AIGP: AN LLM-BASED FRAMEWORK FOR LONG-TERM VALUE ALIGNMENT IN E-COMMERCE PRICING

Chennan Ma, Yanning Zhang, Siqi Hong, Xiuchong Wang*, Fei Xiao, Keping Yang, Bo Zheng
Taobao & Tmall Group of Alibaba, Hangzhou, China
{machennan.mcn, xiuchong.wxc, guren.xf, bozheng}@alibaba-inc.com
{zhangyanning.zyn, hongsiqi.hsq, shaoyao}@taobao.com

ABSTRACT

Traditional dynamic pricing models in large-scale e-commerce suffer from limited interpretability, poor utilization of unstructured information, and misalignment with long-term business objectives such as cumulative Gross Merchandise Value (GMV), Return on Investment (ROI) and milestone achievement. We propose **AIGP**, a novel framework that leverages a Large Language Model (LLM) prompted with domain knowledge, structured data and textual context to make interpretable, knowledge-aware pricing decisions. For efficient deployment while maintaining high-quality outputs, we employ supervised fine-tuning for knowledge distillation. Central to AIGP is the Long-Term Value Estimator (LTVE), trained via offline reinforcement learning on historical data, which serves as a reward model to score candidate pricing actions and select preference pairs for Direct Preference Optimization (DPO), thereby aligning the pricing mechanism with long-term business objectives. Extensive offline evaluations and large-scale online A/B tests on a major e-commerce platform demonstrate that AIGP achieves significant improvements: **+13.21%** in GMV, **+7.59%** in ROI, and **+8.20%** in milestone achievement rate over 14 days compared to the production baseline, while simultaneously providing interpretable and transparent pricing rationales.

1 INTRODUCTION

Dynamic pricing is a critical driver of market efficiency across industries, with established applications in airlines, hotels, and ride-hailing. In large-scale e-commerce platforms managing extensive Stock Keeping Units (SKUs) and high-frequency daily interactions Chen et al. (2016b;a), effective dynamic pricing mechanisms are essential for responding to rapidly changing market conditions and consumer preferences.

Traditional solutions fall into two categories: rule-based systems built upon handcrafted heuristics, and data-driven models leveraging price–sales elasticity estimation and mathematical optimization. However, both approaches suffer from limited decision transparency, poor utilization of unstructured information (e.g., product descriptions, user reviews), and overemphasis on immediate gains. Aggressive discounting to boost short-term sales can erode margins and constrain future operations, making it difficult to optimize long-horizon objectives such as cumulative Gross Merchandise Value (GMV), Return on Investment (ROI), and milestone achievement.

Reinforcement learning (RL) enables direct optimization for long-term business objectives through value function modeling Liu et al. (2019); Zhu et al. (2024). Yet RL inherits interpretability issues and unstructured data limitations, while introducing challenges like distributional shift Kumar et al. (2020); Fujimoto et al. (2019) and reward sparsity Andrychowicz et al. (2017); Pathak et al. (2017), particularly for cold-start or long-tail products Yin et al. (2012); Ji et al. (2021).

Recent advances in large language models (LLMs) offer new possibilities for transparent and knowledge-rich decision making Touvron et al. (2023); Wang et al. (2022); Bubeck et al. (2023); Casper et al. (2023). LLMs' chain-of-thought reasoning provides auditable explanations for pricing

*Corresponding Authors.

decisions, while their ability to process structured and textual inputs enables richer context integration. Moreover, their reasoning over text and analogical knowledge can improve robustness for cold-start and out-of-distribution products. However, generic LLM agents lack platform-specific operational knowledge and supervision from expert demonstrations or long-horizon outcome feedback, leading to suboptimal, short-sighted actions that fail to optimize long-term business objectives.

We propose **AIGP** (**A**rtificial **I**ntelligence **G**enerated **P**ricing), a framework integrating LLM-based reasoning with long-term business value alignment for dynamic pricing. AIGP employs a carefully designed prompt incorporating chain-of-thought Wei et al. (2022) (CoT) reasoning, structured signals, domain knowledge, and textual context to generate interpretable, actionable pricing decisions. For efficient deployment, we use supervised fine-tuning Ouyang et al. (2022); Taori et al. (2023) (SFT) with teacher-student distillation Hinton et al. (2015); Xu et al. (2024), enabling compact models to produce high-quality reasoning while satisfying business constraints. Central to AIGP is the **Long-Term Value Estimator (LTVE)**, trained via offline RL on historical trajectories to model long-horizon business impact beyond immediate sales. By scoring candidate pricing actions, LTVE automates preference pair selection for Direct Preference Optimization Rafailov et al. (2023) (DPO), aligning the pricing mechanism with long-term objectives while preserving interpretability.

We deploy AIGP on a major e-commerce platform and validate its performance through comprehensive offline evaluations and large-scale online A/B tests. AIGP consistently outperforms traditional and RL-based baselines, achieving **+13.21%** in GMV, **+7.59%** in ROI, and **+8.20%** in milestone achievement rate over 14 days, while providing interpretable pricing rationales.

Our main contributions: (1) We propose **AIGP**, the first LLM-based dynamic pricing framework integrating long-horizon business reward modeling for transparent, business-aligned decisions. (2) We develop a preference alignment pipeline leveraging an offline RL-trained LTVE to automate preference pair generation for DPO. (3) We demonstrate robust, scalable deployment via distillation-based SFT, enabling compact models for large-scale production use.

2 Related Work

Traditional and RL-based Pricing Methods. Early dynamic pricing methods rely on rule-based heuristics and demand estimation Talluri & Ryzin (2006); McGill & Ryzin (1999); Gallego & Ryzin (1994); Ferreira et al. (2016), focusing on short-term profit while struggling with cold-start scenarios and unstructured information Phillips (2021); Keskin & Zeevi (2014); Besbes & Zeevi (2009). Reinforcement learning (RL) methods Mnih et al. (2015); Lillicrap et al. (2016); Haarnoja et al. (2018); Chen et al. (2021) model pricing as a Markov decision process, optimizing long-term objectives via value functions Liu et al. (2019); Zhu et al. (2024); Hazenberg et al. (2025); Villarrubia-Martin et al. (2025), but still lacks interpretability and stability under distributional shifts in offline settings Levine et al. (2020); Fujimoto et al. (2019); Kumar et al. (2020).

LLMs for Decision Making and Preference Alignment. Large language models exhibit strong reasoning capabilities through chain-of-thought prompting Wei et al. (2022), tool augmentation Schick et al. (2023), and program synthesis Nye et al. (2021), with successful applications in planning and control Huang et al. (2022); Yao et al. (2022); Ma et al. (2024); Wang et al. (2023). Although supervised fine-tuning (SFT) Ouyang et al. (2022); Taori et al. (2023) improves instruction-following, it does not optimize for long-term outcomes. Direct Preference Optimization (DPO) Rafailov et al. (2023) enables preference alignment, yet scaling high-quality preference data for long-term value remains challenging Ziegler et al. (2019); Stiennon et al. (2020). We address this by integrating offline RL-based value estimation with LLM reasoning for business-aligned dynamic pricing. To our knowledge, this is the first such framework in e-commerce.

3 Methodology

3.1 Preliminary

Dynamic pricing in e-commerce is a sequential decision-making problem where pricing actions influence future market states and cumulative business outcomes. We define this at the SKU level,

where the agent determines product prices based on market dynamics. To ensure ethical compliance, our framework is identity-blind, excluding individual user profiles to preclude price discrimination.

We formulate this problem as a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}, \mathcal{A}$ denote state and action spaces, $P(s_{t+1} \mid s_t, a_t)$ specifies transition dynamics, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. At timestep t, the agent observes $x_t = \phi(s_t, c_t)$ encoding state s_t and context c_t (business goals and operational constraints), and selects pricing action a_t from the admissible set $\mathcal{A}_{\text{safe}}(s_t)$ according to business and market rules. The objective is to maximize the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right], \tag{1}$$

where $\tau = (s_{0:T-1}, a_{0:T-1}, r_{0:T-1})$ is a trajectory generated by policy π.

Long-term value is modeled via state-value function $V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0 = s]$ and action-value function $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0 = s, a_0 = a]$, satisfying the Bellman equations and are central to our Long-Term Value Estimator (LTVE) modeling.

The pricing policy is parameterized by an LLM: $\pi_\theta(a \mid x) = \Psi(\text{LLM}_\theta(x))$, where $\text{LLM}_\theta(x)$ produces chain-of-thought reasoning and a structured decision, and Ψ extracts the pricing action ensuring $a \in \mathcal{A}_{\text{safe}}$ (definition in Appendix A.5). Unlike traditional RL policies outputting only actions, the LLM also generates interpretable reasoning traces explaining decisions.

3.2 Domain-Adaptive Supervised Fine-Tuning

When deploying LLMs as pricing policies, the model generates outputs containing both chain-of-thought (CoT) reasoning processes and final pricing decisions. This dual-output nature enables separate optimization: Supervised fine-tuning (SFT) Ouyang et al. (2022) focuses on improving reasoning quality, instruction-following, and format compliance, establishing a foundation for subsequent decision optimization. The overall workflow is illustrated in Fig. 1.

Training data consists of high-quality instruction–response pairs integrating structured product attributes (sales history, exposure and click metrics, historical actions, competitor statistics), business goals (GMV targets, ROI constraints, milestone achievements), and unstructured context (product descriptions, user reviews, domain knowledge). The prompt guides the model to analyze diagnostic insights, incorporate price-sales predictions, apply domain rules, and execute stepwise CoT reasoning ending with a compliant JSON-formatted action.

To balance inference efficiency and deployment costs, we select a 30B parameter student model (Qwen3-30B-A3B Yang et al. (2025)) for deployment. We adopt teacher-student distillation where high-quality demonstrations are generated by a larger 235B parameter teacher model (Qwen3-235B-A22B Yang et al. (2025)). We design an LLM-as-Judge module (Qwen3-235B-A22B-Instruct) that filters teacher outputs across four dimensions: (1) **Data Accuracy** verifies correct interpretation of numerical features and business constraints; (2) **Content Completeness** evaluates coverage of relevant factors and application of domain rules; (3) **Reasoning Internal Coherence** ensures logical consistency within the CoT process; (4) **Reasoning-Decision Consistency** checks alignment between reasoning and actions. Only responses with perfect scores (5/5) on all dimensions and actions satisfying $\mathcal{A}_{\text{safe}}$ constraints (Appendix A.5) are retained (Fig. 2(a)). Judge reliability is validated via 96.4% agreement with expert annotations on 2000 held-out responses.

Filtered responses are used to fine-tune the student model via parameter-efficient Low-Rank Adaptation (LoRA) Hu et al. (2022). Model outputs must end with a valid JSON action satisfying platform and business criteria. Model optimization minimizes the cross-entropy loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=0}^{|y^\star|-1} \log \pi_\theta(y_t^\star \mid x, y_{<t}^\star). \tag{2}$$

We apply mixed-precision computation, gradient clipping, and early stopping for efficiency and stability. SFT yields a policy that consistently follows instructions, generates transparent reasoning traces, and produces compliant pricing actions. However, SFT alone does not guarantee that the final pricing decisions align with long-term business value; further policy optimization through preference-based alignment is addressed in Section 3.4.

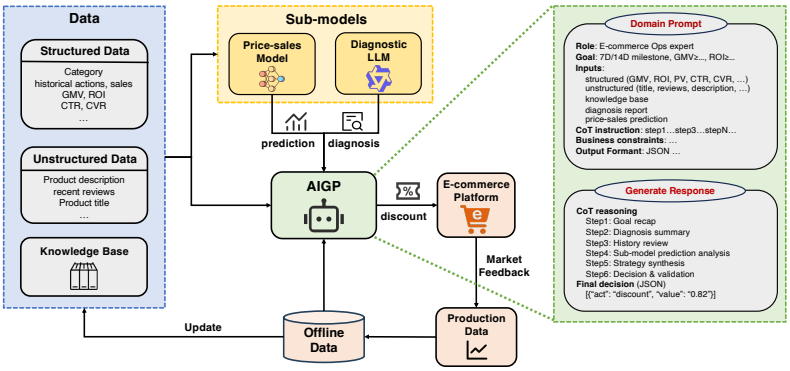

Figure 1: Overall architecture of AIGP.

3.3 Long-Term Value Modeling

Accurately evaluating the long-term business impact of dynamic pricing policies is crucial but challenging due to complex delayed market effects and lack of verifiable correctness criteria. Unlike domains such as mathematical reasoning, pricing lacks deterministic quality metrics, making standard Reinforcement Learning from Human Feedback Ouyang et al. (2022); Dai et al. (2024) (RLHF) or Reinforcement Learning from Verifiable Rewards Guo et al. (2025) (RLVR) difficult to apply. Base LLMs also lack domain-specific business knowledge and alignment with platform objectives.

To address this, we introduce a **Long-Term Value Estimator (LTVE)** Q_ϕ, trained via offline RL on over 5 million historical transitions from 6 months of production logs (60% expert trajectories selected as top 30% by cumulative GMV within comparable product groups, 40% diverse non-expert cases), to estimate expected cumulative reward for action a in state s and enable automatic preference pair construction for Direct Preference Optimization (Section 3.4).

State, Action, and Reward. States s consist of product attributes, sales statistics, inventory, historical ROI, engagement metrics (click-through rate, conversion rate), and temporal features (day of week, promotion indicators). Actions $a_t = d_t - d_{t-1}$ represent the daily discount rate change, constrained by business rules to ensure feasible and safe adjustments. The reward function balances milestone achievement (tier thresholds based on 14-day cumulative GMV) and maintaining healthy ROI. To enable effective offline RL training, we adopt a **category-normalized relative reward**:

$$r_t = \lambda_1(\text{prog}_t - \text{prog}_t^{\text{ref}}) + \lambda_2(\log(1 + \text{ROI}_t) - \log(1 + \text{ROI}_t^{\text{ref}})), \tag{3}$$

where prog_t measures daily contribution toward milestone achievement, reference values are category averages from successful products, and log-transformation handles ROI's heavy-tailed distribution. This relative formulation isolates pricing strategy quality from product characteristics.

Offline RL Training. In real-world e-commerce, online exploration is highly constrained due to business risks and the lack of reliable simulators. We train LTVE using offline RL on logged historical trajectories $\mathcal{D} = \{(s_t, a_t, r_t, \ldots)\}$. A critical challenge is **delayed reward propagation**: pricing decisions on day t influence future exposure, traffic, and sales over multiple days through search and recommendation systems. To model delayed rewards, we adopt a **critic-only architecture with multi-step temporal-difference (TD) targets**, learning action values directly from historical data without explicit policy training. This suits offline pricing logs where limited action coverage makes explicit policy learning prone to extrapolation errors on rarely observed actions. We mitigate out-of-distribution risks by constraining candidates to $\mathcal{A}_{\text{safe}}(s)$ and using double Q-learning and value clipping to reduce overestimation.

We train a state-value network $V_\psi(s)$ via expectile regression on target critics and two critics Q_{ϕ_1}, Q_{ϕ_2} with n-step TD targets:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_\tau(q_{\min}(s,a) - V_\psi(s))\right], \tag{4}$$

$$\mathcal{L}_Q(\phi_i) = \mathbb{E}_{\mathcal{D}}\left[(Q_{\phi_i}(s_t,a_t) - y_t^{(n)})^2\right], \quad y_t^{(n)} = \sum_{k=0}^{n-1}\gamma^k r_{t+k} + \gamma^n V_\psi(s_{t+n}), \tag{5}$$

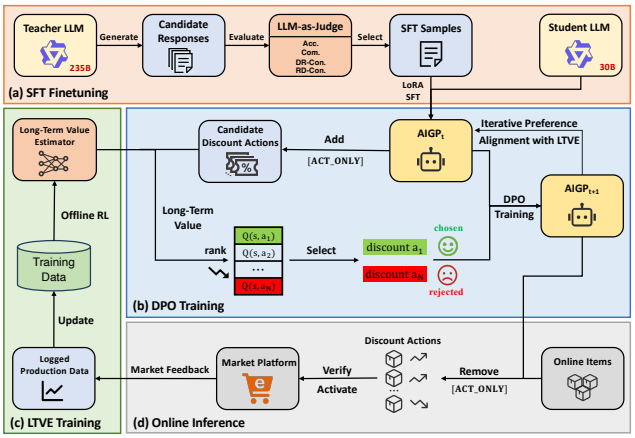

Figure 2: The finetuning and inference pipeline of AIGP.

where $q_{\min} = \min(Q_{\bar{\phi}_1}, Q_{\bar{\phi}_2})$, $L_\tau(u) = |\tau - \mathbb{I}(u < 0)|u^2$ is expectile loss, and target critics are updated softly. Bootstrapping from V_ψ avoids querying actions at future states, improving robustness to out-of-distribution actions. Complete training procedure is listed in Algorithm 1 (Appendix A.3).

3.4 Preference Alignment with Long-Term Value Estimator

While SFT improves instruction following, format compliance, and reasoning quality, it does not explicitly optimize pricing decisions for long-term business objectives. We apply preference-based policy alignment with Direct Preference Optimization (DPO), guided by the Long-Term Value Estimator (LTVE). We use LTVE to curate preference data by ranking sampled candidate actions and selecting reliable chosen-rejected pairs. This reduces the impact of potential value extrapolation errors and provides a robust long-term alignment signal. The overall workflow is shown in Fig. 2.

Decision-only Mode for Preference Learning. To focus preference alignment on action selection, we append a `[ACT_ONLY]` control token during candidate action generation and DPO training, so the model outputs only the structured discount action. This design addresses three challenges: (1) constructing reliable preference pairs from full responses is difficult because reasoning quality is hard to compare and long-form text obscures the action's contribution; (2) comparing full responses weakens the DPO signal with irrelevant tokens, whereas `[ACT_ONLY]` concentrates optimization on action tokens; (3) it avoids post-hoc action extraction, which can introduce distribution mismatch between pair construction and training. Decision-only outputs allow LTVE to rank actions and generate preferences that directly reflect decision quality (Fig. 2(b)). During deployment, the control token is removed to generate full reasoning traces for transparency.

Candidate Action Sampling and Preference Construction. For each prompt x (corresponding to state s), we sample M candidate actions $\mathcal{A}(x) = \{a^{(1)}, \ldots, a^{(M)}\}$ from the decision-only mode using stochastic decoding (temperature sampling with top-k and top-p filtering), validated against business rules to ensure $a \in \mathcal{A}_{\text{safe}}(s)$ (formal definition in Appendix A.5). We score each candidate with LTVE to obtain $Q(s, a)$ as an estimate of long-term business value. Preference pairs are constructed by selecting actions with sufficient separation:

$$Q(s, a^+) - Q(s, a^-) \geq \Delta_Q, \qquad |a^+ - a^-| \geq \Delta_a, \tag{6}$$

where a^+ and a^- denote the chosen and rejected actions, and Δ_Q and Δ_a control pair clarity. We discard ambiguous pairs with small score gaps to reduce sensitivity to estimation noise.

DPO Objective and Deployment. Given a preference pair (x, y^+, y^-), DPO minimizes

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\log \sigma\left(\beta\left(\log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\right)\right)\right]. \tag{7}$$

where π_{ref} is the frozen reference policy (the SFT model) and β is a temperature parameter. Since `[ACT_ONLY]` outputs only the pricing decision, log-probabilities in Eq. equation 7 are computed on action tokens, focusing optimization on action selection. During training, `[ACT_ONLY]` remains

enabled to match the candidate generation distribution. For deployment, we remove the control token and prompt the aligned model to generate full reasoning traces with the final action, preserving transparency. Qualified models are deployed via large-scale A/B testing with real-time monitoring of cumulative GMV, integrated ROI, and long-term milestone completion.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

4.1.1 PLATFORM DEPLOYMENT & DATA DESCRIPTION

Experiments are conducted on a large-scale e-commerce platform with full pricing authority, focusing on optimizing daily discount rates for new products to accelerate sales growth and improve marketing ROI. The experimental dataset comprises millions of SKUs from over 180 days of production operations, including product attributes (category, brand, price, inventory), historical metrics (GMV, ROI, click-through rate, conversion rate), and unstructured context. The system makes offline decisions for about 200,000 products daily (actions effective next day). For offline evaluation, we employ a time-based split to prevent temporal data leakage. Our online A/B tests have been deployed and monitored for over 60 consecutive days, ensuring sustained stability and effectiveness.

4.1.2 IMPLEMENTATION DETAILS

For LTVE, we train a critic-only double Q model using over 5 million transitions from 6 months of production logs. We set the discount factor $\gamma = 0.95$, TD horizon $n = 3$, and soft update rate $\eta = 0.01$. All LTVE networks are optimized with Adam (lr $= 3 \times 10^{-4}$). For the LLM agent, SFT is conducted on 100,000 instruction-response pairs, and DPO uses 50,000 chosen-rejected preference pairs, both via LoRA adaptation. Complete implementation details are in Appendix A.1.

4.1.3 BASELINES

We compare AIGP against three categories of baselines. (1) **LLM Variants**: base student (Qwen3-30B-A3B) and teacher (Qwen3-235B-A22B) models Yang et al. (2025), along with ablated versions **AIGP (SFT-only)** and **AIGP (DPO-only)**. (2) **Online Deployed Policies**: production **Price-Sales Model** (DNN-based elasticity estimator) and **RL-DT** (Decision Transformer Chen et al. (2021)) for sequence-to-sequence long-term goal optimization. (3) **Academic RL Baselines**: **RL-DDPG** Liu et al. (2019) and **RL-SAC** Zhu et al. (2024) using identical historical data and reward functions as LTVE. Detailed architectures and implementation are provided in Appendix A.2.

4.2 EVALUATION METRICS

4.2.1 OFFLINE EVALUATION METRICS

Offline evaluation covers decision quality (Q-values and expert alignment) and reasoning quality (LLM-as-Judge). For decision quality, we use Q-values from LTVE to measure the expected long-term business value (reliability validated in Section 4.3) of pricing decisions. Additionally, we use Expert Action Matching Accuracy (EAMA) to evaluate alignment with expert strategies:

$$\text{EAMA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\left|\pi\left(s_i\right) - a_i^{\text{expert}}\right| < \epsilon\right), \tag{8}$$

where $\pi(s_i)$ is the model's recommended discount, a_i^{expert} is the expert action, and ϵ is a small threshold. For reasoning quality, we employ the LLM-as-Judge module (Section 3.2) on four dimensions: Data Accuracy (Acc.), Content Completeness (Com.), Reasoning Internal Coherence (Int-Coh.), and Reasoning-Decision Consistency (RD-Con.).

4.2.2 ONLINE EVALUATION METRICS

Business impact is evaluated over 7-day and 14-day horizons after product launch. Key metrics include cumulative Gross Merchandise Value (GMV), integrated Return on Investment (ROI), and

Table 1: Comprehensive evaluation of LTVE: offline reliability and 7-day online A/B test.

	Offline Metrics			**Online A/B (7D)**		
n-step TD target	MAE	EAMA	CDA	MAR	GMV	ROI
$n = 1$	0.045	87.7%	91.2%	–	–	–
$n = 3$	**0.027**	**90.7%**	**97.5%**	**+2.02%**	**+8.3%**	**+7.2%**
$n = 5$	0.056	86.4%	89.7%	–	–	–

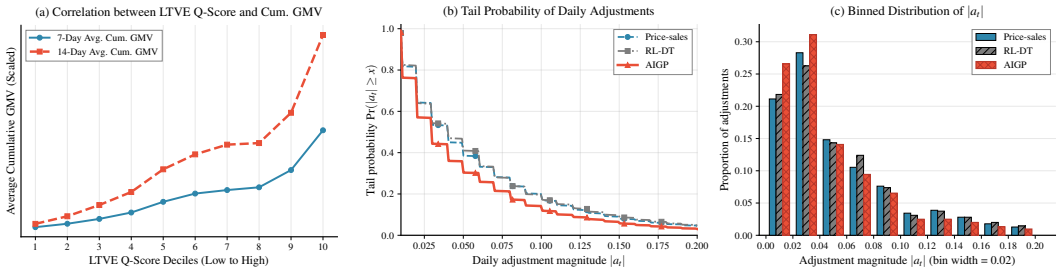

Figure 3: **Comprehensive evaluation of LTVE effectiveness and AIGP pricing stability.** (a) Positive correlation between Q-score deciles and average cumulative GMV over 7-day and 14-day horizons. (b) Tail probability $\Pr(|a_t| \geq x)$ of daily price adjustments. (c) Binned distribution of adjustment magnitudes $|a_t|$ with a bin width of 0.02.

Milestone Achievement Rate (MAR). MAR measures the proportion of SKUs reaching joint long-term GMV and ROI targets, ensuring immediate sales growth is balanced with financial sustainability. Significant gains in MAR demonstrate the model's capacity for stable long-term growth.

4.3 Evaluation of LTVE

4.3.1 Offline Evaluation Results

We evaluate LTVE reliability on a held-out expert set of 30,000 samples. For operational fidelity, we use Expert Action Matching Accuracy (EAMA, Section 4.2.1). To measure the ability to distinguish optimal decisions, we compute Counterfactual Discrimination Accuracy (CDA):

$$\text{CDA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(Q(s_i, a_i^{\text{expert}}) > Q(s_i, a_i^{\text{cf}})\right), \tag{9}$$

where Q is the LTVE value function, a_i^{expert} is the expert action, and a_i^{cf} is a counterfactual variant.

n-step TD ablation. We ablate the multi-step TD horizon $n \in \{1, 3, 5\}$. As shown in Table 1, $n = 3$ achieves the best overall performance (MAE **0.027**, EAMA **90.7%**, CDA **97.5%**). $n = 1$ provides insufficient credit assignment for delayed pricing effects, while $n = 5$ amplifies off-policy estimation errors under limited action coverage in offline logs. $n = 3$ strikes the best balance between long-term impact and training stability, and is used as the default.

Correlation with long-term outcomes. To validate LTVE's predictive reliability, we partition 50,000 samples into 10 equal-sized Q-score deciles and track realized cumulative GMV over subsequent 7-day and 14-day periods, excluding the top and bottom 5% of products within each decile to ensure robustness. Fig. 3(a) shows strong monotonic correlation between Q-score deciles and realized GMV across both horizons. Quantitatively, LTVE scores exhibit Spearman correlation coefficients of $\rho = 0.9306$ with 7-day cumulative GMV and $\rho = 0.9378$ with 14-day cumulative GMV. These high correlations confirm that LTVE effectively captures long-term pricing consequences and enables reliable preference pair construction for policy alignment (Section 3.4).

Table 2: Comprehensive Offline Evaluation: Decision Quality and Reasoning Quality.

Model	Decision Quality			Reasoning Quality (LLM-as-Judge)				
	Q-Score	MAE	EAMA	Acc.	Com.	Int-Coh.	RD-Con.	Total
Price-Sales	2.564	0.084	71.08%	-	-	-	-	-
RL-DT	2.694	0.078	73.89%	-	-	-	-	-
RL-DDPG	2.364	0.110	56.91%	-	-	-	-	-
RL-SAC	2.494	0.097	63.95%	-	-	-	-	-
Qwen3-30B-A3B	2.585	0.085	69.22%	4.52	4.70	4.70	4.21	18.13
Qwen3-235B-A22B	2.556	0.089	67.07%	**4.96**	**5.00**	**4.99**	**4.85**	**19.80**
AIGP (SFT-only)	2.554	0.086	68.98%	4.76	4.97	4.91	4.60	19.24
AIGP (DPO-only)	2.794	0.067	79.26%	4.74	4.92	4.86	4.59	19.11
AIGP (SFT+DPO)	**2.836**	**0.062**	**82.51%**	4.88	4.98	4.95	4.71	19.52

4.3.2 Online A/B Test

We evaluate LTVE in online deployment by integrating it with a production Decision Transformer (RL-DT) model. For each product, RL-DT model generates multiple candidate discount actions. In the control group, DT's built-in critic selects the discount. In the experimental group, LTVE replaces the critic and chooses the discount with the highest estimated long-term value. As shown in Table 1, LTVE improves business metrics over 7-day horizon: +2.02% in MAR, +8.3% in GMV, and +7.2% in ROI, demonstrating effectiveness for real-world optimization.

4.4 Evaluation of AIGP

4.4.1 Offline Evaluation Results

We conduct comprehensive evaluations of AIGP across three dimensions:

Long-term Value Alignment. Table 2 shows full AIGP (SFT+DPO) achieves the highest Q-score of **2.836**, significantly exceeding production baselines (Price-Sales: 2.564; RL-DT: 2.694) and LLM variants. Academic RL baselines (RL-DDPG, RL-SAC) yield lower scores, demonstrating AIGP's superiority in aligning decisions with long-term business value.

Operational Fidelity. We assess operational fidelity by measuring alignment with expert trajectories using Mean Absolute Error (MAE) and Expert Action Matching Accuracy (EAMA). Table 2 shows full AIGP achieves the lowest MAE (**0.062**) and highest EAMA (**82.51%**), indicating outputs closely match expert practices. DPO-only's strong performance confirms the importance of preference optimization, while SFT-only and base models show higher errors.

Reasoning quality. Text generation quality is assessed via LLM-as-Judge on data accuracy, content completeness, reasoning internal coherence (Int-Coh.), and reasoning-decision consistency (RD-Con.). As shown in Table 2, while Qwen3-235B scores highest (**19.80**), reflecting scaling laws Kaplan et al. (2020); Hoffmann et al. (2022), SFT and DPO progressively enhance the 30B model. SFT alone improves total score from 18.13 to 19.24, and combining SFT with DPO achieves **19.52**, approaching the 235B model. This indicates that targeted alignment effectively distills specialized pricing knowledge into compact architectures, closing the size-induced performance gap.

4.4.2 Online A/B Test

We conducted large-scale A/B tests by deploying AIGP alongside baselines on the online platform. Approximately 200,000 SKUs were randomly assigned to mutually exclusive groups via a Hash-based stable partitioning (Appendix A.6), maintaining policy consistency throughout each SKU's lifecycle and ensuring balance across exposure, category, and price levels. The experiment ran continuously for over 60 days with daily monitoring to verify stability. Performance is measured over 7-day and 14-day horizons on MAR, cumulative GMV, and integrated ROI. Table 3 reports relative improvements over the Price-Sales baseline. AIGP (SFT+DPO) achieves strong and sustained performance: **+6.92%** MAR, **+11.04%** GMV, and **+8.23%** ROI over 7 days, and **+8.20%** MAR, **+13.21%** GMV, and **+7.59%** ROI over 14 days, demonstrating substantial business value at scale.

Table 3: Online A/B Test Results at 7-Day and 14-Day Horizons.

Model	7D-MAR	7D-GMV	7D-ROI	14D-MAR	14D-GMV	14D-ROI
RL-DT	+1.33%	+4.02%	+5.71%	+0.88%	+9.04%	+3.46%
Qwen3-30B-A3B	+0.88%	+2.22%	+2.29%	−2.98%	+4.89%	+4.33%
Qwen3-235B-A22B	+5.69%	+8.67%	**+8.52%**	+7.37%	+10.50%	+6.62%
AIGP (SFT-only)	+5.59%	+4.82%	+8.00%	+3.34%	+7.85%	+5.29%
AIGP (SFT+DPO)	**+6.92%**	**+11.04%**	+8.23%	**+8.20%**	**+13.21%**	**+7.59%**

4.4.3 Pricing Adjustment Stability

Beyond business outcomes, we analyze daily discount adjustments $a_t = d_t - d_{t-1}$ to verify AIGP's stability, which is critical for new products where traditional models often exhibit high pricing volatility. Fig. 3(b) shows AIGP yields lower tail probabilities at larger thresholds, indicating fewer extreme discount jumps. Fig. 3(c) reveals AIGP assigns higher proportion to small-magnitude bins, demonstrating more controlled adjustments. This pricing stability, achieved by integrating unstructured context and analogical reasoning, avoids abrupt price swings while providing merchants with predictable dynamics for inventory planning. Moreover, AIGP's explicit natural-language rationales enable post-hoc auditing, supporting safer large-scale deployment.

4.4.4 Ablation Studies

We ablate SFT and DPO across four variants: Qwen3-30B base, SFT-only, DPO-only, and full AIGP. Table 2 shows SFT mainly improves reasoning quality, while DPO significantly boosts long-term value alignment and expert matching. Specifically, DPO-only achieves higher Q-score and expert matching (2.794, 79.26%) than SFT-only (2.554, 68.98%), while SFT yields better reasoning scores. Combining SFT and DPO delivers the best overall performance: reasoning quality (19.52), long-term value (Q-value 2.836), and operational fidelity (EAMA 82.51%). Online testing (Table 3) confirms full AIGP achieves superior business performance. These results highlight that SFT enables interpretable reasoning while DPO optimizes long-term outcomes. Their combination allows compact models to match larger LLM's performance with full explainability.

4.4.5 Case Study

We present two scenarios demonstrating AIGP's advantages: (1) leveraging unstructured information, and (2) cold-start generalization. As detailed in Appendix A.4, **Case 1** shows AIGP pricing a children's desk by integrating reviews (Excellent quality) and seasonal context (back-to-school). CoT reasoning identifies that seasonal demand increases willingness-to-pay while positive reviews reduce price sensitivity, yielding a discount rate of 0.85 versus baseline's 0.71, preserving 14% margin. **Case 2** shows AIGP handling a cold-start 3-layer steamer with minimal data. Leveraging the knowledge base (*1-layer steamers priced at ¥38, 2-layer at ¥44*), AIGP applies analogical reasoning to recommend 0.76 versus baseline's 0.68. Both cases highlight: (1) **Unstructured data utilization**, (2) **Cold-start robustness**, and (3) **Interpretability** via explicit CoT traces.

5 Conclusion & Future Work

We propose **AIGP**, a novel framework integrating LLM-based reasoning with offline RL for interpretable, business-aligned dynamic pricing. By combining supervised fine-tuning with Direct Preference Optimization guided by a Long-Term Value Estimator (LTVE), AIGP enables domain-specific knowledge condensation, allowing compact, deployable models to achieve decision quality and reasoning depth competitive with much larger LLMs while ensuring policy transparency. Extensive evaluations on a large-scale e-commerce platform demonstrate that AIGP significantly outperforms established baselines in cumulative GMV, ROI and milestone achievement, validating its ability to drive sustainable growth at scale. Future work will explore: (1) cross-domain extensibility of AIGP to broader e-commerce decision tasks; and (2) refined alignment strategies for unified frameworks that simultaneously enhance reasoning and decision-making.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Le Chen, Alan Mislove, and Christo Wilson. An empirical analysis of algorithmic pricing on amazon marketplace. *Proceedings of the 25th International Conference on World Wide Web*, 2016a. URL https://api.semanticscholar.org/CorpusID:9570936.

Le Chen, Alan Mislove, and Christo Wilson. An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th international conference on World Wide Web*, pp. 1339–1349, 2016b.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Neural Information Processing Systems (NeurIPS)*, pp. 15084–15097, 2021.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing and Service Operations Management*, 18(1):69–88, 2016.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Guillermo Gallego and Garrett J. Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pp. 1856–1865, 2018.

Thomas Hazenberg, Yao Ma, Seyed Sahand Mohammadi Ziabari, and Marijn van Rijswijk. Multi-agent reinforcement learning for dynamic pricing in supply chains: Benchmarking strategic agent behaviours under realistically simulated market conditions, 2025. arXiv preprint arXiv:2507.02698.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pp. 9118–9147. PMLR, 2022.

Luo Ji, Qi Qin, Bingqing Han, and Hongxia Yang. Reinforcement learning to optimize lifetime value in cold-start recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 782–791, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

N. Bora Keskin and Assaf Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

Jiaxi Liu, Yidong Zhang, Xiaoqing Wang, Yuming Deng, and Xingyu Wu. Dynamic pricing on e-commerce platform with deep reinforcement learning: A field experiment, 2019. arXiv preprint arXiv:1912.02572.

Weiyu Ma, Qirui Mi, Yongcheng Zeng, Xue Yan, Runji Lin, Yuqiao Wu, Jun Wang, and Haifeng Zhang. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *Advances in Neural Information Processing Systems*, 37:133386–133442, 2024.

Jeffrey I. McGill and Garrett J. Van Ryzin. Revenue management: Research overview and prospects. *Transportation Science*, 33(2):233–256, 1999.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models, 2021. arXiv preprint arXiv:2112.00114.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Robert L. Phillips. *Pricing and Revenue Optimization*. Stanford University Press, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 68539–68551, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Kalyan T. Talluri and Garrett J. Van Ryzin. *The Theory and Practice of Revenue Management*. Springer, 2006.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. arXiv preprint arXiv:2304.04487.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Enrique Adrian Villarrubia-Martin, Luis Rodriguez-Benitez, David Mu noz Valero, Giovanni Montana, and Luis Jimenez-Linares. Dynamic pricing in high-speed railways using multi-agent reinforcement learning, 2025. arXiv preprint arXiv:2501.08234.

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. *arXiv preprint arXiv:2305.16291*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

AA. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. Challenging the long tail recommendation. *arXiv preprint arXiv:1205.6700*, 2012.

Chenyao Zhu, Caiqian Cheng, and Sisi Meng. Drl pricepro: A deep reinforcement learning framework for personalized dynamic pricing in e-commerce platforms with supply constraints. *Spectrum of Research*, 4(1), 2024.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019.

A Appendix

A.1 Detailed Hyper-parameters

Table 4 lists the detailed configurations for the LTVE training, SFT, and DPO alignment stages.

Table 4: Comprehensive Hyper-parameters for AIGP.

Category	**Parameter / Value**
Long-Term Value Estimator (LTVE)	
Discount Factor γ	0.95
n-step Horizon	3
Expectile τ	0.8
Learning Rate	3×10^{-4} (Adam)
Soft Update Rate η	0.01
Batch Size	512
Reward Milestone Weight λ_1	0.8
Reward ROI Weight λ_2	0.2
LLM Supervised Fine-Tuning (SFT)	
Learning Rate	5×10^{-5} (Cosine Scheduler)
LoRA Target	o_proj, q_proj, k_proj, v_proj
Cutoff Length	8192
Epochs	2
LLM Preference Alignment (DPO)	
Learning Rate	1×10^{-5}
DPO β	0.1
LoRA Target	o_proj, q_proj, k_proj, v_proj
Cutoff Length	4096
Epochs	3

A.2 Baseline Implementation and Descriptions

Price-Sales Model. This is the current production baseline deployed online. It utilizes a deep neural network to model price-sales elasticity, specifically predicting the uplift of sales relative to price adjustments. However, it lacks the ability to incorporate unstructured context (like user reviews) and fails to optimize for long-term objectives such as milestone achievement. Due to commercial confidentiality, further technical specifics of the neural architecture are omitted, but it represents a high-performing point-estimate approach widely used in large-scale e-commerce.

RL-DT (Decision Transformer). RL-DT model the daily dynamic pricing task as a sequence-to-sequence decision-making problem via Decision Transformer Chen et al. (2021) architecture. Unlike traditional RL, RL-DT leverages the transformer's ability to process long-range historical sequences, predicting the next pricing action conditioned on the Return-to-go (the expected cumulative reward). This makes it a strong baseline for long-term goal optimization. The model uses the same historical transitions and reward signals as our framework to ensure comparability.

RL-DDPG. We implement a continuous control baseline based on the Deep Deterministic Policy Gradient Liu et al. (2019) algorithm. It follows an Actor-Critic architecture where the actor network outputs deterministic discount rates and the critic estimates the Q-value. To ensure a fair comparison, the dataset and reward construction are strictly identical to those used for training our LTVE.

RL-SAC. We also adapt the Soft Actor-Critic Zhu et al. (2024) algorithm, which incorporates an entropy-regularized Actor-Critic framework. This baseline is designed to improve robustness and exploration by maximizing both the expected reward and the policy entropy. To ensure a fair comparison, the dataset and reward construction are strictly identical to those used for training our LTVE.

Base LLMs. We include base Qwen3-30B-A3B and Qwen3-235B-A22B Yang et al. (2025) models that receive the same structured and unstructured inputs but make pricing decisions based solely on their pre-trained general knowledge. This baseline quantifies the specific gains achieved through our long-term value alignment and supervised domain adaptation.

A.3 LTVE Training Algorithm Details

Algorithm 1 presents the complete training procedure for the Long-Term Value Estimator (LTVE), referenced in Section 3.3.

Algorithm 1 Long-Term Value Estimator (LTVE) Training (Critic-only)

Require: Offline dataset $\mathcal{D} = \{(s_t, a_t, r_{t:t+n-1}, s_{t+n})\}$
Require: Critics Q_{ϕ_1}, Q_{ϕ_2} and target critics $Q_{\bar{\phi}_1}, Q_{\bar{\phi}_2}$
Require: Value network V_ψ
Require: Discount factor γ, expectile τ, soft update rate η, clipping bounds $[Q_{\min}, Q_{\max}]$
Initialize ϕ_1, ϕ_2, ψ; set $\bar{\phi}_1 \leftarrow \phi_1$, $\bar{\phi}_2 \leftarrow \phi_2$
for each training iteration **do**
  Sample batch $\{(s_t, a_t, r_{t:t+n-1}, s_{t+n})\}$ from $\mathcal{D}$
  $q_{\min} \leftarrow \min(Q_{\bar{\phi}_1}(s_t, a_t),\ Q_{\bar{\phi}_2}(s_t, a_t))$
  $\delta \leftarrow q_{\min} - V_\psi(s_t); \quad w \leftarrow |\tau - \mathbb{I}(\delta < 0)|$
  Update ψ by minimizing $\mathbb{E}\big[w \cdot \delta^2\big]$
  $y^{(n)} \leftarrow \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n V_\psi(s_{t+n})$
  $y^{(n)} \leftarrow \mathrm{clip}(y^{(n)}, Q_{\min}, Q_{\max})$ ▷ target clipping
  Update ϕ_1, ϕ_2 by minimizing $\mathbb{E}\big[(Q_{\phi_i}(s_t, a_t) - y^{(n)})^2\big]$
  $\bar{\phi}_1 \leftarrow \eta\phi_1 + (1-\eta)\bar{\phi}_1; \quad \bar{\phi}_2 \leftarrow \eta\phi_2 + (1-\eta)\bar{\phi}_2$
end for
return $Q_{\phi_1}, Q_{\phi_2}, V_\psi$

A.4 Detailed Case Study Analysis

This section provides a visual walkthrough of the two cases referenced in Section 4.4.5. Figure 4 illustrates how AIGP's Chain-of-Thought (CoT) reasoning processes both structured and unstructured inputs to outperform traditional Price-Sales baselines.

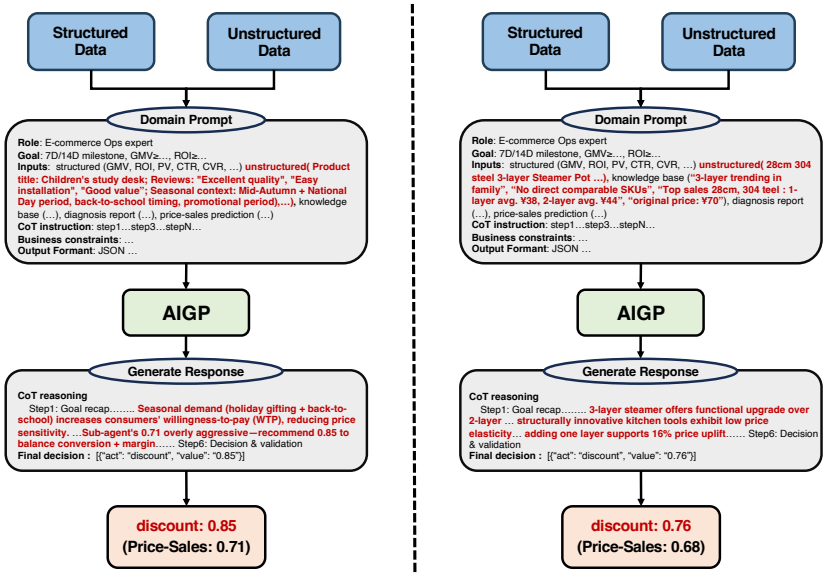

Figure 4: AIGP compared to traditional models. (Left) Case 1: Utilizing unstructured information; (Right) Case 2: Cold-start adaptation.

A.5 Formal Definition of Admissible Action Space $\mathcal{A}_{\text{safe}}$

To ensure business stability and regulatory compliance during online deployment, we formally define the admissible action space $\mathcal{A}_{\text{safe}}(s_t)$. In our framework, a pricing action a_t represents the daily

change in the discount rate, i.e., $a_t = d_t - d_{t-1}$, where d_t is the absolute discount rate applied at time t.

The safety space $\mathcal{A}_{\text{safe}}$ is governed by the following operational constraints:

- **Daily Adjustment Constraints.** To prevent abrupt price swings that may destabilize market expectations or merchant inventory planning, the daily adjustment magnitude is restricted to a fixed range:
$$a_t \in [-0.2, 0.2] \tag{10}$$
This ensures that the discount rate cannot fluctuate by more than 20% within a single 24-hour cycle.
- **Absolute Discount Boundaries.** To maintain brand value and protect minimum profit margins, the resultant absolute discount rate d_t must remain within a predefined sustainable interval:
$$d_t \in [0.5, 1.0] \tag{11}$$
where $d_t = 1.0$ represents the original listing price. Any discount exceeding 50% ($d_t < 0.5$) is prohibited to prevent erosion of Return on Investment (ROI).

Formally, given the previous state's discount d_{t-1}, the set of admissible actions at time t is defined as:
$$\mathcal{A}_{\text{safe}}(s_t) = \{a_t \mid -0.2 \leq a_t \leq 0.2 \text{ and } 0.5 \leq d_{t-1} + a_t \leq 1.0\} \tag{12}$$

As illustrated in our inference pipeline in Fig 2(d), any action a_t generated by the pricing policy that violates these boundaries is projected back to the nearest feasible point in $\mathcal{A}_{\text{safe}}$ before platform execution. This mechanism guarantees that the agent's exploration remains within the safe operational manifold of the e-commerce platform.

A.6 Hash-based Partitioning of Online A/B Test

To ensure the statistical validity of our online A/B tests and maintain policy consistency for each product, we implement a *Hash-based Stable Partitioning* mechanism. The core objective is to map each Stock Keeping Unit (SKU) into mutually exclusive experimental or control groups based on its unique identifier, ensuring that approximately 200,000 SKUs are assigned consistently throughout the testing period. The partitioning process follows these technical steps:

Identifier Selection. We use the unique `SKU_ID` as the primary key for partitioning to ensure that the same product is always processed by the same pricing policy throughout its lifecycle.

Salted Hashing. To prevent potential bucket coupling (where products are always grouped together across different independent experiments), we append a specific `Experiment_Salt` to the `SKU_ID` before hashing. This salted approach also mitigates selection bias that might arise from semi-semantic `SKU_ID` structures by ensuring high entropy in the hash space.

Modulo Mapping. The salted ID is processed using a cryptographic hash function, and the resulting integer is mapped to a set of discrete buckets via a modulo operation. This mapping ensures a deterministic and uniform distribution of SKUs across experimental cohorts.

Group Assignment. Based on pre-defined traffic allocation ratios, SKUs falling into specific numeric buckets are assigned to their respective treatment groups, such as the AIGP (SFT+DPO) cohort or the production baseline.

This method guarantees **policy stability**. Since the platform makes daily decisions for a vast number of products, any fluctuation in group assignment would lead to erratic pricing signals and destabilize merchant inventory planning. By using Hash-based Partitioning, we ensure that experimental groups remain well-balanced across exposure, category, and price levels, thereby isolating the true business impact of the AIGP framework.

A.7 Discussion

Sequential Decision Bias and Confounding. Offline value learning from logged data risks inheriting historical policy biases. We mitigate this through three mechanisms: (1) the admissible action

space $\mathcal{A}_{\text{safe}}$ constrains daily adjustments to $[-0.2, 0.2]$ and absolute discounts to $[0.5, 1.0]$, limiting distribution shift between the logging and learned policies; (2) the category-normalized relative reward (Eq. 3) isolates pricing strategy quality from product-specific confounders by referencing category-level baselines; and (3) online A/B tests (Table 1 and Table 3) provide randomized causal validation independent of offline data assumptions. Additionally, LTVE is retrained monthly on the latest production logs to adapt to evolving market dynamics, and DPO alignment is subsequently updated based on the retrained LTVE to ensure the pricing policy remains consistent with the most current value estimates.

Attribution of LLM's Contribution. We argue that the LLM serves as more than a formatted action generator. First, AIGP (DPO-only) achieves the lowest MAE (0.067) and highest EAMA (79.26%) among all single-stage variants (Table 2), substantially outperforming RL baselines (RL-DT: MAE 0.078, EAMA 73.89%; RL-SAC: MAE 0.097, EAMA 63.95%) in alignment with expert decisions. This indicates that the LLM as a policy backbone provides superior action selection beyond what value guidance alone can explain. Second, case study in Section 4.4.5 and Appendix A.4 demonstrates that AIGP actively leverages unstructured signals from similar products to make decisions that structured-data-only models cannot replicate, particularly for cold-start and long-tail products. We acknowledge that a controlled comparison against an MLP baseline with identical LTVE guidance would strengthen this attribution, and leave it as future work.

Reasoning-Decision Relationship in [ACT_ONLY] Mode. The `[ACT_ONLY]` mode does not decouple reasoning from decision-making in the conventional sense. An alternative approach would generate full chain-of-thought responses and then parse out actions for preference learning, where the optimization signal is diluted across reasoning tokens and the extracted action may not reflect the true generation distribution. In contrast, `[ACT_ONLY]` directs the model to produce only the structured pricing action, so DPO log-probabilities are computed exclusively on action tokens. This ensures that the preference signal directly optimizes the decision itself without interference from reasoning-quality variance. During deployment, the control token is removed and the model generates full CoT reasoning followed by the action, leveraging the reasoning capabilities acquired through SFT. Exploring unified alignment strategies that simultaneously optimize both reasoning quality and decision quality remains a promising direction for future work.

A.8 LLM-AS-JUDGE DETAILED EVALUATION

Figure 5 presents a detailed breakdown of LLM-as-Judge scores across all model variants on the four evaluation dimensions: Data Accuracy (Acc.), Content Completeness (Com.), Reasoning Internal Coherence (Int-Coh.), and Reasoning-Decision Consistency (RD-Con.). The visualization confirms that while the 235B teacher model achieves the highest overall scores, the combination of SFT and DPO enables the 30B student model to approach teacher-level reasoning quality. Notably, AIGP (SFT+DPO) demonstrates balanced performance across all dimensions, achieving 4.88 in Accuracy, 4.98 in Completeness, 4.95 in Internal Coherence, and 4.71 in RD-Consistency, significantly outperforming the base 30B model and approaching the 235B benchmark.

A.9 INFRASTRUCTURE AND DEPLOYMENT EFFICIENCY

Hardware Constraints and Model Selection. To ensure the practical feasibility of the AIGP framework in a large-scale production environment, we conducted all model alignment and experimental evaluations using a cluster of 16 GPUs. While the 235B teacher model exhibits superior general-purpose reasoning, its parameter scale poses prohibitive computational demands for fine-tuning within our available infrastructure. Specifically, the 16-GPU cluster is insufficient to support even parameter-efficient fine-tuning (e.g., LoRA) for a 235B-scale model due to severe memory bottlenecks and excessive training latency. In contrast, the 30B student model (Qwen3-30B-A3B) allows for high-quality SFT and DPO alignment under these hardware constraints, effectively internalizing domain-specific pricing knowledge.

Operational Pipeline and Inference Capacity. The AIGP system serves a vertical marketplace with approximately 200,000 active SKUs per day. Instead of real-time price adjustments, our framework follows a daily offline execution pipeline. Decisions are computed in batch during nightly processing windows, with the resulting pricing strategies synchronized to the production environment for execution on the subsequent day (T+1). This operational design prioritizes total throughput

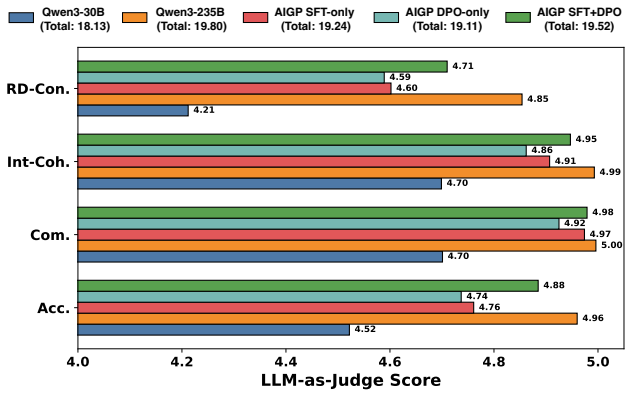

Figure 5: Detailed LLM-as-Judge scores of reasoning quality across model variants. The bar chart shows individual scores on four dimensions (Acc., Com., Int-Coh., RD-Con.) for each model, with AIGP (SFT+DPO) approaching the performance of the much larger 235B teacher model.

for massive SKU volumes over immediate response latency. Utilizing the 30B model significantly optimizes this pipeline:

- **Total Inference Duration**: The average end-to-end inference time, including comprehensive Chain-of-Thought reasoning traces, is approximately 60s per SKU. By utilizing a batch size of 32 across the 16-GPU cluster, the system can process the entire catalog of 200,000 SKUs in approximately 7 hours.
- **Efficiency Comparison**: Under the same hardware constraints, the 235B model would require nearly $6\times$ the computing time to achieve equivalent throughput, far exceeding the allocated 7-hour production window and hindering the feasibility of daily T+1 updates.

Task-Specific Superiority. Our results confirm that the 30B model, after SFT and DPO alignment, outperforms the 235B base model in pricing accuracy (EAMA 82.51% vs. 67.07%) and business value alignment (Q-Score 2.836 vs. 2.556). This confirms that for specialized e-commerce dynamic pricing task, a task-aligned compact model is more effective and resource-efficient than a massive general-purpose base model.