AlphaG-OPD: Reliability-Gated Sibling Counterfactuals for On-Policy Distillation in Symbolic Alpha Factor Discovery

Anonymous authors
Paper under double-blind review

Abstract

Symbolic alpha factor discovery can score a completed expression, but it provides no direct label for the structural decisions that produced it. Generative flow networks (GFlowNets) preserve a diverse, reward-proportional distribution over complete expressions, yet their trajectory-level objective does not compare unchosen sibling actions at an intermediate state. We introduce AlphaG-OPD, a structural on-policy distillation framework that turns terminal factor evaluations into local action guidance. Its design separates three decisions. Component I determines *where to teach* by exposing grammar-valid siblings at partial abstract-syntax-tree (AST) states visited by the current forward policy. Component II determines *what is reliable enough to teach*: it evaluates three supported siblings under four shared suffixes and admits a KL-bounded target only when their matched comparisons exhibit sufficient winner agreement and a positive empirical lower confidence bound (LCB). Component III determines *how strongly and for how long to teach* by consolidating accepted targets through bounded replay, score-indexed expiry, and forward-gradient balancing, without additional factor evaluations. Terminal reward, Trajectory Balance, the backward policy, grammar, and factor-pool rules remain unchanged. An equal-physical-score four-arm ablation tests paired teaching, reliability gating, and consolidation. Across China's CSI300, CSI500, and CSI1000 and the U.S. S&P 500, the complete method delivers strong cross-market performance over multiple random seeds.

1 Introduction

Symbolic alpha discovery searches a combinatorial program space, yet a factor is scored only after its expression is complete. Genetic programming and neural symbolic regression explore this space effectively (Zhang et al., 2020; Cui et al., 2021; Landajuela et al., 2022; Biggio et al., 2021), while REINFORCE and proximal policy optimization (PPO)-style methods directly optimize expected terminal quality (Yu et al., 2023; Zhao et al., 2025; Schulman et al., 2017). GFlowNets provide a complementary strength: rather than return one maximizer, they learn a reward-proportional distribution and can preserve multiple useful, nonredundant factor modes (Bengio et al., 2023; Chen et al., 2025). These methods solve search and exploration; they do not by themselves identify which alternative action at a partial expression caused a better terminal outcome.

This distinction between global quality and local credit is consequential in AST search. Two trajectories may share the same partial expression but close it with different operands, operators, or normalization choices. Their terminal score entangles the action under study with all of those later decisions. A larger search budget can reduce uncertainty about which complete expressions are valuable, yet it still does not identify which legal action was preferable at the shared prefix. Direct local supervision therefore requires a paired comparison that changes the action while controlling the reachable completion.

On-policy distillation (OPD) supplies precisely this missing interface. The evolving student policy provides the states, and a teacher supplies a local action target at those same states. This reduces the state-distribution mismatch of static distillation and turns feedback on the student's current behavior into a direct probability update (Gu et al., 2024; Agarwal et al., 2024; Song & Zheng, 2026). In

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.

our setting, TB remains the global search objective; OPD is helpful as a complementary local-credit channel that can say which next construction should receive more mass on an AST the current GFlowNet actually visits.

The difficulty is that an unfinished symbolic expression has no external teacher. Reusing its eventual terminal score merely repeats TB, independently completing sibling actions confounds action quality with continuation difficulty, and repeatedly distilling an unstable preference can be harmful. Recent OPD studies similarly find that teacher compatibility, local teachability, and supervision horizon determine whether on-policy guidance helps or hurts (Li et al., 2026a; Fu et al., 2026; Armandpour et al., 2026; Liu et al., 2026; Liang et al., 2026). Structural OPD therefore needs matched action comparisons, an explicit reliability rule, and a bounded optimization budget.

We introduce ALPHAG-OPD, to our knowledge the first structural OPD framework for GFlowNet-based symbolic alpha factor discovery. ALPHAG-OPD converts terminal factor evaluations into selective local action supervision by comparing grammar-compatible sibling actions under matched completions at states visited by the current GFlowNet, then consolidates only accepted credit under a fixed evaluator budget without replacing Trajectory Balance. The framework is organized around three questions: *where may OPD intervene, which local preference is reliable enough to teach, and how strongly and for how long should accepted credit influence the policy?* A reward-blind structural interface, a paired counterfactual teacher, and bounded consolidation answer these questions in sequence. Figure 2 summarizes the division of labor.

Our contributions are:

- **Where to teach: a structural on-policy interface.** We expose grammar-valid sibling decisions at partial AST states visited by the evolving forward policy. State selection is reward-blind, and local teaching is restricted to legal actions already supported on the current search frontier.
- **What to teach: a reliability-gated paired preference.** We evaluate three siblings under four common completions and admit a KL-bounded target only when the matched samples agree on the preferred action and yield a positive empirical advantage lower bound. This creates a self-generated teacher without an external model or a held-out verification-score budget.
- **How accepted credit persists: budget-neutral bounded consolidation.** We reuse immutable accepted targets through score-indexed expiry and measured forward-gradient balancing. Replay spends no additional factor evaluations, while TB remains the global objective.

2 RELATED WORK

2.1 SYMBOLIC ALPHA DISCOVERY

Formulaic alpha mining has progressed from genetic programming and evolutionary search (Zhang et al., 2020; Cui et al., 2021; Schmidt & Lipson, 2010) to neural symbolic regression (Landajuela et al., 2022; Biggio et al., 2021; Shojaee et al., 2024; Mundhenk et al., 2021; Valipour et al., 2021; Petersen et al., 2021) and sequential decision methods. AlphaGen formulates expression construction as a Markov decision process (MDP) and searches for synergistic factor sets (Yu et al., 2023); QuantFactor REINFORCE targets lower-variance policy optimization (Zhao et al., 2025); and AlphaForge couples factor generation with dynamic combination (Shi et al., 2025). AlphaAgent regularizes LLM-driven exploration to counteract factor homogenization and decay (Tang et al., 2025) and serves as a cross-paradigm LLM-agent reference in our comparison. AlphaSAGE combines a structure-aware encoder with GFlowNet training and is our primary generative backbone (Chen et al., 2025). Relational graph convolutions, attention, and graph transformers provide complementary structure-aware encoders (Schlichtkrull et al., 2018; Vaswani et al., 2017; Yun et al., 2019; Oono & Suzuki, 2019), while relation graphs have also been used directly for stock forecasting (Xu et al., 2021a;b). Non-symbolic comparisons include tree ensembles, latent factor models, and dynamic allocation policies (Chen & Guestrin, 2016; Ke et al., 2017; Hochreiter & Schmidhuber, 1997; Chung et al., 2014; Duan et al., 2022; Wang et al., 2021; Zhang et al., 2017). These methods operate over

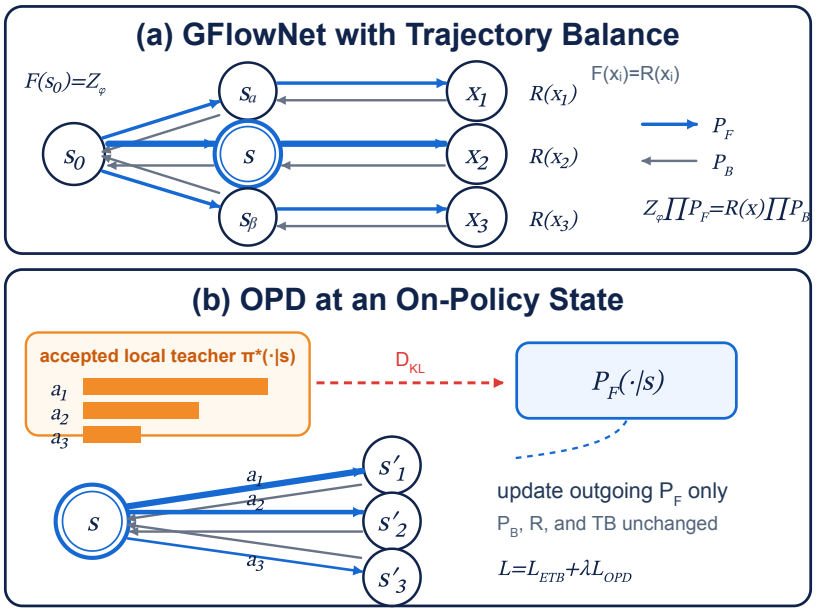

Figure 1: Trajectory-level TB and local OPD. The conceptual state flow F satisfies $F(s_0) = Z_\phi$ and $F(x) = R(x)$. TB balances P_F, P_B, and terminal reward; at visited s, ALPHAG-OPD distills an accepted local sibling teacher into outgoing P_F only. P_B and R remain unchanged.

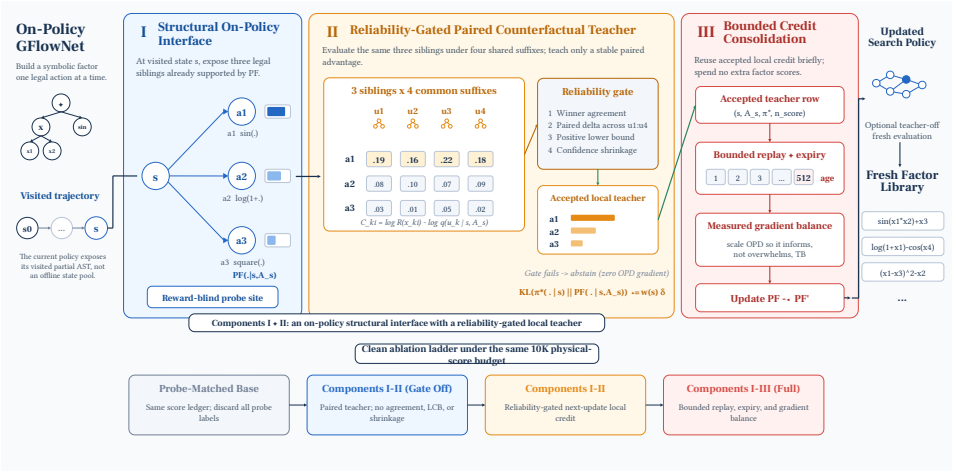

Figure 2: Overview of ALPHAG-OPD through its three questions. Component I determines where OPD may teach by exposing supported structural actions at an on-policy partial AST. Component II determines what is reliable enough to teach by comparing siblings under four common suffixes. Component III controls how strongly and for how long accepted credit persists through bounded replay and measured forward-gradient balancing. All variants use the same physical score budget.

large operator libraries such as Alpha101 (Kakushadze, 2016), where many syntactically different trajectories can encode correlated economic behavior.

2.2 GFLOWNETS AND LOCAL CREDIT

GFlowNets sample terminal objects proportionally to a nonnegative reward (Bengio et al., 2023). TB enforces a trajectory-level flow identity and provides a strong terminal-reward objective (Malkin et al., 2022). Unlike conventional expected-return policy optimization and constrained-action implementations (Schulman et al., 2017; Huang & Ontañón, 2020), this distributional objective supports

multiple high-reward modes and has been useful in diversity-sensitive generation settings such as biological sequence design (Linder et al., 2020). Subtrajectory balance supplies additional constraints on partial trajectories (Madan et al., 2023); forward-looking detailed balance uses partial reward structure to obtain denser updates (Pan et al., 2023); and guided or replay-based approaches prioritize informative trajectories (Shen et al., 2023). Backward-policy optimization can also improve the credit signal (Gritsaev et al., 2024). These objectives change where a flow constraint is applied. Our focus is complementary: we test whether an auxiliary action preference contains information beyond the terminal objective and whether its direction remains positive across repeated matched completions.

Figure 1 separates the signals: TB remains global, while OPD only reallocates mass among legal siblings at the visited state.

2.3 On-Policy Distillation for Structured Search

Policy distillation transfers a teacher distribution to a student, usually through KL or cross-entropy (Rusu et al., 2015). MiniLLM derives reverse-KL on-policy optimization for generative models (Gu et al., 2024), while GKD queries a teacher on student-generated sequences and unifies on- and off-policy mixtures with multiple divergences (Agarwal et al., 2024). Later formulations extend the interface to context-conditioned teachers (Ye et al., 2026) and organize OPD by feedback source and loss granularity (Song & Zheng, 2026). On-policy self-distillation instead constructs a privileged-context teacher from the same model (Zhao et al., 2026c). Objective analyses further separate prefix source from KL direction and revisit the advantage used by the update (Zhao et al., 2026a;b).

OPD is not automatically beneficial. Teacher–student compatibility and student-prefix reliability can reverse its effect (Li et al., 2026a; Fu et al., 2026; Armandpour et al., 2026). Selective variants restrict teacher influence by rollout, action region, token, or horizon: best-of-N and step-wise weighting (Zhang et al., 2026; Zhong et al., 2026); asymmetric and trust-region updates (Jia et al., 2026; Xing et al., 2026); trajectory filtering and prefix pruning (Li et al., 2026b; Yang et al., 2026); and horizon-aware supervision (Liu et al., 2026; Liang et al., 2026). These findings support selective, state-matched guidance rather than an unconditional auxiliary KL.

They still do not supply a teacher for symbolic search. We retain the on-policy interface—the evolving search policy supplies the states—but replace token logits with structural counterfactuals. Shared suffixes make sibling actions comparable; agreement and a paired advantage LCB suppress unstable targets; and bounded replay determines whether accepted credit can persist through subsequent TB updates.

3 Method

3.1 Problem Formulation and Design Overview

Let $\mathcal{X}$ be the set of valid factor expressions in a typed domain- specific language. Expression construction forms a directed acyclic graph with initial state s_0, partial-expression states $s \in \mathcal{S}$, legal actions $a \in \mathcal{A}(s)$, and deterministic transition $\mathsf{T}(s, a)$. A trajectory $\tau = (s_0, a_1, s_1, \ldots, a_L, s_L = x)$ terminates at expression $x \in \mathcal{X}$. The forward policy $P_F^\theta(a \mid s)$ generates the trajectory; the backward policy $P_B(s \mid s')$ defines reverse flow.

For nonnegative terminal reward $R(x)$, forward-policy parameters θ, and a separately learned partition function Z_ϕ, TB minimizes

$$\mathcal{L}_{\mathrm{TB}}(\tau) = \left[\log Z_\phi + \sum_{t=1}^{L} \log P_F^\theta(a_t \mid s_{t-1}) - \log R(x) - \sum_{t=1}^{L} \log P_B(s_{t-1} \mid s_t) \right]^2 . \tag{1}$$

The evaluator computes $r_{\mathrm{IC}}(x)$, the train-only mean cross-sectional Pearson information coefficient (IC). Its nonnegative reward is $R(x) = \max\{|r_{\mathrm{IC}}(x)|, \exp(-10)\}$. The implemented backbone is

Entropy-TB, $\mathcal{L}_{\mathrm{ETB}} = \mathbb{E}_\tau[\mathcal{L}_{\mathrm{TB}}(\tau)] - \eta_H \mathbb{E}_\tau[\sum_{t=0}^{L-1} H(P_F^\theta(\cdot \mid s_t))]$, where H is categorical entropy, $\eta_H = 0.01$, and the entropy temperature is one. Pool admission, eviction, backward-policy semantics, this reward, and the Entropy-TB objective remain unchanged by ALPHAG-OPD.

Three-stage design. At a visited partial state s, consider a supported sibling set $\mathcal{A}_s = \{a_1, \ldots, a_m\} \subseteq \mathcal{A}(s)$. A local teacher must estimate how changing only a changes reachable terminal quality while controlling the completion that follows. ALPHAG-OPD answers three questions in sequence. *Where may OPD intervene?* Component I identifies a legal, on-policy decision unit. *What is reliable enough to teach?* Component II constructs a paired local preference and abstains when its empirical evidence is weak. *How much and how long should it teach?* Component III consolidates only accepted credit under explicit lifetime and gradient bounds. This decomposition separates eligibility, evidence, and persistence. Components I–II form the one-use teacher; the full method activates all three components. Only P_F receives the auxiliary gradient; terminal reward, P_B, TB, and factor-pool rules are unchanged.

3.2 WHY AN ON-POLICY DISTILLATION CHANNEL

Existing search objectives remain useful, but their supervision has the wrong granularity for the local question considered here. In terminal-reward symbolic search, expected-return methods apply a terminal advantage to the actions sampled along one trajectory, with PPO changing how that trajectory-level signal is clipped and reused. TB uses a different, flow-matching objective, but it likewise observes the sampled trajectory and its terminal reward. A high-scoring expression therefore reinforces the actions that occurred; it does not reveal whether an unchosen legal sibling would have been better at the same partial expression. Applying balance constraints to shorter subtrajectories can densify optimization, but without an additional comparison it does not create that missing action label.

OPD provides exactly this additional interface. It queries a teacher at states visited by the evolving student and updates a distribution over several legal actions, rather than assigning one terminal scalar to the chosen path. A single KL update can consequently increase one sibling and decrease another while the global TB objective continues to preserve reward-proportional, multimodal search. The on-policy state source is important: a fixed offline teacher over states that the current miner rarely visits would reintroduce a state-distribution mismatch and spend supervision away from the active search frontier.

The interface alone is not sufficient because symbolic search has no external teacher. Component II builds a local target from matched terminal counterfactuals and abstains when the paired evidence is unreliable; Component III gives accepted sparse targets enough, but not unlimited, optimization strength. The division of labor is therefore explicit: TB determines which complete expressions deserve global probability mass, while OPD redistributes local forward probability only among compared siblings at a state the current policy actually encounters.

3.3 COMPONENT I: ON-POLICY STRUCTURAL INTERFACE

Question 1: Where may OPD intervene? The current forward policy samples a syntactically valid trajectory. Candidate states are partial ASTs from that trajectory, selected without consulting the probe's terminal scores. At state s, ALPHAG-OPD retains the sampled action and two alternatives whose current probabilities exceed 10^{-4} and whose Reverse Polish Notation (RPN) transition signatures match. In particular, operator siblings share family and arity, while feature, constant, and lag actions are compared only within their own classes. This preserves unresolved stack obligations so one completion grammar can remain legal after every sibling. A probe is formed only when such a triple exists; otherwise the method continues ordinary Entropy-TB without spending a counterfactual score. The resulting interface is

$$\mathcal{I}(s) = \left(s, \mathcal{A}_s, \{p_i\}_{i=1}^3\right), \qquad \mathcal{A}_s = \{a_1, a_2, a_3\}, \qquad p_i = \frac{P_F^\theta(a_i \mid s)}{\sum_{a_j \in \mathcal{A}_s} P_F^\theta(a_j \mid s)}. \tag{2}$$

Answer and invariant. Component I restricts teaching to states the student actually visits and actions it can legally take. The full partial state matters because the same token may satisfy different

typed AST obligations in different contexts. Component I does not inspect counterfactual scores or factor-pool contents, change admission rules, or decide which sibling is better. It determines only where a local comparison is well defined.

Design trade-off. On-policy, type-matched probes reduce coverage but ensure that each comparison is reachable and grammar-compatible. Broader probing produces more rows at the risk of mixing unreachable states or incomparable actions. Component I chooses fewer, well-defined teaching contexts.

3.4 Component II: Reliability-Gated Paired Counterfactual Teacher

Question 2: What is reliable enough to teach? The structural interface supplies candidate actions but no preferred action. Component II creates that preference from matched train-only evaluations and uses an empirical reliability rule to decide whether any OPD update is warranted.

Matched completion matrix. For each interface $\mathcal{I}(s)$, a reward-blind shared proposal samples $K = 4$ unique suffixes $U = \{u_1, \ldots, u_K\}$ that validly close every sibling. At each suffix step, the proposal is the normalized geometric mean of the three current branch policies on their common legal support. A fixed logit bias $b_{\text{exit}} = 1.5$ is added when exit is common-legal. Let $q(u_k \mid s, \mathcal{A}_s)$ be the product of these stepwise proposal probabilities and $x_{ki} = x(s, a_i, u_k)$ the resulting terminal expression. The implemented credit is

$$\mathbf{C}_{ki} = \log R(x_{ki}) - \log q(u_k \mid s, \mathcal{A}_s). \tag{3}$$

This is the log importance contribution of one sampled suffix to the terminal flow estimator; we use it as a relative sibling credit, not as an unbiased estimate of log flow. The proposal term is common to all siblings in row k, so it cancels from both the row winner and any probability redistribution whose coefficients sum to zero. A suffix invalid for any sibling is discarded for the entire row, keeping the $K \times 3$ matrix rectangular. The paired design therefore holds prefix, continuation, proposal draw, and reachable grammar fixed while changing only the structural action.

Let p_i be the sibling-normalized probability in Eq. (2), $\bar{C}_i = K^{-1} \sum_k \mathbf{C}_{ki}$, and $c_i = \bar{C}_i - \frac{1}{3} \sum_j \bar{C}_j$. A provisional teacher is anchored to the current policy:

$$\widetilde{\pi}_i(\alpha) = \frac{p_i \exp(\alpha c_i / \tau_{\text{teach}})}{\sum_j p_j \exp(\alpha c_j / \tau_{\text{teach}})}, \qquad D_{\text{KL}}(\widetilde{\pi} \| p) \le \delta, \tag{4}$$

where a one-dimensional search chooses the nonnegative α whose Kullback–Leibler (KL) divergence matches the target radius up to numerical tolerance, $\tau_{\text{teach}} = 1$, and $\delta = 0.03$. Here D_{KL} denotes KL divergence. Anchoring preserves current-policy support instead of replacing it with a standalone softmax over noisy action credits.

Paired reliability gate. The teacher is used only when the four matched comparisons agree. Winner agreement and the per-suffix paired improvement are

$$\gamma(s) = \frac{1}{K} \max_i \sum_{k=1}^{K} \mathbf{1}\left[i = \arg\max_j \mathbf{C}_{kj} \right], \tag{5}$$

$$\Delta_k(s) = \sum_i (\widetilde{\pi}_i - p_i)\, \mathbf{C}_{ki}. \tag{6}$$

$$\begin{aligned} \bar{\Delta}(s) &= \frac{1}{K} \sum_{k=1}^{K} \Delta_k(s), \\ \text{SE}(\Delta) &= \sqrt{\frac{\sum_{k=1}^{K} (\Delta_k - \bar{\Delta})^2}{K(K-1)}}, \\ \text{LCB}(s) &= \bar{\Delta}(s) - z\, \text{SE}(\Delta). \end{aligned} \tag{7}$$

Here $\mathbf{1}[\cdot]$ is the indicator function. The main method fixes $K = 4$, $z = 1$, and $\gamma_{\min} = 0.75$. This LCB is an empirical abstention rule over four paired construction samples, not a 95% confidence

Table 1: Main comparison by method family. Literature-reference rows are from AlphaSAGE (Chen et al., 2025); GFlowNet rows report mean (SD) over three random seeds. Bold marks the better completed result within each matched GFlowNet pair.

Dataset	Family	Method	IC (%)↑	ICIR↑	RankIC (%)↑	RankICIR↑	AR (%)↑	MDD (%)↓	SR↑
CSI300	*LLM Agent*	AlphaAgent	5.10	0.325	5.60	0.329	2.16	26.9	0.65
	Conventional	MLP	2.00	0.158	1.90	0.142	3.54	20.9	0.68
		LightGBM	1.10	0.124	0.60	0.064	2.61	18.5	0.53
		XGBoost	3.10	0.243	3.30	0.248	5.40	17.5	1.26
		GP	2.60	0.215	2.80	0.216	6.80	17.6	1.55
	Neural/RL	AlphaGen	5.80	0.414	5.70	0.360	4.00	22.6	0.76
		AlphaQCM	4.30	0.262	4.20	0.246	1.95	24.8	0.36
		AlphaForge	4.10	0.259	5.20	0.306	3.90	21.9	0.88
	GFlowNet	AlphaSAGE	5.81 (1.03)	0.4177 (0.0686)	7.50 (1.69)	0.5195 (0.0803)	6.146 (3.787)	15.140 (1.397)	1.413 (0.987)
		ALPHAG-OPD (ours)	**6.73** (0.74)	**0.4479** (0.0270)	**8.78** (1.13)	**0.5408** (0.0283)	**8.382** (2.222)	**14.184** (1.562)	**2.080** (0.615)
CSI500	*LLM Agent*	AlphaAgent	5.30	0.396	6.50	0.495	1.82	22.4	0.36
	Conventional	MLP	1.70	0.185	2.00	0.233	1.56	24.3	0.27
		LightGBM	2.40	0.305	2.10	0.264	4.61	17.5	0.89
		XGBoost	3.90	0.365	5.20	0.528	5.50	17.1	1.15
		GP	1.40	0.238	2.20	0.233	3.04	19.4	0.56
	Neural/RL	AlphaGen	3.20	0.270	3.10	0.230	1.15	32.4	0.19
		AlphaQCM	4.80	0.378	7.30	0.546	4.06	24.0	0.75
		AlphaForge	5.30	0.345	8.30	0.600	4.18	16.7	0.93
	GFlowNet	AlphaSAGE	1.97 (1.16)	0.1721 (0.0950)	2.72 (1.98)	0.2352 (0.1691)	−0.778 (2.020)	25.040 (3.286)	−0.139 (0.361)
		ALPHAG-OPD (ours)	**4.71** (1.15)	**0.3369** (0.0547)	**6.81** (1.81)	**0.4671** (0.1114)	**2.968** (1.308)	**18.535** (1.531)	**0.577** (0.237)
CSI1000	*LLM Agent*	AlphaAgent	7.20	0.579	8.90	0.712	5.51	20.5	1.01
	Conventional	MLP	4.80	0.384	6.90	0.621	3.22	25.7	0.47
		LightGBM	6.70	0.501	8.30	0.656	4.98	22.7	0.98
		XGBoost	6.20	0.498	8.60	0.695	4.72	23.5	0.91
		GP	5.80	0.474	7.90	0.657	4.32	24.7	0.67
	Neural/RL	AlphaGen	7.10	0.540	9.20	0.713	5.27	24.0	0.92
		AlphaQCM	6.50	0.453	10.70	0.682	7.12	20.6	1.31
		AlphaForge	7.10	0.537	9.50	0.742	6.07	21.1	1.06
	GFlowNet	AlphaSAGE	**6.04** (0.28)	0.4781 (0.0268)	**8.44** (1.15)	0.6088 (0.0404)	**4.011** (0.256)	**24.804** (3.883)	**0.669** (0.036)
		ALPHAG-OPD (ours)	5.89 (0.33)	**0.4894** (0.0347)	7.74 (0.86)	**0.6298** (0.0715)	3.155 (0.363)	25.271 (2.513)	0.534 (0.081)
S&P 500	*LLM Agent*	AlphaAgent	4.80	0.479	3.30	0.315	18.66	5.7	6.27
	Conventional	MLP	3.50	0.287	2.00	0.143	12.85	5.6	3.35
		LightGBM	2.30	0.196	1.80	0.165	11.11	5.1	4.22
		XGBoost	1.60	0.159	2.60	0.168	13.25	8.3	3.61
		GP	3.20	0.308	0.20	0.016	13.39	13.0	3.15
	Neural/RL	AlphaGen	4.40	0.396	1.30	0.127	10.31	5.5	3.96
		AlphaQCM	3.80	0.262	1.00	0.071	13.86	13.0	3.30
		AlphaForge	3.90	0.422	3.10	0.324	17.24	5.0	6.30
	GFlowNet	AlphaSAGE	**0.89** (1.26)	**0.0921** (0.1314)	1.19 (0.93)	**0.1103** (0.0950)	9.213 (1.809)	26.136 (0.367)	1.795 (0.319)
		ALPHAG-OPD (ours)	0.56 (0.73)	0.0519 (0.0677)	**1.20** (0.74)	0.1043 (0.0619)	**9.398** (0.627)	**25.530** (1.418)	**1.882** (0.107)

Table 2: CSI500 additive ablation under the same 10K physical-score budget and frozen Validation-Sign Equal-50 evaluator. Validation Pearson IC fixes each factor's sign at magnitude $1/50$ before the untouched test.

Method	I: Interface	Paired Teacher	Reliability Gate	III: Consolidation	IC (%)↑	ICIR↑	RankIC (%)↑	RankICIR↑	AR (%)↑	SR↑	MDD (%)↓
Base Model	–	–	–	–	2.175	0.182282	3.200	0.277667	-1.038	-0.209	24.571
Components I–II (Gate Off)	✓	✓	–	–	4.040	0.342000	4.850	0.395800	0.920	0.164	22.860
Components I–II (Gate On)	✓	✓	✓	–	3.730	0.327100	5.520	0.453500	1.320	0.249	21.820
ALPHAG-OPD (I–III)	✓	✓	✓	✓	**6.018**	**0.399784**	**8.822**	**0.595058**	**4.384**	**0.835**	**17.417**

guarantee. Component II abstains when credit is constant, $\gamma < \gamma_{\min}$, or the LCB is nonpositive. Otherwise it computes

$$w(s) = \min\left(1, \frac{\gamma(s)\,\mathrm{LCB}(s)}{\max\{|\bar{\Delta}| + \mathrm{SE}(\Delta), \epsilon\}}\right), \qquad \epsilon = 10^{-12}, \tag{8}$$

and rebuilds Eq. (4) with target KL $w(s)\delta$ to obtain $\pi^*(\cdot \mid s)$. The implementation recomputes Δ_k and the LCB for this shrunken target; reported accepted-row LCBs refer to that final target. The target is strongest when the preferred action is stable and its paired improvement is distinguishable from completion noise.

Answer and invariant. Component II calls a preference teachable only when the matched rows identify a stable winner and the candidate redistribution has a positive empirical LCB. The same four construction suffixes form the teacher and its reliability gate; the main method uses no held-out verification suffixes. Each scheduled probe therefore consumes exactly $3 \times 4 = 12$ physical scores, whether the gate accepts or abstains. The clean "Components I–II (Gate Off)" ablation keeps the state, siblings, suffixes, temperature, and KL radius fixed while removing winner agreement, positive-LCB abstention, and confidence shrinkage. It asks whether reliability gating adds value beyond paired evaluation alone.

Design trade-off. Independent completions cover more continuations but confound action quality with suffix difficulty. Four shared suffixes spend coverage to obtain paired contrasts and an uncertainty estimate. Abstention then exchanges update frequency for a smaller set of directionally reliable local targets.

For Algorithm **??**, let n_{score} denote the cumulative physical terminal-score count and let ρ denote the target ratio between the auxiliary and Entropy-TB forward-policy gradient norms. Component III fixes $\rho = 0.10$; Components I–II use a one-update target and do not apply gradient balancing.

3.5 COMPONENT III: BOUNDED CREDIT CONSOLIDATION

Question 3: How much, and for how long? Components I–II place each accepted teacher in a one-use buffer consumed at the next optimizer update; multiple accepted rows can therefore share one PF update. This pending buffer uses the same 512-row capacity as the replay queue. Component III instead stores the state, sibling actions, target probabilities, and creation score count n_{score} in a queue of at most 512 rows. Rows expire after 1,000 physical terminal evaluations; replay consumes no new factor score and never recomputes the stored target from a later policy.

For active rows $\mathcal{B}$, define the current sibling-normalized student as $P_F^\theta(a \mid s, \mathcal{A}_s) = P_F^\theta(a \mid s) / \sum_{a' \in \mathcal{A}_s} P_F^\theta(a' \mid s)$. The auxiliary objective is forward KL from the teacher to this distribution:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s, \mathcal{A}_s, \pi^*) \in \mathcal{B}} D_{\mathrm{KL}}\big(\pi^*(\cdot \mid s) \,\|\, P_F^\theta(\cdot \mid s, \mathcal{A}_s)\big). \tag{9}$$

The KL is computed after both distributions are renormalized on the stored sibling set. It therefore changes the relative probabilities of the compared actions without assigning labels to unrelated legal actions at the same state. Forward KL is appropriate here because the accepted local teacher, not a new sample from the student, defines which supported actions must retain probability mass. The original TB loss remains responsible for states and actions that have no active accepted row.

Sparse supervision can otherwise be numerically irrelevant compared with Entropy-TB. At each optimizer update, ALPHAG-OPD measures the forward-policy gradient norms $g_{\mathrm{ETB}} = \|\nabla_\theta \mathcal{L}_{\mathrm{ETB}}\|_2$ and $g_{\mathrm{OPD}} = \|\nabla_\theta \mathcal{L}_{\mathrm{OPD}}\|_2$, then sets

$$\lambda = \min\left(\lambda_{\max}, \rho \frac{g_{\mathrm{ETB}}}{g_{\mathrm{OPD}} + \epsilon}\right), \qquad \mathcal{L} = \mathcal{L}_{\mathrm{ETB}} + \lambda \mathcal{L}_{\mathrm{OPD}}, \tag{10}$$

where $\rho = 0.10$ is the target OPD-to-Entropy-TB gradient ratio on P_F, $\lambda_{\max} = 10^4$, and $\epsilon = 10^{-12}$. The auxiliary gradient is exactly zero when no accepted row is active. All other auxiliary guides are disabled in the registered arms, so the reference gradient in Eq. (10) is the Entropy-TB forward-policy gradient.

Replay age is measured in terminal score calls rather than optimizer steps. This choice makes freshness comparable across arms even when probing changes the number of updates between ordinary trajectories. The stored target is not silently recomputed from the later policy: doing so would change the teaching claim after acceptance. Expiration instead makes the intervention explicit— the row teaches the reliability-gated preference for a bounded period and then returns control entirely to current TB learning.

Answer and invariant. Component III controls duration through score-indexed expiry and strength through measured gradient balancing. It cannot reverse or recompute the accepted preference, request another factor evaluation, or label actions outside the stored sibling set. When no accepted row is active, its auxiliary gradient is exactly zero; when replay is active, Eq. (10) bounds its influence relative to TB. Thus TB remains the global search objective throughout training.

Design trade-off. One-use supervision can be too weak relative to TB, whereas unlimited replay becomes stale and loses its connection to the current policy. Capping row age by physical score count preserves freshness, while balancing the gradient that actually reaches P_F gives each accepted preference a controlled amount of influence. The Components I–II versus Components I–III comparison isolates this consolidation increment because both arms use the same teacher-construction rule; their realized states and accepted rows may differ as their policies diverge.

3.6 Ablation Controls and Physical Score Accounting

A sibling probe evaluates every supported action under every common suffix. For m actions and K shared suffixes, its direct evaluator cost is

$$N_{\mathrm{score}}^{\mathrm{probe}} = mK, \tag{11}$$

which is $3 \times 4 = 12$ in Component II. Every call enters the physical-score ledger even when the gate abstains; the same ledger indexes training, checkpoints, and replay expiry. Here $N_{\mathrm{score}}^{\mathrm{probe}}$ is the per-probe call count, whereas n_{score} is the cumulative ledger position stored with an accepted row.

The registered 10K ablation holds the total physical-score budget fixed. BASE MODEL spends that budget on ordinary generation. Each OPD arm spends $8{,}008$ ordinary and $1{,}992$ probe evaluations. COMPONENTS I–II (GATE OFF) introduces the structural interface and 4×3 paired score matrix without agreement, LCB abstention, or confidence shrinkage. COMPONENTS I–II keeps the same allocation and restores those gates, then consumes accepted targets once at the next update; COMPONENTS I–III changes only their persistence and strength. Thus Base Model versus Gate-Off tests

the cost-aware paired-teaching intervention, while the two later adjacent contrasts isolate reliability gating and bounded consolidation.

4 EXPERIMENTS

We evaluate reliability, factor-pool transmission, and portfolio conversion on the Chinese A-share CSI300/500/1000 universes and the U.S. S&P 500. Following Qlib (Yang et al., 2020), the target is the 20-day forward close return `Ref(close,-20)/close-1`. China uses 2010–2020/2021/2022–2024 and the U.S. uses 2010–2016/2017/2018–2020 for train/validation/test. Matched arms use multiple random seeds, 10K physical scores, and otherwise frozen rules. Table 1 uses Validation-Sign Equal-50: validation Pearson IC fixes each factor's sign at magnitude $1/50$ before one untouched test; S&P 500 portfolio metrics are long-only. Table 2 uses the same frozen Validation-Sign Equal-50 evaluator for the CSI500 additive ablation.

IC/RankIC are Pearson/Spearman correlations in percent; ICIR/RankICIR are unitless. AR, MDD, and SR denote annualized return, positive maximum drawdown magnitude, and Sharpe ratio. Parentheses report sample standard deviation across multiple random seeds, and bold marks the better completed value within each matched GFlowNet pair.

5 RESULTS AND ANALYSIS

Main comparison. Across the completed multi-seed results, ALPHAG-OPD improves all seven CSI300/500 means over AlphaSAGE. CSI300 IC/RankIC rise from $5.81/7.50\%$ to $6.73/8.78\%$, AR/Sharpe from $6.146\%/1.413$ to $8.382\%/2.080$, and MDD falls from 15.140% to 14.184%; CSI500 return moves from -0.778% to 2.968%. Figure **??** shows persistent direction transfer and stronger frozen OOS quality in every test year. On CSI1000, ALPHAG-OPD improves ICIR and RankICIR, demonstrating stronger temporal consistency in a broader universe. On S&P 500, it improves RankIC, annualized return, Sharpe, and drawdown. Together, these outcomes connect stronger ranking consistency with stronger portfolio conversion across distinct market scales. The multi-seed evidence demonstrates gains at both factor and portfolio levels.

Component ablation. Table 2 shows the staged effect of the three components on the fixed CSI500 ablation test. Components I–II with Gate Off already improve the Base Model. Enabling the reliability gate shifts the gain toward the decision-relevant metrics, further improving RankIC, return, Sharpe, and MDD. Adding Component III then improves all seven metrics over Gate On, yielding 8.822% RankIC, 4.384% return, 0.835 Sharpe, and 17.417% MDD. The ablation therefore supports the intended progression from paired local credit, through selective teaching, to bounded consolidation. This confirms cumulative component contributions.

6 CONCLUSION

ALPHAG-OPD converts terminal factor evaluations into reliable local structural guidance while preserving the global TB objective. Its structural interface, paired teacher, reliability gate, and bounded consolidation progressively strengthen decision-relevant factor quality under the same 10K evaluator budget. Across four Chinese and U.S. equity markets, the method delivers comprehensive CSI300/500 gains, stronger CSI1000 information-ratio consistency, and improved S&P 500 risk-adjusted performance. This establishes budget-efficient structural alpha discovery. By adding action-level supervision exactly where the current policy searches, ALPHAG-OPD turns sparse terminal feedback into reusable structural knowledge. The same design operates across multiple universes without changing the grammar, terminal reward, factor-pool construction, or global flow objective. Because the reliability gate abstains on weak or inconsistent paired evidence, the teacher intervenes only where the terminal signal is discriminative, limiting the risk of distilling evaluator noise. The physical-score accounting keeps every comparison budget-neutral, so the reported gains are not purchased with extra evaluator calls. Future work can attach the same interface to richer grammars and longer horizons without altering the flow constraint. Code and configuration follow the attached reproduction contract.

AI USE STATEMENT

Generative AI tools assisted with literature review, mathematical checks, method implementation, experiment design and execution, result interpretation, figure preparation, and manuscript drafting and editing. AI-assisted code and claims were checked through executable tests, exact numerical audits, source and experiment receipts, and manual review of the final manuscript. The authors take responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The experimental protocol, implementation details, physical-score accounting, and derivations are documented in the main text and appendix. Tables retain the reported per-market metrics and matched ablation settings; the accompanying source package contains the figures, tables, bibliography, and build files.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *International Conference on Machine Learning*, pp. 936–945. Pmlr, 2021.

Binqi Chen, Hongjun Ding, Ning Shen, Jinsheng Huang, Taian Guo, Luchen Liu, and Ming Zhang. Alphasage: Structure-aware alpha mining via gflownets for robust exploration. *arXiv preprint arXiv:2509.25055*, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Can Cui, Wei Wang, Meihui Zhang, Gang Chen, Zhaojing Luo, and Beng Chin Ooi. Alphaevolve: A learning framework to discover novel alphas in quantitative investment. In *Proceedings of the 2021 International conference on management of data*, pp. 2208–2216, 2021.

Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. Factorvae: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 4468–4476, 2022.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Timofei Gritsaev, Nikita Morozov, Sergey Samsonov, and Daniil Tiapkin. Optimizing backward policies in gflownets via trajectory likelihood maximization. *arXiv preprint arXiv:2410.15474*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.08543.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. *arXiv preprint arXiv:2006.14171*, 2020.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026.

Zura Kakushadze. 101 formulaic alphas. *Wilmott*, 2016(84):72–81, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Mikel Landajuela, Chak Shing Lee, Jiachen Yang, Ruben Glatt, Claudio P Santiago, Ignacio Aravena, Terrell Mundhenk, Garrett Mulcahy, and Brenden K Petersen. A unified framework for deep symbolic regression. *Advances in Neural Information Processing Systems*, 35:33985–33998, 2022.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026a.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Huangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation. *arXiv preprint arXiv:2606.02684*, 2026b.

Kun Liang, Chenming Tang, Clive Bai, Weijie Liu, Saiyong Yang, and Yunfang Wu. ADWIN: Adaptive windows for horizon-aware on-policy distillation. *arXiv preprint arXiv:2605.28396*, 2026.

Johannes Linder, Nicholas Bogard, Alexander B Rosenberg, and Georg Seelig. A generative neural network for maximizing fitness and diversity of synthetic dna and protein sequences. *Cell systems*, 11(1):49–62, 2020.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Bing Wang, Rongxiang Weng, and Jieping Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation. *arXiv preprint arXiv:2605.13643*, 2026.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning gflownets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pp. 23467–23483. PMLR, 2023.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

T Nathan Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P Santiago, Daniel M Faissol, and Brenden K Petersen. Symbolic regression via neural-guided genetic programming population seeding. *arXiv preprint arXiv:2111.00053*, 2021.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Ling Pan, Nikolay Malkin, Dinghuai Zhang, and Yoshua Bengio. Better training of gflownets with local credit and incomplete trajectories. In *International Conference on Machine Learning*, pp. 26878–26890. PMLR, 2023.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Cláudio Prata Santiago, Sookyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=m5Qsh0kBQG.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Michael D Schmidt and Hod Lipson. Age-fitness pareto optimization. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pp. 543–544, 2010.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Max W Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho, and Tommaso Biancalani. Towards understanding and improving gflownet training. In *International conference on machine learning*, pp. 30956–30975. PMLR, 2023.

Hao Shi, Weili Song, Xinting Zhang, Jiahe Shi, Cuicui Luo, Xiang Ao, Hamid Arian, and Luis Angel Seco. Alphaforge: A framework to mine and dynamically combine formulaic alpha factors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12524–12532, 2025.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsen Zheng, Keze Wang, Jinrui Chen, and Liang Lin. Alphaagent: LLM-driven alpha mining with regularized exploration to counteract alpha decay. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2813–2822, 2025. doi: 10.1145/3711896.3736838.

Mojtaba Valipour, Bowen You, Maysum Panju, and Ali Ghodsi. Symbolicgpt: A generative transformer model for symbolic regression. *arXiv preprint arXiv:2106.14131*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. Deeptrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 643–650, 2021.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Wentao Xu, Weiqing Liu, Lewen Wang, Yingce Xia, Jiang Bian, Jian Yin, and Tie-Yan Liu. Hist: A graph-based framework for stock trend forecasting via mining concept-oriented shared information. *arXiv preprint arXiv:2110.13716*, 2021a.

Wentao Xu, Weiqing Liu, Chang Xu, Jiang Bian, Jian Yin, and Tie-Yan Liu. Rest: Relational event-driven stock trend forecasting. In *Proceedings of the Web Conference 2021*, pp. 1–10, 2021b.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. Qlib: An ai-oriented quantitative investment platform. *arXiv preprint arXiv:2009.11189*, 2020.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-OPD: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 2026.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5476–5486, 2023.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

Ke Zhang, Yunjie Tian, Dongdi Zhao, Yijiang Li, Yuanye Liu, Vishal M. Patel, and Di Fu. On-policy distillation with best-of-n teacher rollout selection. *arXiv preprint arXiv:2605.09725*, 2026.

Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2141–2149, 2017.

Tianping Zhang, Yuanqi Li, Yifei Jin, and Jian Li. Autoalpha: An efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv preprint arXiv:2002.08245*, 2020.

Anhao Zhao, Haoran Xin, Yingqi Fan, Junlong Tong, Wenjie Li, and Xiaoyu Shen. Decoupling KL and trajectories: A unified perspective for SFT, DAgger, offline RL, and OPD in LLM distillation. *arXiv preprint arXiv:2605.16826*, 2026a.

Hanyang Zhao, Haoxian Chen, Han Lin, Genta Indra Winata, David Yao, and Wenpin Tang. OPD+: Rethinking the advantage design for on-policy distillation. *arXiv preprint arXiv:2606.01039*, 2026b.

Junjie Zhao, Chengxi Zhang, Min Qin, and Peng Yang. Quantfactor reinforce: mining steady formulaic alpha factors with variance-bounded reinforce. *IEEE Transactions on Signal Processing*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026c.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.