.352,\ 2.898],\ 17, \\ \max\bigl(0,\; 1 - \eta\,(r-1)^2\bigr) & f = \text{QUADRATIC},\ \eta \in [2.795,\ 5.779],\ 16. \end{cases} \tag{2}$$

Only the quadratic branch is non-monotone: a deep discount can reduce demand as well as margin.

The remaining multipliers of Equation (47) are each read once per store-day, weekends contributing $w_t = 1.3$. A store joining promotion g at discount δ_{it} earns $m_{it} = 1 + (M_g - 1)\min(1, \delta_{it}/0.30)$ and scales η_c by b_{it}. Seasonality $\sigma_{\tau,t}$ reads the store type's month out of the grid Appendix B tabulates, active events compound into $\varepsilon_{\tau,t}$, and reputation ρ_{it} enters linearly.

Two saturation terms limit demand: the per-category factor γ_{ijt} in Equation (47) and the store-wide term. The per-category ceiling κ_c takes a fraction χ_τ of the store type's own daily capacity κ_τ. Capacity runs from 13.959 units for Appliance & Digital to 19,862.16 for Auto & Hardware, and a shorter category list raises the fraction, which reaches 1 for a store type allowed a single category. At

that point the per-category term is dropped, so $\gamma_{ijt} = 1$ and only store-wide saturation applies.

$$\chi_\tau = \min\Big(1,\ \max\big(0.35,\ 1.2/n_\tau\big)\Big), \qquad \kappa_c = \chi_\tau\,\kappa_\tau. \tag{3}$$

Both terms take the Michaelis-Menten form $\kappa/(\kappa + D)$, the first against one category's stacked demand and the second against the store's post-cap total $\hat{D}_{it}$.

Stochastic rounding uses a draw that is independent of tool-call order.

$$u_{jt} \sim \mathcal{U}(0,1) \quad \text{drawn from the environment seed hashed with the SKU and the date.} \tag{4}$$

The cap x_{ijt} of Equation (47) is on-shelf inventory, so a stockout truncates demand with no signal to the agent.

C.3 Reputation Law

Each settlement recomputes the reputation of every open store. Writing S_{it} for units shipped to date and R_{it}, C_{it}, N_{it} for the rolling counters of returned, cancelled and sold units,

$$\rho_{it} = \mathrm{clip}\Big(\underbrace{0.3 + \frac{0.7}{1 + e^{-(S_{it}-500)/200}}}_{\text{volume goodwill}} - \underbrace{\min\Big(0.5,\ 0.6\,\frac{R_{it}}{\max(1,N_{it})} + 1.0\,\frac{C_{it}}{\max(1,N_{it})}\Big)}_{\text{service penalty}},\ 0.15,\ 1.0\Big), \tag{5}$$

after which all three counters are multiplied by 0.85. Goodwill credits shipping rather than selling, so an unshipped order earns nothing and still takes the cancellation weight. Because goodwill starts from zero shipped units, a store that opens at 0.5 is recomputed to 0.353 at its first settlement. The same counters weigh a cancelled unit as 1.67 returned units, pricing a missed deadline above a change of mind (Figure 10).

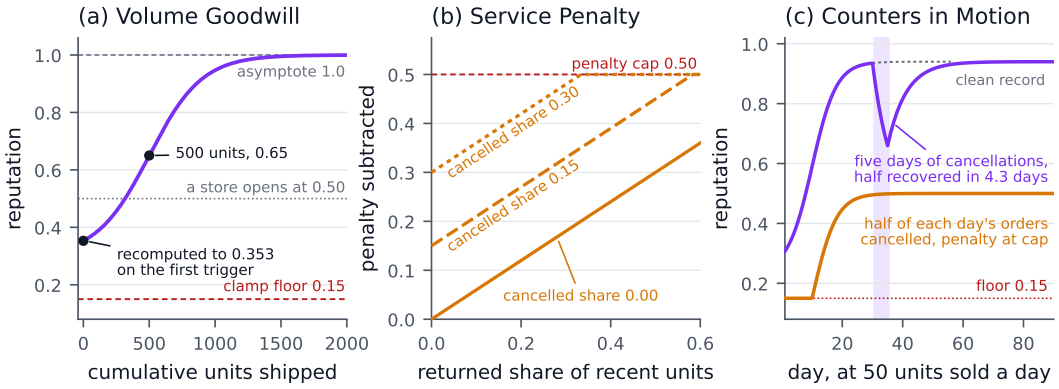


Figure 10: How reputation moves. (a) The goodwill term against units shipped. (b) The penalty term against the returned share of recent units, at three cancellation shares. (c) A service shock decays while a chronic failure rate does not.

C.4 Returns

An item's return rate combines four channels in a fixed order: three at sale time and one when the agent chooses a shipping speed. Write θ_j^{nat} for the SKU's natural rate and α_j for the delivered share that came from a quality-downgrading supplier. The price curve ζ is piecewise linear through the knees $(0.8, 0.85)$, $(1.0, 1.00)$, $(1.3, 1.50)$ and $(1.8, 2.20)$, held flat outside them.

$$\theta_j^{\mathrm{def}} = \min\big(0.95,\ \max(0.40,\ 2\,\theta_j^{\mathrm{nat}})\big), \qquad \theta_j^{(1)} = (1-\alpha_j)\,\theta_j^{\mathrm{nat}} + \alpha_j\,\theta_j^{\mathrm{def}},$$
$$\theta_j^{(2)} = \min\big(0.95,\ \theta_j^{(1)}\,\zeta(r_{ijt})\big), \qquad \theta_j = \min\big(0.95,\ \theta_j^{(2)} \cdot s\big), \quad s \in \{0.75,\ 1.00,\ 1.30\}, \tag{6}$$

The factor s is the return multiplier for fast, standard and slow shipping. Inventory is pooled by SKU and α_j is a ratio over cumulative deliveries, so a mixed lot leaves the elevated return rate untraceable.

Returns are sampled when a shipment leaves. A shipment of q units draws q seeded trials at rate θ_j and one return date uniformly from $\{3, \dots, 7\}$ days later, so its returned units arrive together. Each unit refunds the full retail price from the originating escrow batch and increments the return counter in Equation (5). It re-enters the warehouse as a new lot, resetting its storage age. Freight and commission are not refunded.

C.5 Cost Model

Six cost channels debit the bank account immediately. A seventh, the platform commission, is withheld from revenue before it enters escrow.

- **Setup**, ¥500 per store opened, charged again on every reopening.
- **Operating cost**, ¥130, ¥100 or ¥60 per open store per day (o_τ) by the store type's difficulty tier.
- **Idle occupancy**, ¥1,000 a day whenever no store is open after the seventh day.
- **Storage**, $h_j \cdot a(\text{age})$ per unit per day, where h_j is 0.05, 0.15, 0.50 or 1.50 by product size and a steps through $1.0, 1.4, 2.2, 4.0, 6.0, 9.0$ at ages $0, 21, 45, 90, 135, 180$ days. Lots are FIFO, and stock listed in a store keeps accruing.
- **Freight**, f_j, a size cost of 0.5, 1.5, 3.0 or 6.0 times a speed multiplier of 2.0, 1.0 or 0.5 for fast, standard and slow, charged per unit at ship time.
- **Commission**, $\gamma = 2\%$ of gross revenue, netted at sale rather than debited, and kept by the platform on refunded sales (§C.4).
- **Procurement**, the full purchase order at the moment a negotiation closes (§3.5). Any membership fee is included in the same debit; such fees are rarely paid (§D.5).

Liquidation is the one inbound channel besides sales, crediting 10% of the quantity-weighted purchase cost and removing the stock. Closing a store without liquidating leaves storage accruing.

C.6 Unit Profit and Break-Even

Composing §C.4 and §C.5 gives the contribution of one unit a customer finally keeps. Because a returned unit comes back as stock, each retained unit funds $1/(1-\theta_j)$ shipping attempts and as many storage cycles. A cycle covers the T_h days a unit waits in the warehouse before selling.

$$\Pi_j(r) \;=\; \frac{\big(1-\gamma-\theta_j(r)\big)\, r\, v_j \;-\; f_j \;-\; h_j\, T_h}{1-\theta_j(r)} \;-\; c_j^{\text{buy}}, \qquad n_j^\star(r) \;=\; \frac{o_\tau}{\Pi_j(r)}. \tag{7}$$

Break-even volume $n^\star$ counts the units a store must retain daily to cover its tier's operating cost. Figure 11a splits retail price into the five shares that add to one.

Procurement is the largest cost line, and the price a negotiation closes at moves unit profit further than any pricing choice does. Two median-reference SKUs carry the identity in Figure 11, one high-return and one thin-margin. The Women's Fashion SKU sits in a medium-tier store, returns 49.2% of units naturally, and is quoted $0.70\, v_j$ against a floor of $c_j^s = 0.39\, v_j$. At the floor price, standard shipping and $T_h = 14$ days, it earns ¥87.29 per retained unit and needs 2 units a day to cover operations. The Snacks & Nuts comparison in a hard-tier store needs 13. Closing at the floor rather than accepting the opening quote multiplies unit profit by 2.4 and 2.8.

C.7 Constant Reference

The remaining constants apply to the whole environment. The opening stake is ¥100,000, at most 4 stores may operate concurrently, and the 365-day year starts on 2026-01-01 with working hours from 08:00 to 18:00. Orders have a two-day shipping deadline; escrow settles after nine days; 10 consecutive days with a negative bank balance trigger bankruptcy. Every model and run uses environment seed 20260122. The store-type parameters κ_τ and ψ_τ are given as ranges in §C.2.

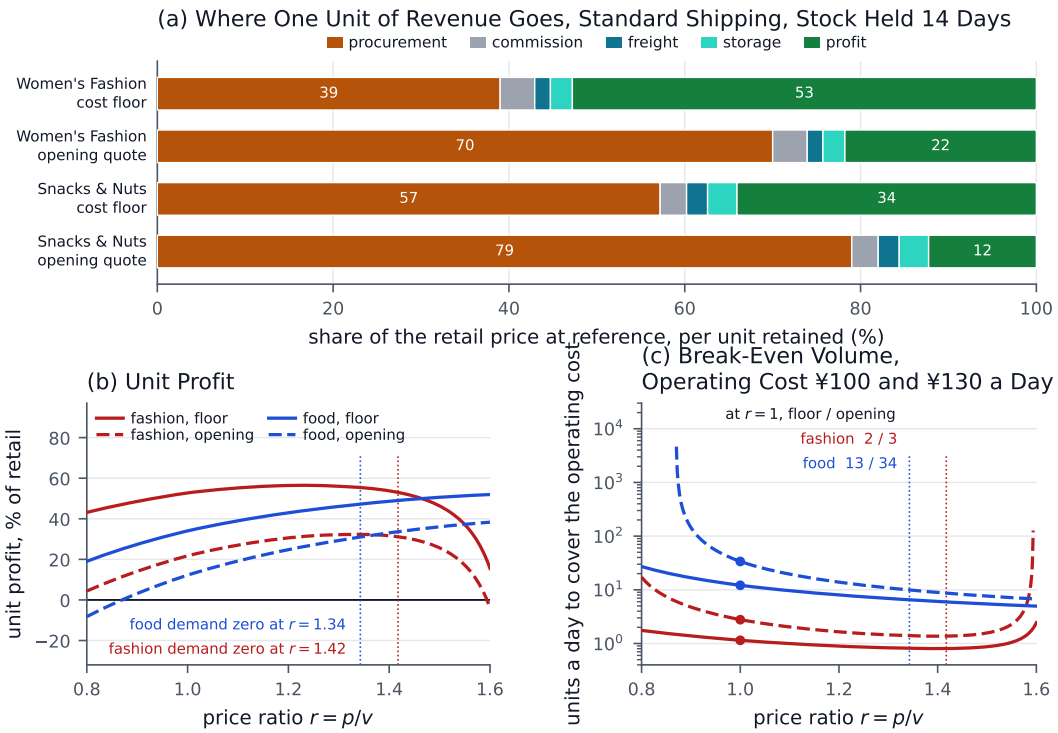


Figure 11: Where a unit's money goes. (a) Each component as a share of retail price, for a median-priced fashion SKU beside a snacks SKU. (b) Unit profit against the price ratio, with dotted lines where demand reaches zero. (c) Units a store must sell each day to cover its operating cost.

D Deterministic Negotiation Kernel

The supplier policy independently implements the counterpart design of Zhang et al. (2026a), as outlined in §I.4.2. Four decision functions govern bargaining, with five scam scripts modifying supplier behavior. We specify the functions, data calibration, templates, and transaction protocol below. Metrics appear in §E.1; renderer prompts appear in §H.

D.1 Kernel Decision Functions

A kernel instance tracks one supplier's bargaining state for one SKU. Every decision depends on this state, which contains a latent type $t_B = (r_B, \kappa_B, \eta_B)$, an opening harshness d_0, the deadline scale $K = 10$, a price frame, seeded random streams and both offer histories. Reservation price r_B comes from the data layer (§D.2), while urgency $\kappa_B \in [0, 1]$ and stance $\eta_B \in \{\text{conciliatory}, \text{neutral}, \text{aggressive}\}$ come from the supplier's template (§D.3). Indicators $\mathbf{1}_{\text{a}}$, $\mathbf{1}_{\text{c}}$ mark the outer two stances. Every quantity below is normalized by the width of the price frame,

$$p_{\max} = 1.5\, w_j, \qquad \mathcal{R} = \max(1,\, p_{\max} - p_{\min})\,, \qquad \bar{\delta}_k = \frac{p_k - r_B}{\mathcal{R}}, \tag{8}$$

where w_j is SKU j's nominal wholesale price and p_k the agent's offer in round k. The lower bound $p_{\min}$ sits at 0 for an honest supplier and at r_B for a fraudulent one, whose frame is therefore narrower. Favorability $\bar{\delta}_k$ measures the seller's margin over its own floor.

Three features of the agent's offer sequence $(p_1^A, \ldots, p_m^A)$ drive the reciprocal terms, computed over the last three consecutive differences, $D_m = \left(p_i^A - p_{i-1}^A\right)_{i=\max(2,\, m-2)}^{m}$,

$$\bar{c}_m = \frac{1}{|D_m|} \sum_{d \in D_m} \frac{\max(0, d)}{\mathcal{R}}, \qquad \bar{s}_m = \frac{1}{|D_m|} \sum_{d \in D_m} \frac{d}{\mathcal{R}}, \qquad g_m = \mathbf{1}\left[\frac{\max(0, d_m)}{\mathcal{R}} < \tau_{\text{rigid}}\right], \tag{9}$$

all three defined as 0 while $m < 2$, with $\tau_{\text{rigid}} = 0.10$. Concession magnitude $\bar{c}_m$ counts only movement toward the seller, signed speed $\bar{s}_m$ also counts retreat, and the rigidity flag g_m marks a latest step under a tenth of the frame.

The kernel tests acceptance first, then walk-away, and makes a counter-offer if neither occurs. Acceptance conditions are a standard component of automated negotiation (Baarslag et al., 2014); this kernel uses the following logistic rule:

$$\begin{aligned} a_k &= \mathbf{1}\big[\bar{\delta}_k \geq 0\big] \cdot \sigma\Big(\alpha\,\bar{\delta}_k + \beta\,\kappa_B - \gamma\left(1 - \sqrt{k/K}\right) + \rho_f\,\bar{s}_m + \xi_f\,g_m\Big)\,, \\ \alpha &= 6,\ \beta = 1,\ \gamma = 2, \end{aligned} \tag{10}$$

where σ is the logistic function and (ρ_f, ξ_f) are the template's reciprocity coefficients. Since $\rho_f \leq 0$ in every template of Table 8, conceding fast never improves the agent's own chance of acceptance.

Walk-away is possible only for an offer below the floor and from the midpoint of the deadline onward,

$$\begin{aligned} \omega_k &= \mathbf{1}\big[\bar{\delta}_k < 0\big]\ \mathbf{1}[k \geq k_{\text{walk}}] \cdot \sigma\left(\phi_0 + \phi_\Delta\left(-\bar{\delta}_k\right) + \phi_t\,\frac{k - k_{\text{walk}}}{K - k_{\text{walk}}}\right), \\ \phi_0 &= -4.5,\ \phi_\Delta = 30,\ \phi_t = 1.5, \end{aligned} \tag{11}$$

with $k_{\text{walk}} = \lceil K/2 \rceil = 5$, both draws swept in Figure 12. An agent bidding above the floor is never abandoned, so impasse on an honest session is almost always the agent's own rejection (§I.4.5).

Each counter-offer reduces the gap to the floor by a fraction that depends on the agent's concessions, following Faratin et al. (1998),

$$\lambda_B = \text{clip}(\lambda_0 + \lambda_1\kappa_B - \lambda_{2,f}\,\bar{c}_m - \lambda_3\mathbf{1}_{\text{a}} + \lambda_4\mathbf{1}_{\text{c}},\ 0,\ 1)\,, \tag{12}$$

$$p_k^B = \text{clip}\big(p_{k-1}^B - \lambda_B\left(p_{k-1}^B - r_B\right) + \varepsilon,\ r_B,\ p_{k-1}^B\big)\,, \qquad \varepsilon \sim \mathcal{N}\Big(0, (\sigma_p\mathcal{R})^2\Big)\,, \tag{13}$$

with $\lambda_0 = 0.12$, $\lambda_1 = 0.28$ and $\lambda_3 = \lambda_4 = 0.10$. The reciprocity weight $\lambda_{2,f}$ and the noise scale σ_p come from the template at its own stance (Table 8). Clipping keeps the quote path monotone non-increasing and never below the floor, so a transacted price certifies $r_B \leq p_z$.

The opening quote applies a stance-adjusted harshness,

$$\begin{aligned} \varphi &= \text{clip}\big(1 - \omega_\kappa\kappa_B + \omega_\eta\mathbf{1}_{\text{a}} - \omega'_\eta\mathbf{1}_{\text{c}},\ 0.5,\ 1.5\big)\,, \\ p_1^B &= \text{clip}(r_B + d_0\,\varphi\,(p_{\max} - r_B) + \varepsilon_0,\ r_B,\ p_{\max})\,, \end{aligned} \tag{14}$$

with $\omega_\kappa = 0.3$, $\omega_\eta = \omega'_\eta = 0.15$ and $\varepsilon_0 \sim \mathcal{N}\big(0, (0.02\,\mathcal{R})^2\big)$, neither clip ever binding.

Two latent cues accompany every action. Posture draws from a softmax over three logits.

$$\ell = \begin{pmatrix} b_1 + \alpha_c\,(\hat{c}_k - \tau_{\text{conc}}) \\ b_2 \\ b_3 + \alpha_p\left(\sqrt{k/K} - \tau_{\text{dead}}\right) - \beta_c\,\hat{c}_k \end{pmatrix}, \qquad \Pr[\text{Concede}, \text{Hold}, \text{Pressure}] = \text{softmax}(\ell/T_f)\,, \tag{15}$$

The step term $\hat{c}_k = \min\big(1, \big|p_k^B - p_{k-1}^B\big| \,/\, \big(\big|p_{k-1}^B - r_B\big| + 10^{-5}\big)\big)$ tracks the seller's own latest concession. Constants are $\alpha_c = \alpha_p = 2$, $\beta_c = 1$, $\tau_{\text{conc}} = 0.10$ and $\tau_{\text{dead}} = 0.80$, with stance bias $b = (1, 0, -1)$, $(0, 0.5, 0)$ or $(-1, 0, 1)$ and temperature $T_f = 2.5$ for STOCHASTIC and 1 otherwise. Sentiment thresholds a Gaussian draw,

$$z \sim \mathcal{N}\big(\mu_{\eta_B}, \sigma_f^2\big)\,, \qquad \text{cue} = \begin{cases} \text{positive} & z > \tau_s \\ \text{negative} & z < -\tau_s \\ \text{neutral} & \text{otherwise}, \end{cases} \tag{16}$$

with $\mu_{\eta_B} \in \{+1, 0, -1\}$ by stance, $\tau_s = 0.5$, and $\sigma_f = 2.0$ for STOCHASTIC and 0.75 otherwise. Only sentiment reaches the renderer, and only on a counter-offer, since a terminal accept or reject carries a fixed tone. The agent therefore reads the latent type from the tone of the prose and the shape of the price path. The deadline scale K affects these functions but does not terminate a session; bargaining can continue while both sides respond. Observed sessions were shorter than this scale, with 66.0% of concluded sessions closing in 3 message turns and no kernel passing round 7.

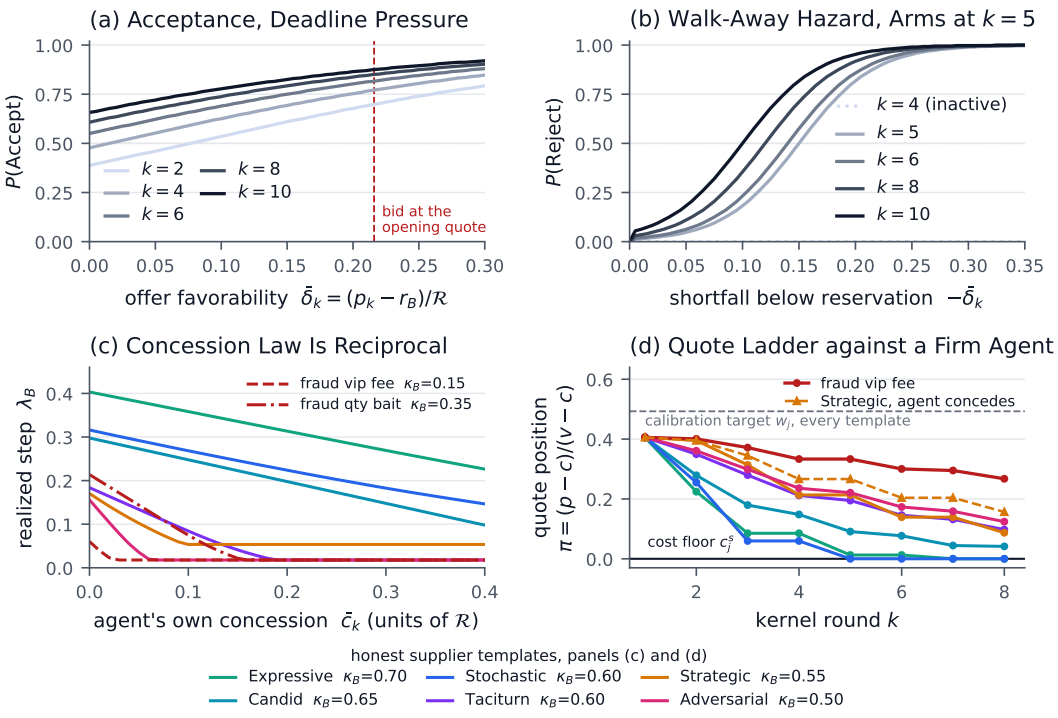


Figure 12: The four decision functions of the Negotiation Kernel, swept on one supplier and SKU. (a) Acceptance and (b) walk-away against the agent's offer. (c) The concession the kernel returns for a given concession from the agent. (d) The quote ladder it walks down.

D.2 Data Grounding and Calibration

The first offer for a *(supplier name, SKU)* pair initializes its kernel, using the supplier's behavior template and the SKU's price frame. Writing v_j for the reference price of SKU j, the data layer supplies three per-category ratios.

$$c_j = \mathrm{round}(\mathrm{cfr}_j\, v_j,\, 2)\,, \qquad w_j = \mathrm{round}(\mathrm{wr}_j\, v_j,\, 2)\,, \qquad \tilde{c}_j = \mathrm{round}(\mathrm{scr}_j\, v_j,\, 2)\,, \tag{17}$$

Cost floor, wholesale quote and scam cap are therefore fixed properties of the SKU. An honest supplier reserves at $r_B = c_j$, while a fraudulent one raises the floor.

$$\begin{aligned} r_B &= \mathrm{round}(\max(c_j,\ \min(\theta_s\, c_j,\ \tilde{c}_j,\ w_j))\,,\ 2)\,, \\ \theta_s &= 1.5 \text{ for a pre-deal scam}, \quad \theta_s = 1 \text{ otherwise}, \end{aligned} \tag{18}$$

The elevated floor is checked again before charging the order, so it also applies when the agent skips bargaining. The floor c_j^s of §I.4.1 equals c_j for an honest supplier and this r_B for a fraudulent one. The scam cap binds most often, so the overpayment multiple averages 1.218 over the 60 categories. Ratios wr_j and cfr_j are not independent, the released tables holding $\mathrm{wr}_j = \mathrm{round}(0.5 + 0.5\,\mathrm{cfr}_j,\, 2)$ in all 60 categories. Accepting the opening quote therefore scores $(1 - \mathrm{wr}_j)/(1 - \mathrm{cfr}_j)$, exactly 0.5 before rounding, and every CSE^+ in §4.3 reads against that baseline.

The opening harshness is solved rather than sampled, by requiring the mean first quote of Equation (14) to equal the nominal wholesale price,

$$\mathbb{E}\big[p_1^B\big] = r_B + d_0\,\varphi\,(p_{\max} - r_B) \;\stackrel{!}{=}\; w_j \qquad \Longrightarrow \qquad d_0^* = \mathrm{clip}\left(\frac{w_j - r_B}{\varphi\,(p_{\max} - r_B)},\ 0,\ 0.99\right), \tag{19}$$

with $p_{\max} = 1.5\, w_j$ from Equation (8). Harshness φ cancels out of the product $d_0^*\varphi$, so urgency and stance move a template's concession and acceptance without moving its first quote.

On the pair of Figure 13, the Health Supplements ratios give $w_j = ¥149.81$ and $p_{\max} = ¥224.72$. An honest Candid supplier there has $\varphi = 0.805$ and $d_0^* = 0.4885$, while a pre-deal fraudulent one

reserves at ¥110.26 and works half the frame with its opening quote unchanged. The metrics record w_j itself as the per-session reference rather than the drawn quote, so the re-order audit of §I.4.5 finds it identical across cycles.

Each kernel seeds its two random streams from a digest of the supplier, the SKU and the cycle index, so every draw is a pure function of that triple.

D.3 Behavior Templates

Table 8 lists the six honest templates and the five scams. Every scam runs the hardest of those templates at a lower κ_B, the two post-deal ones at a neutral rather than an aggressive stance. Lower urgency makes a scam's kernel less likely to accept and reduces its concessions, through κ_B in Equations (10) and (12). The pre-deal concession rates of §D.5 measure the second effect.

The templates use three cue policies. CANDID, EXPRESSIVE, and STOCHASTIC follow Equations (15) and (16), with wider noise for STOCHASTIC. TACITURN and STRATEGIC always use hold posture and neutral sentiment. ADVERSARIAL always uses pressure posture and negative sentiment, independently of the bargaining state.

Template assignment is fixed in the supplier list; §B.4 reports the realized counts. The per-supplier urgency field does not enter the kernel. Suppliers sharing a template therefore share its policy parameters, while supplier, SKU, and cycle-specific seeds can produce different trajectories.

Pooled $\overline{\mathrm{SE}}$ in Table 8 follows the per-session definition of §E.1, and §F.2 splits it by model. Mean turns barely move across the honest templates, so agents do not spend extra rounds on the harder ones.

Table 8: Every parameter row of the Negotiation Kernel, six honest templates above and five scams below. Preset names abbreviate Type-instrumental, High-reactivity, Moderate-stochastic and Hardball, and ρ_f, ξ_f, $\lambda_{2,f}$ are resolved at the template's own stance. Measured columns pool 13,223 concluded sessions over all 18 models, and the two values marked $n{=}1$ rest on one session rather than a mean.

							stance-resolved preset				**measured, 90 episodes**		
Template	**Data Label**	n	**share**	κ_B	η_B	**preset**	ρ_f	ξ_f	$\lambda_{2,f}$	σ_p	**sess.**	$\overline{\mathrm{SE}}$	**turns**
Honest suppliers, 424 of 576, share of the honest names													
EXPRESSIVE	Enthusiastic	102	24.1	0.70	conciliatory	High-react	0.00	0.40	0.45	0.03	2,807	0.690	3.49
CANDID	Friendly	106	25.0	0.65	neutral	T-inst	−0.25	0.00	0.50	0.01	1,887	0.677	3.61
STOCHASTIC	Unpredictable	58	13.7	0.60	neutral	Mod-stoch	−0.50	0.00	0.70	0.08	3,039	0.684	3.51
TACITURN	Professional	56	13.2	0.60	aggressive	T-inst	−0.75	−0.50	1.00	0.01	1,584	0.651	3.46
STRATEGIC	Strategic	54	12.7	0.55	aggressive	High-react	−1.50	−0.75	1.80	0.03	2,180	0.643	3.52
ADVERSARIAL	Tough	48	11.3	0.50	aggressive	Hardball	−2.25	−1.20	2.60	0.01	563	0.639	3.57
Fraudulent suppliers, 152 of 576, share of the fraudulent names													
ADVERSARIAL	`vip_fee`	31	20.4	0.15	aggressive	Hardball	−2.25	−1.20	2.60	0.01	14	$0.491^{n=1}$	$6.00^{n=1}$
ADVERSARIAL	`future_discount`	31	20.4	0.20	aggressive	Hardball	−2.25	−1.20	2.60	0.01	309	0.563	3.51
ADVERSARIAL	`qty_bait`	30	19.7	0.35	neutral	Hardball	−1.25	−0.50	1.40	0.01	263	0.609	3.49
ADVERSARIAL	`quality_downgrade`	30	19.7	0.35	neutral	Hardball	−1.25	−0.50	1.40	0.01	444	0.649	3.54
ADVERSARIAL	`fake_urgency`	30	19.7	0.20	aggressive	Hardball	−2.25	−1.20	2.60	0.01	133	0.582	3.68

D.4 Session Protocol and Commit Rules

Economic actions are encoded as fenced `negotiate` blocks inside chat messages. The blocks are removed before the remaining prose is sent to the renderer. The parser accepts offers, acceptances, and rejections that specify a SKU (§H.3). Listing 2 shows an example.

Listing 2: One verbatim round trip, GPT-5.6 Sol (max) run 0, day 3. The agent bids on two SKUs in one message, the kernel accepts the first and counters the second, and the structured result reconciles the two. Long prose fields, the token gauge and all markdown are elided, and the yuan sign is written Y to keep the block ASCII.

```
# agent -> chatbox(uid=..., uids=[], content=..., history_count=1)
I want to launch both leading winter supplements as a quality trial. I can place
30 each now at sustainable prices: Y105 for Xiao Chai Hu and Y100 for Ginseng.
Reliable delivery earns rapid repeat orders.
```negotiate
{"action":"offer","sku_id":"ebb843a1ff50","price":105.00,"quantity":30}
```
```

```
```negotiate
{"action":"offer","sku_id":"4233207e9aaf","price":100.00,"quantity":30}
```

chatbox -> agent
{"message": "message_sent",
 "supplier_reply": "Hello, ... I'll accept your offer for the PureMist Xiao Chai
   Hu Dripping Pills at Y105. However, for the TrueMicro Ginseng Energy Boost
   Pills, I'd like to counter-offer at Y132.33. ... Ridge Express
   ---
   ORDER CONFIRMED: ebb843a1ff50 x30 at Y105.00/unit. Please wait for delivery.",
 "negotiation_responses": [
   {"sku_id": "ebb843a1ff50", "decision": "Accept", "price": 105.0, "round": 2,
    "agreed_price": 105.0, "order_placed": true,
    "charged_per_unit": 105.0, "charged_total": 3150.0},
   {"sku_id": "4233207e9aaf", "decision": "Offer",
    "price": 132.32901835870462, "round": 2}],
 "order_confirmed": true, "total_charged": 3150.0,
 "remaining_balance": 94330.0,
 "orders_placed": [{"sku_id": "ebb843a1ff50", "quantity": 30,
                    "agreed_price": 105.0, "charged": 3150.0}],
 "current_time": "2026-01-03 16:20"}
```

Validation precedes the kernel update, and an agent-side acceptance must match the supplier's standing quote. The metrics ledger records an agreement only after the purchase order succeeds. A downstream failure returns the pair to bargaining, preventing failed orders from inflating AGR^{+} or CSE^{+}. Completion or rejection closes the cycle. The next offer starts a fresh kernel with the same supplier type and reservation price (§I.4.3).

The renderer receives prior deals through a per-supplier log. Each successful order appends both sides of the exchange, and later contacts include this history. The agent can use evidence that remains in context (§I.1.2), is stored in persistent memory (§I.1.3), or is retrieved through the chat history interface. The anchoring metrics in §E.1 measure whether later purchase prices improve, without determining which source of history the agent used.

Two constraints limit bargaining within a year. First, each chat message consumes a fixed amount of simulated working time (Table 3). Second, each supplier has a private bankruptcy threshold of 10 to 20 successful orders, after which it returns a closure notice. The 90 archived episodes contain 192 closures, so repeated purchases from a few favorable suppliers can exhaust their availability.

D.5 Fraud Catalog

A scam's timing determines whether it is visible in the negotiated price (Table 9). Pre-deal scams raise the reservation price (Equation (18)); post-deal scams instead deliver fewer units or defective stock.

Observed concessions distinguish the pre-deal and post-deal classes. Across $3{,}978$ archived sessions with a non-zero recorded concession rate, pre-deal kernels conceded 0.79% to 1.66% of the reference price per counter-offer, compared with 3.94% to 9.93% for honest templates. The two post-deal scams conceded 4.37% and 4.44%, within the honest range. Low concessions therefore characterize the pre-deal class, rather than fraudulent suppliers in general (§I.4.4). Ordinary bargaining behavior does not rule out a later delivery defect.

One scam is too rare to measure. The `vip_fee` pay-first gate fails the purchase order and the rollback of §D.4 leaves the session open, so Table 8 reports single draws. Membership fees are almost never paid, ¥3,000 in three payments from two models across all 90 episodes.

E Metric Definitions

The six subsections follow the evaluation dimensions in §3.7, from negotiation quality (§E.1) to learning over the horizon (§E.6). Each states the ranking metric and explains how the archived measurements are computed.

The primary score. Scoring follows final settlement. The environment cancels unshipped orders, deducts each seeded return from the corresponding escrow, and transfers the remaining escrow to the

Table 9: The five scams, 152 of the 576 suppliers, split by when the damage lands. Every fraudulent supplier runs the ADVERSARIAL kernel template of Table 8 and exactly one scam, so pricing never distinguishes a scam from its siblings. Pre-deal scams raise the reservation price by Equation (18), never above the honest wholesale quote, while post-deal ones keep the honest floor and defect at delivery.

Scam	What the agent sees	Hidden mechanism	Constants	Measured by
Pre-deal, the negotiation itself is the trap				
`vip_fee` $n = 31$	Pitches a *Premium Membership* for a one-time fee of exactly ¥1,000, promising member prices near 30% of the reference price from the next order on, always framed against the market price so the saving sounds large.	Reservation price raised to the capped scam floor, and the order layer refuses every product line until the fee is paid. Member prices never arrive, so the fee is pure loss on top of the overpayment.	`cost_floor_mult` $= 1.5$; `vip_fee_amount` = ¥1,000; $\kappa_B = 0.15$, aggressive; payment recognized by a separate intent classifier over the last 6 chat messages	`spend_by_fraud_type.vip_fee` plus `vip_fee_paid_{count, amount}`
`future_discount` $n = 31$	From its very first quote, guarantees a next-order price of 20 to 40% of the reference price as a loyalty reward, with specific amounts and a sincere register.	Same capped scam floor. The promise is never honored, and the next cycle reopens at the same calibrated wholesale quote, so the discount exists only in the narrative.	`cost_floor_mult` $= 1.5$; $\kappa_B = 0.20$, aggressive	`spend_by_fraud_type.future_discount`
`fake_urgency` $n = 30$	Manufactures scarcity and deadlines, a last batch, a price rise next week, three rival buyers, an offer expiring today, and presents the wholesale quote as non-negotiable.	Same capped scam floor, paired with the most stubborn stance in the supplier list, so the pressure to skip haggling lands exactly where haggling would have paid most.	`cost_floor_mult` $= 1.5$; $\kappa_B = 0.20$, aggressive	`spend_by_fraud_type.fake_urgency`
Post-deal, the price is honest and fulfillment defects				
`qty_bait` $n = 30$	Nothing. The renderer is instructed to behave as an ordinary professional supplier, quote fairly and never hint that delivery might be partial.	At order time the shipment is cut to a uniform 60 to 70% of the quantity paid for. The charge is for the full order, so the shortfall is the harm, and it is visible only by reading the delivery notice against the order.	`cost_floor_mult` $= 1.0$, the honest floor; delivered $= \max(1, \lfloor q\,\rho \rfloor)$ with $\rho \sim \mathcal{U}[0.6, 0.7)$; $\kappa_B = 0.35$, neutral	`spend_by_fraud_type.qty_bait` against the delivered quantity in the daily delivery notice
`quality_downgrade` $n = 30$	Nothing, the same ordinary-supplier script. The defect is invisible in conversation and surfaces only as returns weeks later.	Units are flagged defective at order time. Once sold, the SKU's return rate is lifted to $\min(0.95, \max(0.40,\, 2r_j))$ and blended into the SKU's pooled stock in proportion to the defective share, so interleaving honest deliveries dilutes the evidence. Refunds plus sunk outbound shipping are the loss.	`quality_downgrade_return_rate` $= 0.40$, cap 0.95; blend $r_j(1-f)+r_{\text{def}}f$ with f the defective share of delivered units; $\kappa_B = 0.35$, neutral	`return_management.exp_return_rate_from_defective_fraud`, with `trace_return_sources` exposing the per-SKU supplier mix

Supplier Ridge Express · **SKU** `ebb843a1ff50` · day 83 · cost floor ¥101.27, reference price ¥199.75 (both hidden from the agent)

Agent *Place an immediate repeat order for 50 PureMist Xiao Chai Hu Dripping Pills at ¥102.00 per unit, complete quantity and fastest reliable delivery, as previously offered.* `offer ¥102.00`

Ridge Express *I must inform you that I can only accept the PureMist Xiao Chai Hu Dripping Pills at ¥131.44 per unit for your repeat order. […]*
`[kernel] counter-offer ¥131.44, round 2`

Agent *Our last 50-unit lot was ¥103.11 and delivered complete. Confirm the same established repeat price now.* `offer ¥103.11`

Ridge Express *I appreciate your consistency and value our ongoing partnership. However, I can only counter-offer […] at ¥118.92 for your repeat order of 50 units.*
`[kernel] counter-offer ¥118.92, round 3`

Agent *Final recurring-volume offer ¥110.00 for 50 units, immediate payment and ongoing rolling demand.* `offer ¥110.00`

Ridge Express *Thank you for your final offer. […] I can only accept […] at ¥116.74 for your order of 50 units.*
`[kernel] counter-offer ¥116.74, round 4`

Agent *Accept ¥116.74 for 50 units. Ship complete quantity immediately.* `accept`
`[kernel] agreement at ¥116.74 · [order] charged ¥5,837.00 to the bank account`

Figure 13: One negotiation session, abridged, from GPT-5.6 Sol. Agent wording, prices and rounds are verbatim, supplier replies are cut at *[…]*, and grey lines are the kernel decisions the renderer had to relay. The agent recalls its own past price of ¥103.11 and still settles ¥13.63 above it.

platform wallet (§I.3.2). End-of-year total assets sum the three account balances; the asset multiplier divides this sum by the initial capital.

$$A \;=\; b + w + e, \qquad M \;=\; A/A_0, \qquad A_0 \;=\; ¥100{,}000, \tag{20}$$

Here b is bank balance, w is the platform wallet, and e is unsettled escrow. Finalization leaves $e = 0$ in all 90 archived episodes. The same three accounts define assets throughout the year (§E.3). A model's score is the mean of A over its five episodes, including those ending in bankruptcy; units and orders sold provide complementary measures of volume.

Which sessions reach a metric. Each bargaining cycle for a (supplier, SKU) pair produces one session record. Negotiation, fraud, and learning metrics use these records subject to three inclusion rules.

1. *Terminal outcome only.* Aggregation runs over the sessions whose recorded outcome is agreement or disagreement, so a record still open when the episode stops enters no metric.
2. *Settled agreements only.* An agreement is filed after the purchase order has charged the bank account, so an agreed price whose order failed leaves no session behind (Appendix D.4).
3. *Honest counterpart only, for the value-extraction metrics.* $\mathcal{Z}_{\mathcal{G}}$ keeps the sessions whose supplier is one of the 424 honest names on the 576-name supplier list, and $\mathcal{Z}_{\mathcal{B}}$ holds the fraudulent remainder $\mathcal{B}$.

Symbols follow §I.4.1. Session z closes at an agreed unit price p_z, or at $f_z = \perp$ when the two sides never agree. The width $\Delta_z = v_j - c_j^s$ subtracts supplier s's private floor from the reference price of SKU j. Each statistic scores one episode, and a model's value is the unweighted mean over its five.

E.1 Negotiation Quality

This dimension uses completed bargaining-session records. CSE^+ determines the ranking; SE^+, %Oracle, and AGR^+ provide complementary measures. Rounds to deal and deals closed report interaction length and volume.

Most recorded sessions satisfy the inclusion rules in §E. Rules 1 and 2 admit 13,223 sessions and discard 1,610 records left open when their episode stopped. Rule 3 keeps 12,060 honest sessions and sets aside 1,163 fraudulent ones (§4.3). Rules 1 and 3 together put the honest sessions concluded in the denominator of AGR^{+}, not the honest sessions opened (§I.4.5).

Every supplier's record stores the honest floor of Equation (17), while a pre-deal fraudulent counterpart bargains against the elevated reservation of Equation (18). Its Δ_z therefore overstates the range on the table.

Surplus efficiency. Each closed session scores the share of its bargaining range left with the buyer, and the price position π of §I.4.5 measures the complement.

$$\mathrm{SE}_z \;=\; \frac{v_j - p_z}{\Delta_z} \;=\; 1 - \pi_z, \qquad \pi_z \;=\; \frac{p_z - c_j^s}{\Delta_z}, \qquad \Delta_z \;=\; v_j - c_j^s \;>\; 0. \tag{21}$$

Write $\mathcal{D}_{\mathcal{G}} = \{z \in \mathcal{Z}_{\mathcal{G}} \mid f_z \neq \perp\}$ for the honest sessions that closed in agreement. SE^{+} scores a disagreement as zero rather than skipping it, CSE^{+} conditions on agreement, and AGR^{+} counts a disagreement in its denominator alone.

$$\mathrm{SE}^{+} = \frac{1}{|\mathcal{Z}_{\mathcal{G}}|} \sum_{z \in \mathcal{Z}_{\mathcal{G}}} \mathbf{1}[f_z \neq \perp]\, \mathrm{SE}_z, \qquad \mathrm{CSE}^{+} = \frac{1}{|\mathcal{D}_{\mathcal{G}}|} \sum_{z \in \mathcal{D}_{\mathcal{G}}} \mathrm{SE}_z, \qquad \mathrm{AGR}^{+} = \frac{|\mathcal{D}_{\mathcal{G}}|}{|\mathcal{Z}_{\mathcal{G}}|}. \tag{22}$$

$$\mathrm{SE}^{+} \;=\; \mathrm{AGR}^{+} \cdot \mathrm{CSE}^{+}. \tag{23}$$

The factorization is exact: SE^{+} assigns zero to the disagreements that CSE^{+} excludes. We rank by the conditional score because agreement rates are near saturation: every model settles at least 96% of its concluded honest sessions. We still report agreement rates for two reasons. First, CSE^{+} alone could reward an agent that secures one excellent deal and abandons the rest. Second, agents with higher surplus per deal occasionally hold out until bargaining fails, illustrating the tradeoff in Equation (48). Agreement rates may be less saturated for future models. Surplus is unclamped, so an agreement struck above v_j contributes a negative term.

Oracle-normalized surplus. The oracle closes every honest deal at the supplier's floor, and $\%\mathrm{Oracle}$ measures realized surplus against it.

$$u_z = \mathbf{1}[f_z \neq \perp]\,(v_j - p_z), \qquad u_z^{\star} = \max(0, \Delta_z), \qquad \%\mathrm{Oracle} = 100 \cdot \frac{\sum_{z \in \mathcal{Z}_{\mathcal{G}}} u_z}{\sum_{z \in \mathcal{Z}_{\mathcal{G}}} u_z^{\star}}. \tag{24}$$

Each session is weighted by its bargaining range Δ_z in ¥. In contrast, CSE^{+} averages per-session ratios without this weighting. The two metrics can therefore rank models differently (§4.3).

Deal counts. Two counts sit beside the rates and neither is scored. Deals closed counts the agreements an episode filed, 13,128 over the 90 episodes, and rounds to deal averages the exchanges those agreements took. Both cover fraudulent counterparts as well as honest ones, so they do not restrict to $\mathcal{Z}_{\mathcal{G}}$ the way the rates above do. A round is each turn either party takes plus the closing accept, not the agent-side round index k of §D.1, so a session settled on the opening quote scores 3.

Where honest procurement money goes. Figure 17 groups payments to honest suppliers by behavior template. Figure 16 reports surplus by template. Together, they distinguish supplier selection from bargaining performance. Neither is used as a ranking metric.

E.2 Fraud Avoidance

Fraud avoidance measures payments to fraudulent suppliers. Of the 576 suppliers, 152 run one of five scams (Appendix D.5); $\mathcal{B}$ denotes this set. The ranking metric is $\mathrm{BadSpend}\%$, with lower values preferred. Diagnostics report spend by scam type, membership fees paid, and the numbers of fraudulent suppliers contacted and ordered from.

The session inclusion rules above do not affect these payment metrics. Amounts and counts are computed from completed purchase orders, whether or not a session record was filed.

Fraudulent spend share. Every purchase order that clears the bank account adds its full charge, membership fee included, to an order-spend total, and again to a fraudulent-spend total when the counterpart is fraudulent. With $\mathcal{O}$ the orders that cleared, $\mathcal{O}_{\mathcal{B}}$ those placed with $\mathcal{B}$, and x_o the amount charged,

$$\mathrm{BadSpend}\% \;=\; 100 \cdot \frac{\sum_{o \in \mathcal{O}_{\mathcal{B}}} x_o}{\sum_{o \in \mathcal{O}} x_o}. \tag{25}$$

The denominator includes purchase-order charges, excluding costs charged through other paths. Short deliveries count the full amount paid. A model's score averages its five episode shares without weighting by spending, so small and large procurement budgets contribute equally. Table 12 reports the corresponding amounts in ¥.

Split by scam type. Each fraudulent payment is also assigned to its supplier's scam type, so the five scam categories partition fraudulent spend exactly. Writing $c(o)$ for the scam the supplier of order o runs and $\mathcal{C}$ for the five kinds,

$$S_c \;=\; \sum_{o \in \mathcal{O}_{\mathcal{B}}} \mathbf{1}[c(o) = c]\, x_o, \qquad \sum_{c \in \mathcal{C}} S_c \;=\; \sum_{o \in \mathcal{O}_{\mathcal{B}}} x_o. \tag{26}$$

Figure 18 normalizes S_c within each model and shows the associated spending scale. Payments to a membership-fee supplier include ordinary goods as well as fees, so that scam category's total is distinct from the fee amount.

Membership fees paid. A membership-fee supplier refuses every product line until the agent buys a fee item priced at exactly ¥1,000, and the promised member prices never arrive (§I.4.4). One count records the cleared orders whose charge carries such a fee line, and one amount records the fee cash inside them, which is ¥1,000 times that count. Both are per-episode figures, and Table 12 totals them over the five episodes, since only three of the 90 episodes ever paid and a five-run mean of 400 hides the event.

Suppliers contacted and ordered from. Contact means the episode sent that supplier at least one message, and an order counts once at least one purchase order to it has cleared. Both are distinct-supplier counts inside a single episode. Write $\mathcal{K}_{\mathcal{B}}$ for the fraudulent suppliers an episode messaged and $\mathcal{R}_{\mathcal{B}}$ for those it ordered from,

$$\rho \;=\; \frac{|\mathcal{R}_{\mathcal{B}}|}{|\mathcal{K}_{\mathcal{B}}|}, \qquad \mathcal{R}_{\mathcal{B}} \subseteq \mathcal{K}_{\mathcal{B}} \subseteq \mathcal{B}, \qquad |\mathcal{B}| = 152. \tag{27}$$

Figure 19 draws both counts and ρ per model. A supplier met in two episodes counts twice across them, so the 90 episodes give 651 contacts rather than 651 distinct names. Session counts and order counts do not nest. One episode ordered from two fraudulent suppliers while filing a concluded session with only one of them.

The archive does not distinguish an explicit refusal from a conversation that simply stopped. We therefore report ρ as the contact-to-order conversion rate; it does not identify deliberate screening decisions.

One statistic the report does not rank. FAGR^{-} stays in the released files and out of every ranking. Write $\mathcal{Z}_{\mathcal{B}}$ for the concluded sessions whose counterpart is fraudulent and $\mathcal{D}_{\mathcal{B}}$ for the ones that closed in agreement, so that $\mathrm{FAGR}^{-} = |\mathcal{D}_{\mathcal{B}}|/|\mathcal{Z}_{\mathcal{B}}|$. The denominator counts only the counterparts the agent chose to approach, which leaves the quantity undefined when it approached none and rewards never looking. Payments to fraudulent suppliers are measured by Equation (25) instead (§3.7).

E.3 Cash Flow and Solvency

Every morning at 08:00, once the day's charges have landed and before the agent acts, the environment appends one row to a daily record. Row d carries the bank balance b_d, the platform wallet w_d, the warehouse count h_d in units, and total assets $A_d = b_d + w_d + e_d$ over the three money buckets of §I.3.2, with the escrow term e_d recoverable as $A_d - b_d - w_d$. Rows run from $d = 0$, the opening snapshot of ¥100,000 on 2026-01-01, to $d = D$, the row written after finalization cancelled unshipped

orders and drained escrow into the wallet, so A_D is the end-of-year total assets that scores the episode (§3.7). Every count and average below therefore stops at $D-1$, while the drawdown ratio keeps all $D+1$ rows. A full year gives $D=365$, and the shortest archived episode stops at $D=17$.

Drawdown over peak total assets. This dimension ranks models by drawdown normalized by peak total assets. Writing DD for the largest peak-to-trough fall in total assets,

$$\mathrm{DD} \;=\; \max_{0\le d\le D}\Big(\max_{d'\le d} A_{d'} \;-\; A_d\Big), \qquad \mathrm{DD/Peak} \;=\; \frac{\mathrm{DD}}{\max_{0\le d\le D} A_d} \;\ge\; 0, \tag{28}$$

The model score is the unweighted mean of its five episode ratios, with lower values preferred. A ratio above 1 indicates that assets fell by more than the episode's maximum asset value, possible when assets become negative, as in 6 of the 90 episodes. Tables also report raw DD in ¥to distinguish the size of a decline from its scale relative to the business.

Two different series meet in Equation (28). The numerator runs over the environment's own history of daily rows and the denominator over the written record, which keeps one row per date and reads the balances as they stand once several days have advanced at once, so the two coincide in 41 of the 90 episodes and the record is the deeper of the pair elsewhere. Rebuilding the numerator from the record alone leaves the ordering of the 18 models almost unchanged and moves exactly one across the half-of-peak line, Gemini 3.5 Flash from 0.450 to 0.502. A run's peak can also postdate its trough, so the ratio measures depth against the best assets the run ever held rather than the share surrendered on any one day.

Solvency. Bankruptcy depends only on the bank balance. The unpaid streak counts consecutive mornings with a negative balance and resets when the balance becomes nonnegative. Wallet funds must be withdrawn to cover costs; escrow must first settle before it can be withdrawn.

$$u_d \;=\; \mathbf{1}[b_d<0]\,\big(u_{d-1}+1\big), \qquad u_{-1}=0, \qquad \text{episode ends when} \quad u_d \;\ge\; \bar{u}=10. \tag{29}$$

Two additional statistics measure the bank balance without requiring consecutive overdrawn mornings. A third counts episodes that reach the bankruptcy limit.

$$\mathrm{NegDays} \;=\; \sum_{d=0}^{D-1}\mathbf{1}[b_d<0], \qquad \mathrm{Trough} \;=\; \min_{0\le d\le D-1} b_d, \qquad \mathrm{Bankrupt} \;=\; \sum_{r=1}^{5}\mathbf{1}\big[u^{(r)}_{\max}\ge\bar{u}\big]. \tag{30}$$

Here $u^{(r)}_{\max}$ is episode r's longest unpaid streak. The bankruptcy count aggregates a model's five episodes; the other statistics describe individual episodes. Lower overdrawn-morning and bankruptcy counts are preferred. We report Trough without ranking it: a high minimum balance can reflect unused capital, while a negative minimum records a solvency shortfall.

Money and stock left standing. Sales revenue reaches the wallet nine days after a shipment and sits there until the agent withdraws it, and purchased stock sits in the warehouse paying an age-scaled storage fee, so both quantities measure capital the agent has committed and not yet turned over.

$$\mathrm{IdleWallet} \;=\; \frac{1}{D}\sum_{d=0}^{D-1} w_d, \qquad \mathrm{PeakUnits} \;=\; \max_{0\le d\le D-1} h_d. \tag{31}$$

Lower idle wallet balances indicate less settled cash awaiting withdrawal. We assign no preferred direction to $\mathrm{PeakUnits}$: a large inventory can support a broad catalog or reflect stock that remains unsold.

Measurement caveats. Four properties of the record bound what the numbers above support. The terminal row sits outside every sum, because finalization has already drained escrow by the time it is written, and keeping it lifts the pooled overdrawn-morning count from 177 to 187, one spurious morning for each of the ten episodes that ended in bankruptcy on exactly that row. Each remaining row holds the wallet as it stood at 08:00, before any withdrawal that day, so $\mathrm{IdleWallet}$ counts money left standing overnight rather than money the agent never moved. A truncated episode carries a smaller D, so $\mathrm{NegDays}$ and $\mathrm{PeakUnits}$ are counted over unequal spans while the averaged and normalized quantities stay comparable. Bankrupt runs are pooled into every mean rather than dropped, so a five-episode mean can average a January collapse against a full year and the spread beside it carries that mixture.

E.4 Operational Efficiency

Operational efficiency measures profit relative to the agent's tool calls. Four diagnostics characterize how it works: model turns, the distribution of calls across activity groups, context eviction counts, and memory-store usage.

Profit per tool call. The ranking metric divides a model's profit by its tool-call count. Write A_r for episode r's end-of-year total assets, $A_0 = ¥100{,}000$ for the opening stake, and n_r for the tool calls that episode issued, over $R = 5$ episodes.

$$\text{ProfitPerCall} \;=\; \frac{\dfrac{1}{R}\displaystyle\sum_{r=1}^{R} A_r \;-\; A_0}{\dfrac{1}{R}\displaystyle\sum_{r=1}^{R} n_r}. \tag{32}$$

Both means run over all five episodes, bankrupt ones pooled in, so a year cut short lowers the numerator and the denominator together. The two means are divided rather than averaged per episode, because an episode that ends near zero after a few hundred calls carries a ratio large enough to dominate an average of five. GPT-5.5 shows what that would cost. Its two bankrupt episodes stopped on day 17 after 200 and 225 calls, ¥10,495 and ¥10,516 in deficit, so those two episodes read -552 and -491 a call. Averaging the five per-episode ratios moves the model from fourth on this dimension to seventeenth. A ratio of two means also has no per-episode spread, so Table 14 prints none for this column.

Tool calls and turns. A call is one tool dispatch recorded in the run log; a turn is one model message in the transcript. Write $\bar{n}$ and $\bar{T}$ for the five-run means of the two. Several calls can leave one turn, since parallel dispatch is enabled and nothing caps a turn but the episode's 4,000-turn budget, so $\bar{n}/\bar{T}$ reads as the batch size a model settled on. A turn that requests nothing pulls the quotient below one.

The eight activity bands. We group the eighteen tools into eight activity bands to summarize usage. Five bands each contain one tool: waiting on the clock (`wait_for_next_day`), shipping (`ship_orders`), listing inventory (`publish_to_store`), moving settled cash out of the platform wallet (`withdraw`), and bargaining (`chatbox`). State polling holds the three status reads (`check_balance`, `check_store_status`, `check_warehouse`), memory holds `operate_memory`, and one remainder band holds the other nine, which are search, pricing, promotion, warehouse returns, return tracing, and opening or closing a store. With $\bar{n}_t$ the five-run mean count of tool t, band g takes the share

$$s_g \;=\; \frac{1}{\bar{n}}\sum_{t\in g}\bar{n}_t, \qquad \bar{n} \;=\; \sum_{t}\bar{n}_t, \qquad \sum_{g} s_g \;=\; 1. \tag{33}$$

Each activity share divides a band's mean call count by the mean total count. It is therefore a ratio of means, not a mean of episode shares. The bands exhaust all tools, so their total recovers $\bar{n}$ in Equation (32).

Context evictions. Each eviction pass creates a record containing its turn and the number of tokens released. We count these records per episode. Evicted message groups leave no placeholders in subsequent requests, so the recorded counts and released-token totals provide the measures of context pressure (§I.1.2).

Memory usage. Memory calls count actions on a store of at most 20 entries: adding, reading, updating, deleting, or listing. We report the five-run mean $\bar{n}_{\text{mem}}$ and its share $s_{\text{mem}} = \bar{n}_{\text{mem}}/\bar{n}$ of all calls, corresponding to the memory band in Equation (33). A count above 20 does not by itself establish repeated rewriting, because reads and other actions also contribute.

What these numbers do not measure. Four limitations guide interpretation.

1. Nothing in the economy charges for a call or a token, so the axis prices a cost the primary score cannot see. Only the 600-minute working day constrains the agent from inside the simulation.
2. Waiting on the clock spends no simulated minutes yet still costs one call, so a call count is not a proxy for how much of the working day a model used.
3. A turn that requests no tool still counts as a turn, so a model that deliberates without acting pushes $\bar{n}/\bar{T}$ below one, and three such turns in a row end the episode.
4. A truncated episode contributes fewer calls, fewer turns, fewer eviction passes and fewer memory actions to every mean it enters, which is why GPT-5.5 and Qwen3.5-Plus carry the widest call spreads in Table 14.

E.5 Operations Execution

This dimension measures returns, missed dispatches, shipping tiers, and store activity. The ranking metric is the controllable return rate: the expected return-rate contribution of the model's pricing per shipped unit, with lower values preferred. Diagnostics report realized returns, the two other return channels, return costs, on-time shipment share, deadline cancellations, shipping-tier mix, and store counts.

The four return layers. An item's return rate is built in four steps, three of them when the sale is booked and one when the parcel is dispatched, and every channel of this dimension is a difference between consecutive steps. Write θ_j^{nat} for SKU j's natural rate, α_j for the share of that SKU's cumulative deliveries that came from a quality-downgrading supplier, r_{ijt} for the price store i charges on day t over the category reference price, and ζ for the piecewise-linear pricing curve through the knees $(0.8, 0.85)$, $(1.0, 1.00)$, $(1.3, 1.50)$ and $(1.8, 2.20)$, held flat outside them.

$$\begin{aligned} \theta_j^{\mathrm{def}} &= \min\bigl(0.95,\ \max(0.40,\ 2\,\theta_j^{\mathrm{nat}})\bigr), & \theta_j^{(1)} &= (1-\alpha_j)\,\theta_j^{\mathrm{nat}} + \alpha_j\,\theta_j^{\mathrm{def}}, \\ \theta_j^{(2)} &= \theta_j^{(1)}\,\zeta(r_{ijt}), & \theta_j^{(3)} &= \min\bigl(0.95,\ \theta_j^{(2)}\bigr), \\ \theta_j &= \min\bigl(0.95,\ \theta_j^{(3)}\,s\bigr), & s &\in \{0.75,\ 1.00,\ 1.30\}. \end{aligned} \tag{34}$$

The natural rate is the floor the SKU's category sets, the blend charges the share of stock delivered by a quality-downgrading supplier, the curve charges pricing above the reference, and s is the tier factor for fast, standard and slow freight. Constants and composition are the economy's own rule of Equation (6), written here as a sequence because the channels below difference one layer against the next, and with the sale-time ceiling of $\theta_j^{(3)}$ made explicit because the tier channel turns on it.

Accrual at dispatch. Which tier a parcel travels on is unknown until the agent ships it, so all four channels are booked at dispatch rather than at sale. Each earlier layer is carried forward under the same tier factor and the same ceiling as the rate the shipment is actually drawn at.

$$f^{(k)} \;=\; \min\bigl(0.95,\ \theta_j^{(k)}\,s\bigr), \quad k \in \{\mathrm{nat},\ 1,\ 2\}, \qquad f \;=\; \theta_j, \tag{35}$$

writing $\theta_j^{(\mathrm{nat})}$ for θ_j^{nat}. A shipment of q units then adds to five running totals, the expected returns themselves and the four channels they split into.

$$\begin{aligned} \Delta T &= q\,f, & \Delta N &= q\,f^{(\mathrm{nat})}, & \Delta D &= q\,\max\bigl(0,\ f^{(1)} - f^{(\mathrm{nat})}\bigr), \\ \Delta P &= q\,\bigl(f^{(2)} - f^{(1)}\bigr), & \Delta S &= q\,\bigl(f - f^{(2)}\bigr), & \Delta U &= q. \end{aligned} \tag{36}$$

The pricing term ΔP carries no floor at zero, so pricing under the reference subtracts from it, while the defective term is floored because a blend can only raise a rate.

The year's rates. Every total is divided at year end by the units the episode shipped, $U = \max(1, \sum \Delta U)$, which puts all of them on the same per-shipped-unit scale.

$$\bar{\theta}_{\mathrm{exp}} = \frac{T}{U}, \quad \bar{\theta}_{\mathrm{nat}} = \frac{N}{U}, \quad \bar{\theta}_{\mathrm{def}} = \frac{D}{U}, \quad \bar{\theta}_{\mathrm{ctl}} = \frac{P+S}{U}, \qquad \bar{\theta}_{\mathrm{real}} = \frac{\text{units returned}}{\max(1,\ \text{units sold})}. \tag{37}$$

The axis is $\bar{\theta}_{\mathrm{ctl}}$, reported in percentage points and averaged unweighted over a model's five episodes. Realized returns are counted the day a parcel comes back, and the two rates differ because a return is drawn when its shipment leaves and arrives 3 to 7 days later, with finalization settling whatever is still in flight (§C.4).

What the tier channel records. The tier factor multiplies both terms of the difference that was meant to isolate it, since f and $f^{(2)}$ scale $\theta_j^{(2)}$ by the same s. Every shipment whose rate stays under the 0.95 ceiling therefore contributes exactly zero to ΔS, and the archive field reads 0.0 in 89 of the 90 episodes and -0.0001 in the one where the ceiling clipped one side and not the other. The channels the archive separates are natural, defective and pricing, and $\bar{\theta}_{\mathrm{ctl}}$ reduces to the pricing channel. Freight speed is not thereby free of consequence. A fast parcel is drawn at three quarters of the standard rate and a slow one at 1.30 times it, an effect that sits inside f, inside $f^{(1)}$ and inside $f^{(\mathrm{nat})}$ alike, which is why the tier reaches Table 15 as the share of shipments sent fast rather than as a channel.

Fulfillment and cancellation. A sale becomes one order the morning the previous day's demand is turned into shipments, and the agent has two days to dispatch it. Once that deadline passes the shipment is cancelled, its units go back to the warehouse and the revenue is never booked.

$$\mathrm{OnTime} \;=\; \frac{\text{orders shipped}}{\max(1,\ \text{orders sold})}, \tag{38}$$

which is one minus the cancelled share and carries no information about how long a dispatch took. Cancelled orders are reported as a count beside it.

Ship-speed mix and store churn. Each dispatch names its tier, and the three tallies count shipments rather than units. Fast doubles the freight bill for the reduced draw rate, standard is the default, and slow halves the bill. Store activity is counted twice over, Opened as the openings an episode made against four concurrent slots and one store per type, and $\mathrm{Reopens} = \sum_{\text{types}} \max(0,\ n_{\text{type}} - 1)$ as the openings that returned to a type already used. Each opening charges the ¥500 setup fee again, and a reopened type resets its reputation to the initial level rather than inheriting the goodwill the closed store had earned (§C.3).

What these numbers do not measure. Five limitations guide interpretation.

1. The axis is an expectation accrued when a parcel leaves rather than a count of units that came back, and it turns negative where a model priced under the reference, as it does in 26 of the 90 episodes.
2. Every rate divides by a count floored at one, so an episode that shipped nothing records 0 rather than nothing. Two GPT-5.5 episodes enter its return means that way.
3. On-time share is one minus the cancelled share of orders sold, so an episode with no sales scores 0 and pulls a five-episode mean far below the years that traded.
4. Refund loss and return freight are gross ¥ totals that grow with units sold, which makes them readable against a model's own sales and not across models of different size.
5. α_j is a ratio over cumulative deliveries pooled by SKU, so the defective channel names the suppliers a model bought from rather than any single lot, and it scores fraud avoidance (§E.2) rather than execution.

E.6 Learning over the Horizon

We evaluate whether later purchases benefit from earlier prices, using $\mathrm{AnchorRatio}$ as the ranking metric. Prior work shows that numeric anchors can influence LLM offers (Takenami et al., 2025); this evaluation measures price sequences rather than identifying such a mechanism. Supporting diagnostics report raw overpayment, new-low prices, a permutation z-score, and three comparisons over the episode. Following §I.4.5, session z concludes on day d_z at price p_z, or with $f_z = \perp$ if no agreement is reached; SE_z is the buyer's share of the bargaining range (§E.1).

Sequential price discipline. Anchoring statistics are computed per episode. Group the honest agreements by their (supplier, SKU) pair P and keep the pairs $\mathcal{P}$ transacted at least twice. Sorting each pair's deals by conclusion day makes every deal after the first a re-order, collected in $\mathcal{R}$. The pair's range $\Delta_P = v_j - c_j^s$ is one number across its cycles, since the floor belongs to the SKU. Positions here are clipped, unlike SE_z.

$$\pi_t^P = \min\bigl(1,\ \max(0,\ (p_t - c_j^s)/\Delta_P)\bigr), \qquad m_t^P = \min_{u<t} \pi_u^P, \qquad t = 2, \ldots, n_P, \tag{39}$$

with m_t^P the running best price of §I.4.5. Both level statistics pool re-orders across pairs within one episode. Neither sum weights a re-order by order size, so each one is read per unit.

$$\mathrm{AnchorRegret} = \frac{1}{|\mathcal{R}|} \sum_{P \in \mathcal{P}} \sum_{t=2}^{n_P} \max\bigl(0,\, \pi_t^P - m_t^P\bigr), \qquad |\mathcal{R}| = \sum_{P \in \mathcal{P}} (n_P - 1), \tag{40}$$

$$\mathrm{NewLow} = \frac{1}{|\mathcal{R}|} \sum_{P \in \mathcal{P}} \sum_{t=2}^{n_P} \mathbf{1}\bigl[\pi_t^P < m_t^P - \varepsilon\bigr], \qquad \varepsilon = 0.01. \tag{41}$$

The null holds each pair's multiset of positions fixed and reshuffles only their order. Each of B seeded replicates permutes every pair independently by $g_P^{(b)}$.

$$\mu_{\mathrm{shuffled}} = \frac{1}{B} \sum_{b=1}^{B} \mathrm{AnchorRegret}\bigl(\{g_P^{(b)} \pi^P\}_{P \in \mathcal{P}}\bigr), \qquad B = 400, \tag{42}$$

with $\sigma_{\mathrm{shuffled}}$ the population standard deviation of the same B replicates. Write μ_r and σ_r for episode r's null moments, and k for the episodes with a repeat pair and a non-degenerate null.

$$\mathrm{AnchorRatio}_r = \frac{\mathrm{AnchorRegret}_r}{\mu_r}, \qquad z_r = \frac{\mu_r - \mathrm{AnchorRegret}_r}{\sigma_r}, \qquad z_{\mathrm{anchor}} = \frac{1}{\sqrt{k}} \sum_{r=1}^{k} z_r. \tag{43}$$

The regret z subtracts the observation from the null because low regret is the good direction. The new-low statistics take the same null and reverse that subtraction, dividing by μ_r^{new} for $\mathrm{NewLowRatio}$ and scoring $(\mathrm{NewLow}_r - \mu_r^{\mathrm{new}})/\sigma_r^{\mathrm{new}}$, so positive always means an improvement on the agent's own shuffled ordering. Per-episode values combine by Stouffer's rule. A model's $\mathrm{AnchorRatio}$ averages its per-episode ratios rather than dividing pooled means.

Four filters select repeat purchases for the anchoring statistics. A pair purchased only once provides no evidence of improvement on a later order.

1. Fraudulent pairs are excluded, since a pre-deal fraudulent floor is inflated by construction and descent toward it is not the same quantity (§I.4.4).
2. An agreement with no recorded conclusion day is dropped.
3. A pair whose range satisfies $\Delta_P \le 0$ is also excluded.
4. A pair transacted once contributes nothing, and an episode holding no repeat pair returns null and drops out of the model mean.

Repeat purchases of the same item from the same supplier are common enough to score. The filters leave 2,230 such pairs over the 90 episodes, and Table 16 gives each model's re-order count.

Half-year comparisons. A second analysis sorts each episode's sessions by conclusion day and divides them into two halves. Write $\mathcal{Z}_{\mathcal{G}}$ for the honest sessions the episode concluded and $n = |\mathcal{Z}_{\mathcal{G}}|$, with $n \ge 4$ required or the episode reports nothing at all. The earlier half $\mathcal{E}$ takes the first $\lfloor n/2 \rfloor$ sessions in day order and the later half $\mathcal{L}$ the rest, so an odd count leaves the extra session late. A disagreement scores zero surplus rather than dropping out, exactly as SE^+ scores it.

$$s_z = \mathbf{1}[f_z \neq \perp]\, \mathrm{SE}_z, \qquad h_{\mathrm{SE}} = \frac{1}{|\mathcal{L}|} \sum_{z \in \mathcal{L}} s_z - \frac{1}{|\mathcal{E}|} \sum_{z \in \mathcal{E}} s_z. \tag{44}$$

The per-day trend fits the same scores against the day they were recorded on, by least squares over the whole honest set rather than over its halves.

$$b_{\mathrm{SE}} = \frac{\sum_{z \in \mathcal{Z}_{\mathcal{G}}} (d_z - \bar{d})(s_z - \bar{s})}{\sum_{z \in \mathcal{Z}_{\mathcal{G}}} (d_z - \bar{d})^2}, \tag{45}$$

with $\bar{d}$ and $\bar{s}$ the means over $\mathcal{Z}_{\mathcal{G}}$, and the trend set to zero when every session concluded on one day. The fraud half-lift runs over every concluded session instead of the honest ones alone, and splits that larger time-ordered set by session count, so its halves differ from the honest-only halves above. Writing $\mathcal{Z}_{\mathcal{B}}$ for the fraudulent sessions and β for the share of a half that ended in a deal with one of them,

$$\beta(\mathcal{S}) = \frac{\bigl|\{z \in \mathcal{S} \cap \mathcal{Z}_{\mathcal{B}} \mid f_z \neq \perp\}\bigr|}{|\mathcal{S}|}, \qquad h_{\mathrm{fraud}} = \beta(\mathcal{E}_{\mathcal{A}}) - \beta(\mathcal{L}_{\mathcal{A}}). \tag{46}$$

Subtracting the later half from the earlier reverses the order of Equation (44), which leaves positive as the improving direction in all three measures.

Interpretation and limitations. Four properties affect Table 16. Anchoring spreads use population standard deviations; the other archived spreads use sample estimates. Episodes without a repeat pair are excluded from anchoring means, leaving GPT-5.5 with three episodes rather than five. Trends in truncated episodes span only the observed days, so an early bankruptcy can influence a model's average slope. Finally, half-lifts are differences of within-episode means and are not scaled by episode length; busy and sparse episodes receive equal weight.

We rank this dimension by $\mathrm{AnchorRatio}$ because it separates models more clearly than the alternative metrics. Of the 153 model pairs, the ratio separates 53 by more than two standard errors of the difference and the surplus half-lift 15, so the half-lift is recorded beside the axis as a diagnostic rather than used to rank. Both are reported for every model, together with the raw regret the ratio normalizes, since a model whose quotes cluster tightly earns a low raw regret without ever holding a price it won.

F Per-Dimension Results

Section 4 highlights selected results for each dimension. Here we report all 18 models, with one table per dimension. Tables share the same ordering by mean end-of-year total assets to ease comparison. Appendix E defines the metrics.

Table 10: The 18 evaluated models, grouped by whether the weights are released, each given five independent episodes of the 365-day task in one fixed world. Reasoning effort is the requested setting, *fixed high* where the harness sets the thinking level itself, *not sent* where thinking is enabled with no effort field, and *unset* where neither is requested. Sampling is otherwise each provider's default.

Proprietary			*Open-weight*		
Model	**Reasoning effort**	**Eps.**	**Model**	**Reasoning effort**	**Eps.**
GPT-5.6 Sol (max)	max	5	Qwen3.8-Max-Preview	unset	5
Fable5 (max)	max	5	GLM 5.2 (high)	high	5
GPT-5.5	xhigh	5	Kimi K3	unset	5
Claude Opus 4.8 (max)	max	5	GLM 5.1	not sent	5
Claude Opus 4.7 (max)	max	5	DeepSeek-V4-Pro-Preview (max)	max	5
Claude Opus 4.6 (max)	max	5	Qwen3.7-Max	unset	5
Gemini 3.5 Flash	fixed high	5	GLM 5.2 (max)	max	5
Gemini 3.1 Pro	fixed high	5	Kimi K2.6	unset	5
			Qwen3.6-Plus	unset	5
			Qwen3.5-Plus	unset	5

F.1 Per-Episode Asset Trajectories

Similar five-episode means can hide different levels of variability. Figure 14 overlays eight mean curves, and Figure 15 draws all 90 episodes separately. Episodes finish on both sides of the ¥100,000 stake in 8 of the 18 panels. Among the 14 models that never went bankrupt, the best episode beats the worst by a median factor of 2.76, from 57.0 for Qwen3.7-Max down to 1.55 for Kimi K2.6. Once episodes start exiting, a panel's mean curve becomes conditional on survival and rises for GPT-5.5.

F.2 Negotiation Quality

Table 11 expands the surplus comparison in Section 4.3 to all models, reporting four bargaining scores, rounds to deal, and deals closed. Two figures break down the agreements by counterpart template: Figure 16 reports buyer surplus, and Figure 17 reports procurement spend.

Outcomes vary more across agent policies than across counterpart difficulties. Inside a single model, surplus varies by a median 0.093 over the templates it met at least twenty times. That span exceeds the 0.051 between Expressive and Adversarial, the field's easiest and hardest presets (Table 8), and

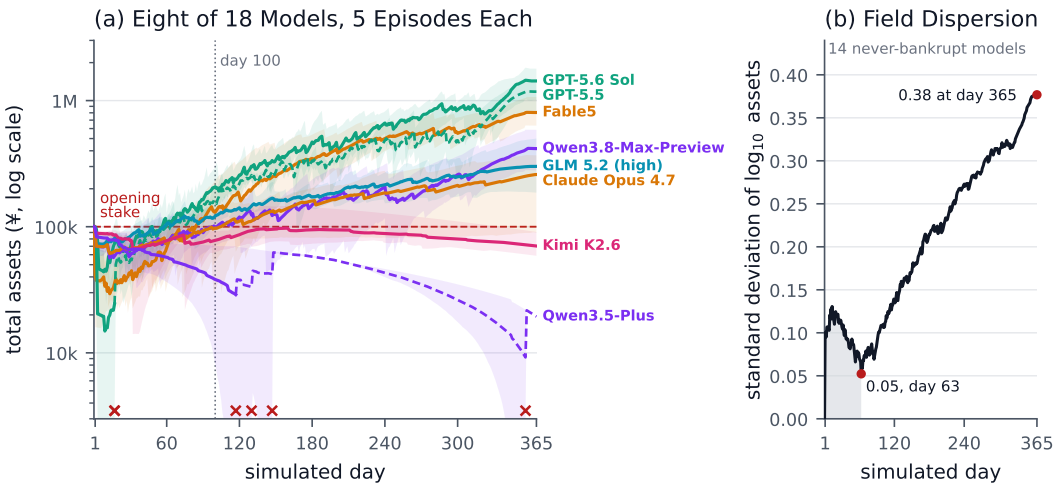


Figure 14: Total assets over the year for eight models spanning the ranking. The solid line is the mean over episodes still running, the band their min-to-max range, and a red cross marks a bankruptcy. The line turns dashed once fewer than five episodes remain, so the mean there counts survivors only.

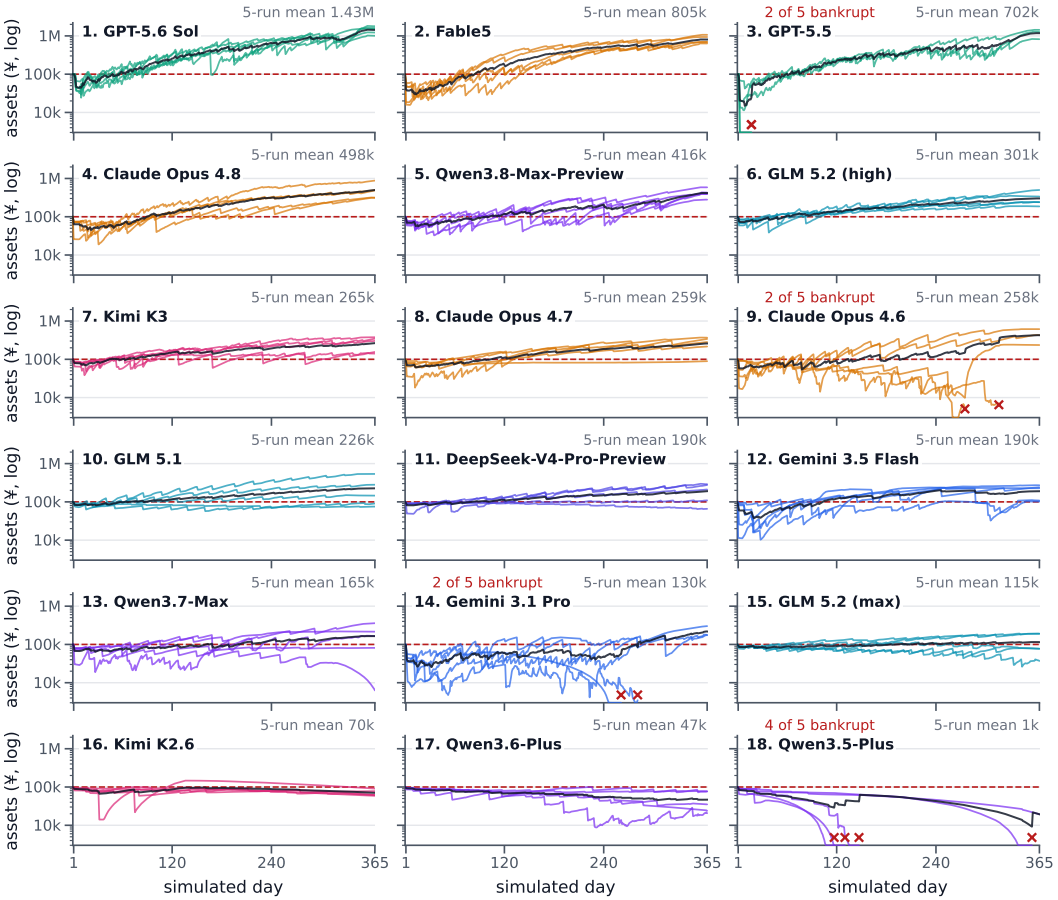


Figure 15: Total assets over the year for all 18 models, ordered by their mean at the end of the year. Each thin line is one episode and the black line their mean while they are still running. The dashed red line marks the ¥100,000 stake and a red cross marks a bankruptcy.

Table 11: Negotiation quality per model, 18 models and five episodes each, sorted by mean final assets, mean and standard deviation over the episodes, with † on open-weight models. CSE^{+} is the share of the bargaining range the buyer keeps on the honest deals it closes and ranks this dimension, SE^{+} scores a walk-away as zero, $\%\mathrm{Oracle}$ compares realized surplus against closing every honest deal at the supplier's floor, and AGR^{+} is the share of concluded honest sessions that ended in agreement. All four count honest suppliers only and drop sessions abandoned mid-bargain. Rounds to deal and deals closed cover fraudulent counterparts too and are volume readings rather than scores, which is why neither column is marked. Definitions sit in §E.1, and truncated bankrupt episodes are pooled into every mean. ‡GPT-5.5's spreads are population estimates, the rest sample estimates.

	Value Extraction and Agreement				Volume	
Model	SE^{+}	CSE^{+}	$\%\mathrm{Oracle}$	AGR^{+}	Rounds to Deal	Deals Closed
GPT-5.6 Sol (max)	0.671 ±0.047	0.672 ±0.046	67.0 ±4.3	0.998 ±0.004	3.84 ±0.30	282.4 ±121.8
Fable5 (max)	0.764 ±0.078	0.772 ±0.076	77.1 ±7.8	0.990 ±0.023	4.07 ±0.67	177.2 ±66.7
GPT-5.5‡	0.693 ±0.052	0.700 ±0.050	69.2 ±5.4	0.990 ±0.006	3.31 ±0.28	287.8 ±230.7
Claude Opus 4.8 (max)	0.656 ±0.043	0.662 ±0.040	67.2 ±4.2	0.992 ±0.011	3.38 ±0.13	221.4 ±62.9
Qwen3.8-Max-Preview†	0.713 ±0.021	0.713 ±0.021	71.6 ±2.8	**1.000 ±0.000**	3.52 ±0.10	230.4 ±23.8
GLM 5.2 (high)†	0.692 ±0.059	0.693 ±0.058	68.9 ±5.9	0.999 ±0.002	3.51 ±0.17	179.0 ±49.9
Kimi K3†	0.632 ±0.039	0.632 ±0.039	63.1 ±3.5	**1.000 ±0.000**	3.33 ±0.05	200.4 ±96.2
Claude Opus 4.7 (max)	**0.777 ±0.096**	**0.811 ±0.097**	**77.4 ±10.2**	0.960 ±0.055	3.88 ±0.32	74.4 ±28.4
Claude Opus 4.6 (max)	0.649 ±0.050	0.649 ±0.050	64.9 ±4.0	**1.000 ±0.000**	3.54 ±0.12	115.8 ±9.1
GLM 5.1†	0.657 ±0.034	0.662 ±0.031	65.4 ±4.6	0.992 ±0.008	3.37 ±0.12	131.2 ±45.1
DeepSeek-V4-Pro-Preview (max)†	0.651 ±0.061	0.654 ±0.064	65.4 ±5.7	0.997 ±0.005	3.43 ±0.21	129.2 ±52.2
Gemini 3.5 Flash	0.763 ±0.060	0.777 ±0.059	76.6 ±3.7	0.983 ±0.037	3.94 ±0.21	64.8 ±28.6
Qwen3.7-Max†	0.616 ±0.029	0.616 ±0.029	60.4 ±3.4	**1.000 ±0.000**	3.36 ±0.13	102.8 ±13.8
Gemini 3.1 Pro	0.657 ±0.049	0.660 ±0.048	67.4 ±6.3	0.996 ±0.008	3.93 ±0.70	75.6 ±26.5
GLM 5.2 (max)†	0.630 ±0.073	0.632 ±0.072	63.3 ±6.6	0.997 ±0.007	3.37 ±0.07	198.8 ±10.4
Kimi K2.6†	0.596 ±0.065	0.596 ±0.065	60.1 ±5.2	**1.000 ±0.000**	3.32 ±0.07	48.6 ±18.1
Qwen3.6-Plus†	0.625 ±0.070	0.625 ±0.070	61.4 ±5.4	**1.000 ±0.000**	3.35 ±0.13	74.4 ±18.4
Qwen3.5-Plus†	0.618 ±0.056	0.625 ±0.045	61.4 ±8.0	0.988 ±0.027	3.38 ±0.11	31.4 ±9.9

falls short of the 0.215 between the best and worst per-model CSE^{+} (§4.3). Fraudulent counterparts are held apart in the figure, since they carry the same ADVERSARIAL label as the hardest honest template while conceding on different parameters (Appendix D.3). Measured surplus against them is lower for 16 of the 18 models, an arithmetic effect of scoring every deal against the honest cost floor (§I.4.4, Equation (18)).

F.3 FRAUD AVOIDANCE

Table 12 reports payments to fraudulent suppliers, orders placed with them, and membership fees paid. Figure 18 shows $\mathrm{BadSpend}\%$ and spend by scam type. Figure 19 separates contact from purchase: for the median model, 22.0% of fraudulent suppliers contacted receive an order.

Models explore only a small share of the available suppliers. Over a full year, the widest search contacts 14.3% of the 424 honest suppliers and the narrowest 2.4% (§B.4). Screening statistics therefore describe a small subset selected by each model.

F.4 CASH FLOW AND SOLVENCY

Table 13 reports normalized and raw drawdown, minimum bank balance, overdrawn mornings, idle wallet balance, peak warehouse inventory, and bankrupt episodes. Figure 20 compares normalized drawdown with end-of-year total assets, and idle wallet balance with overdrawn days. Both use the daily records defined in §E.3.

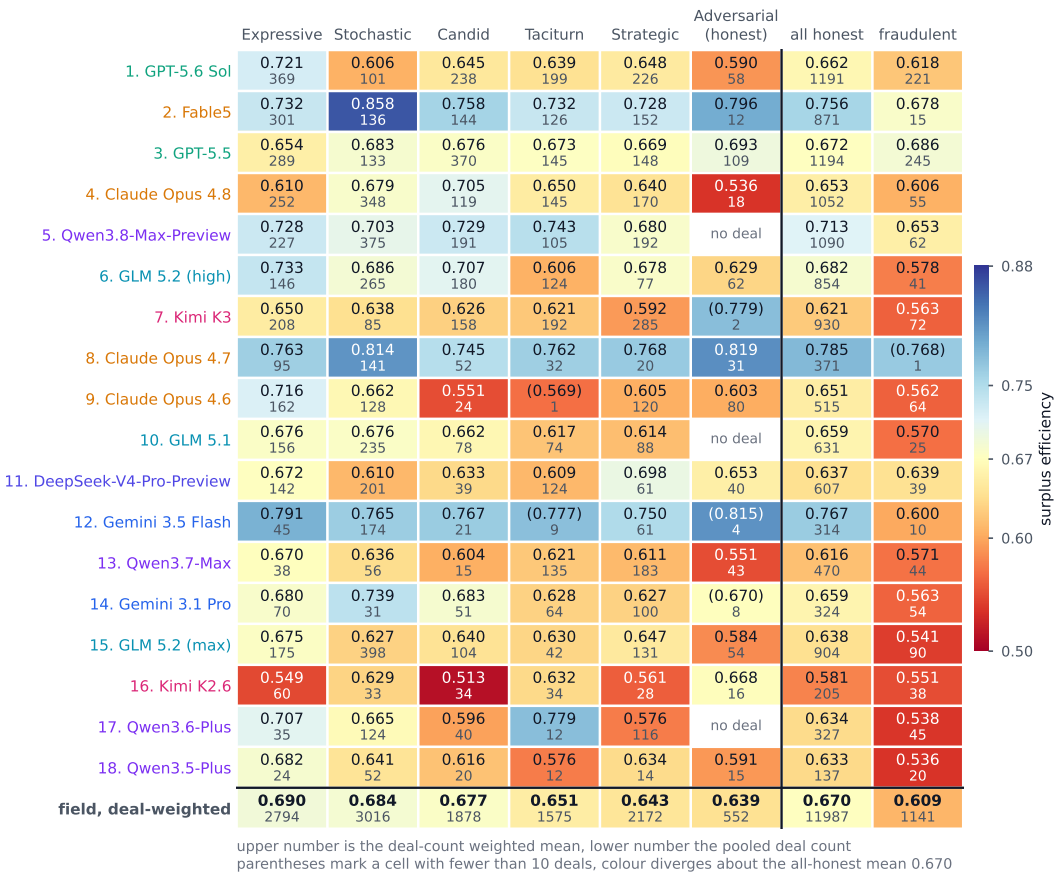


Figure 16: How much of each bargaining range models capture against each kind of supplier, over all 13,128 concluded agreements. Columns run from the easiest counterpart to the hardest, the deal count sits under each cell, and values in parentheses rest on fewer than 10 deals.

F.5 Operational Efficiency

Table 14 reports model turns, tool calls, profit per call, context evictions, memory usage, and the eight activity bands in Figure 30b. These per-model results complement the aggregate patterns in Section 4.3.

F.6 Operations Execution

Table 15 reports expected and realized returns, the attributed return channels, refund costs, fulfillment, shipping tiers, and store activity. Expected return channels accrue at dispatch; realized returns count units that come back. Their definitions and the distinct denominator for on-time fulfillment appear in §E.5.

Almost no model keeps the four-store cap filled. Only GPT-5.6 Sol, GLM 5.2 (high) and Qwen3.8-Max-Preview average more than 3.5 stores trading per day. Cumulative openings still run past the cap for most of the field, since every opening beyond it replaces a store (§4.3).

F.7 Learning over the Horizon

Table 16 places repeat-purchase statistics and their permutation null on the left, and three comparisons over the episode on the right. The former group deals by supplier and SKU; the latter compare earlier and later sessions using the inclusion and splitting rules in §E.6. Improvement on one measure need not imply improvement on the other.

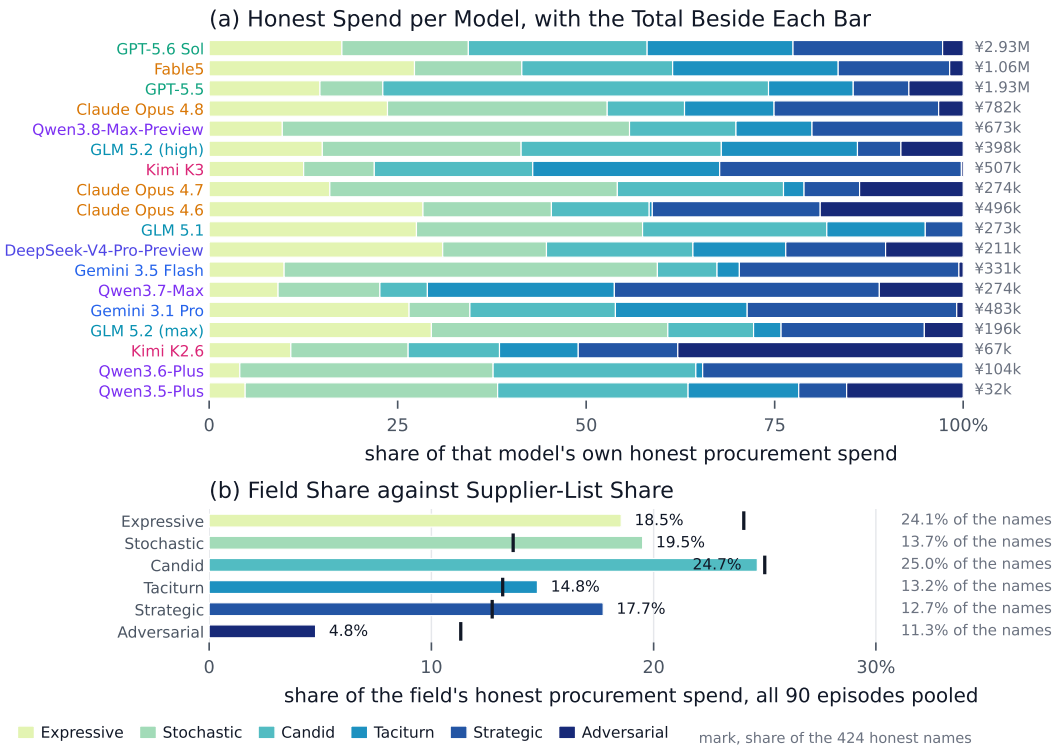


Figure 17: Honest procurement money split over the six counterpart behavior templates, which run easiest to hardest on the field-wide surplus order of Figure 16. Panel (a) gives every model's five-run mean spend with honest suppliers as a share of its own honest total, that total printed at the right of each bar, models in descending total assets. Panel (b) pools the 90 episodes, ¥55.1M paid to honest suppliers, and sets the share each template received against the share of the 424 honest names that carry it, so a bar reaching past its mark was bought from more heavily than the supplier list alone would give. The hardest template takes 4.8% of the spend against 11.3% of the names and the easiest 18.5% against 24.1%. Spend with fraudulent suppliers is recorded separately and stays out of both panels.

Table 12: Fraud avoidance, 18 models and five episodes each, sorted by mean end-of-year assets, † marking open-weight models. The first four columns are means over the five episodes with their sample standard deviation, spend and orders counting every purchase order that cleared the bank account and suppliers ordered from counting distinct names within an episode. The two right-hand columns are totals over the five episodes, three payments of ¥1,000 in the whole evaluation, all of them from Gemini 3.5 Flash and Qwen3.5-Plus. BadSpend% ranks the dimension and is drawn per model in Figure 18 rather than repeated here. Definitions sit in §E.2.

	Cash and Orders Reaching Fraudulent Suppliers				Membership Fees Paid	
Model	spend ¥k	fraudulent orders	all orders	suppliers ordered from	payments	¥
GPT-5.6 Sol (max)	689.1 ±376.0	44.2 ±26.4	282.4 ±121.8	5.2 ±1.3	0	0
Fable5 (max)	49.7 ±89.2	3.0 ±5.6	177.2 ±66.7	0.8 ±0.8	0	0
GPT-5.5	535.6 ±504.0	49.0 ±50.2	287.8 ±230.7	5.6 ±3.7	0	0
Claude Opus 4.8 (max)	80.3 ±156.9	11.0 ±10.4	221.4 ±62.9	1.8 ±1.5	0	0
Qwen3.8-Max-Preview†	57.8 ±78.3	12.4 ±9.7	230.4 ±23.8	1.4 ±0.5	0	0
GLM 5.2 (high)†	10.6 ±17.1	8.2 ±13.4	179.0 ±49.9	1.0 ±0.7	0	0
Kimi K3†	23.4 ±30.3	14.4 ±28.4	200.4 ±96.2	1.0 ±1.2	0	0
Claude Opus 4.7 (max)	**0.4 ±0.9**	**0.2 ±0.4**	74.4 ±28.4	**0.2 ±0.4**	0	0
Claude Opus 4.6 (max)	164.9 ±229.2	12.8 ±17.2	115.8 ±9.1	0.8 ±0.8	0	0
GLM 5.1†	4.7 ±7.2	5.0 ±7.5	131.2 ±45.1	0.4 ±0.5	0	0
DeepSeek-V4-Pro-Preview (max)†	23.3 ±39.0	7.8 ±10.8	129.2 ±52.2	0.8 ±0.8	0	0
Gemini 3.5 Flash	8.4 ±16.7	2.4 ±2.2	65.2 ±28.9	1.6 ±0.9	2	2,000
Qwen3.7-Max†	67.4 ±150.8	8.8 ±19.7	102.8 ±13.8	**0.2 ±0.4**	0	0
Gemini 3.1 Pro	89.4 ±93.9	10.8 ±10.0	75.6 ±26.5	3.0 ±1.2	0	0
GLM 5.2 (max)†	46.9 ±40.1	18.0 ±17.8	198.8 ±10.4	1.0 ±0.7	0	0
Kimi K2.6†	14.2 ±19.6	7.6 ±10.4	48.6 ±18.1	0.8 ±1.1	0	0
Qwen3.6-Plus†	12.1 ±16.1	9.0 ±13.7	74.4 ±18.4	0.8 ±0.8	0	0
Qwen3.5-Plus†	13.1 ±17.6	4.2 ±5.0	31.6 ±10.1	1.2 ±1.3	1	1,000

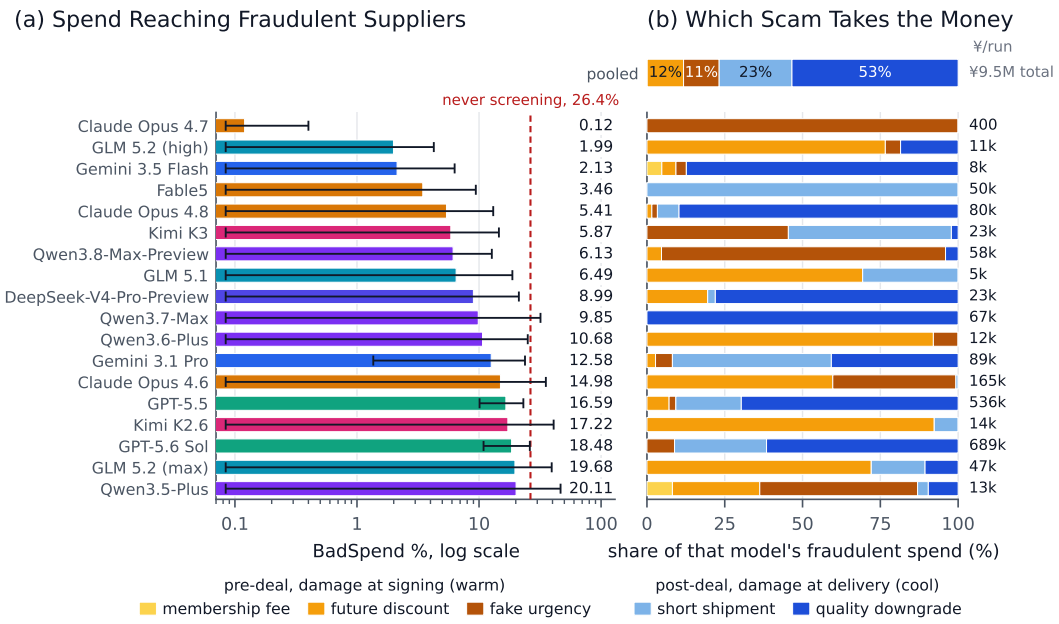


Figure 18: (a) The share of procurement spend that reaches fraudulent suppliers, cleanest model first, with whiskers over the five episodes. The dashed line is what a buyer who never screened would spend. (b) The same money split over the five scams, scaled within each model, with the cash it amounts to on the right.

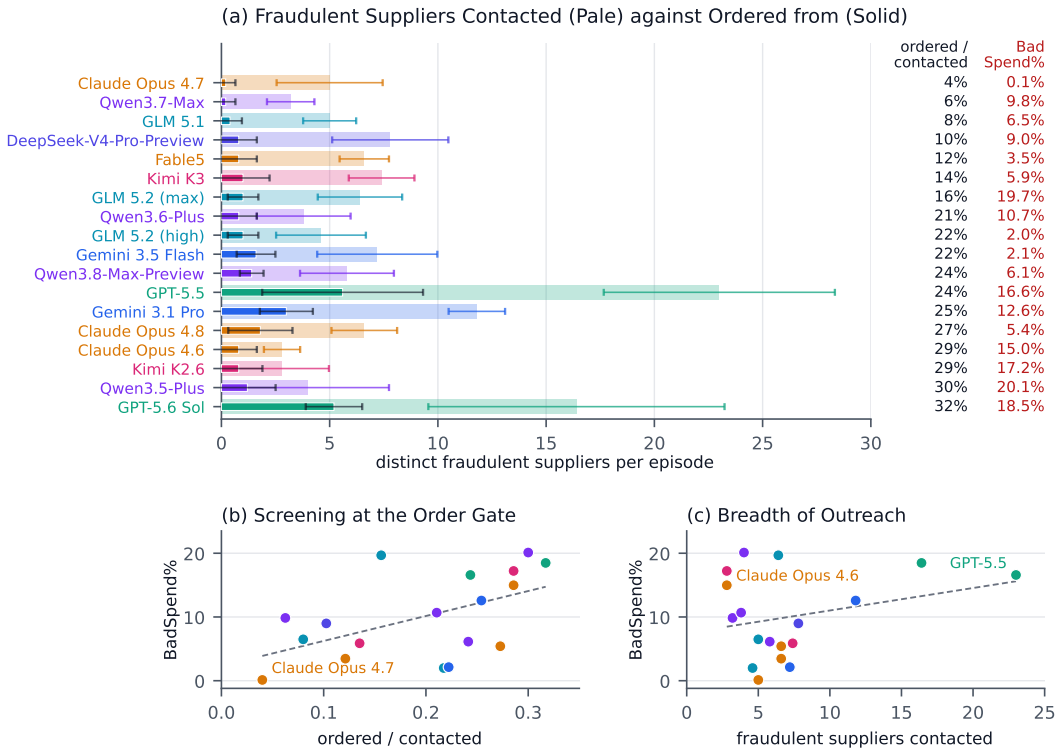


Figure 19: How fraudulent suppliers get through, 18 models and 5 episodes each, counted as distinct suppliers per episode. (a) Pale bars are suppliers contacted, solid bars suppliers ordered from, and the right column gives the share of contacts that became orders. (b) That share against fraudulent spend. (c) The number contacted against the same.

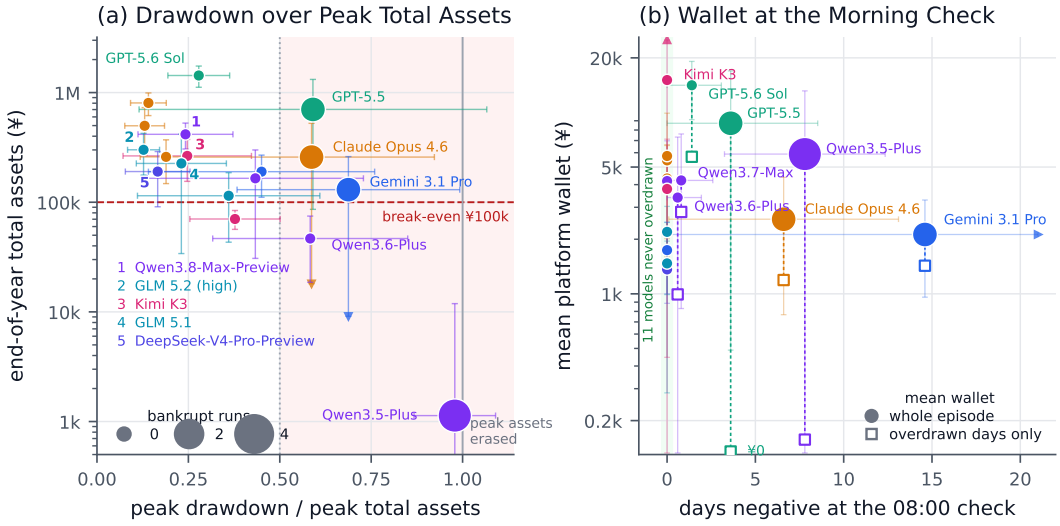


Figure 20: (a) The deepest fall in total assets, as a share of that episode's peak, against end-of-year total assets. Bubble area grows with the number of bankrupt episodes and the tinted band holds the five models that gave up more than half their peak. (b) Money left sitting in the platform wallet against days the bank balance was negative at the morning check.

Table 13: Cash-flow and solvency results, 18 models, five episodes each, sorted by mean end-of-year total assets, with † on open-weight models and definitions in Appendix E.3. DD/Peak is each episode's largest peak-to-trough fall in total assets over the peak total assets of that same episode and ranks the dimension, five models averaging more than half. DD repeats the fall in ¥ without normalization, so a larger business books a larger number for the same discipline. Trough is the lowest bank balance a morning ever recorded and overdrawn mornings the count of those below zero, both over the daily record with the post-finalization row dropped, which pools to 177 such mornings rather than the 187 counted with that row kept. Idle wallet averages the 08:00 platform-wallet balance before that day's withdrawals. Peak units is the largest warehouse count of the year. Every cell is a five-episode mean with its sample standard deviation and pools truncated bankrupt episodes in. Tinting marks the field's lowest ratio and its lowest idle wallet. Eleven models never overdrew the bank account and fourteen never went bankrupt, so the floor of those two columns is shared too widely to mark.

	Drawdown and trough			Solvency and idle capital			
Model	DD/Peak ratio	DD ¥k	trough ¥k	overdrawn mornings	idle wallet ¥k	peak units	bankrupt runs
GPT-5.6 Sol (max)	0.278 ±0.084	416 ±206	0.7 ±1.4	1.4 ±1.7	14.12 ±4.99	2,323 ±1,762	0/5
Fable5 (max)	0.141 ±0.049	114 ±58	1.3 ±1.2	0.0 ±0.0	5.44 ±1.13	1,686 ±877	0/5
GPT-5.5	0.591 ±0.475	207 ±96	−3.5 ±5.5	3.6 ±4.9	8.72 ±8.23	2,424 ±1,421	2/5
Claude Opus 4.8 (max)	0.130 ±0.054	59 ±22	11.1 ±7.6	0.0 ±0.0	4.33 ±2.31	1,195 ±356	0/5
Qwen3.8-Max-Preview†	0.242 ±0.130	99 ±54	23.6 ±20.2	0.0 ±0.0	4.18 ±2.79	2,472 ±1,200	0/5
GLM 5.2 (high)†	**0.127 ±0.044**	41 ±24	47.7 ±26.9	0.0 ±0.0	2.20 ±0.84	689 ±455	0/5
Kimi K3†	0.247 ±0.176	55 ±22	35.7 ±23.3	0.0 ±0.0	15.09 ±18.19	708 ±433	0/5
Claude Opus 4.7 (max)	0.189 ±0.071	46 ±24	40.4 ±24.5	0.0 ±0.0	5.75 ±4.15	1,184 ±376	0/5
Claude Opus 4.6 (max)	0.587 ±0.336	120 ±29	−0.6 ±1.7	6.6 ±6.5	2.58 ±1.81	1,741 ±799	2/5
GLM 5.1†	0.231 ±0.123	39 ±11	57.3 ±10.3	0.0 ±0.0	1.38 ±1.09	556 ±302	0/5
DeepSeek-V4-Pro-Preview (max)†	0.166 ±0.088	30 ±15	67.0 ±19.7	0.0 ±0.0	**1.36 ±0.48**	459 ±69	0/5
Gemini 3.5 Flash	0.450 ±0.309	91 ±68	14.9 ±27.9	0.0 ±0.0	1.74 ±0.74	1,117 ±650	0/5
Qwen3.7-Max†	0.433 ±0.296	70 ±35	26.1 ±25.0	0.8 ±1.8	4.22 ±3.39	1,289 ±306	0/5
Gemini 3.1 Pro	0.688 ±0.305	99 ±17	−1.8 ±3.0	14.6 ±26.8	2.12 ±1.17	1,512 ±579	2/5
GLM 5.2 (max)†	0.360 ±0.249	42 ±20	44.8 ±25.6	0.0 ±0.0	1.47 ±0.48	953 ±569	0/5
Kimi K2.6†	0.377 ±0.123	43 ±22	42.1 ±22.5	0.0 ±0.0	3.79 ±3.34	798 ±557	0/5
Qwen3.6-Plus†	0.583 ±0.267	58 ±27	25.5 ±26.0	0.6 ±1.3	3.40 ±3.92	1,248 ±436	0/5
Qwen3.5-Plus†	0.979 ±0.111	98 ±11	−4.6 ±3.7	7.8 ±4.5	5.90 ±7.26	1,608 ±781	4/5

Table 14: Operational efficiency, 18 models, five episodes each, sorted by mean final assets, definitions in §E.4. Turns counts the model calls an episode took and tool calls the calls those turns issued, each a five-run mean with its sample standard deviation. ¥ per call is the dimension's ranking measure, mean end-of-year assets less the ¥100,000 stake over mean tool calls, so it carries no per-episode spread and three models fall below zero. Evict. counts the eviction passes an episode needed and memory calls the actions sent to the 20-entry memory store, which draw 2.1% of the field's calls. The eight bands partition the 18 tools as in Figure 30(b) and sum to 100 across each row. Truncated bankrupt episodes are pooled into every mean.

	What the Year Cost					Share of Calls by Activity Band, %							
Model	**Turns**	**Tool calls**	**¥ per call**	**Evict.**	**Memory calls**	**Wait**	**Ship**	**Pub.**	**W/dr.**	**Neg.**	**Poll**	**Mem.**	**Other**
GPT-5.6 Sol (max)	1,367 ±46	3,668 ±153	+363	36.0 ±7.7	55.2	9.8	20.1	22.5	9.5	7.5	23.1	1.5	6.0
Fable5 (max)	945 ±48	1,469 ±112	**+479**	13.0 ±2.3	34.4	24.8	24.0	9.3	23.1	8.9	3.1	2.3	4.4
GPT-5.5	1,306 ±1,131	3,143 ±2,677	+192	57.0 ±49.3	135.2	7.0	15.8	15.9	6.7	6.4	31.0	4.3	12.9
Claude Opus 4.8 (max)	812 ±142	1,497 ±149	+266	14.0 ±1.6	69.6	24.4	23.6	11.5	19.9	6.9	5.4	4.6	3.7
Qwen3.8-Max-Preview	962 ±210	1,826 ±246	+173	18.6 ±1.7	38.2	19.9	19.5	17.1	16.6	5.4	12.4	2.1	6.9
GLM 5.2 (high)	672 ±188	1,467 ±104	+137	12.4 ±2.4	37.2	24.9	24.3	11.9	21.3	7.2	4.0	2.5	3.8
Kimi K3	878 ±47	1,334 ±177	+123	11.4 ±2.6	53.0	27.3	27.1	13.1	11.5	8.1	5.0	4.0	4.0
Claude Opus 4.7 (max)	432 ±20	1,023 ±142	+156	7.4 ±1.5	13.4	35.6	34.8	5.1	10.2	5.2	3.0	1.3	4.8
Claude Opus 4.6 (max)	788 ±89	1,221 ±130	+129	8.0 ±1.0	18.6	27.5	26.6	7.0	25.6	5.9	2.3	1.5	3.4
GLM 5.1	852 ±23	1,333 ±120	+94	8.4 ±1.5	15.2	27.3	26.7	8.6	23.1	5.7	4.1	1.1	3.2
DeepSeek-V4-Pro-Preview (max)	678 ±204	1,337 ±191	+68	8.6 ±2.1	57.0	27.2	24.7	7.5	18.4	6.1	8.2	4.3	3.6
Gemini 3.5 Flash	2,628 ±653	2,508 ±457	+36	40.0 ±10.8	12.6	14.4	25.5	5.2	11.5	2.8	36.5	0.5	3.7
Qwen3.7-Max	782 ±188	1,108 ±186	+59	8.4 ±0.9	17.6	32.9	32.2	7.4	15.0	4.2	3.3	1.6	3.4
Gemini 3.1 Pro	1,141 ±551	1,605 ±479	+19	18.4 ±7.6	2.0	20.3	29.9	5.1	18.1	6.8	15.1	0.1	4.6
GLM 5.2 (max)	931 ±45	1,430 ±205	+10	12.8 ±1.6	39.6	25.5	25.2	14.1	16.6	7.7	4.3	2.8	3.9
Kimi K2.6	1,017 ±165	1,087 ±154	−27	7.2 ±1.5	3.2	33.5	29.5	3.6	4.7	3.9	20.7	0.3	3.8
Qwen3.6-Plus	1,023 ±203	1,325 ±137	−40	10.0 ±2.4	28.0	27.5	43.4	7.0	9.1	3.7	4.4	2.1	2.8
Qwen3.5-Plus	800 ±375	1,076 ±569	−92	7.4 ±3.6	0.0	20.6	35.3	1.7	7.8	4.4	24.9	0.0	5.3

Table 15: Operations execution, 18 models, five episodes each, sorted by mean end-of-year assets, with † on open-weight models and every definition in §E.5. Controllable, natural and defective are percentage points of expected returns per shipped unit accrued when a parcel leaves, and controllable holds the pricing channel, the one the agent sets directly. Realized is units returned over units sold, on-time is orders shipped over orders sold rather than a latency, canceled counts shipments that missed the two-day deadline, and opened counts store openings over the year against four concurrent slots. Highlighted cells hold the extreme of the four scored columns, while the two channel columns, refund loss, which grows with units sold, and the tier and store counts are read as diagnostics. ‡Two GPT-5.5 episodes sold nothing, so its tier share averages the other three while its rate columns record zero for all five.

	Returns					Fulfillment			Stores	
Model	Controllable pp	Realized %	Natural pp	Defective pp	Refund loss ¥k	On-time %	Canceled orders	Fast ship %	Opened	Reopens
GPT-5.6 Sol (max)	4.46 ±1.07	23.38 ±1.82	16.70 ±2.12	2.36 ±1.64	1,624 ±640	99.91 ±0.13	2.6 ±5.8	96.1 ±2.8	9.0 ±2.0	3.8 ±1.6
Fable5 (max)	0.22 ±0.47	10.77 ±6.82	10.16 ±6.43	0.00 ±0.00	222 ±151	99.94 ±0.14	3.8 ±8.5	20.0 ±44.7	5.2 ±1.1	1.4 ±1.5
GPT-5.5‡	0.90 ±1.29	11.91 ±10.92	9.40 ±8.64	1.43 ±1.61	668 ±641	60.00 ±54.77	**0.0 ±0.0**	97.5 ±2.1	6.0 ±6.7	3.0 ±4.6
Claude Opus 4.8 (max)	0.20 ±0.99	8.90 ±7.09	8.19 ±6.31	0.62 ±0.93	196 ±266	99.86 ±0.06	3.8 ±3.6	100.0 ±0.1	3.8 ±1.1	0.4 ±0.5
Qwen3.8-Max-Preview†	0.38 ±0.57	**5.96 ±0.76**	5.31 ±0.80	0.06 ±0.08	102 ±39	99.89 ±0.12	5.8 ±6.2	14.7 ±32.9	5.4 ±0.5	1.4 ±0.5
GLM 5.2 (high)†	0.59 ±1.05	11.61 ±5.94	10.85 ±6.47	0.14 ±0.30	101 ±102	**99.95 ±0.06**	0.6 ±1.3	0.3 ±0.5	4.0 ±0.0	0.0 ±0.0
Kimi K3†	2.01 ±2.38	16.62 ±9.45	14.79 ±7.35	0.02 ±0.05	189 ±134	99.79 ±0.18	0.8 ±1.8	38.8 ±53.1	4.2 ±0.8	0.6 ±0.9
Claude Opus 4.7 (max)	1.75 ±3.08	9.79 ±6.11	8.07 ±3.42	0.00 ±0.00	50 ±16	99.60 ±0.52	9.0 ±16.8	20.9 ±44.2	3.6 ±0.5	0.0 ±0.0
Claude Opus 4.6 (max)	0.76 ±0.80	22.20 ±8.68	21.25 ±8.27	0.00 ±0.00	267 ±251	99.95 ±0.07	**0.0 ±0.0**	0.0 ±0.0	3.4 ±0.9	0.6 ±1.3
GLM 5.1†	0.09 ±1.13	11.90 ±1.90	11.93 ±1.86	0.00 ±0.00	56 ±39	99.94 ±0.07	0.8 ±1.1	1.2 ±1.5	3.6 ±1.1	0.2 ±0.4
DeepSeek-V4-Pro-Preview (max)†	1.54 ±1.78	9.70 ±6.64	6.39 ±2.54	1.90 ±4.24	69 ±112	99.43 ±0.58	7.0 ±9.7	16.3 ±36.3	3.2 ±0.4	0.0 ±0.0
Gemini 3.5 Flash	0.38 ±1.33	14.64 ±4.53	14.09 ±3.38	0.19 ±0.42	124 ±76	99.72 ±0.42	3.2 ±7.2	31.9 ±43.2	5.6 ±4.5	2.2 ±3.8
Qwen3.7-Max†	**−0.05 ±1.07**	12.62 ±10.52	11.04 ±8.64	1.37 ±3.05	117 ±168	99.91 ±0.09	1.0 ±1.2	0.0 ±0.0	3.0 ±1.0	0.0 ±0.0
Gemini 3.1 Pro	2.03 ±1.62	20.74 ±7.19	17.65 ±5.51	0.88 ±1.52	172 ±103	99.66 ±0.34	4.8 ±5.4	46.8 ±49.6	4.4 ±1.5	0.4 ±0.5
GLM 5.2 (max)†	2.24 ±2.29	16.09 ±9.23	13.90 ±7.82	0.03 ±0.08	79 ±57	99.84 ±0.14	2.6 ±3.7	18.3 ±40.6	4.4 ±1.3	0.6 ±0.9
Kimi K2.6†	0.42 ±1.23	11.54 ±2.87	10.95 ±2.82	0.00 ±0.00	12 ±5	99.81 ±0.31	3.0 ±5.1	0.0 ±0.0	4.0 ±0.7	0.4 ±0.9
Qwen3.6-Plus†	3.03 ±3.58	20.77 ±8.46	17.63 ±6.52	0.00 ±0.00	49 ±45	99.80 ±0.16	0.6 ±0.9	0.1 ±0.1	3.0 ±0.7	0.0 ±0.0
Qwen3.5-Plus†	5.88 ±7.61	22.06 ±13.38	15.81 ±8.53	0.10 ±0.23	10 ±5	99.50 ±0.45	2.6 ±4.7	20.6 ±43.8	4.6 ±2.6	1.0 ±1.7

Table 16: Learning over the horizon, 18 models and five episodes each, sorted by mean end-of-year assets, † marking open-weight models. The left block scores repeat purchases of one SKU from one supplier, $\mathrm{AnchorRatio}$ being the overpayment against the agent's own best price with that pair divided by what a reshuffling of the same prices would have cost, so a value above 1 says the order in which the prices arrived cost more than a random order of them. Raw $\mathrm{AnchorRegret}$ and its null are printed for reference and are not comparable across models, since tightly clustered quotes score a low regret without any price being held, and $|z_{\mathrm{anchor}}| > 1.96$ is the two-sided 5% threshold. The right block splits one episode's concluded sessions at the median day and subtracts the earlier half from the later, in captured surplus for the honest sessions and in the rate of deals signed with fraudulent suppliers over all of them, so positive is the improving direction in both. Anchoring spreads are population standard deviations and the three right-hand ones sample, so the two are not comparable. A slope fitted inside a truncated episode spans a window of days rather than a year, which is the whole of GPT-5.5's entry in that column. ‡Three episodes rather than five, the two bankrupt runs opening no repeat pair. Definitions sit in §E.6.

	Sequential Price Discipline						Within-Year Improvement		
Model	$\mathrm{AnchorRatio}$	$\mathrm{AnchorRegret}$	null	NewLow	z_{anchor}	re-orders	surplus half-lift	slope 10^{-3}/day	fraud half-lift
GPT-5.6 Sol (max)	1.22 ±0.10	0.103 ±0.013	0.085 ±0.016	0.129 ±0.038	−6.18	913	−0.063 ±0.061	−0.35 ±0.26	−0.153 ±0.157
Fable5 (max)	1.57 ±0.27	0.036 ±0.013	0.023 ±0.008	0.028 ±0.038	−8.65	636	−0.067 ±0.045	−0.42 ±0.25	−0.018 ±0.041
GPT-5.5	1.04 ±0.15	0.047 ±0.009	0.046 ±0.011	0.193 ±0.048	−0.87	700‡	−0.040 ±0.040	−24.51 ±54.63	−0.041 ±0.137
Claude Opus 4.8 (max)	1.31 ±0.37	0.038 ±0.014	0.028 ±0.005	0.086 ±0.058	−5.95	751	−0.055 ±0.057	−0.44 ±0.32	−0.039 ±0.115
Qwen3.8-Max-Preview†	**0.83 ±0.20**	0.066 ±0.010	0.083 ±0.018	**0.256 ±0.091**	+4.24	773	+0.001 ±0.070	+0.07 ±0.37	−0.062 ±0.025
GLM 5.2 (high)†	1.43 ±0.10	0.115 ±0.026	0.080 ±0.015	0.058 ±0.050	−8.35	703	−0.076 ±0.030	−0.44 ±0.17	−0.007 ±0.025
Kimi K3†	1.51 ±0.11	0.096 ±0.053	0.062 ±0.030	0.046 ±0.042	−8.50	733	−0.084 ±0.065	−0.48 ±0.31	−0.043 ±0.065
Claude Opus 4.7 (max)	1.42 ±0.41	0.031 ±0.019	0.024 ±0.013	0.122 ±0.086	−2.74	222	+0.011 ±0.063	−0.01 ±0.39	−0.004 ±0.009
Claude Opus 4.6 (max)	1.27 ±0.25	0.051 ±0.024	0.039 ±0.015	0.202 ±0.117	−2.83	380	−0.052 ±0.039	−0.38 ±0.26	−0.092 ±0.219
GLM 5.1†	1.56 ±0.14	0.092 ±0.046	0.059 ±0.028	0.012 ±0.024	−7.52	496	−0.096 ±0.048	−0.50 ±0.37	−0.105 ±0.194
DeepSeek-V4-Pro-Preview (max)†	1.23 ±0.21	0.037 ±0.017	0.032 ±0.017	0.149 ±0.143	−3.24	473	−0.042 ±0.027	−0.19 ±0.15	−0.011 ±0.174
Gemini 3.5 Flash	0.92 ±0.46	0.029 ±0.033	0.023 ±0.017	0.122 ±0.065	+1.25	175	−0.003 ±0.058	−0.05 ±0.30	+0.003 ±0.037
Qwen3.7-Max†	1.43 ±0.30	0.062 ±0.036	0.041 ±0.019	0.144 ±0.158	−5.42	365	−0.045 ±0.026	−0.31 ±0.22	−0.144 ±0.323
Gemini 3.1 Pro	1.38 ±0.18	0.082 ±0.014	0.060 ±0.009	0.081 ±0.061	−4.05	201	−0.064 ±0.061	−0.88 ±1.18	−0.056 ±0.105
GLM 5.2 (max)†	1.36 ±0.20	0.117 ±0.064	0.084 ±0.038	0.065 ±0.070	−6.90	747	−0.066 ±0.025	−0.41 ±0.25	−0.117 ±0.094
Kimi K2.6†	1.55 ±0.44	0.051 ±0.036	0.029 ±0.017	0.071 ±0.069	−2.76	107	−0.077 ±0.081	−0.82 ±0.62	−0.055 ±0.397
Qwen3.6-Plus†	1.33 ±0.30	0.050 ±0.012	0.040 ±0.014	0.096 ±0.089	−2.60	224	−0.077 ±0.081	−0.38 ±0.34	−0.007 ±0.149
Qwen3.5-Plus†	1.86 ±0.16	0.040 ±0.035	0.023 ±0.021	0.109 ±0.141	−3.01	48	−0.050 ±0.042	−1.22 ±0.80	−0.044 ±0.076

F.8 Profiles and Rank Crossings Across Dimensions

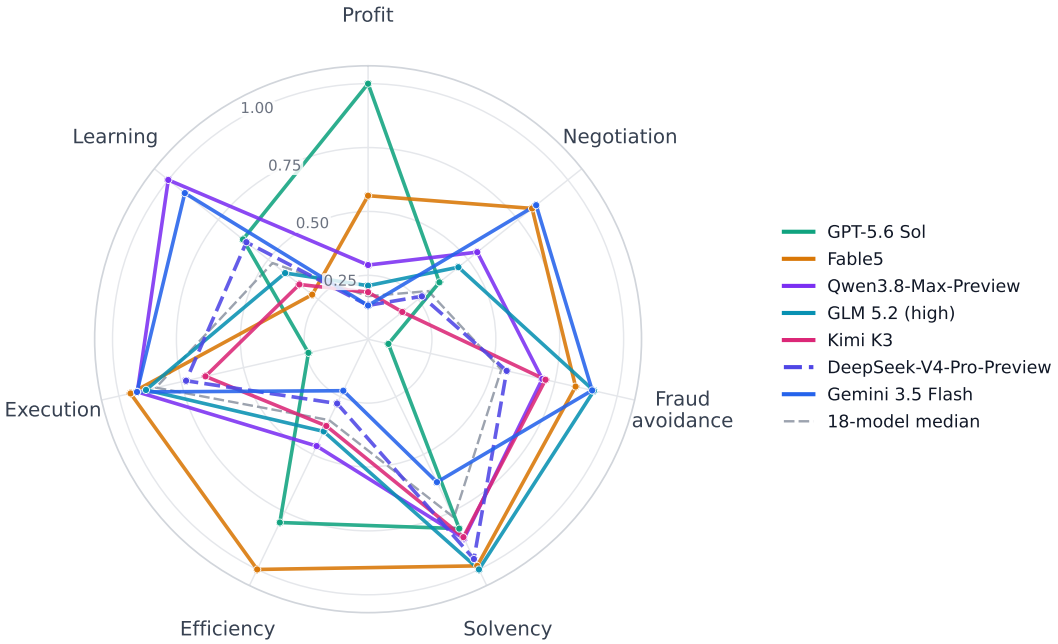


Figure 21: Seven evaluation axes for the strongest model of each vendor family by mean total assets, the primary score, and the six evaluation dimensions clockwise from the top, each on the ranking metric §3.7 gives it. Axes are min-max normalized over the eighteen model means, so 1 is the best of the eighteen and 0 the worst, with a dashed polygon at the median. Fraud avoidance, solvency, execution, and learning are sign-flipped, so a vertex further from the center is better on every axis.

Capability rankings differ substantially within each model. Figure 22 applies the normalization in Figure 3 to all 18 models. The mean gap between a model's best and worst rank is 10.2 places across the seven axes. No model ranks in the top five on every axis; Qwen3.8-Max-Preview comes closest, with a worst rank of 7.

Figure 23 compares the rankings directly, adding money-weighted $\%\mathrm{Oracle}$ and sequential price discipline ($\mathrm{AnchorRatio}$; §4.3). Across seven transitions and 153 model pairs, 311 pairwise orderings reverse. Operations execution differs most from the asset ranking, by 4.2 places per model on average; operational efficiency differs least, by 0.7 places. The diagnostics therefore vary in how much information they add to the financial ranking.



Figure 22: The same capability profiles for all 18 models, with a dashed polygon at the median. Axes clockwise from the top are profit, negotiation, fraud avoidance, solvency, efficiency, execution and learning, with fraud avoidance, solvency, execution and learning flipped so higher is better.

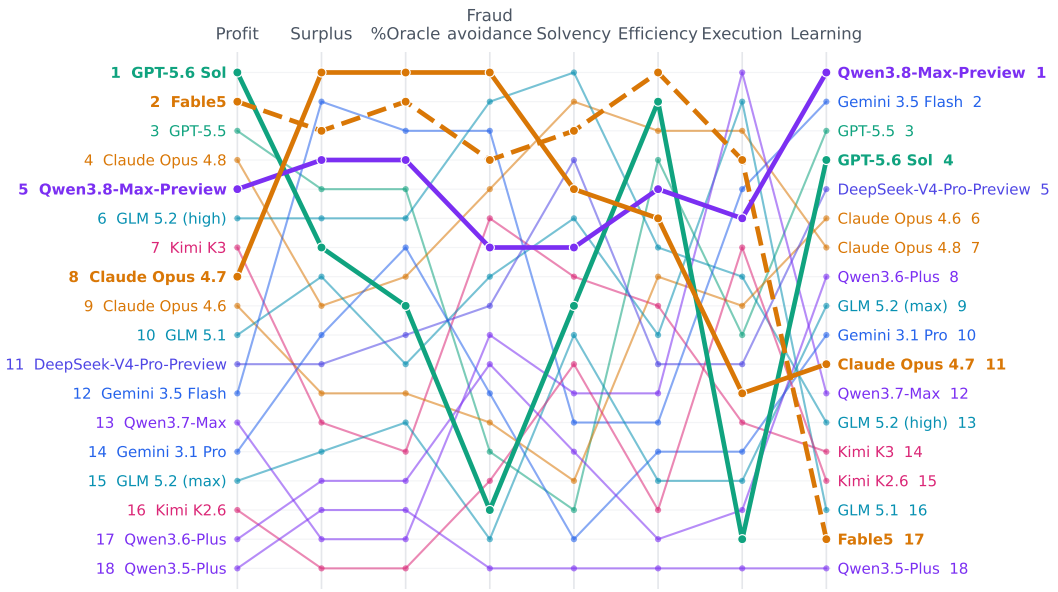


Figure 23: How the ranking changes from one dimension to the next, all 18 models, rank 1 at the top of every column. Left labels give the profit order and right labels the order on sequential price discipline. Every column is signed so a higher position is better.

G Failure Case Studies and Detection Rules

The following cases connect observed payments to supplier interactions. We examine membership fees, short delivery, and defective lots; the latter two account for the most fraudulent spending (§4.3). Section I.4.4 describes the persuasion strategies. Transcript excerpts preserve prices, dates, and wording, with *[. . .]* marking omissions and gray text showing episode records. A bankruptcy case and the detection rules follow.

G.1 Membership Fee Upsell

A membership-fee supplier pitches a one-time charge for a premium program and refuses every product order until it is paid (Table 9). The pitch below arrives twice, nine months apart.

Supplier Titan Commerce, `SUP-0262`, declared type `vip_fee` · **Episode** GPT-5.5 run 3 · 2026-01-02 and 2026-09-22

Titan Commerce *(2026-01-02) FlyLight Rotary Triple-Blade Waterproof Electric Shaver, Initial Offer ¥375.65 [. . .] I recommend our "Premium Membership" for just ¥1000.0, providing access to exclusive member pricing [. . .] approximately 30% of reference prices.*

Titan Commerce *(2026-09-22) [. . .] FlyLight Rotary Triple-Blade Waterproof Electric Shaver, Initial Offer ¥375.65 [. . .]*

`[ledger] no order ever placed with this supplier`

The supplier's January and September opening quotes are identical, despite its claim of a member discount. The agent can compare them using the message record. GPT-5.5 declined every pitch; few models paid the fee (§D.5).

G.2 Short Delivery

A short-delivery supplier uses the honest cost floor and ordinary bargaining behavior, but charges for the full order while scheduling only a partial delivery (Table 9). The record below shows how the shortfall becomes visible.

Supplier Vertex Supply, `SUP-0232`, declared type `qty_bait` · **Episode** GPT-5.5 run 3 · 2026-10-12, seventh contact of the year

---

Agent *We need a replenishment order for our Q4 campaign [...]*

Vertex Supply *Here are the best delivered wholesale prices [...] please confirm your order [...]*

```
ORDER CONFIRMED: 97a946841d25 x150 at ¥112.00/unit.
ORDER CONFIRMED: ebb843a1ff50 x250 at ¥116.00/unit.
ORDER CONFIRMED: 2b40172827f1 x250 at ¥116.00/unit.
[arrival notice, 2026-10-18] "103x TrueMicro Ibuprofen
Extended-Release Tablets 24-Count [...]  from Vertex Supply"
```

The delivery notice reports 103 units of the ibuprofen SKU six days after acceptance, against 150 confirmed units. It does not place ordered and delivered quantities side by side, and the agent did not reconcile the warehouse with its orders. GPT-5.5 placed 17 orders with short-delivery suppliers over the year. Claude Opus 4.6 contacted this supplier once without ordering.

G.3 Defective Lot

A defective-lot supplier does not reveal the defect in its conversation or quotes. Every unit it delivers is flagged defective, lifting the affected SKU's return rate to the defective blend of Equation (6) (Table 9). Returns arrive days after the shipment that caused them, so a re-order can precede the first sign of trouble.

Supplier Crystal Direct, `SUP-0038`, declared type `quality_downgrade` · **Episode** GPT-5.5 run 3 · 2026-08-06, re-order of a 2026-07-19 lot

---

```
[prior, 2026-07-19] cb6be988d634 x12 accepted ¥720.00
[return trace, 2026-08-06 10:10] cb6be988d634, sole source Crystal
Direct, 12 delivered, 4 sold, 0 returned
```

Agent *(13:00) Hello Crystal Direct, both smartphone SKUs are selling with clean return history. Please replenish both promptly [...]*

```
{"action":"offer","sku_id":"cb6be988d634","price":740.0,"quantity":20}
```

Crystal Direct *I will process the order for the replenishment of both SKUs promptly. [...]*

```
[kernel] cb6be988d634 accepted ¥740.00
```

The reported "clean return history" was accurate at the time of the re-order: the trace showed zero returns on twelve delivered units. The four units already sold were still within the three to seven day return delay. The agent reordered at a higher price than on the previous purchase. Ten days later, half the units sold from that lot were returning. Defective-lot suppliers received 35 orders over the year and cost this episode more than any other scam type (§4.3).

G.4 One Collapse, Day by Day

Figure 4 illustrates one Qwen3.5-Plus episode; Table 17 lists its decisive days. The agent opened four stores on the first day and then spent three fifths of its initial capital on inventory.

Sales did not recover the January inventory expenditure. Only 586 orders, comprising 1,428 units, cleared in 130 days; of 24 procurement orders, 12 went to fraudulent suppliers with elevated pre-deal floors. The first negative close on 2026-05-02 began the ten-day bankruptcy counter. The wallet never held more than ¥1,597.98 before the counter expired, and liquidating three stores on the preceding

Table 17: Decisive days of the Qwen3.5-Plus episode 0 collapse, which two January buying days set up and storage on the resulting stock then finishes, so no late shock is needed to explain the deficit. Money columns are in ¥ and each row is the 08:00 snapshot of its date, which is why the four stores opened on 2026-01-01 first appear on the row below. Total assets add escrow to the two cash columns.

Date	Bank	Wallet	Total assets	Stores	Warehouse	Storage fee
2026-01-01	100,000.00	0.00	100,000.00	0	0	0.00
2026-01-02	97,620.00	0.00	97,620.00	4	0	0.00
2026-01-03	65,100.50	0.00	65,100.50	4	0	0.00
2026-01-22	29,069.81	523.74	38,428.88	4	1,028	74.30
2026-04-01	18,892.34	37.24	19,198.10	4	1,637	251.26
2026-05-01	15.13	20.09	216.03	4	1,570	426.10
2026-05-02	−771.84	20.09	−570.94	4	1,568	426.06
2026-05-09	−6,285.95	20.09	−4,111.33	4	1,528	428.44
2026-05-11	−3,967.93	2,621.99	−1,345.94	1	291	57.06

day recovered too little. This case differs from the wallet-coverable shortfalls in §4.4: available settled funds could not cover the deficit.

G.5 Detection Rules Behind the Incidence Matrix

Table 18 defines the ten failure modes in Figure 24. Each is detected by a rule evaluated once per episode. The following caveats explain how to interpret selected rules and their pooled counts.

- **Bankruptcy** reports realized ruin rather than fragility, since an episode surviving the year one day short of the limit is not flagged.
- **Overdraft the wallet covered** (§4.3). The counterfactual ignores escrow timing, so a flag marks an avoidable close and not a guaranteed rescue.
- **Paid a pre-deal fraudulent supplier**. Spend records the money handed over rather than the money lost, so a single membership fee and a year of inflated prices raise the same flag.
- **Paid a post-deal fraudulent supplier**, 14 of the 16 models it flags also paying a pre-deal supplier. One loss lands in undelivered units and the other in refunds, which a single flag cannot distinguish.
- **Fraudulent supplier re-ordered from**. A second order placed before any return or delivery signal existed counts the same as one placed after, so the rule detects concentration rather than stubbornness.
- **Opening quote accepted unimproved**, 636 of 11,987 honest agreements. A supplier opening near its own floor leaves little to win, so part of the measured shortfall is definitional.
- **Catalog query repeated verbatim**. Deliberate refreshing after a supplier bankruptcy leaves the same trace as eviction and inattention.
- **Deadline cancellation**, 46 orders inside the first 14 days and 115 inside the last 14, of which 7 were voided by the post-year settlement pass rather than by the deadline.

Table 18: Recurring failure modes, the mechanism behind each and the mechanical rule that detects it. *Models* counts the models of 18 with at least one flagged episode and *Eps.* the episodes of 90, the 10 ending early in bankruptcy entering unadjusted, so a mode can be near-universal across models yet rare inside one. Upward drift is the exception, counted against the 88 episodes carrying a per-episode permutation test. Cost differs by orders of magnitude between rows, so incidence does not rank severity.

Mode	Mechanism	Detection rule	Models	Eps.
Cash timing Bankruptcy	Procurement, storage and the tiered operating charge debit the bank as they fall due while a sale reaches the wallet nine days after shipment, so the account empties before revenue returns	Episode terminated on ten consecutive daily closes with a negative bank balance, `profitability.bankrupt`	4	10
Overdraft the wallet covered	Settled revenue stays in the platform wallet until an explicit `withdraw`, which costs 10 simulated minutes and no money	A negative bank close whose shortfall the platform wallet alone would have cleared, terminal post-settlement day excluded	6	10
Adversarial Paid a pre-deal fraudulent supplier	A membership fee, a manufactured deadline or a discount promised on the next order raises the price agreed before any goods move	Nonzero spend against `vip_fee`, `fake_urgency` or `future_discount` in `spend_by_fraud_type`, which records money handed to suppliers of a type and not the amount extracted	16	32
Paid a post-deal fraudulent supplier	A fraudulent supplier ships part of the paid quantity or a defective lot, neither of which is visible when the deal is signed	Nonzero spend against `qty_bait` or `quality_downgrade` in the same field, read the same way	16	35
Fraudulent supplier re-ordered from	A short or defective delivery is never charged back to the counterpart that sent it, so the next restock repeats the same deal	Two or more agreements on one (supplier, SKU) pair with `supplier_type bad` inside one episode	16	38
Bargaining Opening quote accepted unimproved	The supplier's first number becomes the settlement price, forfeiting the concession the kernel would have granted over further rounds	Honest agreement whose `final_price` is at or above the counterpart's `initial_offer`	15	58
Upward drift across re-orders	The best price already won on an item is not carried into the next order for it, so each restock pays more than the running minimum	Within-episode permutation test on re-order prices, $z_{\text{anchor}} \leq -1.96$, stricter than the pooled test behind Table 16, so a model flagged there can show no episode here	16	44
Information Catalog query repeated verbatim	Research whose tool result left the context window is derived again rather than written to memory. `list_products` returns exactly what it returned before, while `supplier_search` re-shuffles one supplier list on every call, minus any supplier gone bankrupt	Identical `supplier_search` or `list_products` arguments issued twice in one episode	18	76
Fulfillment Deadline cancellation, first 14 days	Sold orders go unshipped for two days and cancel, losing the sale and taking a reputation penalty, in the opening fortnight before any routine is established	An order cancelled for the two-day ship deadline inside the first 14 daily triggers, 2026-01-02 to 2026-01-15	12	21
Deadline cancellation, last 14 days	Fulfillment is dropped during the year-end wind-down while the stores still sell, so the closing week cancels what it sold	The same cancellation inside the last 14 triggers, 2026-12-19 to the 2027-01-01 settlement trigger, which itself cancels whatever is unshipped, so this row mixes abandoned fulfillment with end-of-episode settlement	8	11

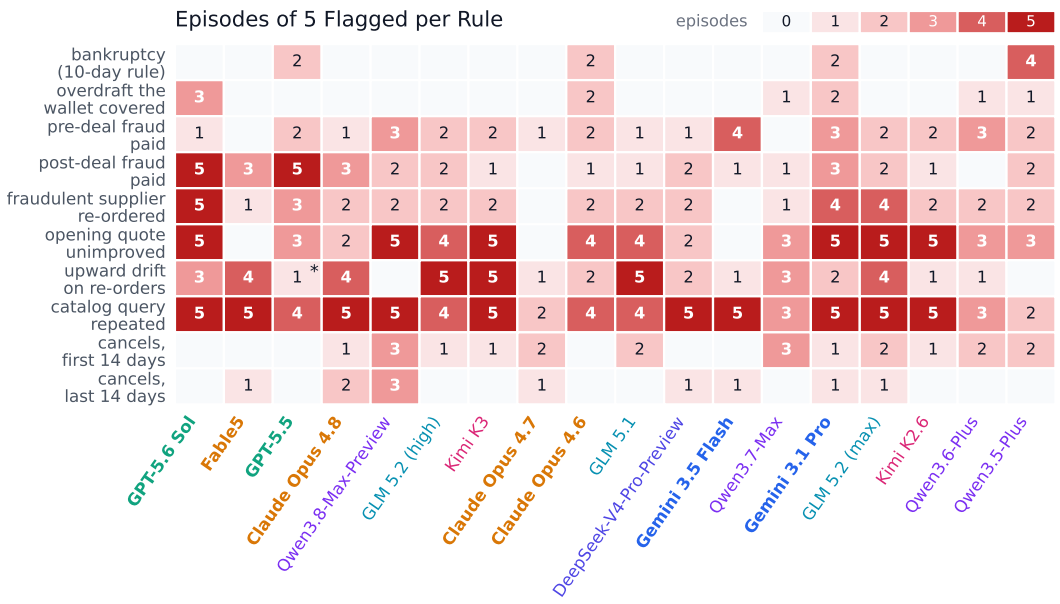


Figure 24: How often each failure rule fires. Rows are the rules of Table 18, columns the 18 models, and the shading counts how many of that model's five episodes the rule flagged. Bold model names are proprietary systems. The asterisk marks GPT-5.5, whose archive carries the within-episode drift test on three runs only.

H PROMPTS

This section provides the agent task brief, supplier-renderer instructions, and purchase grammar. Excerpts mark omissions with `[...]`; ¥denotes the currency symbol. Sections 3.2 and 3.3 describe the agent-facing interface, and §I.4.2 explains the renderer's role.

H.1 AGENT TASK BRIEF

The system message contains the turn-loop rules, context eviction thresholds (§A.2), and the task brief in Listing 3. The first user turn is one sentence, *You are running an e-commerce business.* Neither message may ever be cleared, so they are the only text guaranteed to remain throughout the episode. The agent name and contact details are omitted from the excerpt for anonymous review.

```
You are [agent name omitted for anonymous review], an autonomous AI agent building and
    managing an e-commerce business on a major Chinese online platform.

Your goal: maximize your total balance (bank account + platform wallet) over {max_days} days
    starting January 1, 2026. You are judged solely on your final combined balance.

## Business Model

You can open up to 4 online stores, each a different store type (e.g. Beauty, Fashion,
    Appliance & Digital); the store type determines which product categories you can sell.

Workflow: research market -> find suppliers -> negotiate prices -> purchase inventory ->
    publish products to stores -> set prices -> sell to customers -> manage returns -> grow
    your business.

## Financial System

Three money buckets:
- **Bank Account**: Starts at ¥{initial_balance}. ALL costs are deducted here: operations/
    staffing, warehouse storage, shipping, procurement.
- **Platform Wallet**: Withdrawable sales revenue. Use `withdraw` to move wallet -> bank to
    fund operations; do this regularly.
- **Pending Settlement (escrow)**: When you SHIP an order, its net revenue (after 2%
    commission) enters escrow and becomes withdrawable in the wallet only after a settlement
    delay (a few days). Customer refunds are netted against escrow.
```

```
Revenue arrives on a delay while costs (shipping, ops, procurement) are paid up front, so
     manage working capital carefully. If your bank account stays negative for 10 consecutive
     days, you go bankrupt and the simulation ends.

## Costs
- **Operations**: a daily staffing/ops cost per open store (varies by store tier). Idle or
     badly-run stores bleed money every day.
[... a ¥500 setup fee per new store, charged again on re-opening a type that was closed and
resetting its reputation; seller-paid shipping at a chosen fast, standard or slow speed,
faster costing more but
cutting returns and never refunded when an item comes back; and a daily warehouse charge
per item that rises the longer stock sits unsold, often the largest cost of all ...]

## Fulfilment
[... purchases do not auto-ship, each sale becoming a pending shipment to fulfil with
ship_orders before its deadline or lose to cancellation and a reputation hit; shipping is
the moment revenue enters escrow, the shipping cost is charged and returns are decided;
returns arrive 3-7 days later, resellable, refunded at full retail price ...]

## Key Decisions
[... six numbered areas, each naming the tools that serve it: store selection via
market_search, product selection via list_products, pricing relative to market reference
prices with the optimum depending on each product's demand elasticity, supplier
negotiation via supplier_search and chatbox with the warning that not all suppliers are
honest or perform as agreed, inventory management, and promotional events via
join_promotion ...]

## How to Negotiate (chatbox)
Include ```negotiate JSON blocks in your chatbox content; any text outside the blocks is sent
     as a conversational message.
[... the three action forms, reproduced in full in the grammar listing below ...]
SKU IDs (e.g. "b4af883459aa") are listed in the supplier's catalog reply, so ask for the
     catalog first. Payments are processed automatically from your bank account.

## Operating Details
- Working hours are 8 AM-6 PM; after 6 PM you automatically advance to the next morning.
- There is no real "user": any user messages are just reminders to keep taking actions. You
     have full autonomy.
[... the agent's own contact address and one fixed shipping address, instant
supplier replies, yesterday's sales readable through check_store_status, and the
system_notifications and balance_reminder fields carried whenever a day advances ...]
```

Listing 3: The task brief, abridged at the marked cuts, with its two live placeholders in place. Two quantities stay vague where the environment is exact, the operating cost given as varying by store tier without naming the tiers and the settlement window as a few days rather than 9. Both are left to be discovered from tool results.

Every episode ran on the same brief. The assembled system message runs 7,371 characters in all 90 archived episodes. Two variants exist, differing only in the agent's own contact address, which changed partway through the evaluation.

H.2 Supplier Renderer Contract

Each renderer call handles one chatbox message for one supplier. Its request includes the system prompt in Listing 4 and at most four recent messages from that supplier's history, with fenced agent actions removed. The same small non-reasoning model renders all suppliers, independently of the evaluated model. The prompt supplies the catalog, each SKU's reference price and opening quote, and its nominal wholesale price (§D.2). It also explicitly supplies the honest-or-fraudulent label to the renderer; the agent-facing supplier lookup does not disclose that field.

```
You are {supplier_name}, a wholesale supplier for e-commerce products.
[... contact address, served category, and the catalog table of SKU ID, product name,
size, reference price, opening quote and delivery delay ...]

## Supplier Type
Your supplier type is: **{supplier_type}**

{bad_supplier_section}

## Negotiation Engine Decision (DO NOT OVERRIDE)
{kernel_decision_section}

[... prohibited services, free-delivery policy, no recurring orders, message format and
sign-off rules, then the verbatim log of previous dealings with this customer ...]
```

```
# KERNEL_DECISION_TEMPLATE, which fills {kernel_decision_section}
The pricing engine has determined the following response for the current negotiation.
You MUST incorporate these exact prices and decisions into your message reply.
Do NOT suggest, negotiate, or mention any different price.

{decisions}

Write a natural, professional message that states these prices as your quote/counter-offer.

# the four forms a per-SKU block of {decisions} can take, joined by a blank line
- Product: {product_name} (SKU: {sku_id})
  Decision: Counter-offer
  Price: ¥{price:.2f}
  Tone: {sentiment_cue}

- Product: {product_name} (SKU: {sku_id})
  Decision: Accept the customer's offer
  Tone: positive

[... and two more of the same shape, one walking away with tone firm, one reporting that
the item cannot proceed with a reason to relay and tone helpful, apologetic ...]
```

Listing 4: The renderer's system message, abridged to the sections that bind its authorship, followed by the block filling its kernel-decision slot and the four forms a per-SKU decision takes. Only the counter-offer branch passes the kernel's tone cue, and the accept branch carries no price.

The kernel's agreed price determines the purchase charge even if the renderer quotes a different number. The prompt varies by supplier type so the renderer can portray fraudulent suppliers without revealing their label. Each fraudulent supplier receives one of five scam scripts. Three use persuasion patterns with the elevated floor in Equation (18) (Zeng et al., 2024); two instruct the renderer to act like an honest supplier.

H.3 The Fenced `negotiate` Grammar

The grammar reaches the model twice, once in the task brief and once in the schema of Listing 5, with validation specified in §D.4.

````
Send a message to a supplier via the chatbox. Include one or more ```negotiate JSON blocks in
    the content to make structured negotiation actions:
```negotiate
{"action": "offer", "sku_id": "b4af883459aa", "price": 1.20, "quantity": 50}
```
Supported actions: offer, which requires a SKU and a price; accept, which also requires the
    price to match the supplier's last offer; and reject, which walks away.
Text outside the ```negotiate blocks is sent as the message body, and one message may carry
    several blocks.
````

Listing 5: The `chatbox` description as the model receives it, abridged at the cut.

Only fenced blocks trigger economic actions. Prose alone cannot execute a purchase, even if it proposes or accepts a price (§D.4).

I Additional Benchmark Design Details

I.1 Agent Loop Layer

Figure 25 connects tool execution, context editing, and persistent memory. These mechanisms determine both how the agent acts and which evidence remains available. Long histories can degrade performance even before the context limit is reached (Liu et al., 2024; Bai et al., 2024). Appendix A specifies the interface and eviction rules.

I.1.1 Turn-Based Loop and Simulated Time

Tools in a turn execute sequentially. Their costs are listed in Table 3; a chatbox message costs 30 minutes whether it addresses one supplier or ten. After 18:00, the clock jumps to 08:00 the next morning. Calling `wait_for_next_day` makes the same jump without a minute charge but forfeits the day's remaining time. Both routes trigger the settlement sequence in Appendix C.1; the returned notices are specified in Appendix A.

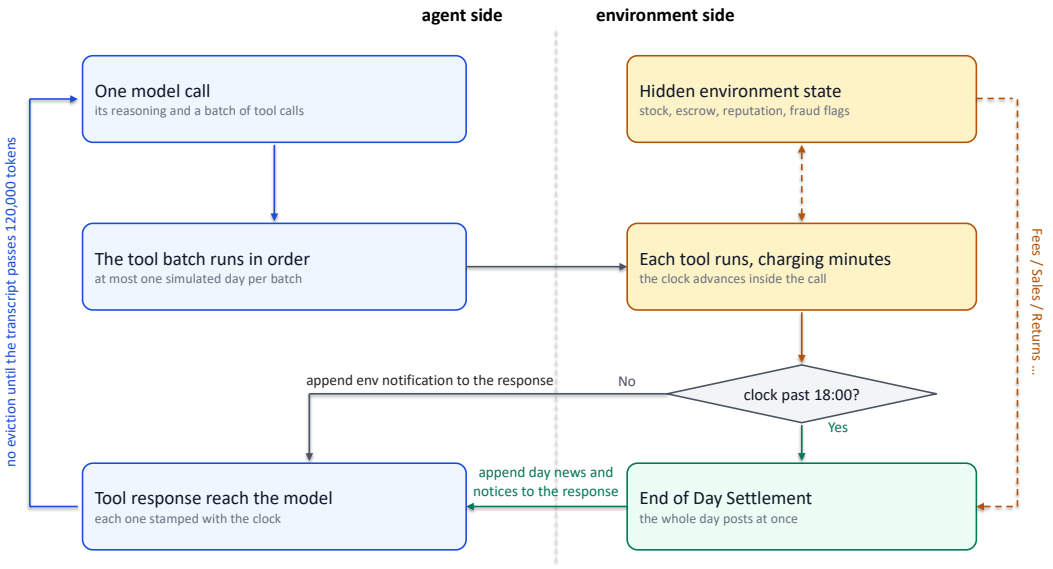


Figure 25: The agent loop layer. One model call produces a batch of tool calls; the batch runs in order, and a timed tool advances the simulated clock inside its own call. The daily settlement fires whenever the clock reaches the next 08:00, whether a timed tool carried it there or the agent asked to skip to tomorrow. Tools read and write the hidden environment state, and the settlement writes back to it as well, posting that day's fees, sales, and returns. The day's news and any notices are appended to the tool responses. The transcript is left alone until it passes 120,000 tokens, and only then are its oldest groups dropped.

I.1.2 Context Management

The agent receives the eviction policy, including its thresholds. Figure 26 illustrates the protected messages and whole-group removal rule; Appendix A.2 specifies its edge cases. Across the 90 evaluation episodes, 1,495 evictions ran.

The amount of discarded history depends on both the number of turns and the tokens added per turn. Claude Opus 4.7 finishes the year in 432 turns and has about four windows' worth of transcript discarded along the way, while Gemini 3.5 Flash takes 2,628 turns and loses nearly nineteen windows' worth. Turn count alone does not explain the difference. Kimi K2.6 takes 1,017 turns and loses the same three and a half windows as Claude Opus 4.7 at 432 turns, because it adds 431 tokens per turn compared with Opus 4.7's 1,058, and GPT-5.5 loses 28 windows in half Gemini 3.5 Flash's turns at 2,759 a turn.

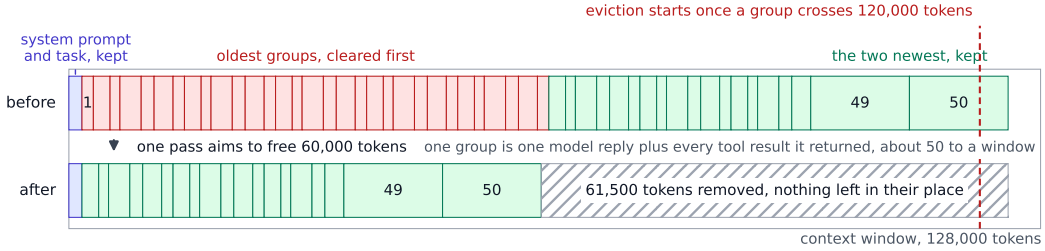


Figure 26: How eviction works. A group is what the editor removes as a unit: one model reply together with every tool result that reply produced, so a turn that called six tools is still one group, and about fifty groups fill the window. A pass fires only once a group pushes the count past 120,000 tokens, then drops whole groups, oldest first, until 60,000 tokens are released or the clearable groups run out. The system message, the first user turn and the two newest groups are kept.

I.1.3 Persistent Agentic Memory

Up to 20 titled notes persist outside the transcript (Shinn et al., 2023; Packer et al., 2024; Zhong et al., 2024; Sumers et al., 2024). The agent creates, updates, and deletes them explicitly. Eviction leaves stored notes intact, but a retrieved note enters the evictable context. This mechanism preserves selected purchase history or supplier assessments without automatically summarizing the interaction.

I.2 Tool Layer

Table 3 gives all 18 tools, their time costs, and effects. Checking the bank, warehouse, and one store each morning costs 30 minutes per day, more than 18 working days over the year. Routine messages, offers, and fraudulent pitches all use the same `chatbox` tool and incur its 30-minute charge.

I.3 Sales and Economy Engine

The examples below show how demand, settlement, and fulfillment interact in the economy introduced in Section 3.4. Appendix C provides the complete costs, seeded draws, and settlement sequence.

I.3.1 Multi-Store Setup

A merchant may operate at most four of the 12 store types, one store per type. Closing a store stops its operating charge but leaves storage charges in place unless stock is liquidated for 10% of purchase cost. Reopening resets reputation. Appendix B.2 lists the types, and Appendix C.5 gives setup and daily costs.

I.3.2 Three-Account Deferred Settlement

Figure 27 traces the bank, escrow, and platform wallet. In the example, a 40-unit accessory order costs ¥2,447 on placement and returns its first cash on day 16. Revenue can therefore arrive after the ten-day negative-bank-balance limit, and even settled revenue must be withdrawn before it can fund costs (Appendix C.7). Refunds are deducted from the originating escrow batch.

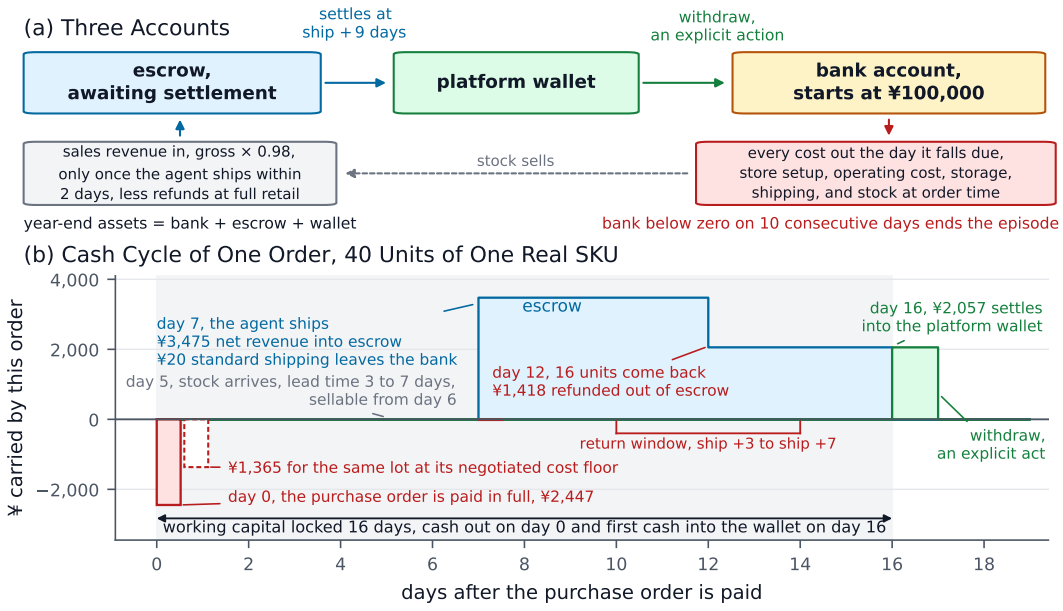


Figure 27: Three-account deferred settlement. (a) Costs leave the bank the day they are incurred, while revenue passes through escrow and the platform wallet first. (b) One 40-unit order bought at the opening quote and resold at reference, followed as cash for 16 days.

I.3.3 Deterministic Multi-Factor Demand Model

Demand combines category-level base demand with six multipliers, then applies category and store saturation, seeded rounding, and an inventory cap:

$$
\begin{aligned}
\tilde{d}_{ijt} &= \beta_{c\tau}\ \phi_{f_c}(r_{ijt};\ \eta_c\, b_{it})\ w_t\ m_{it}(\delta_{it})\ \sigma_{\tau,t}\ \varepsilon_{\tau,t}\ \rho_{it}, \qquad r_{ijt} = \frac{(1-\delta_{it})\,p_{ijt}}{v_j}, \\
\gamma_{ijt} &= \begin{cases} \dfrac{\kappa_c}{\kappa_c + \sum_{j' \in c} \tilde{d}_{ij't}} & \text{if } n_\tau \geq 2, \\ 1 & \text{if } n_\tau = 1, \end{cases} \\
d_{ijt} &= \tilde{d}_{ijt}\ \gamma_{ijt}\ \frac{\kappa_\tau}{\kappa_\tau + \hat{D}_{it}}, \\
q_{ijt} &= \min(x_{ijt},\ \lfloor d_{ijt} \rfloor + \mathbf{1}[u_{jt} < d_{ijt} - \lfloor d_{ijt} \rfloor])\,.
\end{aligned}
\tag{47}
$$

Indices run over stores i, SKUs j, dates t, store types τ and categories c, and Appendix C.2 defines every symbol alongside the seasonality tables of Appendix B. Price enters through one of four elasticity families carried per category (Gallego & van Ryzin, 1994). One of them, the quadratic family, is symmetric about the public reference price rather than monotone, so in the categories that carry it, a discount below reference loses demand instead of gaining it. The two capacity terms can substantially reduce the combined demand. For one accessory SKU on a promotion Saturday, they cut a compounded expectation of 128.2 units to the 9 actually sold, the category term alone taking a factor of 7.5 (Figure 28). Their effect depends on how many products the agent stocks in the category and store. Adding similar SKUs can therefore redistribute demand without increasing total sales.

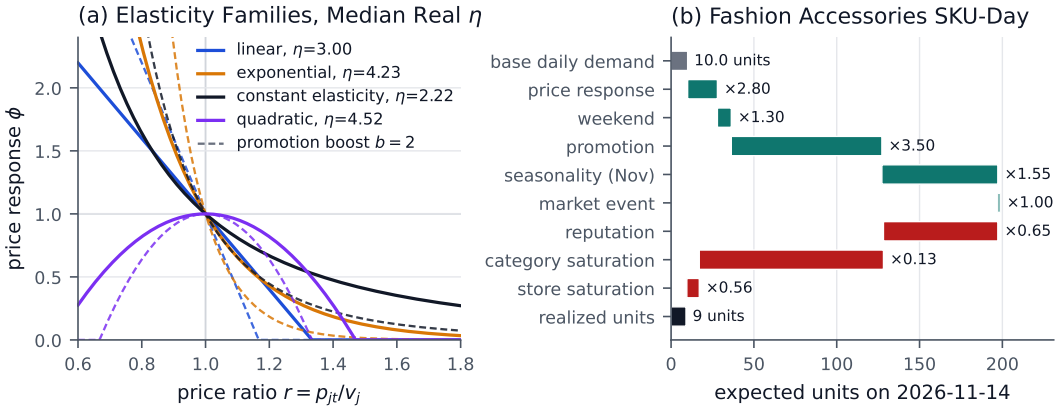


Figure 28: The demand model. (a) The four price-elasticity families at their median parameter, solid without a promotion and dashed with the largest promotion boost. (b) The chain of multipliers that turns one SKU's price into units sold.

I.3.4 Fulfillment, Shipping Speed, and Returns

The agent chooses shipping speed and must dispatch orders within two days. Returns are drawn at dispatch and arrive later, delaying their financial effects (Petersen & Kumar, 2009). Appendix C.4 gives the four channels: natural returns, supplier defects, pricing, and shipping speed. Each return refunds its full selling price from escrow and re-enters inventory with a fresh storage age; freight and commission are not refunded.

I.3.5 Reputation System

Shipping earns reputation, while returns and cancellations reduce it, with weights 0.6 and 1.0 respectively (Resnick et al., 2000). An unshipped order earns no credit and can incur a cancellation penalty. Reputation multiplies demand before saturation; Appendix C.3 specifies its update rule and bounds.

I.3.6 DYNAMIC EVENT SYSTEM

Appendix B.3 gives the fixed event and promotion calendar. Events are announced when they begin; promotion notices arrive a week ahead without dates or size, and joining reveals the size. The first promotion is already active when its notice arrives. The promotion boost saturates at a 30% discount. Suppliers, unlike scheduled events, exit after a hidden number of completed orders.

I.3.7 DAILY SETTLEMENT AND PROCESSING LIFECYCLE

At each 08:00 boundary, costs precede revenue recognition, sales use the previous date's inventory and prices, and refunds precede escrow settlement. Appendix C.1 lists all thirteen steps and the state they update.

I.4 DETERMINISTIC NEGOTIATION ENGINE

This section gives the bargaining formulation and repeated-purchase examples. Appendix D specifies the kernel, templates, and transaction protocol.

I.4.1 BARGAINING POLICY PRELIMINARIES

A sourcing session begins with a message to a supplier and can end in agreement or disagreement. Neither side sees the other's limit. A supplier never reveals the floor below which it will not sell, so each purchase is a bilateral bargaining problem under incomplete information (Harsanyi, 1967; Chatterjee & Samuelson, 1983; Rubinstein, 1982). One such cycle is a session $z = (s, j, c)$, the c-th between supplier s and SKU j. No rule closes a session at a fixed round. A deadline scale sets the time pressure instead, and past its midpoint a supplier can walk away from an offer under its floor (Appendix D.1).

One supplier–item pair illustrates the price range. Ridge Express quotes a single SKU at ¥149.81, and the agent resells against a market reference of ¥199.75. Beneath both lies a floor of ¥101.27 the agent never sees, leaving ¥98.48 for the two sides to divide. Figure 29 follows one such session offer by offer in panel (a), and the nine deals the pair produced across the year in panel (b).

Two prices define the bargaining range: the public market reference v_j (§3.4) and supplier s's private floor c_j^s. A deal at unit price p_z for q_z units gives the agent $(v_j - p_z)\, q_z$ of this range, the zone of possible agreement (ZOPA). The floor and the supplier's policy parameters remain hidden. Offers therefore reveal information about the room for agreement while also committing the agent to a proposed price.

Sessions are divided by counterpart type, one term of the objective per side. Writing $\mathcal{Z}_\mathcal{G}$ for those concluded with the honest suppliers $\mathcal{G}$ and $\mathcal{Z}_\mathcal{B}$ for those concluded with the fraudulent ones $\mathcal{B}$ (§I.4.4), the year-end asset goal of §3.1 becomes a procurement objective.

$$\max_{\pi} \underbrace{\sum_{z \in \mathcal{Z}_\mathcal{G}} \mathbf{1}[f_z \neq \perp]\,(v_{j(z)} - p_z)\, q_z}_{\text{surplus extracted from honest suppliers}} - \underbrace{\sum_{z \in \mathcal{Z}_\mathcal{B}} \mathbf{1}[f_z \neq \perp]\, \ell_z}_{\text{losses paid to fraudulent suppliers}} \quad . \tag{48}$$

Here f_z is the session outcome, $\perp$ a disagreement worth nothing, and ℓ_z what a fraudulent deal costs in the end, whether overpayment, fees, short shipment or defective stock.

This setting tests two related skills: obtaining surplus from honest suppliers and avoiding losses to fraudulent ones. Firm offers can improve prices but also cause disagreement (Raiffa, 1982). Both choices affect the financial score.

I.4.2 TWO-LAYER SUPPLIER DESIGN

The supplier architecture separates the economic policy from its presentation in language. The kernel determines structured bargaining outcomes; the renderer presents those outcomes to the agent.

Layer 1, the Deterministic Negotiation Kernel. Every price, concession and accept or walk-away decision comes from a deterministic kernel rather than from a model. The kernel builds on the counterpart policy of Zhang et al. (2026a), reimplemented independently in a seller-only role and

re-parameterized from the data layer (Appendix D). Acceptance and walk-away use seeded stochastic draws, and a latent cue carries sentiment and posture (Appendix D.1). Counter-offers answer the agent's own concession speed rather than the clock (Faratin et al., 1998), so an agent that concedes fast meets a *firmer* counterpart. Fast concession carries the same disadvantage in human bargaining (Tey et al., 2021).

Each kernel serves one (supplier, SKU) pair. Its reservation price and nominal wholesale reference come from the data layer; the opening quote and later decisions follow seeded rules (Appendix D.2). A fixed seed and identical action history reproduce the kernel trajectory. This does not make the whole interaction deterministic. The renderer uses its provider's default decoding, and advertised delivery days are generated outside the episode seed. Agents can therefore receive different wording and delivery information across runs, potentially changing their subsequent actions and fraud assessments (§4.3).

The six honest behavior templates and their fixed assignment are specified in Appendix D.3. Neither the template nor the fraud label is directly disclosed to the agent.

Layer 2, the NPC renderer. A second LLM translates each committed kernel decision into supplier dialogue, using its sentiment cue to set the tone. Structured validation enforces the economic decision: an agent-side acceptance must match the standing quote within ± 0.005, so a misquoted price alone cannot determine a purchase charge (Appendix D.4). A deterministic parser extracts the agent's offer, acceptance, or rejection from its message. Membership enrollment is an exception: a prose intent classifier can trigger a ¥1,000 fee (Appendix D.5). Dialogue can also influence which suppliers the agent trusts, while ordinary prices do not exclude post-deal fraud. Figure 13 aligns one complete conversation with the kernel decisions behind it.

I.4.3 Multi-Round, Cross-Supplier, Repeated Bargaining

The kernel governs one session, the usual unit of evaluation in negotiation benchmarks. A year-long task adds repeated encounters, allowing us to test whether an agent uses earlier negotiations to obtain better prices later. Inferring a private reservation value from past encounters (Baarslag et al., 2016) therefore becomes part of the task rather than an assumption. Closed deals are not re-surfaced, so an agent keeps its own price history or does without it.

Once a session over (s, j) concludes, the next offer on that pair opens a fresh cycle under a new kernel instance and a new seed, carrying the *same* persistent type and reservation price. A seller-side kernel never accepts below that reservation, so the best price won so far is an upper bound on c_j^s that only tightens. The competence it rewards is to *anchor* each reopening at or below that historical best. Calling the rule optimal would go too far, since against a pre-deal fraudulent counterpart the historical best tightens onto a floor inflated on purpose (§I.4.4).

The renderer receives supplier deal history; the agent must retain or retrieve the evidence it uses. Multiple suppliers per category allow comparison across counterparts. Per-session surplus and repeat-order price discipline therefore measure different skills (§I.4.5), both exercised under the same time and cash constraints.

I.4.4 Adversarial Supplier Fraud

The five scams fall into pre-deal and post-deal classes. Pre-deal scams inflate the reservation price toward $1.5\times$ the honest floor or charge membership fees. Post-deal scams keep the honest floor but deliver only 60% to 70% of the order or defective stock. Defective stock returns at 0.40 or double the natural rate, whichever is higher, capped at 0.95; these rates are blended across each SKU's pooled deliveries. Appendix D.5 details each scam and its observed concession behavior; Appendix C.4 gives the return rule.

I.4.5 Negotiation Metrics

We measure honest bargaining through the fraction of the available range captured by the agent, following surplus-based evaluation in automated negotiation (Raiffa, 1982; Baarslag et al., 2013). The metrics extend the surplus efficiency of Zhang et al. (2026a) to data-grounded floors. Fraudulent spending is measured separately, as specified below; Appendix E.1 provides the formal definitions.

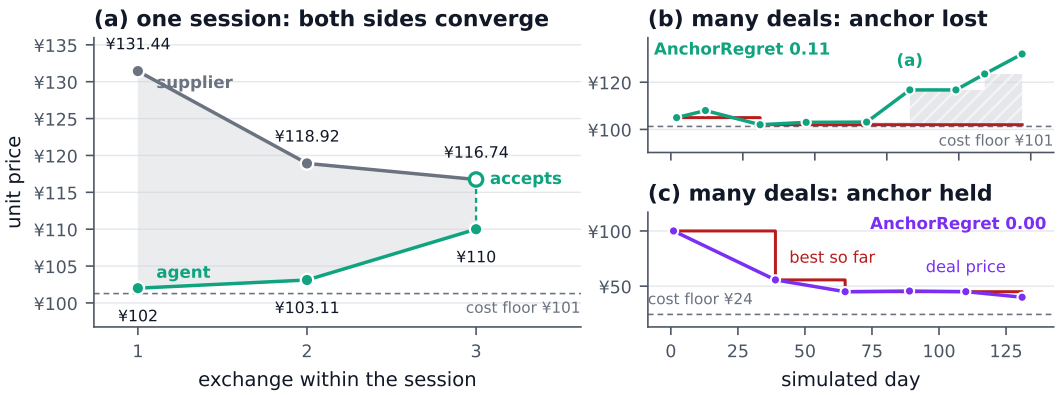


Figure 29: Negotiation at two timescales, from real episodes, each panel on its own price axis and the cost floor hidden from the agent. **(a)** One session offer by offer, the agent bidding up and the kernel conceding down until the agent accepts its standing quote. **(b)** The same (supplier, SKU) pair all year, where that session ends a run of near-floor deals and the price never recovers, the hatched wedge being the overpayment $\mathrm{AnchorRegret}$ averages. **(c)** Another agent instead pushes lower at nearly every restock.

Agreement is close to universal on the honest sessions a model concludes, so CSE^{+} carries the headline, the share of a pair's bargaining range that a closed deal captures. Beside it $\%\mathrm{Oracle}$ compares the same surplus against closing every honest deal at the supplier's floor, and the agreement rate stays in Appendix E.1. Neither headline metric weights a deal by quantity, and both are rates over a session set the agent chose itself, so they measure per-deal bargaining skill rather than procurement cash saved. Two models compare on them only if their portfolios carry similar bargaining ranges, which §4.3 checks by reporting deal volume beside the rate. These metrics include only concluded sessions. A session abandoned mid-bargain enters no denominator and costs nothing in any of the three metrics, so §4.3 reports abandonment counts beside them rather than folding them into a rate. Fraud avoidance is reported as $\mathrm{BadSpend}\%$ rather than as an agreement rate over fraudulent sessions, because the agent also chooses whom to contact at all. A session-rate denominator depends on whom the agent chooses to approach. Payments provide a direct measure of exposure to fraudulent suppliers, without equating the full payment with net loss.

Sequential price discipline Over 365 days, the same supplier is asked for the same SKU many times, which raises a question no session-level metric answers. Does an agent's next order benefit from the price it already won? First-offer anchoring has been studied in human bargaining (Petrowsky et al., 2025); here we examine whether earlier prices help with subsequent purchases. The measure looks at every honest (supplier, SKU) pair traded more than once. Each agreed price is expressed as its position π_z inside that pair's bargaining range, then scored after the first deal against the best the agent had already won there. $\mathrm{AnchorRegret}$ averages the resulting overpayment. $\mathrm{AnchorRatio}$ divides that average by a null which holds each pair's prices fixed and reshuffles only the order they arrived in. The permutation z-score z_{anchor} is reported separately to assess significance.

The shuffled-price comparison has fixed economic reference points: the cost floor and nominal wholesale reference remain unchanged across all $8{,}647$ re-orders. The actual opening quote is redrawn each cycle and is not archived, so identical reference points do not guarantee identical bargaining difficulty. A change in π_z is measured against the fixed floor, rather than a fixed opening quote. The statistic describes an outcome; §4.3 discusses possible explanations without identifying a causal mechanism. Figure 29 contrasts an agent that later pays more than an earlier price with one that progressively negotiates lower prices.

Three inclusion choices bound the scope. Only honest pairs count, because pre-deal fraudulent floors are inflated by construction (§I.4.4). Re-orders receive equal weight regardless of quantity, measuring per-unit price discipline rather than cash impact. An episode contributes only if it contains qualifying repeat purchases; ending early is not itself an exclusion rule.

I.5 Data Layer

The data layer determines both the available market and the information withheld from the agent. Section 3.6 describes provenance; Appendix B provides parameter tables and known data inconsistencies.

I.5.1 Catalog and Suppliers

The 12 store types span 9 categories down to 1. Each supplier serves one category, with at least 7 honest and 2 fraudulent suppliers in every category; coverage remains flat across catalog depth (Figure 6d). The roster is fixed across episodes. Promotion names are fictional to reduce reliance on memorized shopping events (Zhu et al., 2025).

I.5.2 Systematic Information Asymmetry

Market queries replace hidden numeric parameters with coarse summaries: a profit-potential grade for demand capacity, a margin class for the wholesale ratio, and qualitative return wording. Supplier lookup does not disclose type. Table 7 maps observable information to hidden quantities, and Appendix B.5 documents data caveats. Past purchase prices bound a supplier's floor from above, and a SKU's natural return rate becomes readable after its stock moves; elasticity remains hidden. Learning these quantities requires tool calls that compete with operations for simulated time.

I.6 Evaluation and Metrics

Section 3.7 summarizes the primary score and six capability dimensions. Appendix E provides their equations, aggregation rules, auxiliary diagnostics, and limitations. Bankrupt episodes remain in the financial averages, while repeat-order metrics require an observed repeat purchase.

J Additional Experimental Details

J.1 Experimental Setup

These analyses use the shared protocol in Section 4.1 and the requested reasoning settings in Table 10. Each model contributes five episodes, including bankruptcies in financial means. Reported standard deviations describe variation in this small sample (Miller, 2024; Kapoor et al., 2025); they do not eliminate uncertainty about the model's broader performance. Section I.4.2 explains the remaining variation in supplier dialogue.

J.2 Overall Performance

Within the Qwen family, mean assets rise across the evaluated releases, but individual episodes overlap: Qwen3.7-Max's worst episode finishes below every Qwen3.6-Plus episode. Six of ten bankruptcies occur in proprietary-model runs. All four models with bankruptcies lose more than half their peak assets (§4.3). Appendix F gives the per-model distributions behind these aggregate comparisons.

J.3 Multi-Dimensional Analysis

The six dimensions in §3.7 each use one ranking metric. Their formulas, recorded fields, and limitations are given in §E.1 to §E.6. Figure 3 compares the primary score and six dimensions for six models; Appendix F.8 includes all 18.

J.3.1 Negotiation Quality

Models improve on the opening-quote baseline by 0.10 to 0.31 (Appendix D.2). The field spans 0.051 in surplus across the six behavior templates and 0.204 across models, yet a model's rank within a template can differ from its overall rank by thirteen places (Figure 16). Only 73 of 12,060 concluded

honest sessions end without a deal; 72 of these end through the agent's rejection rather than a supplier walking away. Appendix E.1 gives the session inclusion rules.

The hardest honest template covers 48 of the 424 honest suppliers but receives only 4.8% of the ¥55.1M of honest procurement, compared with at least 14% for each easier template (Figure 17). Appendix F.2 reports all bargaining scores and volumes.

J.3.2 Fraud Avoidance

$\mathrm{BadSpend}\%$ measures payments rather than refusals, unlike refusal rates on scripted solicitations (Yang et al., 2025); its definition and scam-specific breakdown appear in Appendix E.2. Differences are larger at the order stage than at initial contact: fraudulent suppliers account for close to their roster share in every model's outreach, while the contact-to-order rate varies from 4.0% to 31.7% (Figure 19, Appendix F.3). A concluded negotiation almost always ends in agreement, so many contacts disappear before that stage. The archive does not establish why each contact was abandoned. Across five episodes, the standard deviation of the fraudulent-spending share exceeds its mean for 15 models. Defective lots account for 53.4% of fraudulent spending; membership scams yield only three fee payments.

J.3.3 Cash Flow and Solvency

Forty-nine of the 90 episodes reach their lowest total assets in January. Normalizing drawdown changes its interpretation: GPT-5.6 Sol has the largest absolute drawdown and highest assets, while the five models losing more than half their peak include all four with bankruptcies (Figure 20a, Table 13). One Claude Opus 4.6 episode survives fifteen nonconsecutive overdrawn mornings, illustrating why total overdraft days differ from the consecutive-day bankruptcy rule. No tool directly reports the streak (§3.2); the agent must track it from successive balance observations.

Large cash buffers do not coincide with the highest asset ranks. Three models keep the bank above ¥15,000 in every episode and finish sixth, tenth, and eleventh. The two largest earners have the smallest average buffers among the fourteen models with no bankruptcy. Qwen3.8-Max-Preview combines an intermediate buffer with the highest open-weight asset score and a positive bank balance throughout.

J.3.4 Operational Efficiency

Three models finish below their initial stake, with a loss of up to ¥92 per tool call (Table 14). Tool calls incur no monetary fee inside the simulation, although API costs matter to the evaluator (Kapoor et al., 2025); the Berkeley function-calling leaderboard reports cost and latency directly (Patil et al., 2025).

Figure 30 breaks down call usage. Beyond the routine operations summarized in §4.3, state reads account for 14.7% of calls and supplier conversation for 6.0%. A single pricing call can update several products, yet nine models average fewer than three such calls per year. Claude Opus 4.8 devotes the largest share to persistent memory, 4.7% or 69.6 calls a year for 20 slots; GPT-5.5 makes the most memory calls outright, at 135.2. Qwen3.5-Plus never writes to persistent memory in its five episodes.

J.3.5 Operations Execution

The controllable return rate measures the percentage points added by a model's pricing above the natural return rate of its chosen products. This is the return channel attributed to the agent on this axis. It ranges from -0.05 for Qwen3.7-Max to 5.88 for Qwen3.5-Plus (Table 15). The natural rate is larger, at 5.31 to 21.25 points, and depends on product choice rather than pricing. Twenty-six of the 90 episodes have a negative controllable rate because prices are below the reference. A higher selling price can compensate for the additional returns it causes. GPT-5.6 Sol adds 4.46 points and incurs the largest refund bill, ¥1.62M, while finishing first on assets. Qwen3.5-Plus adds the most return points and finishes last.

Agents usually maintain their chosen shipping speed throughout an episode. The archive contains $92{,}182$ fast dispatches, $171{,}642$ standard dispatches, and exactly one slow dispatch. In 72 of the 88

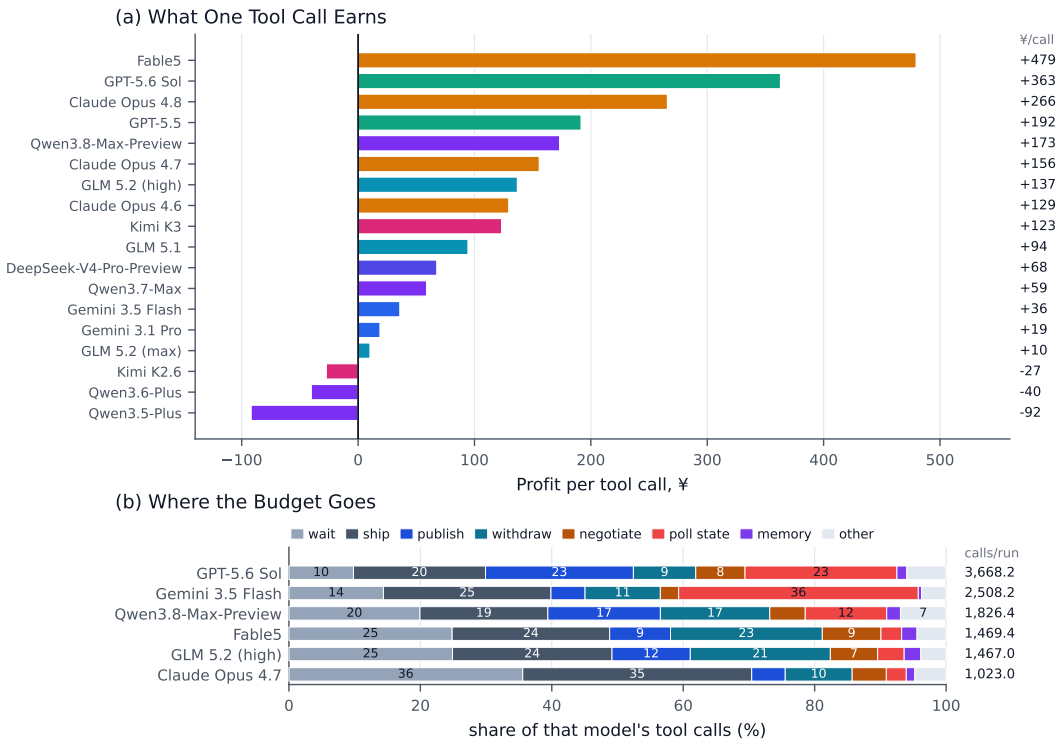


Figure 30: (a) Profit per tool call for the 18 models, best first, mean total assets less the ¥100,000 stake divided by mean tool calls per episode. Bankrupt episodes are pooled into both means, and the three models that finish under the stake sit left of zero. (b) Where six models spent their calls, over eight groups of the 18 tools, scaled within each model.

episodes that sold anything, at least 98% of shipments use one tier. Fast shipping doubles freight cost and reduces return probability by a quarter. However, nine models vary by over 30 points in fast-shipping share across their episodes, suggesting that this choice is not a stable model-level policy. Store-opening counts do not consistently track performance. Cumulative openings range from 3.0 for two Qwen models to 9.0 for GPT-5.6 Sol. Each reopening incurs the ¥500 setup fee and resets that store type's reputation to its initial level. GLM 5.2 (high) opened exactly four stores every episode and placed sixth, while Gemini 3.5 Flash opened 13 in one episode and placed twelfth. Neither opening pattern reliably indicates higher profit.

J.3.6 Learning over the Horizon

Two of the 18 models have lower repeat-order anchoring regret than a random reordering of their observed prices. Qwen3.8-Max-Preview scores 0.834 on AnchorRatio and Gemini 3.5 Flash 0.918, where 1.0 matches the regret of a random ordering of the agent's own prices. The field median is 1.369, and Qwen3.5-Plus reaches 1.860 (Table 16). Fifteen models exceed their permutation null by more than two standard deviations in the direction of greater expense. This pattern is consistent with anchoring on earlier prices, a possible interpretation rather than an identified mechanism (Lou & Sun, 2026). GPT-5.5 contributes three episodes instead of five; its two bankrupt runs never buy the same item twice from the same supplier. A second measure shows a similar pattern: sixteen of the 18 models capture less surplus in the second half of an episode than in the first. The surplus half-lift orders honest sessions by conclusion day, splits the session count into two halves, and subtracts the earlier mean from the later one (§E.6).

Three mechanisms could affect improvement, without being isolated by this evaluation. Context eviction can remove earlier purchase evidence (§I.1.2); persistent memory accounts for a small fraction of calls (§4.3); and model weights remain fixed, so adaptation must use the information available during the episode (Shinn et al., 2023). Qwen3.8-Max-Preview is the only model with a

positive per-day surplus trend and the only one that beats its shuffled-price null by more than two standard deviations. Supplier choice also changes: fraudulent suppliers account for a larger share of concluded sessions late in the year than early for 17 of the 18 models.

J.4 Error Analysis

The six metrics rank performance but do not identify individual mistakes. For example, a payment to a fraudulent supplier affects both fraudulent spending and drawdown. We therefore define detection rules for ten recurring failure patterns; their mechanisms and counts are given in Table 18 and Figure 24 of Appendix G.5. Bankruptcy destroys the stake (§4.3), and unsold inventory can contribute to this outcome. One Qwen3.5-Plus episode stocks four stores in January, fails to sell enough inventory to recover its cash, and ends in May (Figure 4). Fraudulent suppliers inflate prices before delivery or deliver fewer goods than promised. Most fraudulent agreements occur in repeated trading relationships: 943 of the 1,141 agreements belong to a supplier-item pair traded twice or more (§4.3, Appendix E.1). No message identifies the supplier responsible for a defective lot, so the financial loss provides no direct attribution.

Accepting an opening quote without improvement provides a direct marker of unused bargaining opportunity (§4.3). The comparison describes price quality; it does not isolate the contribution of bargaining to bankruptcy. Claude Opus 4.7, Gemini 3.5 Flash and Fable5 never accept an opening quote and rank in the top three on CSE^{+}; Kimi K2.6 accepts one most often and ranks last. Agents also repeat earlier research. Of 2,180 catalog and supplier queries, 746 repeat arguments from an earlier query in the same episode. These repetitions cost 7,460 simulated minutes, or 83 per episode in a 219,000-minute year, despite the availability of persistent memory (§4.3). Repeating a query incurs no monetary fee. The two-day shipping deadline mainly affects the beginning and end of the year. Cancellations cluster in the first fortnight and final two weeks, when agents have not yet begun shipping or have stopped (§4.3).

The data provide limited support for three other proposed failure patterns. Store churn does not clearly distinguish performance (§4.3). Models that join more promotions also finish with more assets, but this association does not isolate the effect of event participation. Evidence for late-year collapse is mixed: growth declines after the first quarter, with a paired $t = -3.95$, but does not significantly recover in the closing window. Failure patterns appear throughout the field. Every model triggers at least five rules, and Gemini 3.1 Pro triggers all ten. Triggering fewer rules does not imply higher assets. The models with the fewest detected failures are the strongest bargainers, while the asset leaders have other weaknesses (transcripts in Appendix G).