# SkillNet: Create, Evaluate, and Connect AI Skills

Yuan Liang
Ruobin Zhong
Haoming Xu
Zhejiang University
China

Chen Jiang
Southeast University
China

Jia-Chen Gu
UCLA
United States

Yang Liu
NTU
Singapore

Zhaopeng Tu
Tencent
China

Jeff Z. Pan
The University of Edinburgh
United Kingdom

Guilin Qi
Southeast University
China

Haofen Wang
Tongji University
China

Ningyu Zhang*
Huajun Chen
Zhejiang University
China

## Abstract

Current AI agents can flexibly invoke tools and execute complex tasks, yet their long-term advancement is hindered by the lack of systematic accumulation and transfer of skills. Without a unified mechanism for skill consolidation, agents frequently "reinvent the wheel", rediscovering solutions in isolated contexts without leveraging prior strategies. To overcome this limitation, we introduce **SkillNet**[1], an open infrastructure designed to create, evaluate, and organize AI skills at scale. SkillNet structures skills within a unified ontology that supports creating skills from heterogeneous sources, establishing rich relational connections, and performing multi-dimensional evaluation across *Safety*, *Completeness*, *Executability*, *Maintainability*, and *Cost-awareness*. Our infrastructure integrates a repository of over 400,000 skills, an interactive platform[2], and a versatile Python toolkit. Experimental evaluations on ALFWorld, WebShop, and ScienceWorld demonstrate that SkillNet significantly enhances agent performance, improving average rewards by 40% and reducing execution steps by 30% across multiple backbone models. By formalizing skills as evolving, composable assets, SkillNet provides a robust foundation for agents to move from transient experience to durable mastery[3].

***CCS Concepts:*** • **Computing methodologies** → **Intelligent agents**; *Ontology engineering*; Planning and scheduling; • **Software and its engineering** → Software libraries and repositories.

***Keywords:*** AI Agent, Skills, Large Language Models, Resource, Infrastructure

*Corresponding author

[1]https://github.com/zjunlp/SkillNet

[2]http://skillnet.openkg.cn/

[3]The full author list is given in Section 7 due to space constraints.

*Agent Skills '26, San Jose, CA, USA*
2026.

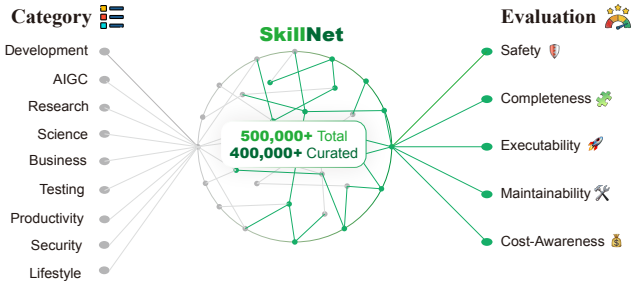


**Figure 1.** Overview of **SkillNet**. SkillNet organizes large-scale agent skills into a structured skill network, modeling rich relations (e.g., similarity, composition, and dependency), supporting multi-dimensional evaluation, and providing unified interfaces for skill discovery, creation, and analysis.

## 1 Introduction

As Richard S. Sutton noted, "we are in the era of experience" [29]. Intelligence is increasingly grounded not in *ab initio* knowledge acquisition, but in the efficient retrieval and adaptive reuse of heuristics distilled from prior experience. This transition marks the advent of the agent era, where AI moves beyond static question answering to orchestrate long-horizon, executable tasks [4, 10, 24, 36, 41, 50, 53]. Despite recent progress in agentic systems, a critical yet underexplored question remains: **How can agents systematically consolidate episodic experience into durable, transferable mastery?**

Current AI methods largely rely on manual engineering or transient in-context learning [7, 8, 26]. While humans excel at consolidating episodic experience into reusable schemas [12], for instance when a programmer internalizes the logic of an algorithm rather than merely memorizing its syntax, contemporary AI systems struggle to bridge the divide between transient context and long-term capability. Without a unified mechanism for skill consolidation and

sharing, agents repeatedly "reinvent the wheel" in isolated contexts, and even well-established strategies rarely inform future tasks. This gap highlights a broader, longstanding tension in AI between explicit structure and scalable representation, a limitation that becomes especially evident when examined through the historical evolution of knowledge engineering [33]:

- **The Symbolic Era**: Systems relied on rigid symbolic logic, offering interpretability but suffering from brittleness and limited scalability [2, 13, 15].
- **The Deep Learning Era**: Knowledge became parametric, taking the form of high-dimensional weight matrices. While powerful, such representations are opaque and difficult to modularize or reuse [11, 14].
- **The Agentic Era (Current Frontier)**: We are witnessing a convergence toward agent skills, which serve as simple, transferable units that provide agents with new capabilities and expertise while separating intelligence from monolithic parameter spaces [16, 27, 49, 51].

This historical trajectory suggests a natural progression toward skills as modular, externalized units of knowledge that reconcile structural interpretability with scalable representation. Consequently, the central challenge in the agent era is no longer merely learning from experience, but transforming fragmented experience into durable, composable skill units that support generalizable intelligence. However, current AI approaches fall short in two fundamental respects. First, there is no unified mechanism to acquire and consolidate skills from experience. Valuable expertise exists at scale across open-source repositories, academic publications, and agent execution traces [3, 45], yet it remains largely unstructured and disconnected. Unlike human learners, who continuously internalize external information into organized knowledge schemas, AI agents cannot automatically distill these resources into reusable, executable capabilities. As a result, skill acquisition remains a manual and episodic process rather than an autonomous, cumulative one [41]. Second, there is no principled framework to validate and maintain skill quality at scale. Without intrinsic and systematic evaluation, skill repositories are prone to "pollution," where executability, safety, and robustness are assessed only indirectly through downstream task success [25, 30, 31]. Such stochastic and opaque validation introduces technical debt, undermining long-term capability growth and preventing agents from leveraging accumulated experience in a reliable and scalable manner. Together, these gaps highlight the need for a structured framework that not only captures and organizes skills, but also rigorously evaluates and interconnects them to enable dependable, large-scale knowledge reuse.

To address these issues, we introduce **SkillNet** (Figure 1), an open infrastructure for creating, evaluating, and organizing AI skills at scale. We conceptualize a skill as a unified knowledge representation that bridges unstructured language understanding with structured, machine-executable logic. SkillNet constructs a comprehensive Skill Ontology with three interconnected layers: a taxonomic layer for functional categorization, a relational layer encoding dependencies and composition, and a skill-package layer for modular deployment. To ensure the reliability of these skills, we propose a multi-dimensional evaluation framework that assesses skills based on **Safety**, **Completeness**, **Executability**, **Maintainability**, and **Cost-awareness**.

Our approach transforms fragmented experience from diverse sources, including execution trajectories, open-source repositories, and documentation, into a structured network of over 400,000 curated skills. Experimental results across three text-based simulated environments (ALFWorld, WebShop, and ScienceWorld) demonstrate that agents augmented with SkillNet achieve substantial gains. For instance, our approach improves the average reward by 40% while reducing interaction steps by 30% across various backbone models (e.g., DeepSeek V3, Gemini 2.5 Pro, and o4 Mini), validating that systematic skill accumulation effectively enhances agent competence cumulatively rather than episodically.

Our contributions are summarized as follows:

1. We introduce SkillNet, a unified framework that transforms fragmented agent experience into a structured network of modular, composable skills with rich relational modeling, serving as a scalable foundation for actionable knowledge engineering.
2. We establish a rigorous skill evaluation protocol that quantitatively measures safety, completeness, executability, maintainability, and cost-awareness, ensuring the reliability of large-scale skill repositories.
3. We release an open-source ecosystem, including a repository of over 400,000 curated skills, a Python toolkit, and comprehensive benchmarks, which empirically demonstrate significant performance improvements in agent planning and execution tasks.

## 2 SkillNet

### 2.1 Overview

Figure 2 illustrates the overall architecture of SkillNet, which systematically creates, evaluates, and organizes high-quality skills for agent systems. SkillNet is designed to transform fragmented agent experiences and human knowledge into reusable and verifiable skill entities, enabling scalable and reliable capability growth. SkillNet consists of three core modules:

***Skill Creation.*** This module analyzes diverse inputs, including user trajectories, office documents, GitHub projects,

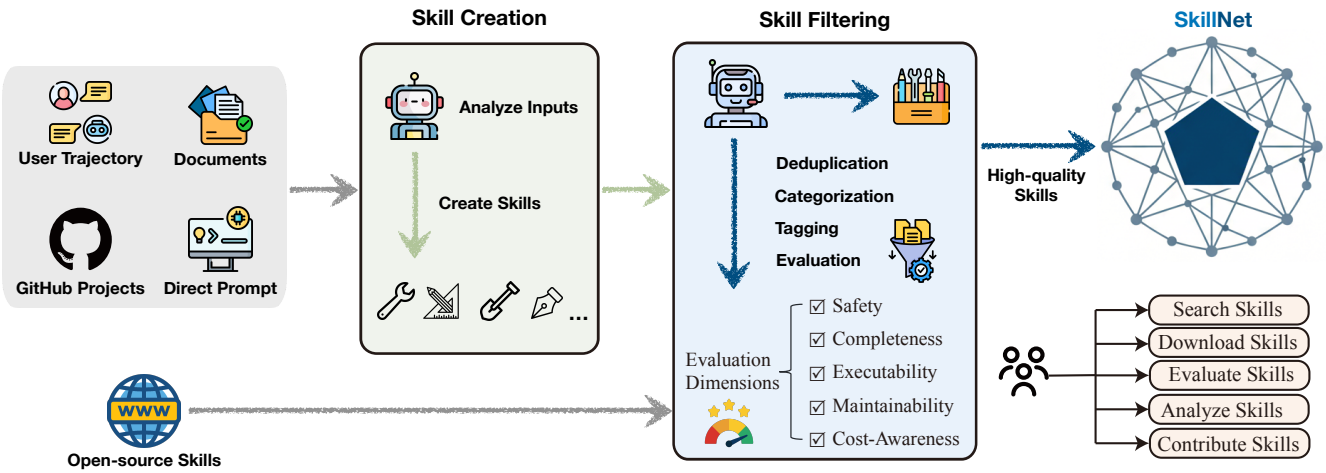


**Figure 2.** End-to-end Pipeline of SkillNet. SkillNet transforms heterogeneous user inputs and open internet resources into executable skills through automated skill creation and multi-dimensional evaluation, and organizes high-quality skills into a structured network to support search, download, analysis, and contribution.

direct prompts, and open internet resources. Based on these inputs, SkillNet generates new skills by extracting executable patterns and structuring them as reusable capabilities.

***Skill Evaluation.*** Generated skills are filtered and evaluated along multiple principal dimensions, including safety, completeness, executability, maintainability, and cost-awareness. This evaluation ensures that only high-quality skills are retained, mitigating redundancy, brittleness, and potential risks.

***Skill Analysis.*** Beyond isolated skills, SkillNet automatically analyzes structural and functional relationships among skills, constructing large-scale skill graphs that capture similarity, hierarchy, composition, and dependency patterns. This structured representation enables global reasoning over skill repositories, supporting efficient retrieval, composition, and workflow synthesis.

***Open Resources.*** SkillNet organizes curated, high-quality skills within a structured repository, offering versatile toolkits that enable users and agents to efficiently search, download, create, evaluate, analyze, and contribute skills. By providing these standardized interfaces for skill interaction, SkillNet facilitates seamless skill reuse and collaborative evolution across tasks, domains, and agent populations.

## 2.2 Skill Ontology

The Skill Ontology organizes individual skills into a structured, composable network, as illustrated in Figure 3. The architecture is organized into three progressive layers:

***Skill Taxonomy.*** This layer organizes skills into a multi-level hierarchy using `category` and `tag` relations. It categorizes broad domains (e.g., Development, AIGC, Science ...) into fine-grained tags (e.g., frontend, llm, physics), providing a high-level semantic skeleton.

***Skill Relation Graph.*** This layer instantiates abstract tags into specific *Skill Entities* (e.g., Matplotlib, Playwright). It defines the core interaction logic through multi-relational edges: `similar_to`, `compose_with`, `belong_to`, and `depend_on`, forming the backbone for reasoning and planning.

***Skill Package Library.*** The bottom layer represents the physical organization of skills. Individual skills are encapsulated into *Skill Packages* (e.g., `data-science-visualization`) via the `packaged_in` relation, facilitating modular release and deployment.

Note that the skill ontology is dynamic, modeling relations between skills. Tags can be continually added from the taxonomy, and LLMs infer relations from them, enabling the instantiation of a skill relation graph.

## 2.3 Skill Creation

SkillNet conceptualizes skills as structured intermediate capability units that bridge abstract, declarative knowledge and concrete, executable programs. Specifically, we develop an automated creation pipeline that systematically transforms heterogeneous information sources into standardized, reusable, and modular agent skills.

### 2.3.1 From Human Experience to Skill Abstraction.

To construct a comprehensive and versatile skill repository, SkillNet abstracts skills from diverse data sources, transforming heterogeneous information into reusable structured agent skills.

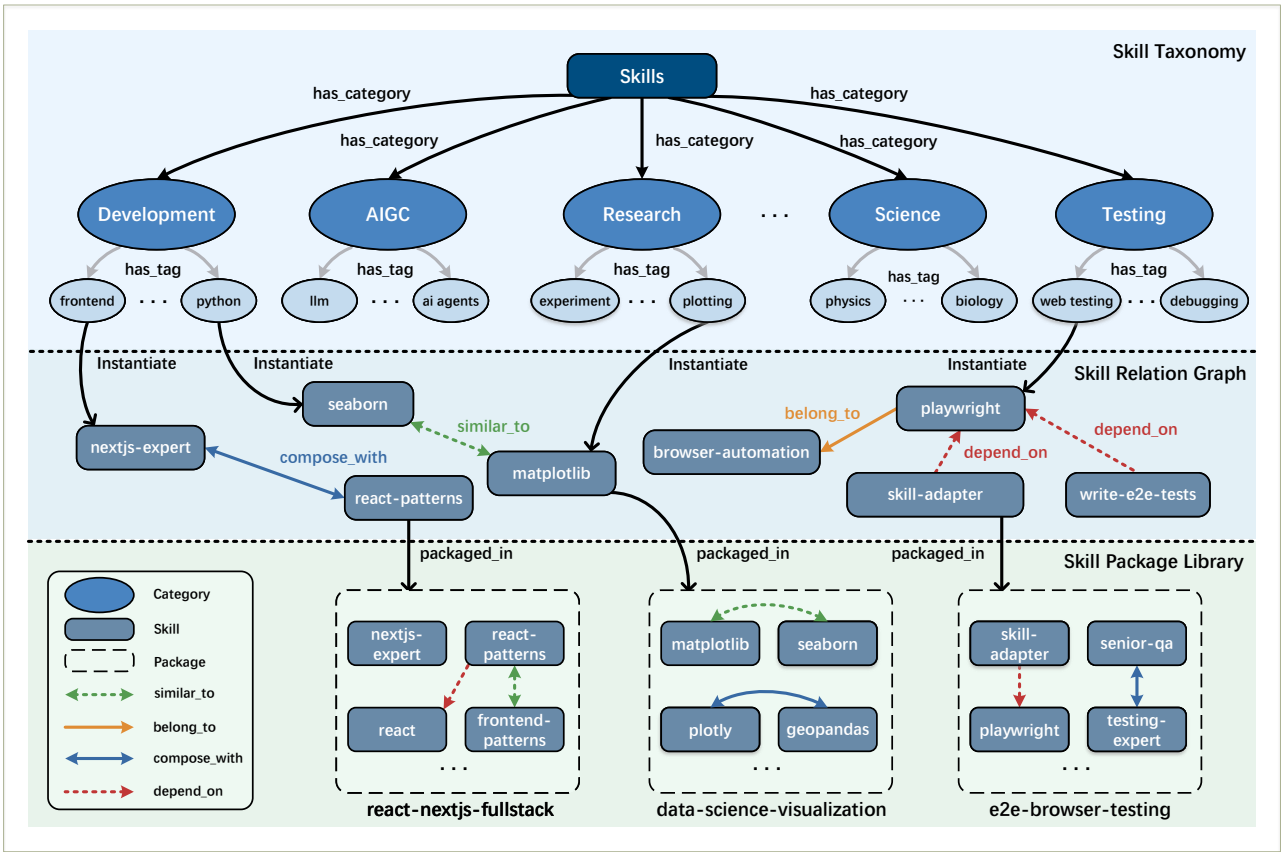


**Figure 3.** The Skill Ontology for SkillNet. It consists of three levels: the **Skill Taxonomy** (top) defines functional categories; the **Skill Relation Graph** (middle) models inter-skill dependencies and semantic associations; and the **Skill Package Library** (bottom) organizes skills into modular, task-oriented bundles.

We design a multi-source automated skill creation pipeline that enables SkillNet to induce agentic skills from heterogeneous human knowledge and experience. Specifically, SkillNet supports four major categories of data sources: (1) execution trajectories and conversational interaction logs, (2) open-source GitHub repositories, (3) semi-structured documents including PDF, PowerPoint, and Word files, and (4) direct natural language prompts provided by users. This process is fully implemented through LLMs, and users can also customize the underlying models. In addition, SkillNet continuously expands its skill pool through open internet resources, in-house development, and community contributions, ensuring sustained scalability and coverage expansion.

#### 2.3.2 Data-Driven Skill Filtering and Consolidation.

Automatic construction of skills does not imply indiscriminate accumulation. Instead, SkillNet introduces a data-driven filtering and consolidation pipeline to ensure the quality of the skill repository. Specifically, SkillNet applies a multi-stage curation process consisting of deduplication, filtering, categorization and tagging, evaluation, and final selective consolidation.

Deduplication is performed by jointly comparing skill directory structures and MD5 hashes of skill markdown files, effectively removing redundant skills. Filtering eliminates low-quality, incomplete, or semantically meaningless skills through rule-based validation and model-based checking. Categorization and tagging classify each skill into one of ten functional categories (Development, AIGC, Research, Science, Business, Testing, Productivity, Security, Lifestyle, Other) and assign fine-grained semantic tags to facilitate retrieval and composition. Then, a multi-dimensional evaluation mechanism is applied to determine whether a skill is admitted into the SkillNet repository. SkillNet automatically establishes inter-skill associations based on ontology-defined relations, ultimately resulting in a structured skill package library.

This structured pipeline enables SkillNet to function as a self-evolving skill ecosystem rather than a static skill collection, continuously improving both skill quality and coverage.

### 2.4 Skill Evaluation

Although skill creation is important, the utility of a skill repository ultimately depends on its reliability. To bridge the

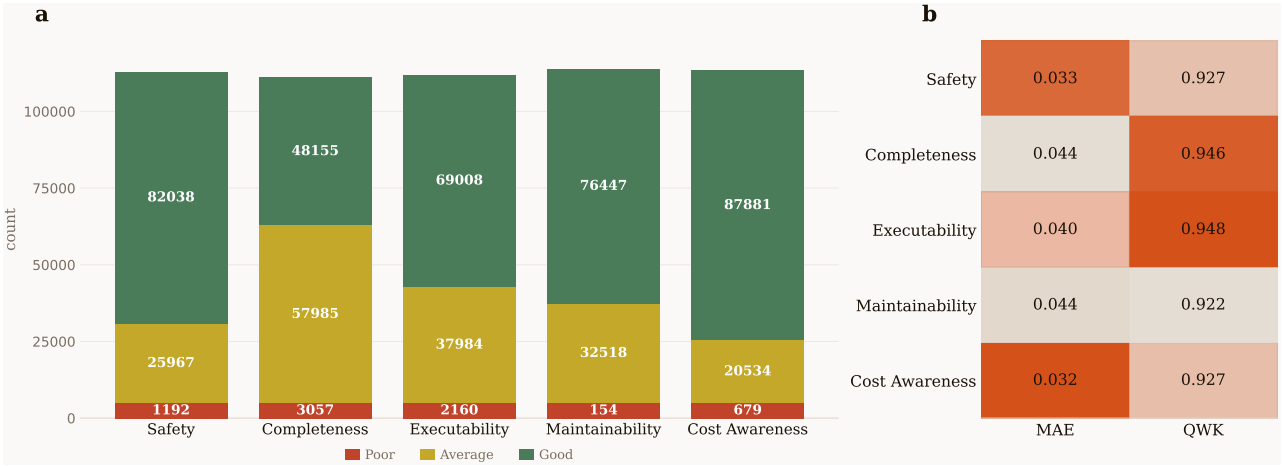


**Figure 4.** Multi-dimensional Skill Evaluation and Reliability Analysis. (a) Quality distribution of curated skills in SkillNet, assessed using a three-level grading scheme, where most skills are rated *Good* or *Average*. (b) Reliability validation of the automated evaluator on 200 randomly sampled skills. The heatmap shows mean absolute error (MAE) and quadratic weighted kappa (QWK) between human annotators and the model evaluator, with near-perfect QWK and low MAE across all dimensions, demonstrating the robustness and scalability of the evaluation framework.

gap left by existing repositories that lack standardized assessment, we propose a multi-dimensional systematic evaluation framework.

We define five core dimensions to quantitatively characterize the quality and readiness of each skill:

- **Safety:** Assesses potential risks, including hazardous system operations (e.g., unauthorized file deletion) and robustness against prompt injection or adversarial manipulation.
- **Completeness:** Evaluates whether the skill encapsulates all critical procedural steps and explicitly defines necessary prerequisites, dependencies, and execution constraints.
- **Executability:** Verifies if the skill can be successfully implemented by agents in sandboxed environments, identifying hallucinated tool calls or ambiguous instructions.
- **Maintainability:** Measures the modularity and composability of skills, ensuring they can be locally updated without disrupting global dependencies or breaking backward compatibility.
- **Cost-awareness:** Quantifies execution overhead, including time latency, computational resource consumption, and API usage costs to support efficiency optimization.

To achieve high-throughput assessment, all dimensions are primarily evaluated via an automated LLM-based evaluator (instantiated with GPT-5o-mini), guided by fine-grained rubrics. For **Executability**, we complement LLM judgment with empirical validation: skills containing code or tool invocations are executed in controlled sandbox environments to verify runtime correctness. Each dimension is categorized into three levels: *Good*, *Average*, and *Poor*.

The quality distribution across the repository is illustrated in Figure 4(a). We observe that while **Safety** and **Maintainability** maintain a high proportion of "Good" ratings, **Executability** presents a greater challenge, with a larger fraction of skills rated as *Average*. This reflects our rigorous filtering criteria, ensuring that only high-fidelity skills are prioritized for complex task execution.

To validate the reliability of this automated pipeline, we randomly sampled 200 skills and recruited three PhD-level computer science annotators for independent blind reviews. As shown in Figure 4(b), the inter-rater agreement between human judgments and LLM-based scores exhibits exceptional consistency. Across all dimensions, the Mean Absolute Error (MAE) remains below 0.03, and the Quadratic Weighted Kappa (QWK) consistently reaches near-perfect levels (above 0.90). These results confirm that our automated evaluator provides a human-aligned, robust, and scalable foundation for managing the SkillNet ecosystem.

### 2.5 Skill Analysis

Beyond isolated skill creation and evaluation, large-scale skill repositories introduce a new challenge: how to systematically understand, organize, and exploit the relations among skills. To address this, SkillNet introduces a dedicated Skill Analysis module that automatically discovers and models structural relations between skills, forming a structured and interpretable skill relation graph. This enables global reasoning over large skill repositories and supports advanced downstream applications such as skill retrieval, composition, dependency resolution, and workflow synthesis.

```
# Install
pip install skillnet-ai

# Search & Download skill
skillnet search "bioinformatics pipeline"
skillnet download https://github.com/repo/skill

# Create new skill
skillnet create ./logs/chat.txt --model gpt-5

# Evaluate skill quality
skillnet evaluate ./my_skills/biopython

# Analyze skill relations
skillnet analyze ./my_skills
```

```
# Initialize
from skillnet_ai import SkillNetClient
client = SkillNetClient(api_key="sk-...")

# Search & Download skill
skills = client.search("bioinformatics pipeline")
client.download(skills[0].skill_url,
target_dir="./my_skills")
# Create new skill
client.create(trajectory_log, output_dir="./my_skills")

# Evaluate skill quality
client.evaluate("./my_skills/biopython")

# Analyze skill relations
client.analyze("./my_skills")
```

**Figure 5.** Practical use examples of skillnet-ai. The package offers a unified functional experience through both a command-line interface (left) for interactive usage and a Python library (right) for seamless integration into AI development pipelines.

***Relation Modeling.*** SkillNet formulates skill analysis as a structured relations discovery problem. Given a large amount of heterogeneous skills, the system automatically identifies and annotates multiple types of semantic and functional relations between skills, including (task-specific relations are omitted here, and users may extend the schema by defining custom relations as needed):

- **similar_to:** Two skills perform functionally equivalent or highly similar tasks and can often be used interchangeably, enabling redundancy detection, replacement, and robustness enhancement.
- **belong_to:** A skill serves as a sub-component or atomic step within a larger composite workflow, capturing hierarchical structures and supporting skill abstraction and modularization.
- **compose_with:** Two skills are frequently co-invoked in workflows, where one typically produces outputs consumed by the other, enabling automatic workflow composition and pipeline generation.
- **depend_on:** A skill cannot execute independently without prerequisite skills, such as environment setup or API initialization, enabling explicit dependency tracking and safe execution planning.

***Automated Skill Relation Graph Construction.*** SkillNet can construct large-scale skill relation graphs through a hybrid pipeline that integrates semantic embedding and LLM-based relational inference. Candidate relations are first generated via similarity matching, dependency extraction, and execution trace alignment, and LLM reasoning, enabling scalable and accurate discovery of structured skill relations. These relations jointly form a directed, typed multi-relational graph, where nodes represent skills and edges encode fine-grained structural dependencies and functional associations.

***Task-Oriented Skill Collection Release.*** To promote reproducibility and practical adoption, SkillNet releases over 20 task-specific skill collections, covering domains such as scientific research automation, software development, data processing pipelines, web automation, etc. In particular, several collections are designed for academic benchmarks, and we conduct quantitative evaluations to assess the effectiveness of this paradigm, as detailed in a later section. These released collections provide structured and reusable infrastructures for studying large-scale agent planning, workflow synthesis, and skill evolution, offering a foundation for future research on structured agent intelligence.

### 2.6 Open Resources

SkillNet provides a comprehensive open infrastructure for creating, evaluating, and organizing AI skills at scale. This includes a large-scale skill repository, a front-end website, an open access API, and a versatile Python toolkit (skillnet-ai), forming a unified ecosystem for skill management and utilization.

The front-end website allows users to seamlessly browse, search, and download skills directly. Users can explore skill categories, view detailed API documentation and Python toolkit usage guidelines, and learn how to apply SkillNet to scientific research or coding scenarios. The platform also supports community contributions, enabling users to submit and share newly constructed skills, thus fostering a collaborative and continuously evolving skill ecosystem.

For API access, SkillNet supports both keyword-based and vector-based searches. Users can search for skills by keywords, categories, or semantic similarity via http://api-skillnet.openkg.cn/v1/search.

| Model | Method | ALFWorld | | | | WebShop | | ScienceWorld | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Seen | | Unseen | | | | Seen | | Unseen | |
| | | R ↑ | S ↓ | R ↑ | S ↓ | R ↑ | S ↓ | R ↑ | S ↓ | R ↑ | S ↓ |
| DeepSeek V3.2 | React | 66.43 | 19.51 | 69.40 | 19.27 | 31.55 | 24.06 | 69.86 | 17.59 | 64.67 | 19.26 |
| | Expel | 67.86 | 18.86 | 76.12 | 17.41 | 29.23 | 24.00 | 74.91 | 15.98 | 74.09 | 17.53 |
| | + SkillNet | **80.60** | **14.54** | **83.57** | **14.81** | **46.18** | **17.84** | **84.87** | **11.89** | **81.31** | **12.48** |
| Gemini 2.5 Pro | React | 60.00 | 18.72 | 61.94 | 19.18 | 31.66 | 22.12 | 58.24 | 18.42 | 56.13 | 19.07 |
| | Expel | 68.57 | 17.88 | 70.15 | 17.04 | 33.12 | 19.31 | 72.76 | 15.01 | 67.37 | 14.91 |
| | + SkillNet | **91.43** | **12.80** | **91.04** | **11.96** | **53.02** | **14.91** | **88.84** | **11.49** | **86.26** | **11.30** |
| o4 Mini | React | 45.71 | 23.31 | 49.25 | 23.33 | 24.19 | 22.02 | 64.89 | 15.12 | 59.93 | 14.99 |
| | Expel | 56.43 | 21.35 | 58.96 | 21.85 | 26.71 | 21.91 | 67.95 | 13.65 | 65.68 | 13.95 |
| | + SkillNet | **68.57** | **18.94** | **73.28** | **17.08** | **36.21** | **18.79** | **73.24** | **13.30** | **71.06** | **12.35** |
| Claude Sonnet 4.6 | React | 88.57 | 12.00 | 89.55 | 12.32 | 41.09 | 21.58 | 79.90 | 13.72 | 78.16 | 14.27 |
| | Expel | 92.86 | 11.82 | 91.04 | 12.29 | 48.63 | 20.07 | 75.99 | 12.99 | 76.24 | 14.88 |
| | + SkillNet | **95.71** | **10.68** | **95.52** | **11.03** | **50.92** | **19.14** | **83.25** | **10.77** | **82.94** | **11.91** |

**Table 1.** Experimental results on ALFWorld, WebShop, and ScienceWorld. R denotes average reward, S denotes average steps (↑ indicates the larger values are better, and ↓ denotes the smaller values are better). The top-performing results are highlighted in **bold**, and the second-best are underlined.

To bridge the gap between static repositories and dynamic execution, we developed `skillnet-ai`, a unified Python library and CLI tool that enables flexible integration with users. Through this toolkit, users can search for skills using keywords or semantic similarity, download skills directly from GitHub into local workspaces, and create structured skills from heterogeneous sources, including execution trajectories, GitHub repositories, office documents, and natural language prompts. The package also supports evaluating skills across multiple dimensions (safety, completeness, executability, maintainability, and cost-awareness) to ensure practical reliability, and analyzing relations between skills to uncover dependencies, hierarchical compositions, collaborations, and functional similarities. The practical use examples are illustrated in Figure 5.

SkillNet has initially aggregated over 500k candidate skills from open internet resources, automated creation pipelines, and community contributions. After multi-stage filtering and evaluating, more than 400k high-quality skills are curated in the final repository (and which is constantly expanding). This scale ensures broad coverage across scenarios while maintaining rigorous quality standards, enabling SkillNet to serve as a reliable skill infrastructure for both scientific investigation and real-world deployment. We actively encourage community contributions, and all community-submitted skills are subject to the same automated quality checks. In addition, we conduct periodic manual audits through random sampling to ensure the ongoing reliability and integrity of the shared skill repository.

# 3 Quantitative Evaluation

## 3.1 Settings

To quantitatively assess the effectiveness of SkillNet, we conduct experiments across three text-based simulated environments. ALFWorld [28] provides an embodied household environment, requiring agents to navigate and manipulate objects to complete daily tasks; WebShop [43] simulates realistic online shopping scenarios, where agents perform product search, comparison, and purchasing under specified constraints; and ScienceWorld [38] presents a virtual scientific laboratory in which agents must conduct experiments and operate scientific instruments. All three environments are formulated as partially observable Markov decision processes (POMDPs), requiring agents to make sequential decisions under incomplete and noisy observations.

We adopt ReAct [44], Expel [49] and a standard Few-Shot approach as the baseline methods. ReAct interleaves reasoning, action, and observation for step-by-step task solving. ExpeL gathers experiences from past tasks to extract natural language insights and retrieves past successful trajectories as in-context examples during inference. The Few-Shot baseline randomly samples a complete expert trajectory to serve as a static in-context exemplar. In contrast, our proposed variant is augmented with SkillNet by leveraging expert trajectories from ETO [32] to synthesize benchmark-specific skill collections (these collections are integrated into SkillNet, and users can extend it to other datasets by creating

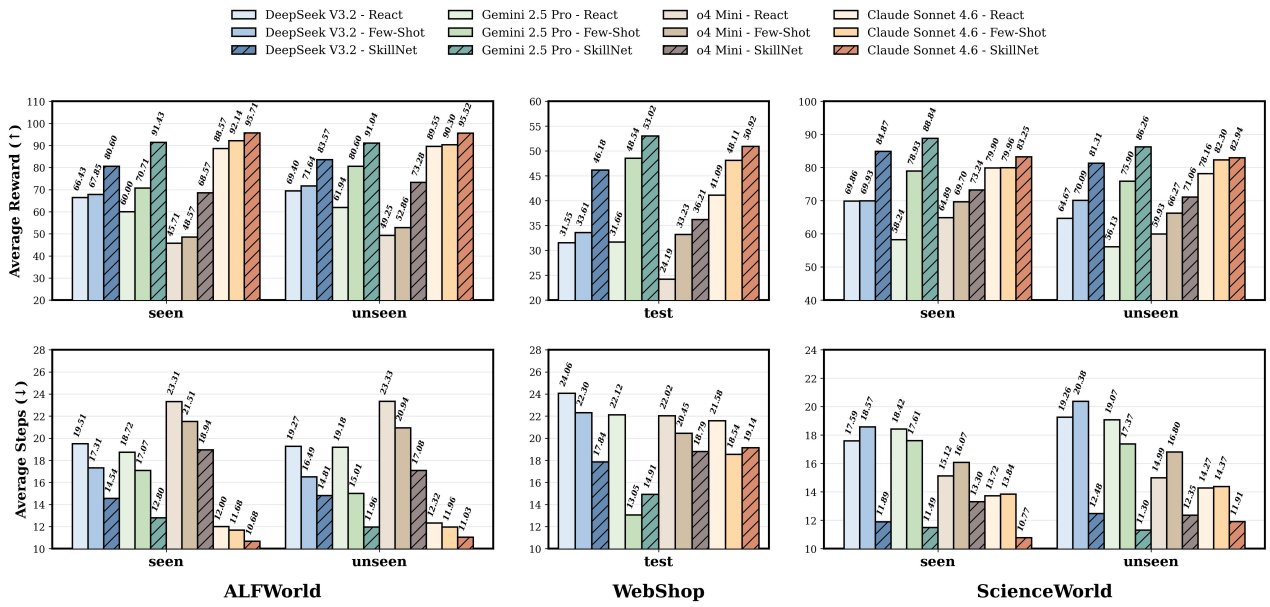


**Figure 6.** Performance comparison across diverse methods and models. The results illustrate that SkillNet consistently outperforms React and Few-shot baselines, achieving significantly higher average rewards (top) and reduced average steps (bottom) across ALFWorld, WebShop, and ScienceWorld.

corresponding skills)[4]. Note that the above experiences and the seen and unseen splits of the test set have no overlap, thereby preventing data leakage. During evaluation, agents are equipped with these collections and can dynamically select, activate, and execute the most relevant skills based on the current state. We employ four representative LLMs: DeepSeek V3.2 [6], Gemini 2.5 Pro [5], o4 Mini [21], and Claude Sonnet 4.6 [1] as backbone models.

### 3.2 Results

Table 1 and Figure 6 present the quantitative results across diverse environments and backbone models. In general, integrating SkillNet consistently yields substantial gains in both task effectiveness and execution efficiency, validating the effectiveness of our framework. Compared with ReAct, SkillNet improves the average reward by 40%, while reducing the number of interaction steps by 30% on average, indicating that agents equipped with SkillNet are able to solve tasks more reliably while executing shorter and more coherent action trajectories. This reflects SkillNet's ability to transform fragmented experience into reusable procedural abstractions, enabling agents to perform structured planning, reduce redundant exploration, and execute complex behaviors in a more principled and systematic manner.

Importantly, the performance improvements remain robust across backbone models of varying capacities, ranging from compact models (+15.7 R for o4 Mini) to large-scale LLMs (+28.5 R for Gemini 2.5 Pro). This demonstrates that SkillNet provides complementary capabilities beyond parametric knowledge, effectively augmenting reasoning with persistent, executable skills. Moreover, the consistent gains observed under both seen and unseen settings highlight SkillNet's strong generalization ability, suggesting that skill abstraction and reuse facilitate knowledge transfer across tasks and environments.

Taken together, these results substantiate the central design principle of SkillNet: by formalizing skills as independent, systematically accumulated, knowledge-grounded capability units, agent competence can be enhanced cumulatively rather than episodically, ultimately fostering domain-specialized and continually self-improving agents.

## 4 Related Work

### 4.1 Experience Consolidation and Skill Abstraction

LLM-based agents have been extended with tools, planning, and memory to perform long-horizon tasks in complex environments [24, 25, 36, 41, 53]. Several works have explored how agents can acquire and refine skills from interaction and experience. Reflexion [27] and Expel [49] investigate how agents can summarize failures and extract corrective feedback. Memory-centric methods aim to accumulate long-term experience to support continual learning [9, 23, 39, 47]. Other studies explore structured skill learning and cross-task reuse, promoting generalization and transfer across domains [3, 34, 37, 40, 42, 46, 48, 52]. Despite these advances, most approaches still implicitly represent skills by encoding them in prompts, latent memories, or loosely organized workflows,

[4] https://github.com/zjunlp/SkillNet/tree/main/experiments/src/skills.

thereby hindering systematic consolidation, evaluation, and reuse, which fundamentally limit the scalability and reliability of autonomous agent ecosystems.

### 4.2 Skill Repositories and Evaluations

Parallel to the development of agentic systems, a number of community-driven skill repositories [22, 30, 31, 35] and evaluations [18–20] have emerged to curate agent skills, aiming to promote reuse, standardization, and ecosystem-level collaboration. For instance, platforms like ClawHub function as npm-like version management hubs, while SkillsMP and Skills.sh act as expansive open-source directories aggregating GitHub repositories. Additionally, SkillHub introduces premium marketplace features with predefined skill stacks and basic rating systems. Recently, [17] introduce SkillsBench, an 86-task, 11-domain benchmark showing that curated agent Skills significantly boost LLM agent performance (+16.2 pp on average) while self-generated Skills offer no gain, revealing that models benefit from consuming but cannot reliably author procedural knowledge. These platforms provide valuable resources for sharing prompts, tool wrappers, and task-solving templates.

Existing platforms predominantly operate as static package managers or marketplaces, facing three critical limitations: First, they largely rely on manual curation and ad hoc quality control, lacking automated mechanisms to generate skills dynamically from agent trajectories or existing codebases. Second, their evaluation practices primarily focus on simple community metrics (e.g., repository stars) or end-task performance, failing to offer comprehensive insights into intrinsic properties such as safety, completeness, executability, maintainability, and cost-awareness. Third, these skill collections often suffer from redundancy, brittleness, and poor composability because skills are treated as isolated entities, restricting their scalability in large agent populations.

These limitations motivate the need for a unified framework for skill organization, verification, and evolution. In contrast to distribution-centric platforms, our proposed **SkillNet** provides a full-lifecycle infrastructure. It addresses the aforementioned bottlenecks through automated skill creation via LLM pipelines, rigorous multi-dimensional evaluation, and relational connectivity that weaves isolated tools into a structured Skill Graph, thereby offering comprehensive foundational support for self-evolving agent ecosystems.

## 5 Conclusion, Discussion and Future Work

In this work, we introduced SkillNet, an open infrastructure for creating, evaluating, and organizing AI skills at scale. By systematically consolidating experience, structuring skills, and providing principled evaluation, SkillNet enables agents to improve cumulatively, perform reliably across tasks, and adapt to complex, open-ended environments. This framework lays the foundation for scalable continual learning and robust skill composition, bridging the gap between raw model capability and sustained, evolving intelligence.

Conceptually, SkillNet is grounded in a unified view of three complementary constraints on an agent's generative capability: workflows, memory, and skills. Workflows impose explicit procedural structure, ensuring reliability but remaining inherently rigid. Memory accumulates contextual experience and associative knowledge, enabling adaptation but lacking operational boundaries. Skills bridge these extremes by packaging reusable capability units that both constrain generation and organize memory into actionable patterns. In this perspective, skills serve as the structured interface through which memory becomes executable and workflows become flexible.

Looking ahead, the emerging paradigm of the "one-person company" or "one-person lab" envisions a single expert orchestrating a society of agents. Through SkillNet, skills become the primary unit of knowledge integration and delegation: individuals curate skill repositories, agents compose them into workflows, and memory continuously refines them through experience. This closed loop transforms isolated automation into cumulative machine expertise, enabling minimal human teams to achieve organizational-scale intelligence. It marks a pivotal step forward in the journey toward truly autonomous organizational intelligence.

***Open-World Skill Evolution.*** Achieving automatic skill discovery, abstraction, and cross-domain transfer in open-world settings remains highly challenging. In domains such as **industrial manufacturing, finance, and scientific research**, this requires the dynamic composition and optimization of complex tasks. An industry-specific, privately curated SkillNet may itself become a foundational component of the agent infrastructure. Moreover, integrating skill evolution mechanisms with online feedback, causal reasoning, and uncertainty modeling holds promise for improving the reliability of skill selection.

***Model-Skill Synergy.*** Although SkillNet provides agents with large-scale executable skills, the synergy between these skills and the underlying model capabilities remains largely unexplored. In particular, the question of how to leverage neuro-symbolic integration and memory mechanisms to enable skill structures to guide model decision paths and to dynamically restructure skill hierarchies and dependencies as model capabilities evolve remains central to systematic study.

***Multi-Agent Collaboration and Knowledge Sharing.*** In multi-agent environments, SkillNet can function as a shared representation and exchange layer, supporting collaborative planning, knowledge transfer, and the accumulation of experience across agents. By continuously consolidating agent behaviors into reusable skills, SkillNet may further supports the emergence of **digital avatar** whose capabilities are

progressively distilled from accumulated skills. More broadly, this paradigm may open a pathway toward collective intelligence, where skills evolve into transferable, composable units of coordination and where digital personas can inherit, share, and extend competencies beyond individual agents.

## 6 Limitations

There remain several limitations in the present work. First, the coverage of skills is inevitably incomplete. Many capabilities in private or specialized domains cannot be incorporated, while low-frequency or highly tacit abilities are difficult to capture and consolidate within the repository, particularly when they resist explicit linguistic description. Second, the quality of self-constructed skills cannot be fully guaranteed. Although our evaluation procedures filter out some problematic cases, a substantial portion of skills still lacks rigorous and systematic assessment. If malicious users contribute "poisoned" or adversarial skills, SkillNet's current Safety evaluation mechanism can detect some of these cases, but cannot fully mitigate them. Third, an end-to-end pipeline that translates natural-language requirements into fully instantiated agents via SkillNet has not yet been established; this remains an important direction for future work. We will continue to curate, maintain, and improve the system.

## 7 Full Author List

Yuan Liang, Ruobin Zhong, Haoming Xu, Chen Jiang, Yi Zhong, Runnan Fang, Jia-Chen Gu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Xin Xu, Tongtong Wu, Kun Wang, Yang Liu, Zhen Bi, Jungang Lou, Yuchen Eleanor Jiang, Hangcheng Zhu, Gang Yu, Haiwen Hong, Longtao Huang, Hui Xue, Chenxi Wang, Yijun Wang, Zifei Shan, Xi Chen, Zhaopeng Tu, Feiyu Xiong, Xin Xie, Peng Zhang, Lei Liang, Zhengke Gui, Jun Zhou, Chiyu Wu, Jin Shang, Yu Gong, Junyu Lin, Changliang Xu, Hongjie Deng, Wen Zhang, Keyan Ding, Qiang Zhang, Fei Huang, Ningyu Zhang, Jeff Z. Pan, Guilin Qi, Haofen Wang, Huajun Chen

## References


[1] Anthropic. 2026. Announcing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-04-26.

[2] Roberto Baldoni, Emilio Coppa, Daniele Cono D'elia, Camil Demetrescu, and Irene Finocchi. 2018. A survey of symbolic execution techniques. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–39. https://dl.acm.org/doi/10.1145/3182657

[3] Tianyi Chen, Yinheng Li, Michael Solodko, Sen Wang, Nan Jiang, Tingyuan Cui, Junheng Hao, Jongwoo Ko, Sara Abdali, Leon Xu, Suzhen Zheng, Hao Fan, Pashmina Cameron, Justin Wagle, and Kazuhito Koishida. 2026. CUA-Skill: Develop Skills for Computer Using Agent. arXiv:2601.21123 [cs.AI] https://arxiv.org/abs/2601.21123

[4] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*. 2778–2788. https://dl.acm.org/doi/10.1145/3485447.3511998

[5] Google DeepMind. [n. d.]. Gemini 2.5 Technical Report and Model Card (Referencing Gemini 2.5 Pro). https://deepmind.google/models/gemini/pro/ Accessed: 2026-02-07.

[6] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. https://github.com/deepseek-ai/DeepSeek-V3. Accessed: 2026-02-07.

[7] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 1107–1128. https://doi.org/10.18653/v1/2024.emnlp-main.64

[8] Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. 2026. CL-bench: A Benchmark for Context Learning. *arXiv preprint arXiv:2602.03587* (2026).

[9] Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2026. Memp: Exploring Agent Procedural Memory. arXiv:2508.06433 [cs.CL] https://arxiv.org/abs/2508.06433

[10] Runnan Fang, Xiaobin Wang, Yuan Liang, Shuofei Qiao, Jialong Wu, Zekun Xi, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2025. SynWorld: Virtual Scenario Synthesis for Agentic Action Knowledge Refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 437–448. https://doi.org/10.18653/v1/2025.acl-short.33

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[12] Douglas L Hintzman. 1986. " Schema abstraction" in a multiple-trace memory model. *Psychological review* 93, 4 (1986), 411.

[13] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)* 54, 4 (2021), 1–37. https://dl.acm.org/doi/10.1145/3447772

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444. https://www.nature.com/articles/nature14539

[15] Clarence Irving Lewis, Cooper Harold Langford, and P Lamprecht. 1959. *Symbolic logic*. Vol. 170. Dover publications New York.

[16] Xiaoxiao Li. 2026. When single-agent with skills replace multi-agent systems and when they fail. *arXiv preprint arXiv:2601.04748* (2026). https://arxiv.org/abs/2601.04748

[17] Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, et al. 2026. SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks. *arXiv preprint arXiv:2602.12670* (2026). https://arxiv.org/abs/2602.12670

[18] Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, Shuyi Wang, Qunhong Zeng, Di Wang, Xuandong Zhao, Yuanli Wang, Roey Ben Chaim, Zonglin Di, Yipeng Gao, Junwei He, Yizhuo He, Liqiang Jing, Luyang Kong, Xin Lan, Jiachen Li, Songlin Li, Yijiang Li, Yueqian Lin, Xinyi Liu, Xuanqing Liu, Haoran Lyu, Ze Ma, Bowei Wang, Runhui Wang, Tianyu Wang, Wengao Ye, Yue Zhang, Hanwen Xing, Yiqi Xue, Steven Dillmann, and Han chung Lee. 2026. SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse

Tasks. arXiv:2602.12670 [cs.AI] https://arxiv.org/abs/2602.12670

[19] George Ling, Shanshan Zhong, and Richard Huang. 2026. Agent Skills: A Data-Driven Analysis of Claude Skills for Extending Large Language Model Functionality. arXiv:2602.08004 [cs.SE] https://arxiv.org/abs/2602.08004

[20] Yi Liu, Weizhe Wang, Ruitao Feng, Yao Zhang, Guangquan Xu, Gelei Deng, Yuekang Li, and Leo Zhang. 2026. Agent Skills in the Wild: An Empirical Study of Security Vulnerabilities at Scale. arXiv:2601.10338 [cs.CR] https://arxiv.org/abs/2601.10338

[21] OpenAI. 2025. OpenAI o4-mini Model Documentation. https://platform.openai.com/docs/models. Accessed: 2026-02-07.

[22] OpenClaw. 2026. ClawHub: The Skill Dock for Sharp Agents. https://clawhub.ai/. Accessed: 2026-02-22.

[23] Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. 2025. ReasoningBank: Scaling Agent Self-Evolving with Reasoning Memory. arXiv:2509.25140 [cs.AI] https://arxiv.org/abs/2509.25140

[24] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (Eds.). ACM, 2:1–2:22. https://doi.org/10.1145/3586183.3606763

[25] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=dHng2O0Jjr

[26] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR* abs/2402.07927 (2024). https://doi.org/10.48550/arXiv.2402.07927

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

[28] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. arXiv:2010.03768 [cs.CL] https://arxiv.org/abs/2010.03768

[29] David Silver and Richard S Sutton. 2025. Welcome to the era of experience. *Google AI* 1 (2025).

[30] SkillHub Team. 2026. SkillHub: A Marketplace for Claude and Agent Skills. https://www.skillhub.club/. Accessed: 2026-02-06.

[31] SkillsMP Team. 2026. SkillsMP: Claude Skills Marketplace. https://skillsmp.com/. Accessed: 2026-02-06.

[32] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and Error: Exploration-Based Trajectory Optimization for LLM Agents. arXiv:2403.02502 [cs.CL] https://arxiv.org/abs/2403.02502

[33] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. 1998. Knowledge Engineering: Principles and Methods. *Data Knowl. Eng.* 25, 1-2 (1998), 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6

[34] Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, Ge Zhang, Jiaheng Liu, Xingyao Wang, Sirui Hong, Chenglin Wu, Hao Cheng, Chi Wang, and Wangchunshu Zhou. 2025. Agent KB: Leveraging Cross-Domain Experience for Agentic Problem Solving. arXiv:2507.06229 [cs.CL] https://arxiv.org/abs/2507.06229

[35] Vercel. 2026. Skills.sh: The Open Agent Skills Ecosystem. https://skills.sh/. Accessed: 2026-02-06.

[36] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Trans. Mach. Learn. Res.* 2024 (2024). https://openreview.net/forum?id=ehfRiF0R3a

[37] Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. 2025. Reinforcement Learning for Self-Improving Agent with Skill Library. arXiv:2512.17102 [cs.AI] https://arxiv.org/abs/2512.17102

[38] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. ScienceWorld: Is your Agent Smarter than a 5th Grader? arXiv:2203.07540 [cs.CL] https://arxiv.org/abs/2203.07540

[39] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025. Agent Workflow Memory. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025 (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR / OpenReview.net. https://proceedings.mlr.press/v267/wang25bx.html

[40] Yue Wu, Tianhao Su, Shunbo Hu, and Deng Pan. 2026. Skill-Based Autonomous Agents for Material Creep Database Construction. arXiv:2602.03069 [cs.DB] https://arxiv.org/abs/2602.03069

[41] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. 2025. The rise and potential of large language model based agents: a survey. *Sci. China Inf. Sci.* 68, 2 (2025). https://doi.org/10.1007/s11432-024-4222-0

[42] Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. 2026. SkillRL: Evolving Agents via Recursive Skill-Augmented Reinforcement Learning. arXiv:2602.08234 [cs.LG] https://arxiv.org/abs/2602.08234

[43] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. arXiv:2207.01206 [cs.CL] https://arxiv.org/abs/2207.01206

[44] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[45] Haoran Ye, Xuning He, Vincent Arak, Haonan Dong, and Guojie Song. 2026. Meta Context Engineering via Agentic Skill Evolution. arXiv:2601.21557 [cs.AI] https://arxiv.org/abs/2601.21557

[46] Simon Yu, Gang Li, Weiyan Shi, and Peng Qi. 2025. PolySkill: Learning Generalizable Skills Through Polymorphic Abstraction. arXiv:2510.15863 [cs.CL] https://arxiv.org/abs/2510.15863

[47] Guibin Zhang, Muxin Fu, and Shuicheng Yan. 2025. MemGen: Weaving Generative Latent Memory for Self-Evolving Agents. arXiv:2509.24704 [cs.CL] https://arxiv.org/abs/2509.24704

[48] Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026. MemSkill: Learning and Evolving Memory Skills for Self-Evolving Agents. arXiv:2602.02474 [cs.CL]

https://arxiv.org/abs/2602.02474

[49] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. ExpeL: LLM Agents Are Experiential Learners. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 19632–19642. https://doi.org/10.1609/aaai.v38i17.29936

[50] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023). https://arxiv.org/abs/2303.18223

[51] Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. 2025. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079* (2025). https://arxiv.org/abs/2504.07079

[52] Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and Yu Su. 2025. SkillWeaver: Web Agents can Self-Improve by Discovering and Honing Skills. arXiv:2504.07079 [cs.AI] https://arxiv.org/abs/2504.07079

[53] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory. arXiv:2305.17144 [cs.AI] https://arxiv.org/abs/2305.17144