SKILL IS NOT ONE-SIZE-FITS-ALL: MODEL-AWARE SKILL ALIGNMENT SELF-EVOLVING FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents increasingly retrieve externally curated *skills*—procedural instructions retrieved at decision time—to improve performance on long-horizon interactive tasks. Existing skill libraries are typically improved in one of two ways: updating model weights—costly, and infeasible for frozen or API-served models—or sharing a single model-agnostic library across backbones with substantially different capacities. However, our controlled experiments across multiple model scales show that skill effectiveness is strongly model-dependent: a skill that benefits one backbone can harm another. Motivated by this, we take a third route—adapting the skills to each model rather than the model to the task—and propose MASA (*Model-Aware Skill Alignment*), which aligns skills with each target backbone while keeping agent weights frozen. MASA lets skills *self-evolve* from environmental reward—without human supervision–in two stages: (1) a hierarchical skill evolution pipeline in which a stronger teacher LLM iteratively rewrites general and task-specific skills using hill climbing and UCB-driven tree search, guided by model capability profiles; and (2) a lightweight model-conditioned skill rewriter, trained on these trajectories, that amortizes this search into a single forward pass. Experiments across three interactive environments and four backbones show that MASA consistently achieves the best overall performance, with gains of up to 25.8 points over the strongest baseline. The learned rewriter further generalizes to unseen tasks and environments without additional search, consistently outperforming a much larger teacher LLM at a fraction of the inference cost. Our code is publicly available at `https://anonymous.4open.science/r/MASA-BE5F/`.

1 INTRODUCTION

LLM agents increasingly solve long-horizon interactive tasks, including web navigation (Ouyang et al., 2026), embodied control (Lu et al., 2026b), and tool use (Schick et al., 2023; Jiang et al., 2026; Wang et al., 2024a; Hsiao et al., 2025). Two routes exist for improving such an agent: updating its weights via fine-tuning or reinforcement learning Zhu et al. (2026); Lu et al. (2026a), or—without touching the weights—retrieving short pieces of procedural knowledge, which we call *skills*, from an external library at each step (Wang et al., 2023; 2024b;c; Chen et al., 2024; Zhao et al., 2024; Ma et al., 2026). The former is expensive, must be repeated for every backbone, and is inapplicable when weights are frozen or served behind an API; the latter keeps the model fixed and has become a popular, lightweight alternative. Existing skill-library systems, whether hand-crafted (Zhu et al., 2023) or distilled from agent trajectories (Zhao et al., 2024; Chen et al., 2024; Xia et al., 2026; Wang et al., 2025a), typically construct a single shared library and reuse it across different LLM backbones. In practice, deployment constraints such as latency budgets, inference cost, and hardware availability mean that real-world agent systems must operate with backbones of vastly different scales rather than simply relying on the strongest available model (Yao et al., 2025; Zheng et al., 2025). This deployment heterogeneity raises a critical question for skill-library design: can a single skill formulation serve models with substantially different capacities equally well?

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.

To examine this, we experiment on ALFWorld (Shridhar et al., 2020) (full setup and analysis in §2): keeping the principles of a skill library fixed, we vary only its granularity and evaluate four Qwen3 backbones (4B–32B) (Yang et al., 2025). As Figure 1 shows, the optimal granularity varies across models; indeed, a skill that boosts one backbone can actively degrade another. A parallel experiment on the Gemma3 family (Appendix E.3) confirms that the same pattern holds across families, and that models of the same size but from different families also prefer different skill formulations. This observation suggests that the effectiveness of a skill library depends not only on what knowledge it encodes, but also on how that knowledge is expressed relative to the target model's capacity: when the expression is misaligned, retrieved skills distract rather than help. A well-designed skill library should amplify the strengths of its target backbone, unlocking capabilities that generic, model-agnostic skills cannot.

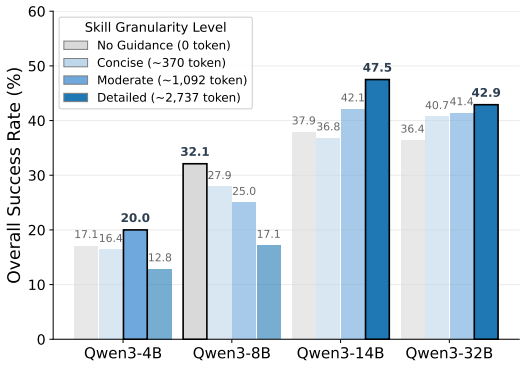

Figure 1: **Skill granularity is not one-size-fits-all.** ALFWorld success rate (%) for four Qwen3 backbones under a `No-Skill` control and three granularity levels (`Concise`, `Moderate`, `Detailed`). The optimal level differs across backbones.

We pursue this goal with **MASA**, **M**odel-**A**ware **S**kill **A**lignment, a framework that adapts a shared skill library to each target backbone without modifying agent weights. Instead of hand-designing a library per model, MASA lets skills *self-evolve* from environmental reward rather than human supervision, treating skill alignment as a hierarchical search. It first runs a *hierarchical model-conditioned skill evolution*: guided by a capability profile of the target model and by environmental reward, a stronger teacher LLM proposes rewrites while hill-climbing over general skills and UCB-driven tree search over task-specific skills select the improvements that the target backbone actually benefits from. To remove the costly teacher and search at deployment, the discovered improvements are distilled into a lightweight model-conditioned skill rewriter that amortizes the entire evolution into a single forward pass—autonomously adapting skills for new models and tasks, while outperforming the teacher at orders-of-magnitude lower cost. Our main contributions are as follows:

- We show, through controlled experiments across model scales and families, that skill effectiveness is strongly model-dependent: the same library that benefits one backbone can actively degrade another. This refutes the one-size-fits-all assumption behind shared skill libraries and motivates aligning skill formulations with each target model's capabilities.
- We propose MASA, a framework that aligns skills with each target backbone without modifying agent weights: skills *self-evolve* from environmental reward through a hierarchical, model-conditioned search, and this search is then amortized into a lightweight rewriter that transforms unaligned skills into model-appropriate ones in a single forward pass.
- We evaluate MASA across three diverse environments and four Qwen3 backbones, achieving the highest success rate with gains up to $+25.8$ points. MASA-Rewriter further generalizes to unseen tasks and environments in a single forward pass, outperforming a much larger teacher LLM at negligible cost.

2 Preliminary Study: One Skill Library Does Not Fit All

Before introducing MASA, we ask whether a single skill library serves all model scales equally. To isolate the effect of skill form from skill content, we keep the underlying *principles* fixed and vary only the *granularity* of their textual expression.

2.1 Setup

We use ALFWorld (Shridhar et al., 2020), a text-based household task suite spanning six task types, and evaluate on the validation split. We compare four Qwen3 (Yang et al., 2025) backbones

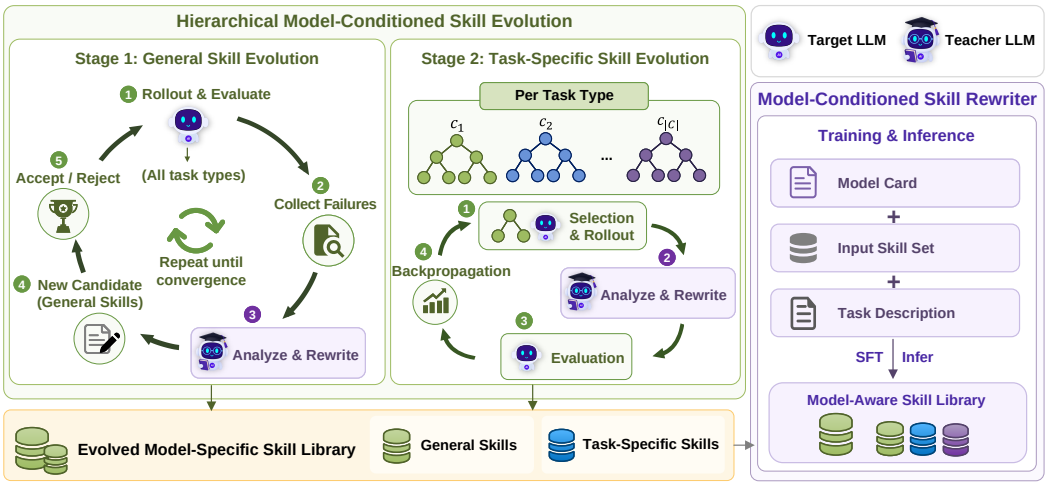

Figure 2: The overall framework of MASA.

(4B/8B/14B/32B) that differ primarily in capacity while sharing the same architecture and training regimen. We design one `No Skill` control and three skill-granularity levels that encode identical behavioral principles but differ in representational depth. Following prior work, we adopt the skill library of Xia et al. (2026) as the `Moderate` variant and construct the `Concise` and `Detailed` variants through controlled rewriting that preserves the underlying principles while adjusting granularity (see Table 8 in Appendix E.1 for side-by-side examples). All three levels use the same retrieval pipeline, ensuring that observed differences are attributable to granularity alone.

2.2 Findings

Figure 1 reports overall ALFWorld success rates.

Finding 1: The optimal skill form is model-dependent, and mismatches can hurt. No single granularity level is uniformly optimal across models. Qwen3-4B performs best with `Moderate skills` while Qwen3-14B and Qwen3-32B achieve their highest scores with `Detailed skills`. Notably, Qwen3-8B performs best under the `No Skill` condition (32.1%) and all three skill variants reduce performance. Importantly, this does not suggest that skills are inherently incompatible with Qwen3-8B. Trajectory inspection reveals that, without external guidance, Qwen3-8B often follows short and effective action chains that directly solve the task. Misaligned skills instead introduce procedural reasoning patterns that override these naturally concise action chains, causing the model to over-explore or deliberate unnecessarily. This suggests that the effectiveness of a skill depends not only on its content but also on whether its expression is compatible with the model's default problem-solving strategy.

Finding 2: The granularity–performance relationship is non-monotonic and defies simple heuristics. It is unclear how skill granularity should scale with model capability: smaller models may benefit from concise guidance due to limited context utilization capacity, yet they may also require more explicit procedural supervision because of weaker reasoning abilities. Our results show that neither direction holds consistently. Qwen3-32B underperforms Qwen3-14B by 4.6 points under `Detailed` despite being twice the size, inverting the usual scaling trend. For Qwen3-4B, performance does not monotonically improve in either direction: `Moderate` outperforms both `Concise` and `Detailed`, indicating that the optimum lies at an intermediate level that cannot be reached by simply "adding more detail" or "stripping to minimum." This complexity necessitates search-based rather than rule-based skill adaptation.

Finding 3: Performance varies sharply across task types. Per-task breakdown (Appendix E.2) reveals that within a given model–granularity pairing, success rates can vary by over 60 points across task types—a spread far exceeding the differences between granularity levels for any single task. For

example, Qwen3-14B with `Concise skills` scores 74.2 on PICK but only 13.7 on COOL. Some task types benefit from detailed skills regardless of model size, while others are insensitive or even harmed. This heterogeneity suggests that global optimization alone is insufficient—skill alignment must also operate at the task-type level to address the distinct demands of each task type.

A parallel experiment on the Gemma3 family (4B/12B/27B) reveals the same scale-dependent trend (Appendix E.3), suggesting the phenomenon generalizes across model families.

Implications. Together, the three findings impose concrete design requirements:

(i) *Model-conditioned:* the optimal skill form varies per backbone, so alignment must be explicitly conditioned on the target model's capacity (Finding 1).

(ii) *Search-based rather than heuristic:* the relationship between skill granularity and performance is non-monotonic and model-specific, ruling out simple alignment rules (Finding 2).

(iii) *Task-type-specific:* within the same backbone, different task types respond differently to the same skills, requiring per-task-type adaptation in addition to global optimization (Finding 3).

We further note that our controlled study varies only one axis of skill form (textual granularity) while holding content fixed. In practice, misalignment can also arise from differences in decision strategy, framing, or format, suggesting that a complete solution must perform open-ended, model-aware rewriting. MASA is designed to address all three requirements.

3 METHOD: MASA

We present MASA, a framework that *conditions* skill evolution on the capability profile of a target backbone, yielding skill libraries specifically adapted to each model rather than relying on a universal, model-agnostic formulation. MASA comprises two complementary components: a search-time skill evolution pipeline (Section 3.2) that evolves skills under explicit capability conditioning provided by a structured *model card*, and a deployment-time skill rewriter (Section 3.3) that learns this model-conditioned rewriting policy and adapts new skills in a single forward pass. An overview of the framework is shown in Figure 2.

3.1 PROBLEM FORMULATION AND SKILL LIBRARY

Agent setup. A frozen LLM agent F interacts with environment $\mathcal{E}$. At each step t, the agent receives observation o_t, retrieves relevant skills from a skill library $\mathcal{S}$, and produces an action a_t :

$$a_t \sim F\Big(\cdot \mid \tau_{<t},\ \hat{\mathcal{S}}_t\Big), \quad \hat{\mathcal{S}}_t = \mathrm{TopK}(\mathcal{S}, o_t, k), \tag{1}$$

where $\tau_{<t} = (o_1, a_1, \ldots, o_{t-1}, a_{t-1})$ is the interaction history and TopK retrieves the k most relevant skills by cosine similarity. The backbone F remains frozen throughout, and the sole optimization variable is the skill library $\mathcal{S}$.

Hierarchical skill library. Following Xia et al. (2026), we structure $\mathcal{S}$ into two levels: *general skills* $\mathcal{S}^G$ (cross-task strategy principles) and *task-specific skills* $\mathcal{S}^T = \{\mathcal{S}^{T_c}\}_{c\in\mathcal{C}}$, where each $\mathcal{S}^{T_c}$ contains action guidelines tailored to task type c and $\mathcal{C}$ is the set of all task types. At inference, a lightweight encoder (Qwen3-Embedding-0.6B) separately retrieves top-k_G general skills and top-k_T task-specific skills for the current observation.

Model card. The key conditioning signal in MASA is the *model card* $\mathcal{M}_F$, a structured profile of a target backbone F. Each card contains: (i) *architecture metadata* (model family, parameter count, layer/attention configuration, context window), (ii) *training provenance* (training data scale, multilingual support), and (iii) *capability profile* (strengths and weaknesses of the backbone). The construction protocol is detailed in Appendix F.

Objective. We define a per-episode adjusted reward R that balances task completion against skill-induced stalling:

$$R(F, \mathcal{S}, e) = \mathrm{SR}(F, \mathcal{S}, e) - \lambda \cdot \mathrm{NHR}(F, \mathcal{S}, e), \tag{2}$$

where e denotes a single episode, $\mathrm{SR} \in \{0, 1\}$ is task success, NHR is the *nothing-happens rate*—the fraction of steps after which the environment state remains unchanged, serving as a proxy for skill-induced stalling (e.g., the agent repeatedly issuing ineffective or invalid actions), and $\lambda \in [0, 1]$ controls the penalty strength. The overall optimization objective seeks the skill library maximizing expected adjusted reward over a set of evaluation episodes $\mathcal{D}$:

$$\mathcal{S}_F^\star = \operatorname*{argmax}_{\mathcal{S}} \ \mathbb{E}_{e\sim\mathcal{D}}[R(F, \mathcal{S}, e)]\,, \tag{3}$$

where $\mathcal{S}_F^\star$ denotes the optimal skill library adapted to backbone F.

3.2 Hierarchical Model-Conditioned Skill Evolution

The skill evolution pipeline is a teacher-driven search over skill texts. A stronger *teacher* LLM T (i) analyzes failure trajectories of F to produce a structured failure attribution and (ii) rewrites skills conditioned on the model card $\mathcal{M}_F$, steering edits toward formulations compatible with F's observed characteristics.

Two-stage optimization. The pipeline optimizes $\mathcal{S}^G$ and $\{\mathcal{S}^{T_c}\}$ in separate stages, motivated by both computational efficiency and conceptual separation. From a computational perspective, a single edit to $\mathcal{S}^G$ requires evaluation over the full task suite, whereas edits to $\mathcal{S}^{T_c}$ affect only a single task type. From a modeling perspective, the two skill levels encode fundamentally different forms of knowledge. General skills capture backbone-level behavioral guidance that is intended to transfer across tasks (e.g., "always verify your action parsed correctly"), while task-specific skills encode domain procedures tailored to particular environments (e.g., "check the fridge before the counter"). Therefore, separating the two stages simplifies credit assignment across the two skill levels while substantially reducing search cost.

Stage 1: General skills via iterative hill climbing. General skills encode high-level behavioral priors that affect agent behavior across many task types. Evaluating a candidate general skill requires running the agent across the full task suite and aggregating feedback over diverse environments, making exhaustive search prohibitively expensive. We therefore optimize $\mathcal{S}^G$ via iterative hill climbing (Russell, 2010), which provides a simple and effective strategy for progressively improving the current skill set under environment feedback.

Each iteration proceeds as follows. *Rollout*: the target model F equipped with the current best general skills is rolled out across all task types to compute the reward across episodes. *Analysis*: the teacher collects failed trajectories from these rollouts and produces a structured failure attribution focusing on general behavioral deficiencies rather than task-specific procedural gaps. *Rewrite*: given the current best skill set, the failure attribution, the model card $\mathcal{M}_F$, and the K highest-reward skill sets from all previous iterations (which help the teacher learn from the full optimization trajectory rather than only the most recent failure), the teacher outputs a revised general skill set. *Accept/Reject*: the new candidate is evaluated on the full task suite and accepted only if it achieves a higher reward than the current best. Search terminates after at most I iterations or after p consecutive iterations without improvement.

Stage 2: Task-specific skills via per-type tree search. Unlike general skills, task-specific skills encode domain procedures where multiple structurally different strategies may be effective for the same task type. This motivates a tree-structured search that can explore diverse branches rather than committing to a single refinement path. We run an independent tree search per task type c, where each node holds a candidate task-specific skill set $\mathcal{S}^{T_c}$ and each edge corresponds to a teacher rewrite.

Each iteration proceeds in four steps. *Selection*: starting from the root, UCB1 (Kocsis & Szepesvári, 2006) is applied recursively to select the most promising leaf node, balancing exploitation of high-reward nodes with exploration of less-visited ones. *Expansion*: the target model F is rolled out on type-c tasks using the selected node's skill set, and the teacher collects failed trajectories, produces a failure attribution, and outputs a revised task-specific skill set—forming a new child node. *Evaluation*: the new child's skill set is evaluated on type-c tasks to obtain its average reward. *Back-propagation*: the reward is propagated from the new node back to the root, updating visit counts and value estimates along the path. Per-type trees are independent and fully parallelizable.

Overall, the two stages run sequentially: $\mathcal{S}_F^{G\star}$ obtained in Stage 1 is held fixed throughout Stage 2, and the final output is a *model-specific* skill library $\mathcal{S}_F^\star = (\mathcal{S}_F^{G\star}, \{\mathcal{S}_F^{T_c\star}\}_{c\in\mathcal{C}})$. The detailed procedures are given in Algorithms 1 and 2, and further details of the two-stage search are provided in Appendix G.

3.3 Model-Conditioned Skill Rewriter

The skill evolution pipeline delivers strong skill libraries but requires substantial compute (hundreds to thousands of full-environment rollouts) and an environment-provided reward signal. MASA-Rewriter addresses this by learning the rewriting policy that the evolution pipeline implicitly executes, enabling cheap adaptation to new domains and tasks without further environment interaction.

Training data. Each training instance maps an input skill set to the corresponding evolved optimum:

$$(\mathcal{M}_F,\ \mathcal{S}_{F_{\text{in}}},\ d) \longrightarrow \mathcal{S}_F^\star, \tag{4}$$

where $\mathcal{M}_F$ is the model card, d is the task description, and $\mathcal{S}_{F_{\text{in}}}$ is an input skill set (either general $\mathcal{S}^G$ or task-specific $\mathcal{S}^{T_c}$). The output $\mathcal{S}_F^\star$ is always drawn from the evolution pipeline's high-scoring skill sets. To ensure the rewriter learns to improve skills regardless of their initial quality, $\mathcal{S}_{F_{\text{in}}}$ is deliberately sampled from sources spanning a wide quality range: (i) *search intermediates* at early, mid, and late stages of the evolution pipeline; (ii) *cross-model transfers*—optimal skills from a different backbone; (iii) *one-shot teacher rewrites* without iterative search; and (iv) *augmented variants* (noisy, partial, or verbose perturbations of existing skills). This diversity exposes the rewriter to the full range of input conditions it may encounter at deployment. Additional details are provided in Appendix H.

Training. We instantiate the skill rewriter with Qwen3-4B, a lightweight backbone chosen to keep deployment cost minimal while retaining sufficient capacity for structured rewriting. The model is trained via supervised fine-tuning (SFT) with cross-entropy loss:

$$\mathcal{L} = -\mathbb{E}_{\mathcal{D}_{\text{train}}}[\log\, p_\theta(\mathcal{S}_F^\star \mid \mathcal{M}_F,\, \mathcal{S}_{F_{\text{in}}},\, d)]\,. \tag{5}$$

Inference. At deployment, given the target backbone's model card $\mathcal{M}_F$, an input skill set $\mathcal{S}_{F_{\text{in}}}$, and the task description d, the skill rewriter produces an adapted skill set in a single forward pass:

$$\mathcal{S}_F^\star = f_\theta(\mathcal{M}_F,\ \mathcal{S}_{F_{\text{in}}},\ d)\,, \tag{6}$$

without requiring any environment interaction or iterative search.

Complementary roles. The skill evolution pipeline provides per-backbone upper bounds via explicit search and produces the skill rewriter's training signal. MASA-Rewriter amortizes this search into a single forward pass, enabling rapid adaptation without environment interaction. The evolution pipeline is preferred when rollout budget permits thorough optimization, whereas the skill rewriter is better suited to compute-constrained deployment scenarios.

4 Experiments

We evaluate whether model-conditioned skill evolution outperforms model-agnostic baselines across diverse environments and backbones, and whether MASA-Rewriter can generalize the learned rewriting policy to held-out task types and unseen environments without additional search.

4.1 Experimental Setup

Environments. We evaluate on three environments spanning distinct action spaces and reasoning demands. (i) ALFWorld (Shridhar et al., 2020) is a text-based embodied environment where agents complete household tasks (e.g., heating, cleaning, picking up objects) by issuing text commands to navigate rooms and interact with objects. It contains six task types with varying difficulty. (ii) WebShop (Yao et al., 2022a) simulates online shopping: agents navigate a realistic web interface, search for products, compare attributes, and make purchase decisions that satisfy natural-language user

		ALFWorld								WebShop		
Model	**Method**	**Pick**	**Look**	**Clean**	**Heat**	**Cool**	**Pick2**	**SR ↑**	**Steps ↓**	**SR ↑**	**Score ↑**	**Steps ↓**
Qwen3-4B	No Skill	20.0	15.4	18.5	18.8	16.0	12.5	17.1	44.6	23.0	42.2	9.5
	+ Base Skill	20.0	30.8	29.6	12.5	20.0	8.3	20.0	42.3	19.4	34.8	11.0
	+ DS-Adapter	**28.6**	30.8	37.0	18.8	**32.0**	12.5	27.1	40.0	19.2	25.7	12.4
	+ MASA	25.7	**53.8**	**40.7**	**37.5**	24.0	**20.8**	**31.4**	**38.4**	**26.4**	**49.1**	**8.4**
Qwen3-8B	No Skill	54.3	**46.2**	29.6	6.2	24.0	20.8	32.1	39.1	4.6	32.7	10.0
	+ Base Skill	17.1	38.5	40.7	31.2	20.0	12.5	25.0	40.5	6.0	32.6	9.3
	+ DS-Adapter	25.7	38.5	44.4	25.0	16.0	16.7	27.1	39.6	4.4	18.2	12.6
	+ MASA	**62.9**	38.5	**70.4**	**75.0**	**56.0**	**37.5**	**57.9**	**29.2**	**28.6**	**60.1**	**4.7**
Qwen3-14B	No Skill	65.7	38.5	25.9	43.8	16.0	29.2	37.9	36.7	2.8	19.9	12.7
	+ Base Skill	68.6	46.2	44.4	25.0	20.0	33.3	42.1	34.1	1.6	14.8	13.5
	+ DS-Adapter	68.6	**53.8**	40.7	18.8	40.0	29.2	44.3	34.8	2.0	12.6	13.6
	+ MASA	**85.7**	**53.8**	**81.5**	**56.2**	**44.0**	**45.8**	**64.3**	**25.7**	**29.2**	**54.4**	**8.0**
Qwen3-32B	No Skill	48.6	**46.2**	44.4	25.0	32.0	16.7	36.4	37.0	6.6	35.2	9.9
	+ Base Skill	48.6	**46.2**	40.7	50.0	44.0	20.8	41.4	35.6	7.2	24.2	12.0
	+ DS-Adapter	51.4	38.5	59.3	37.5	44.0	29.2	45.0	32.2	3.6	14.3	13.3
	+ MASA	**57.1**	**46.2**	**77.8**	**81.3**	**76.0**	**54.2**	**65.7**	**24.3**	**34.6**	**59.9**	**7.3**

Table 1: Performance on ALFWorld and WebShop. ALFWorld reports per-task and average success rate (SR %), and average interaction steps across all task types; WebShop reports average SR (%), score, and average steps. **Bold** marks the best within each backbone.

specifications. (iii) Search-augmented QA requires agents to retrieve and synthesize information from web search results. We include seven benchmarks covering both single-hop (NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023)) and multi-hop reasoning (HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), Bamboogle (Press et al., 2023)).

Backbones and baselines. Target agents are Qwen3-{4B, 8B, 14B, 32B} (Yang et al., 2025); [1] the teacher LLM is DeepSeek-V4-Pro (DeepSeek-AI, 2026). We deliberately hold the backbone family fixed (Qwen3) and vary scale from 4B to 32B, isolating the capability axis under a controlled setting. We compare against three baselines: (1) *No Skill* (the raw backbone without any skill augmentation), (2) *Base Skill* (the initial skill library from SkillRL (Xia et al., 2026), shared across all backbones without model-specific adaptation; it also serves as the initialization for MASA's evolution pipeline), and (3) *DS-Adapter* (a one-shot teacher rewrite that adapts the Base Skill library conditioned on the model card $\mathcal{M}_F$, without iterative search).

We further compare against prior model-agnostic skill-evolution systems—SkillOS (Ouyang et al., 2026), SkillOps (Pu et al., 2026), and SkillX (Wang et al., 2026). As each requires its own evolution/curation run under a distinct evaluation protocol, we defer their setup and results to Appendix B.

4.2 Skill Evolution Evaluation

Table 1 reports ALFWorld and WebShop results across all four backbones.

ALFWorld. MASA achieves the highest average success rate for every backbone: 31.4 (4B), 57.9 (8B), 64.3 (14B), and 65.7 (32B), with gains of +4.3, +25.8, +20.0, and +20.7 over the strongest baseline respectively. We highlight several observations: **(1) Per-task dominance.** Beyond the aggregate, MASA achieves the best per-task SR in most individual task types. For Qwen3-14B and 32B, MASA ranks first on *all six* task types simultaneously, indicating that the evolved skills improve overall performance without sacrificing coverage across tasks. **(2) Model-agnostic skills can hurt.** Base Skill and DS-Adapter exhibit severe performance drops on individual tasks, indicating that generic or one-shot adapted skills can introduce model-specific conflicts. In contrast,

[1] All Qwen3 backbones are used in non-thinking mode. This choice reflects typical deployment scenarios where latency and token budgets are constrained, and ensures that observed performance differences are attributable to skill design rather than reasoning-mode configuration.

Model	Method	Single-Hop QA			Multi-Hop QA				Avg.
		NQ†	TriviaQA⋆	PopQA⋆	HotpotQA†	2Wiki⋆	MuSiQue⋆	Bamboogle⋆	
Qwen3-4B	No Skill	29.4	51.0	37.2	27.7	22.8	6.4	9.3	32.9
	+ Base Skill	34.5	**57.4**	38.2	28.5	24.4	7.8	10.1	35.5
	+ DS-Adapter	33.0	56.5	**41.8**	**28.6**	23.9	9.3	12.9	36.2
	+ MASA	**35.5**	55.3	38.9	27.4	**27.0**	**9.4**	**61.3**	**36.9**
Qwen3-8B	No Skill	19.1	46.5	30.3	24.8	**30.6**	6.7	**68.1**	31.3
	+ Base Skill	34.0	**58.5**	38.8	**28.6**	25.9	6.2	10.1	36.1
	+ DS-Adapter	33.2	57.6	38.7	27.8	22.9	5.6	7.7	35.0
	+ MASA	**36.4**	56.7	**39.0**	**28.6**	25.7	**10.0**	62.5	**37.2**
Qwen3-14B	No Skill	33.8	60.2	40.6	31.7	26.8	7.6	10.5	37.6
	+ Base Skill	35.3	60.5	39.5	32.7	**30.3**	**11.4**	**15.3**	38.5
	+ DS-Adapter	33.9	60.2	39.5	31.6	28.5	9.2	12.5	37.7
	+ MASA	**35.6**	**61.8**	**40.7**	**32.8**	30.0	9.7	8.9	**39.0**
Qwen3-32B	No Skill	29.1	59.8	38.3	32.2	29.3	8.6	64.5	38.1
	+ Base Skill	33.8	61.4	39.3	33.8	26.0	11.7	**67.7**	38.7
	+ DS-Adapter	34.4	61.5	**40.6**	34.0	32.0	11.6	64.1	40.6
	+ MASA	**37.0**	**61.6**	40.0	**34.2**	**35.6**	**11.8**	66.1	**41.5**

Table 2: Search-augmented QA results (success rate %). Skill evolution is conducted on NQ and HotpotQA; † and ⋆ indicate in-domain and out-of-domain datasets, respectively. **Bold** marks the best within each backbone.

MASA avoids these regressions through iterative model-conditioned search. **(3) Scaling behavior.** For 8B and above, the backbones have sufficient capacity to leverage model-specific skills, yielding substantial improvements. The gain on 4B is comparatively modest, likely due to the backbone's inherent capability ceiling limiting how much skill guidance can help. **(4) Inference efficiency.** MASA consistently reduces average interaction steps (e.g., 8B: $39.1 \rightarrow 29.2$; 14B: $36.7 \rightarrow 25.7$), as model-tailored skills steer the agent toward correct action sequences and cut redundant exploration.

WebShop. MASA again achieves the highest success rate and score for every backbone, substantially outperforming all baselines. WebShop reveals a critical challenge for larger Qwen3 models: **(1) Larger models perform worse than 4B without adaptation.** Notably, 8B/14B/32B baselines all underperform 4B on WebShop (e.g., 14B No Skill: 2.8% vs. 4B No Skill: 23.0%). We trace this to excessive chain-of-thought generation: even in non-thinking mode, larger models tend to produce verbose reasoning preambles before each action, inflating action length and exhausting the step budget on deliberation rather than environment interaction (detailed statistics in Appendix I). Since model-agnostic skills are not designed to address this model-specific behavioral pattern, they provide limited benefit and in some cases further degrade performance. **(2) MASA addresses this challenge.** By evolving skills conditioned on each backbone's behavioral profile, MASA guides models toward effective action patterns, achieving SR of 26.4 (4B), 28.6 (8B), 29.2 (14B), and 34.6 (32B) —far surpassing all baselines while using only 7–8 steps (down to 4.7 on 8B) versus 12–13 for baselines that succeed.

Search-augmented QA. Table 2 shows that MASA achieves the highest average SR for every backbone. Skill evolution is conducted only on NQ and HotpotQA, yet the gains generalize strongly to out-of-domain benchmarks (⋆)—e.g., on 4B, MASA improves Bamboogle from 12.9 (best baseline) to 61.3. On the largest backbone (32B), MASA ranks first on 5 out of 7 datasets. These results demonstrate that the evolved skills capture transferable strategies for retrieval and reasoning, rather than overfitting to the datasets used during skill evolution.

4.3 Skill Rewriter Generalization

We evaluate whether MASA-Rewriter can adapt skills for task types not seen during its training, by holding out three ALFWorld task types (Clean, Heat, Cool) and asking MASA-Rewriter to produce task-specific skills for these types. The general skills remain unchanged. We compare against three baselines: *Base Skill*, *4B-Rewrite* (Qwen3-4B used as a rewriter without SFT—i.e., the

same architecture as MASA-Rewriter but without learned rewriting ability), and *DS-Adapter* (one-shot teacher rewrite targeting the specific held-out task). The two MASA-Rewriter variants differ in training data composition:

Cross-environment transfer. MASA-Rewriter is trained exclusively on skill evolution traces from Search and WebShop, then applied to ALFWorld without any in-environment data. Despite the substantial environment gap (different action spaces and observation formats), Cross-env MASA-Rewriter outperforms DS-Adapter on all four backbones (Figure 3), with gains of $+1.5$ (4B), $+3.0$ (8B), $+2.9$ (14B), and $+3.0$ (32B). This demonstrates that the learned rewriting policy captures model-specific adaptation patterns that transfer across environments—the rewriter can produce useful skills even without exposure to the target environment during training.

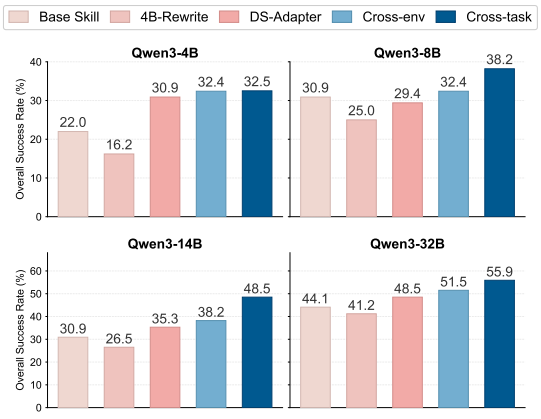

Figure 3: OOD generalization of MASA-Rewriter on held-out ALFWorld task types. Pink bars denote baselines and blue bars denote MASA-Rewriter variants.

Cross-task transfer. Building on Cross-env, this variant additionally uses evolution traces from three other ALFWorld task types (PICK, LOOK, PICK2), excluding the held-out evaluation types. Cross-task MASA-Rewriter achieves substantially larger gains: $+8.8$ (8B), $+13.2$ (14B), and $+7.4$ (32B) over DS-Adapter. The gains over Cross-env are especially large on 8B ($+5.8$) and 14B ($+10.3$), suggesting that even skills from unrelated task types provide valuable supervision for adapting to environment-specific interaction patterns and observation structures.

Notably, MASA-Rewriter (4B parameters) consistently surpasses DS-Adapter powered by DeepSeek-V4-Pro, demonstrating that a small trained rewriter can outperform a much larger teacher at a fraction of the inference cost.

Due to space constraints, additional materials—including extended related work, a comparison with prior skill-evolution systems, ablations, a search-budget efficiency analysis, supplementary validation on Gemma3, WebShop trajectory analysis, qualitative examples, and full model-card, pipeline, and training details—are reported in Appendices A–K.

5 CONCLUSION

We presented MASA, motivated by the observation that the one-size-fits-all assumption in agent skill design breaks down across model scales. MASA adapts the skills to the model: keeping the backbone frozen, it lets skills self-evolve from environmental reward through a hierarchical, teacher-guided search, and amortizes this search into a lightweight model-conditioned rewriter that adapts skills in a single forward pass. Across three environments and four Qwen3 backbones, MASA achieved the best success rate in all settings, with gains up to $+25.8$ points, and the rewriter further generalized to unseen tasks and environments at negligible deployment cost.

More broadly, MASA reframes a skill library as a model-conditioned artifact—aligned to the backbone that consumes it rather than shared uniformly across models of different capacities—and shows that this alignment can be self-evolved from experience rather than hand-engineered. With proper alignment, even compact models can exhibit behaviors traditionally associated with frontier-scale systems, enabling more accessible and resource-efficient deployment. Looking forward, we envision MASA-Rewriter as a lightweight plug-and-play middleware that automatically rewrites existing skill libraries for new backbones, requiring no environment rollouts, retraining, or manual prompt engineering. This positions skill alignment as infrastructure rather than a per-deployment engineering effort.

AI USE STATEMENT

We disclose the use of generative AI tools in the preparation of this paper, consistent with the ICLR 2027 AI Policy for Authors. **(i) Writing aid and polishing.** We used LLM-based assistants to correct grammar, improve clarity, and refine the phrasing of author-written text. No section was generated wholesale by an AI tool; all text was drafted and subsequently verified by the authors. **(ii) Retrieval and discovery.** We used LLM-based assistants to help discover and locate related work (e.g., surfacing candidate references and prior systems). All retrieved references were read, verified, and cited by the authors, and no citation was included solely on the basis of a model suggestion. We did not use generative AI for other purposes requiring disclosure (e.g., generating experimental results, claims, proofs, or analysis). We have reviewed all AI-assisted content and take full responsibility for the final text, claims, and artifacts in this work.

ETHICS STATEMENT

Data and Licensing. MASA does not introduce new data collection from human subjects; all experiments use standard public benchmarks (ALFWorld, WebShop, and open-domain QA datasets) and publicly released models accessed in accordance with their respective licenses.

Safety of Agent Empowerment. By improving the effectiveness of LLM agents through skill adaptation, MASA may also increase the capability of agents operating in interactive environments. The framework should therefore be deployed in safety-critical or high-risk settings only with additional monitoring, policy constraints, and human oversight.

Bias and Reliability of Evolved Skills. The skill evolution pipeline may inherit biases or unsafe heuristics from the trajectories and feedback used during optimization. Evolved skill libraries should therefore be inspected and validated before deployment.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The full skill-evolution pipeline, including both the hill-climbing and UCB-driven tree-search stages, is specified in Section 3.2 and detailed with complete algorithmic procedures (Algorithms 1 and 2) and hyperparameters in Appendix G. The structure and construction of the model cards used for capability conditioning are described in Appendix F, and the skill-rewriter training setup (data composition, optimizer, and augmentation) is given in Appendix H. All environments (ALFWorld, WebShop, Search-augmented QA) are standard public benchmarks, and we report the train/evaluation split protocol in Appendix G; all target backbones use greedy decoding. Anonymized source code is available at `https://anonymous.4open.science/r/MASA-BE5F/` to facilitate replication of the pipeline and evaluation.

REFERENCES

Le Chen, Erhu Feng, Yubin Xia, and Haibo Chen. Skvm: Revisiting language vm for skills across heterogenous llms and harnesses. *arXiv preprint arXiv:2604.03088*, 2026.

Minghao Chen, Yihang Li, Yanting Yang, Shiyu Yu, Binbin Lin, and Xiaofei He. Automanual: Constructing instruction manuals by llm agents via interactive environmental learning. *Advances in Neural Information Processing Systems*, 37:589–631, 2024.

Yuyan Chen, Zhihao Wen, Ge Fan, Zhengyu Chen, Wei Wu, Dayiheng Liu, Zhixu Li, Bang Liu, and Yanghua Xiao. Mapo: Boosting large language model performance with model-adaptive prompt optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3279–3304, 2023.

DeepSeek-AI. Deepseek-v4 technical report. Technical report, DeepSeek-AI, 2026. URL `https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf`.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful

prompt optimizers. In *International Conference on Learning Representations*, volume 2024, pp. 34133–34156, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Vincent Hsiao, Mark Roberts, and Leslie Smith. Procedural knowledge improves agentic llm workflows. *arXiv preprint arXiv:2511.07568*, 2025.

Yanna Jiang, Delong Li, Haiyu Deng, Baihe Ma, Xu Wang, Qin Wang, and Guangsheng Yu. Sok: Agentic skills–beyond tool use in llm agents. *arXiv preprint arXiv:2602.20867*, 2026.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4, 2025.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*, 2026a.

Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*, 2026b.

Ziyu Ma, Shidong Yang, Yuxiang Ji, Xucong Wang, Yong Wang, Yiming Hu, Tongwen Huang, and Xiangxiang Chu. Skillclaw: Let skills evolve collectively with agentic evolver. *arXiv preprint arXiv:2604.08377*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, et al. Skillos: Learning skill curation for self-evolving agents. *arXiv preprint arXiv:2605.06614*, 2026.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Hongji Pu, Xinyuan Song, and Liang Zhao. Skillops: Managing llm agent skill libraries as self-maintaining software ecosystems. *arXiv preprint arXiv:2605.13716*, 2026.

Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

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. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Chenxi Wang, Zhuoyun Yu, Xin Xie, Wuguannan Yao, Runnan Fang, Shuofei Qiao, Kexin Cao, Guozhou Zheng, Xiang Qi, Peng Zhang, et al. Skillx: Automatically constructing skill knowledge bases for agents. *arXiv preprint arXiv:2604.04804*, 2026.

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.

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. Reinforcement learning for self-improving agent with skill library. *arXiv preprint arXiv:2512.17102*, 2025a.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024a.

Yaxuan Wang, Quan Liu, Zhenting Wang, Zichao Li, Wei Wei, Yang Liu, and Yujia Bao. Promptbridge: Cross-model prompt transfer for large language models. *arXiv preprint arXiv:2512.01420*, 2025b.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1894–1907, 2024b.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024c.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Yiheng Xu, Dunjie Lu, Zhennan Shen, Junli Wang, Zekun Wang, Yuchen Mao, Caiming Xiong, and Tao Yu. Agenttrek: Agent trajectory synthesis via guiding replay with web tutorials. In *International Conference on Learning Representations*, volume 2025, pp. 79822–79843, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, volume 2024, pp. 12028–12068, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Zhiwei Yao, Yang Xu, Hongli Xu, Yunming Liao, and Zuan Xie. Efficient deployment of large language models on resource-constrained devices. *arXiv preprint arXiv:2501.02438*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57 (8):1–35, 2025.

Jiapeng Zhu, Jianxiang Yu, Yibo Zhao, Chengcheng Han, Qi Gu, Xunliang Cai, Xiang Li, and Weining Qian. Skill0. 5: Joint skill internalization and utilization for out-of-distribution generalization in agentic reinforcement learning. *arXiv preprint arXiv:2605.28424*, 2026.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

A Related Work

LLM agents and skill libraries. Equipping LLM agents with reusable procedural knowledge is a scalable approach to improve agent performance without modifying model weights. Early efforts such as ReAct (Yao et al., 2022b) and Reflexion (Shinn et al., 2023) leverage textual feedback as in-context skill; Voyager (Wang et al., 2023) maintains a growing skill library for Minecraft; JARVIS-1 (Wang et al., 2024b) and Ghost-in-the-Minecraft (Zhu et al., 2023) cache successful behaviors for replay at inference; AgentTrek (Xu et al., 2025) bootstraps web agents with synthesized trajectories; AutoManual (Chen et al., 2024) induces an *operating manual* from interaction traces; and ExpeL (Zhao et al., 2024) distills cross-trial experiences into a reusable insights library. More recent systems further elevate skills into first-class agent components: SkillRL (Xia et al., 2026) distills trajectories into a hierarchical SkillBank and recursively evolves skills with the agent policy; and SkillOS (Ouyang et al., 2026) learns a long-horizon curator that inserts, updates, and deletes skills in an external SkillRepo. Other systems such as SkillOps (Pu et al., 2026) and SkillX (Wang et al., 2026) similarly construct or curate reusable skill libraries, but all of them share the same skill formulation across backbones rather than conditioning it on the target model. Notably, SkVM (Chen et al., 2026) also identifies the model-skill mismatch problem—reporting that 87% of tasks have at

Method	ALFWorld SR ↑		WebShop SR ↑
	Qwen3-4B	Qwen3-8B	Qwen3-8B
SkillOps (Pu et al., 2026)	23.6	36.4	–
SkillX (Wang et al., 2026)	28.6	45.0	–
SkillOS-Gemini (Ouyang et al., 2026)	–	50.7	13.2
MASA	**31.4**	**57.9**	**28.6**

Table 3: Comparison with prior skill-evolution systems. All systems share the same initial skill library and evolve/curate skills without updating executor parameters. Dashes (–) mark configurations outside each system's evaluated setting. **Bold** marks the best per column.

least one LLM that gains no benefit from the same skill—and addresses it by compiling skills into optimized runtime formats (e.g., code solidification, parallelization) to reduce latency and token cost. MASA shares the same motivation but pursues a complementary direction: rather than compiling skills for execution efficiency, we *rewrite the natural-language expression* of skills to match each backbone's comprehension and reasoning style, directly improving task success rate.

Self-improving and experiential agents. A growing body of work seeks agents that improve from their own experience rather than from human supervision, through self-reflection (Madaan et al., 2023; Shinn et al., 2023), self-generated training data (Zelikman et al., 2022), and distilling repeated inference-time search into the model (Singh et al., 2023). MASA contributes to this trend but places the self-improvement loop at the *context level* rather than the parameter level: skills self-evolve from environmental reward under a *frozen* backbone, and the resulting adaptation is amortized into a lightweight rewriter that adapts a new agent's skills in a single forward pass. This keeps the loop cheap, applicable to frozen or API-served models, and transferable across backbones—in contrast to parameter-level self-improvement, which must be re-run for each model and is unavailable when weights are inaccessible.

Model-aware adaptation and prompt optimization. LLM behavior is highly sensitive to instruction phrasing even under semantically equivalent prompts (Sclar et al., 2024), motivating methods that tailor prompts to specific backbones. Teacher-driven search methods such as OPRO (Yang et al., 2024) and EvoPrompt (Guo et al., 2024) iteratively refine a single instruction for a given task, yet treat the target model as fixed context—the same output applies regardless of backbone. MAPO (Chen et al., 2023) and PromptBridge (Wang et al., 2025b) further account for model identity by optimizing or transferring individual task instructions across backbones, yet they operate on single monolithic prompts in non-agent settings rather than on retrievable multi-entry skill libraries used at agent decision time. MASA differs in two key respects: (i) the optimization target is a dynamically retrieved *skill library* rather than a monolithic prompt, and (ii) the evolutionary search is jointly steered by a structured model capability profile and environment reward signals, explicitly conditioning skill expression on target-model characteristics. Furthermore, we train a lightweight skill rewriter that amortizes the expensive search process into a single forward pass—conceptually related to distilling costly inference-time computation into efficient learned models (Singh et al., 2023)—enabling skill adaptation to new backbones without repeated search.

B Comparison with Prior Skill-Evolution Systems

Setup. These systems construct or curate skill libraries in a model-agnostic manner. All share the same initial skill library as MASA and adapt skills without updating executor parameters. Following its original setting, SkillOS uses Gemini-2.5-Pro as an external curator over a frozen Qwen3-8B executor (reported as SkillOS-Gemini); SkillOps and SkillX are re-implemented from their public designs and evaluated on ALFWorld. Since each system requires its own evolution/curation run under a distinct evaluation protocol, we report these comparisons on the applicable Qwen3-4B/8B settings rather than across all backbones.

Table 3 compares MASA against representative model-agnostic skill-evolution systems on Qwen3-4B and 8B. MASA achieves the best success rate in every evaluated configuration. On ALFWorld-

Variant	4B	8B	14B	32B
ALFWorld				
Full pipeline	**31.4**	**57.9**	**64.3**	**65.7**
w/o Task-specific	25.0 (−6.4)	32.9 (−25.0)	63.6 (−0.7)	50.0 (−15.7)
w/o General	25.0 (−6.4)	50.0 (−7.9)	47.9 (−16.4)	64.3 (−1.4)
WebShop				
Full pipeline	**26.4**	**28.6**	**29.2**	**34.6**
w/o Task-specific	22.4 (−4.0)	25.6 (−3.0)	24.2 (−5.0)	31.8 (−2.8)
w/o General	23.4 (−3.0)	7.2 (−21.4)	10.2 (−19.0)	9.6 (−25.0)

Table 4: Ablation of the two-stage search-based evolution pipeline (SR %). Each stage's evolved skills are replaced with one-shot teacher rewrites in turn. Reference row (full pipeline) highlighted; red values report the SR drop when a stage is removed.

8B, the prior systems form a clear ordering (SkillOps 36.4, SkillX 45.0, SkillOS-Gemini 50.7), yet MASA still improves over the strongest of them (SkillOS-Gemini) by +7.2, despite SkillOS relying on a stronger Gemini curator. The gap is even larger on WebShop-8B, where SkillOS-Gemini reaches only 13.2 versus MASA's 28.6. These results indicate that the gains of MASA do not stem merely from iterative skill construction—which prior systems also perform—but from conditioning skill evolution on the target backbone's capability profile.

C Ablations

MASA rests on two core design choices: a *two-stage search structure* for skill evolution, and *model-aware conditioning* via the model card. We ablate each in turn, organized around three questions: (i) is the two-stage structure necessary (Table 4)? (ii) does model-aware conditioning drive the gains, rather than merely the extra search compute (Table 5)? and (iii) does the model card remain effective once search is amortized into the MASA-Rewriter (Table 6)? Questions (ii) and (iii) both probe the model card—first inside the search-based system, then in the lightweight rewriter that deploys without search.

Does the two-stage structure matter? (Table 4). We ablate the two-stage search structure by replacing each stage's evolved skills with one-shot teacher (DeepSeek-V4-Pro) rewrites, isolating the contribution of each search stage. *w/o Task-specific* retains MASA-evolved general skills but substitutes teacher-written task-specific skills (i.e., Stage 1 only), while *w/o General* retains MASA-evolved task-specific skills but uses teacher-written general skills (i.e., Stage 2 only). Both stages contribute to the full pipeline, but their relative importance is environment- and scale-dependent. On ALFWorld, removing task-specific search causes the largest drops for Qwen3-8B (−25.0) and Qwen3-32B (−15.7), indicating that per-task-type procedural guidance is critical for these backbones. Conversely, removing general search most severely affects Qwen3-14B (−16.4), suggesting that high-level behavioral priors are essential when the model has sufficient capacity to follow them but still benefits from strategic framing. On WebShop, removing general skills is catastrophic for 8B/14B/32B (SR collapses to roughly 7–10), while removing task-specific skills has a comparatively modest effect. This asymmetry reflects the nature of each environment—WebShop demands consistent high-level decision strategies that general skills encode, whereas ALFWorld requires fine-grained procedural sequences that task-specific skills address.

Does model-aware conditioning drive the gains? (Table 5). Because MASA both consults a model card *and* runs iterative search, a natural concern is that its gains come purely from the extra search compute rather than from model-aware conditioning. To rule this out, we hold the search budget *fixed*: all variants share an *identical* two-stage optimization and search configuration, and differ *only* in the model-card information supplied to the teacher during skill evolution, on ALFWorld with Qwen3-4B. Removing the model card (*w/o Model Card*) drops SR from 31.4 to 27.1, and replacing the target card with an unrelated one (*Random Model Card*) drops it further to 23.8—below even the no-card variant. Since all three settings expend the same search budget, this rules

Setting	Avg. SR	Δ vs. MASA
Base Skill	20.0	(−11.4)
Random Model Card	23.8	(−7.6)
w/o Model Card	27.1	(−4.3)
MASA (target card)	**31.4**	–

Table 5: Effect of model-card conditioning under a *matched* search budget on ALFWorld (Qwen3-4B). All variants use the identical two-stage optimization and search configuration, differing only in the model-card information provided during skill evolution: *w/o Model Card* removes the card, and *Random Model Card* replaces the target card with an unrelated one. Red values report the SR drop from MASA with the correct target card.

Variant	4B	8B	14B	32B
Cross-env				
w/ Model Card	**32.4**	**32.4**	**38.2**	**51.5**
w/o Model Card	14.7 (−17.7)	23.5 (−8.9)	33.8 (−4.4)	39.7 (−11.8)
Cross-task				
w/ Model Card	**32.5**	**38.2**	**48.5**	**55.9**
w/o Model Card	8.8 (−23.7)	30.9 (−7.3)	20.6 (−27.9)	42.5 (−13.4)

Table 6: Ablation of model-card conditioning in the MASA-Rewriter (average SR % on three held-out ALFWorld tasks: Clean/Heat/Cool). Reference row (with-card) highlighted; red values report the SR drop when the card is removed, for both training-data variants.

out that MASA's gains stem merely from additional optimization iterations: accurate model-aware conditioning contributes beyond search alone, and an *incorrect* card is actively harmful rather than a neutral source of extra text.

Does the model card transfer to the Rewriter? (Table 6). Having shown the model card matters in the search-based system, we test whether it remains effective once search is amortized into the MASA-Rewriter. We compare performance with and without the target model's capability card, using both training data variants: *Cross-env* (trained on Search + WebShop) and *Cross-task* (trained on Search + WebShop + ALFWorld Pick/Look/Pick2). All results are average SR on three held-out ALFWorld tasks (Clean/Heat/Cool). Removing the model card consistently degrades performance. For Cross-task, the gap is especially pronounced on Qwen3-14B (27.9) and Qwen3-4B (23.7), while 8B and 32B show smaller but still consistent drops, confirming that the card provides a critical conditioning signal. For Cross-env, model card removal also causes substantial drops. Overall, these results suggest that the model card provides useful backbone-specific conditioning signals that help the rewriter generate more appropriate skill adaptations.

D Search Budget and Computational Efficiency

We analyze the trade-off between search budget and performance on ALFWorld using Qwen3-32B, varying only the number of teacher-guided optimization iterations (and the corresponding environment rollouts) while keeping all other settings fixed (Table 7).

We highlight several observations:

(1) Diminishing returns and saturation. Performance improves rapidly with additional search—from 49.3 at a single iteration to 65.7 at eight—but saturates thereafter: increasing the budget from 8 to 10 iterations yields no further improvement. In practice, MASA exploits this by applying an early-stopping strategy (patience p=3; Appendix G), avoiding unnecessary optimization once performance converges.

Setting	Search Iter.	Rollouts	Avg. SR
Base Skill	0	0	41.4
MASA	1	100	49.3
MASA	2	200	54.8
MASA	4	400	61.7
MASA	6	600	64.3
MASA	8	800	65.7
MASA	10	1000	65.7
MASA-Rewriter	–	0	55.9

Table 7: Search budget vs. performance on ALFWorld (Qwen3-32B). *Search Iter.* is the number of teacher-guided optimization iterations and *Rollouts* the corresponding environment interactions; all other settings are held fixed. Performance saturates at 8 iterations (highlighted). MASA-Rewriter attains strong performance with a single forward pass and *zero* search rollouts at deployment.

(2) Offline cost, lightweight deployment. The search budget is incurred only once, offline, during skill evolution. At deployment, MASA-Rewriter requires a single forward pass and *zero* search rollouts, teacher calls, or environment interaction, yet still improves substantially over the Base Skill (55.9 vs. 41.4). The online inference cost is therefore comparable to the original skill-based agent.

E Preliminary Study: Supplementary Details

E.1 Skill Variant Comparison

Table 8 shows concrete examples of the three non-empty ALFWorld skill variants used in the preliminary study. All variants keep the same skill IDs and task coverage; what changes is how much procedural text is exposed to the agent. We use bold text to highlight trigger conditions, executable steps, and failure-prevention cues added by the more detailed variants.

E.2 Per-Task Breakdown: Qwen3

Table 9 expands the overall numbers visualized in Figure 1 into per-task success rates for each (model, skill) cell. The breakdown is computed on the same ALFWorld validation set. Note the large within-condition swings across task types (e.g., Qwen3-14B Concise: 74.2 on Pick vs. 13.7 on Cool; Qwen3-4B Detailed: 1.6 on Pick vs. 46.7 on Look), which are substantially larger than cross-condition differences and motivate the task-specific tree-search stage of the evolution pipeline.

E.3 Supplementary Validation: Gemma3

To verify that the scale-dependent granularity pattern is not unique to Qwen3, we repeat the fixed-granularity sweep on Gemma3 backbones (4B/12B/27B) (Kamath et al., 2025). It supports the motivating conclusion: the best skill form is model-dependent rather than universally transferable. Gemma3-4B and Gemma3-12B are strongest with Concise skills, while Gemma3-27B reaches its best success rate with Detailed skills. Figure 4 shows the overall results and Table 10 gives the per-task breakdown.

Comparing models of the same parameter count across families further isolates the effect of architecture and training from that of scale alone. Gemma3-4B achieves its best performance with `Concise` skills, whereas Qwen3-4B peaks under `Moderate` skills (Figure 1 vs. Figure 4). Despite identical parameter budgets, the two models respond to skill granularity in qualitatively different ways—confirming that the optimal skill form is determined by a model's overall characteristics (architecture, training data, alignment procedure) rather than parameter count alone. This observation reinforces the necessity of conditioning skill adaptation on a rich model profile rather than relying on scale as a proxy.

(a) General Skill: *Systematic Exploration*

Granularity	Skill Text
CONCISE	**Principle:** Search all surfaces and containers once before revisiting.
MODERATE	**Principle:** Search every plausible surface or container exactly once before revisiting; prioritize unopened or unseen locations to cover the whole room methodically. **When to apply:** Anytime the goal object count is not yet met and unexplored locations remain.
DETAILED	**Principle:** When searching for an object, follow these steps exactly: Step 1: Make a mental list of ALL possible locations in the room (countertop 1, countertop 2, shelf 1, drawer 1, cabinet 1, fridge 1, etc.). Step 2: Visit each location one by one using 'go to [location] [number]' (e.g., 'go to countertop 1'). Step 3: For closed containers (drawer, cabinet, fridge, safe, microwave), always use 'open [container] [number]' to check inside. Step 4: Read the observation carefully — look for the exact name of the target object. Step 5: Mark each location as 'checked' mentally and do NOT go back to it. Step 6: Only after checking ALL locations in the room, consider that the object may not be present. EXAMPLE: Looking for a mug → 'go to countertop 1' → check → 'go to countertop 2' → check → 'go to shelf 1' → check → 'open cabinet 1' → check inside → continue until found. **When to apply:** At the VERY START of every task that involves finding or locating any object. This is always your first action — never skip the systematic search.

(b) Task-Specific Skill: *Open Then Heat*

Granularity	Skill Text
CONCISE	**Principle:** Open microwave, put object in, heat it.
MODERATE	**Principle:** Upon reaching the microwave with the target in hand, always open the door, place the object inside, and execute the heat action before leaving. **When to apply:** Immediately after navigating to the microwave with the target object held.
DETAILED	**Principle:** The microwave heating sequence must be executed in this EXACT order: (1) 'go to microwave 1' — navigate to the microwave. (2) 'open microwave 1' — the door must be open to put things in. (3) 'put [object] in/on microwave 1' — place the object inside. (4) 'heat [object] with microwave 1' — execute the heating action. (5) 'open microwave 1' — open the door again to retrieve (if needed). (6) 'take [object] from microwave 1' — take the now-heated object. COMMON MISTAKE: Trying to 'heat' without first putting the object in the microwave → fails. ANOTHER MISTAKE: Forgetting to open the microwave before putting the object in → fails. **When to apply:** When you are holding the target object and ready to heat it. Execute this exact 6-step sequence.

Table 8: Examples from the ALFWorld skill-library variants used in the preliminary study. We show the same cross-task skill and task-specific skill under three granularity levels: CONCISE, MODERATE, and DETAILED. The empty-bank control (NO SKILL) is omitted for brevity.

F MODEL CARD CONSTRUCTION

Each model card is constructed from a fixed rubric combining public documentation and automated analysis:

Model	Skill	Pick	Clean	Heat	Cool	Pick2	Look	Overall
Qwen3-4B	No Skill	20.0	18.5	18.8	16.0	12.5	15.4	17.1
	Concise	3.2	17.9	37.5	9.1	10.7	40.0	16.4
	Moderate	20.0	29.6	12.5	20.0	8.3	30.8	**20.0**
	Detailed	1.6	16.1	9.3	6.8	10.7	46.7	12.8
Qwen3-8B	No Skill	54.3	29.6	6.2	24.0	20.8	46.2	**32.1**
	Concise	32.2	33.9	25.0	11.3	25.0	40.0	27.9
	Moderate	17.1	40.7	31.2	20.0	12.5	38.5	25.0
	Detailed	6.5	17.9	21.9	13.6	10.7	50.0	17.1
Qwen3-14B	No Skill	65.7	25.9	43.8	16.0	29.2	38.5	37.9
	Concise	74.2	26.8	18.8	13.7	30.3	43.4	36.8
	Moderate	68.6	44.4	25.0	20.0	33.3	46.2	42.1
	Detailed	64.5	46.4	18.8	34.1	46.4	66.7	**47.5**
Qwen3-32B	No Skill	48.6	44.4	25.0	32.0	16.7	46.2	36.4
	Concise	56.5	41.0	37.5	18.2	33.9	56.6	40.7
	Moderate	48.6	40.7	50.0	44.0	20.8	46.2	41.4
	Detailed	54.8	46.4	31.2	36.4	28.6	60.0	**42.9**

Table 9: Per-task ALFWorld success rate (%) for the four Qwen3 backbones under each skill granularity condition.

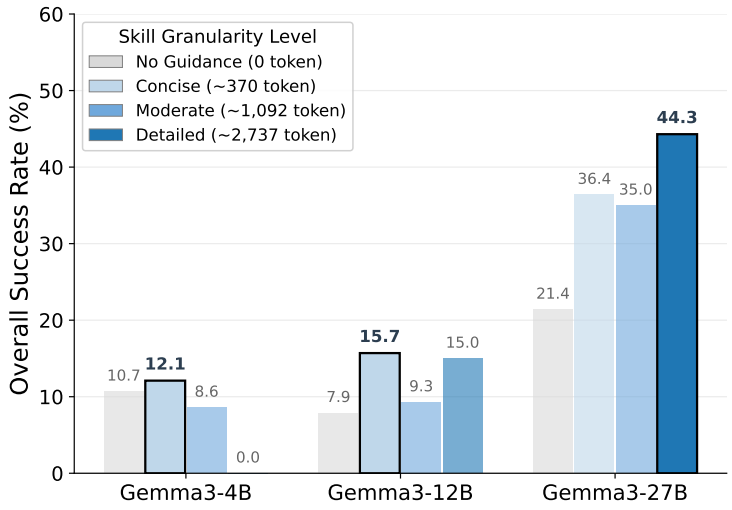

Figure 4: The supplementary validation of Gemma family.

1. *Architecture metadata.* Model family, variant name, parameter count, architecture type, layer/attention configuration, context window, and vocabulary size—sourced directly from the published model card or config files.
2. *Training provenance.* Whether the checkpoint is base or instruction-tuned, the alignment pipeline (e.g., SFT + DPO + GRPO), training data scale, and multilingual support—sourced from official documentation.
3. *Capability profile.* Strengths are extracted from the model's official release notes (e.g., "strong at math and code generation"). Weaknesses are generated by the teacher LLM summarizing behavioral patterns observed during a small set of preliminary rollouts (Section 2), produced automatically without human annotation.

Note that the card does not include any downstream evaluation results (e.g., ALFWorld success rates) or oracle style labels (e.g., *prefers_concise*). The preliminary rollouts used for weakness summarization are disjoint from the evaluation set.

Below is the card for Qwen3-4B; cards for the remaining backbones follow the same template.

Model	Skill	Pick	Clean	Heat	Cool	Pick2	Look	Overall
Gemma3-4B	No Skill	22.6	0.0	0.0	0.0	7.1	40.0	10.7
	Concise	22.6	7.1	6.2	0.0	7.1	33.3	**12.1**
	Moderate	3.2	7.1	0.0	4.5	7.1	40.0	8.6
	Detailed	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gemma3-12B	No Skill	16.1	3.6	0.0	0.0	3.6	26.7	7.9
	Concise	32.3	10.7	12.5	0.0	7.1	33.3	**15.7**
	Moderate	9.7	10.7	0.0	0.0	7.1	33.3	9.3
	Detailed	38.7	21.4	6.2	0.0	3.6	6.7	15.0
Gemma3-27B	No Skill	22.6	14.3	18.8	13.6	25.0	40.0	21.4
	Concise	61.3	21.4	31.2	9.1	42.9	46.7	36.4
	Moderate	51.6	32.1	18.8	4.5	42.9	53.3	35.0
	Detailed	41.9	60.7	31.2	31.8	35.7	66.7	**44.3**

Table 10: Per-task ALFWorld success rate (%) for the three Gemma3 backbones under each skill granularity condition.

```
# Model Card: Qwen3-4B
# Source: https://huggingface.co/Qwen/Qwen3-4B

# === Architecture Metadata ===
family: "Qwen3"
variant: "4B"
parameter_count: "4B"
architecture: "dense-transformer"
num_layers: 36
hidden_size: 2560
num_attention_heads: 32
num_kv_heads: 8
context_window: 32768
vocab_size: 151936

# === Training Provenance ===
base_or_instruct: "instruct"
alignment_method: "SFT + DPO + GRPO"
training_data_size: "36T tokens"
multilingual: true

# === Official Capabilities (from release notes) ===
strengths: "math, code generation, instruction
  following, multilingual, tool use, thinking
  mode support"

# === Observed Weaknesses (teacher-summarized) ===
weaknesses: "limited reasoning depth due to small
  parameter count, may struggle with complex
  multi-step planning"
```

G Skill Evolution Pipeline Details

This appendix provides the full algorithmic procedures (Algorithms 1 and 2) and hyperparameters for the two-stage skill evolution pipeline described in Section 3.2. All evolution experiments use the original training split of each environment for exploration; the evaluation results reported in the main paper are obtained on the held-out test split.

Algorithm 1 Stage 1: General Skill Search (Hill Climbing)

Require: Target model F, model card $\mathcal{M}_F$, teacher T, eval set $\mathcal{D}$ (sampled from training episodes), initial general skills $\mathcal{S}_F^{G_0}$, max iterations I, patience p, history size K
Ensure: Optimized general skills $\mathcal{S}_F^{G\star}$

$\mathcal{S}_F^{G\star} \leftarrow \mathcal{S}_F^{G_0}$ {current best general skill set}
$R^\star \leftarrow \mathrm{Eval}(F, \mathcal{S}_F^{G\star}, \mathcal{D})$ {its average reward across all task types}
$\mathcal{H} \leftarrow \{(\mathcal{S}_F^{G_0}, R^\star)\}$ {search history: (skill set, reward) pairs}
for $i = 1$ **to** I **do**
  // Rollout & Analysis
  $\mathcal{F}_i \leftarrow \mathrm{CollectFailures}(F, \mathcal{S}_F^{G\star}, \mathcal{D})$
  $\mathrm{attr}_i \leftarrow T.\mathrm{Analyze}(\mathcal{F}_i)$ {structured failure attribution}
  // Rewrite
  $\mathcal{S}_F^{G_i} \leftarrow T.\mathrm{Rewrite}(\mathcal{S}_F^{G\star}, \mathrm{attr}_i, \mathrm{TopK}(\mathcal{H}, K), \mathcal{M}_F)$
  // Accept / Reject
  $R_i \leftarrow \mathrm{Eval}(F, \mathcal{S}_F^{G_i}, \mathcal{D})$
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\mathcal{S}_F^{G_i}, R_i)\}$
  if $R_i > R^\star$ **then**
    $\mathcal{S}_F^{G\star} \leftarrow \mathcal{S}_F^{G_i}$; $R^\star \leftarrow R_i$ {accept}
  end if
  if no improvement for p consecutive iterations **then**
    break
  end if
end for
return $\mathcal{S}_F^{G\star}$

Algorithm 2 Stage 2: Task-Specific Skill Search (Per-Type Tree Search)

Require: Target model F, model card $\mathcal{M}_F$, teacher T, fixed general skills $\mathcal{S}_F^{G\star}$, initial task-specific skills $\{\mathcal{S}_F^{T_{c_0}}\}_{c\in\mathcal{C}}$, iterations J
Ensure: Optimized task-specific skills $\{\mathcal{S}_F^{T_c\star}\}_{c\in\mathcal{C}}$

for each task type $c \in \mathcal{C}$ **in parallel do**
  Initialize tree root with $\mathcal{S}_F^{T_{c_0}}$
  for $j = 1$ **to** J **do**
    // Selection
    $n \leftarrow \mathrm{UCB1Select}(\mathrm{root})$ {select leaf via Eq. 7}
    // Expansion
    $\mathcal{F} \leftarrow \mathrm{CollectFailures}(F, \mathcal{S}_F^{G\star}, \mathcal{S}_{F,n}^{T_c}, c)$
    $\mathrm{attr} \leftarrow T.\mathrm{Analyze}(\mathcal{F})$ {failure attribution}
    $\mathcal{S}_F'^{T_c} \leftarrow T.\mathrm{Rewrite}(\mathcal{S}_{F,n}^{T_c}, \mathrm{attr}, \mathcal{M}_F)$
    Add $\mathcal{S}_F'^{T_c}$ as child of node n
    // Evaluation
    $R' \leftarrow \mathrm{Eval}(F, \mathcal{S}_F^{G\star}, \mathcal{S}_F'^{T_c}, c)$
    // Backpropagation
    Update visit counts and value estimates from new node to root
  end for
  $\mathcal{S}_F^{T_c\star} \leftarrow$ skill set of the highest-value node
end for
return $\{\mathcal{S}_F^{T_c\star}\}_{c\in\mathcal{C}}$

G.1 Stage 1: Hill Climbing

Maximum iterations I=10; patience p=3 (early stopping after 3 consecutive iterations without improvement); top-K=5 highest-reward historical skill sets provided to the teacher at each iteration. A candidate general skill set is accepted if and only if its average adjusted reward strictly exceeds the current best.

G.2 Stage 2: UCB-Driven Tree Search

At each iteration, the node n maximizing the following UCB1 score is selected:

$$\mathrm{UCB1}(n) = \bar{R}(n) + C\sqrt{\frac{\ln N_{\mathrm{parent}}}{N_n}}, \tag{7}$$

where $\bar{R}(n)$ is the mean adjusted reward of node n and all its descendants, N_n is the visit count of node n, N_{parent} is the visit count of its parent, and $C{=}1.4$ is the exploration constant. We run $J{=}10$ iterations per task type with $N{=}100$ episodes per node evaluation.

H Skill Rewriter Training Details

We perform full-parameter SFT on Qwen3-4B in BF16 precision. Training uses AdamW (lr $1\mathrm{e}{-5}$, cosine schedule, warmup ratio 0.1, gradient checkpointing), effective batch size 4 (per-device $1\times$ gradient accumulation 4), 5 epochs, and max sequence length 4096. We select the best checkpoint based on training loss convergence.

The training data consists of pairs for in-domain tasks, with data augmentation including noisy inputs (noise ratio 0.3), partial inputs (keep ratio 0.6), and cross-model transfer pairs. We train two rewriter variants: a *Cross-task* rewriter on 769 samples spanning all three environments (ALFWorld Pick/Look/Pick2 only—excluding the held-out types—plus WebShop and Search), and a *Cross-env* rewriter on 499 samples (WebShop + Search only). These correspond to the Cross-task and Cross-env settings reported in Section 4.3.

I WebShop Supplementary Results

I.1 Trajectory Analysis: Why Larger Models Fail

We analyze failed WebShop trajectories to understand why larger Qwen3 models (8B/14B/32B) perform worse than 4B under baseline conditions (Section 4.2).

Table 11 reveals a striking pattern: Qwen3-4B produces concise, action-only outputs (0% steps with chain-of-thought, $\sim$73 chars per action), while 8B/14B/32B prepend extensive reasoning preambles before each action command. Qwen3-14B is the most severe case, with 97% of steps containing verbose reasoning. This behavior exhausts the fixed step budget on deliberation rather than environment interaction—the agent "thinks" through multiple options but never completes enough purchase actions to succeed.

Model	CoT (%)	Action Len.
Qwen3-4B	0	73 chars
Qwen3-8B	57	1,021 chars
Qwen3-14B	97	574 chars
Qwen3-32B	66	491 chars

Table 11: WebShop trajectory statistics. CoT: fraction of steps containing reasoning preambles.

I.2 Per-Category Breakdown

Table 12 provides the full per-category success rate breakdown. Several observations stand out:

- For 8B/14B/32B baselines, most categories have near-zero SR, consistent with the verbose-reasoning bottleneck identified above.
- MASA achieves the best SR in the vast majority of categories across all backbones, with particularly large gains on Other and Electronics.
- The improvement is broad rather than category-specific: MASA does not exploit a single easy category to inflate the average but improves performance across the board.

Model	Method	Per-Category SR (%)							Avg.
		Apparel	Other	Footwear	Home	Elec.	Access.	Beauty	
Qwen3-4B	No Skill	18.6	40.0	11.1	19.0	71.4	20.0	16.7	23.0
	+ Base Skill	16.1	27.0	26.7	19.0	28.6	10.0	16.7	19.4
	+ DS-Adapter	20.3	20.0	13.3	14.3	28.6	10.0	16.7	19.2
	+ MASA	**23.2**	**38.0**	**28.9**	**19.0**	28.6	**20.0**	**16.7**	**26.4**
Qwen3-8B	No Skill	2.6	11.0	0.0	4.8	14.3	**20.0**	0.0	4.6
	+ Base Skill	5.5	9.0	0.0	0.0	14.3	**20.0**	16.7	6.0
	+ DS-Adapter	4.8	5.0	0.0	0.0	0.0	10.0	16.7	4.4
	+ MASA	**27.7**	**41.0**	**13.3**	**14.3**	**57.1**	10.0	**33.3**	**28.6**
Qwen3-14B	No Skill	1.0	6.0	6.7	4.8	0.0	10.0	0.0	2.8
	+ Base Skill	0.3	2.0	4.4	4.8	0.0	20.0	0.0	1.6
	+ DS-Adapter	0.3	6.0	2.2	4.8	0.0	10.0	0.0	2.0
	+ MASA	**32.8**	**33.0**	**6.7**	**9.5**	**28.6**	**30.0**	**16.7**	**29.2**
Qwen3-32B	No Skill	4.2	16.0	2.2	9.5	0.0	10.0	0.0	6.6
	+ Base Skill	4.8	14.0	4.4	9.5	14.3	20.0	0.0	7.2
	+ DS-Adapter	2.3	8.0	2.2	4.8	0.0	10.0	0.0	3.6
	+ MASA	**32.2**	**48.0**	**26.7**	**19.0**	**42.9**	**40.0**	**33.3**	**34.6**

Table 12: WebShop per-category success rate (%). **Bold** marks the best within each backbone.

J SKILL REWRITER OOD: PER-TASK BREAKDOWN

Figure 5 shows the per-task SR breakdown for the OOD generalization experiment (Section 4.3).

Cross-task transfer. Adding ALFWorld PICK/LOOK/PICK2 traces to the training set (dark blue) yields consistent improvements over Cross-env on all three tasks. The gains are most pronounced on COOL (e.g., 14B: 24.0→44.0; 32B: 48.0→52.0) and CLEAN (e.g., 8B: 29.6→51.9; 14B: 40.7→51.9), indicating that in-environment traces help the rewriter learn ALFWorld-specific action patterns such as navigation sequences and object interaction protocols. On HEAT, Cross-task improves for 4B and 14B but is comparable to Cross-env for 8B and 32B, suggesting that Heat-specific patterns are already partially captured by the cross-environment signal.

Cross-environment transfer. Trained only on Search and WebShop traces, the Cross-env rewriter (light blue) shows notable strengths on HEAT across all backbones—particularly 14B (56.2%) and 32B (56.2%)—substantially exceeding DS-Adapter. On CLEAN and COOL, Cross-env performance is more mixed: it matches or slightly exceeds DS-Adapter for 4B and 8B, but falls short on some larger-backbone cells (e.g., 14B CLEAN: 40.7 vs. DS-Adapter 40.7, tied). This suggests that cross-environment transfer is most effective when the target task involves decision patterns (e.g., sequential verification in HEAT) that overlap with those in the training environments.

K QUALITATIVE ANALYSIS: EVOLVED SKILL EXAMPLES

Table 13 presents a case study of how MASA adapts skills differently for each backbone on the same subtask—WebShop's color-matching decision, identified as the highest-failure-rate subtask during skill evolution.

Rather than producing minor wording variations, the evolution pipeline discovers qualitatively distinct strategies tailored to each model's dominant failure mode:

- Qwen3-4B tends to guess visually similar colors. The evolved skill imposes a strict binary rule: match exactly or leave immediately.
- Qwen3-8B abandons products too easily, scoring zero. The evolved skill encourages buying approximate matches for partial credit.

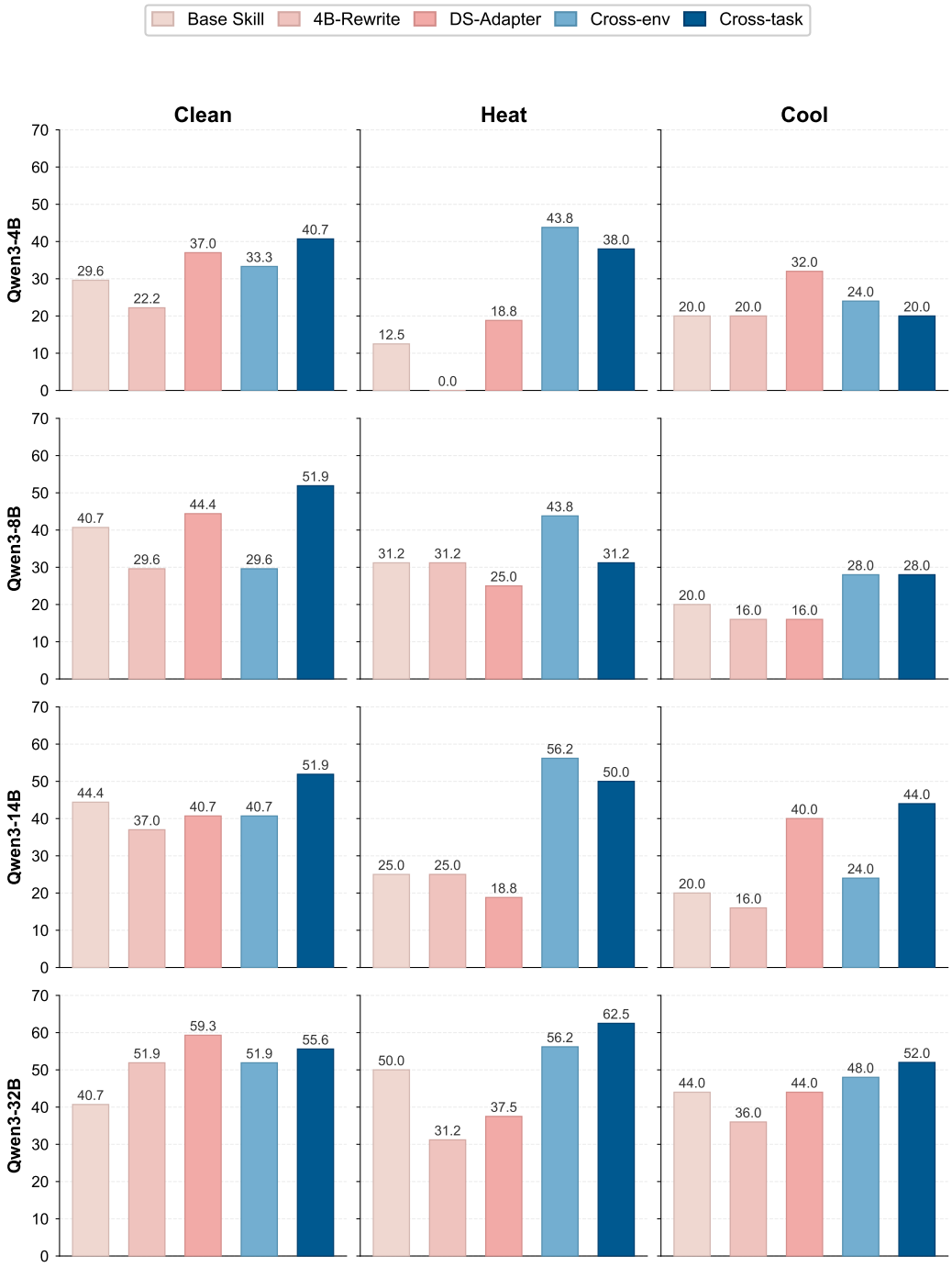

Figure 5: Per-task OOD generalization of MASA-Rewriter. Rows: target backbones (4B–32B). Columns: held-out task types (Clean, Heat, Cool).

- Qwen3-14B wastes steps deliberating on bad products. The evolved skill enforces a one-step fast-fail exit when no exact match exists.

- Qwen3-32B mishandles coded color names (e.g., `d01green`). The evolved skill provides an algorithmic substring-matching procedure with tie-breaking rules.

Model	Evolved Skill Text	Failure Mode	Strategy
4B	When selecting a color variant, match the EXACT string from the task requirement to the available options. 'navy blue' ≠ 'light blue' ≠ 'navy'. 'c3-black' ≠ 'c-black'. If the exact color name is not available in the options list, this product CANNOT satisfy your requirement---**leave immediately**. **Do NOT select an approximate or similar color.**	Picks visually similar colors by guessing	Strict binary match: exact or leave
8B	Scan ALL color options. If the EXACT color name appears, click it. 'c3-black' is NOT the same as 'a6-black'. If the exact color is NOT available but a SIMILAR one exists (e.g., goal says 'green', options have 'e-green'), **select the CLOSEST match and proceed to buy**. A close color match gives partial credit which is better than 0. **Even if the product doesn't perfectly match---BUY IT.**	Abandons products too easily (0 credit)	Flexible match: buy anyway for partial credit
14B	If your required color is NOT in the admissible actions list, **click 'back to search' immediately**. Do not try similar colors. Do not try similar sizes. **One glance at options → if exact match missing → back to search. Takes 1 step, not 5.**	Wastes steps deliberating on bad products	Fast-fail: 1-step exit if no exact match
32B	**Match color by checking if the task's required color name appears as a SUBSTRING** in any admissible action, or vice versa. 'patina green' matches 'patina green' (exact) but NOT 'yellow'. 'green' matches 'a1-green' or 'd01green' (contains). **When multiple options contain the color word, prefer the one that matches more of the full color name.**	Mishandles coded names (e.g. d01green)	Algorithmic: substring matching with preference rule

Table 13: WebShop color-matching skill evolved by MASA for four backbones. Red: rigid rejection rule (4B); Blue: flexible buy-anyway heuristic (8B); Teal: 1-step fast-fail exit (14B); Violet: algorithmic substring matching (32B). Each strategy targets the dominant failure mode of its target model.

This demonstrates that model-conditioned adaptation operates at the level of *decision strategy*—the same problem requires fundamentally different solutions depending on how each backbone fails.