

# SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks

Xiangyi Li<sup>1</sup> Wenbo Chen<sup>2†</sup> Yimin Liu<sup>3</sup> Shenghan Zheng<sup>4</sup> Xiaokun Chen<sup>5</sup> Yifeng He<sup>6</sup> Yubo Li<sup>7</sup>  
 Bingran You<sup>8</sup> Haotian Shen<sup>9</sup> Jiankai Sun<sup>9</sup> Shuyi Wang<sup>9</sup> Qunhong Zeng<sup>10</sup> Di Wang<sup>11</sup> Xuandong Zhao<sup>8</sup>  
 Yuanli Wang<sup>12</sup> Roey Ben Chaim<sup>13</sup> Zonglin Di<sup>14</sup> Yipeng Gao<sup>15</sup> Junwei He<sup>16</sup> Yizhuo He<sup>7</sup> Liqiang Jing<sup>17</sup>  
 Luyang Kong<sup>9</sup> Xin Lan<sup>18</sup> Jiachen Li<sup>19</sup> Songlin Li<sup>5</sup> Yijiang Li<sup>20</sup> Yueqian Lin<sup>21</sup> Xinyi Liu<sup>9</sup> Xuanqing Liu<sup>9</sup>  
 Haoran Lyu<sup>9</sup> Ze Ma<sup>22</sup> Bowei Wang<sup>9</sup> Runhui Wang<sup>9</sup> Tianyu Wang<sup>9</sup> Wengao Ye<sup>23</sup> Yue Zhang<sup>17</sup>  
 Hanwen Xing<sup>9</sup> Yiqi Xue<sup>15</sup> Steven Dillmann<sup>5</sup> Han-chung Lee<sup>9</sup>

## Abstract

Agent Skills are structured packages of procedural knowledge that augment LLM agents at inference time. Despite rapid adoption, there is no standard way to measure whether they actually help. We present SKILLSBENCH, a benchmark of 86 tasks across 11 domains paired with curated Skills and deterministic verifiers. Each task is evaluated under three conditions: no Skills, curated Skills, and self-generated Skills. We test 7 agent-model configurations over 7,308 trajectories. Curated Skills raise average pass rate by 16.2 percentage points(pp), but effects vary widely by domain (+4.5pp for Software Engineering to +51.9pp for Healthcare) and 16 of 84 tasks show negative deltas. Self-generated Skills provide no benefit on average, showing that models cannot reliably author the procedural knowledge they benefit from consuming. Focused Skills with 2–3 modules outperform comprehensive documentation, and smaller models with Skills can match larger models without them.

## 1. Introduction

Large language models (LLMs) have evolved from text generators into autonomous agents capable of executing complex, multi-step tasks in real-world environments (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023;

†This work was conducted outside the author’s role at Amazon.<sup>1</sup>BenchFlow<sup>2</sup>Amazon<sup>3</sup>Ohio State University<sup>4</sup>Dartmouth College<sup>5</sup>Stanford University<sup>6</sup>UC Davis<sup>7</sup>Carnegie Mellon University<sup>8</sup>UC Berkeley<sup>9</sup>Independent<sup>10</sup>Beijing Institute of Technology<sup>11</sup>Foxconn<sup>12</sup>Boston University<sup>13</sup>Zenity<sup>14</sup>UC Santa Cruz<sup>15</sup>USC<sup>16</sup>ByteDance<sup>17</sup>UT Dallas<sup>18</sup>Michigan State University<sup>19</sup>UT Austin<sup>20</sup>UC San Diego<sup>21</sup>Duke University<sup>22</sup>Columbia University<sup>23</sup>University of Oxford. Correspondence to: Xiangyi Li <xiangyi@benchflow.ai>.

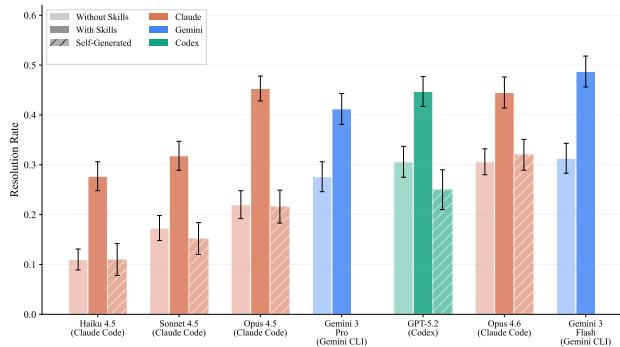
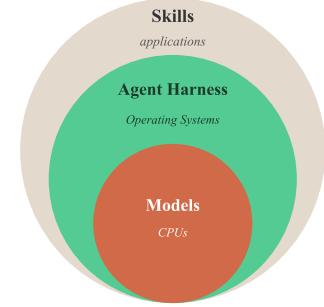


Figure 1. Agent architecture stack and resolution rates across 7 agent-model configurations on 84 tasks. Curated Skills (beige) improve performance by +16.2pp on average; self-generated Skills (amber) provide negligible or negative benefit.

Ouyang et al., 2022; Yao et al., 2022). This evolution is exemplified by agent-centric CLI tools: Claude Code (Anthropic, 2025b) from Anthropic, Gemini CLI (Google, 2025) from Google, and Codex CLI (OpenAI, 2025) from OpenAI enable developers to leverage frontier models as agentic assistants within terminal environments. However, a fundamental tension exists: foundation models provide broad capabilities but lack the procedural knowledge required for domain-specific workflows, while fine-tuning is expensive and sacrifices generality.

Agent Skills offer an emerging solution. A Skill is a struc-

tured package comprising instructions, code templates, resources, and verification logic that augments agent behavior at inference time without model modification (Anthropic, 2025a). Skills encode procedural knowledge: standard operating procedures, domain conventions, and task-specific heuristics that guide agent behavior. This modular approach builds on the options framework for temporal abstraction (Sutton et al., 1999) and cognitive architectures for language agents (Sumers et al., 2023), mirroring successful computing paradigms: foundation models provide base capabilities (analogous to CPUs), agent harnesses orchestrate context and tools (operating systems), and Skills extend competence to specialized domains (applications).

Skills ecosystems have grown rapidly, with community repositories now hosting thousands of user-contributed Skills spanning software engineering, data analysis, and enterprise workflows. Yet despite this proliferation, no benchmark systematically evaluates how and when Skills improve agent performance, what content drives gains, or what design principles distinguish effective Skills from ineffective ones. Figure 1 illustrates this layered architecture and previews our main result: curated Skills consistently improve resolution rates across all 7 model-harness configurations, while self-generated Skills provide negligible benefit. The question is not whether adding task-relevant context helps, but rather: How much do Skills help compared to baseline augmentation? Which Skills components (instructions vs. code vs. examples) contribute most? When do Skills fail despite being present?

Existing agent benchmarks (Liu et al., 2023; Merrill et al., 2026; Jimenez et al., 2024; Zhou et al., 2024b; Xie et al., 2024; Koh et al., 2024; Trivedi et al., 2024; Yang et al., 2023; Chan et al., 2025; Zhuo et al., 2025) evaluate raw model capabilities in isolation, answering “How well can this model perform task X?” but not “How much does Skills Y improve performance on task X?” This gap has practical consequences: practitioners cannot make informed decisions about Skills adoption, and researchers lack empirical grounding for Skills design principles.

To address this, we introduce SKILLSBENCH, the first benchmark that treats Skills as first-class evaluation artifacts, with two core contributions:

- 1. A Skills-centric evaluation framework.** We curate 84 tasks across 11 domains, each executed under three conditions—no Skills, curated Skills, and self-generated Skills—with deterministic verifiers and full trajectory logging. We stratify tasks by difficulty and conduct leakage audits to ensure Skills provide guidance rather than solutions.
- 2. Large-scale empirical evaluation.** We evaluate 7 agent-model configurations across 7,308 trajectories, producing the first systematic evidence on Skills efficacy, varying

ance, and failure modes.

## 2. SKILLSBENCH

We present SKILLSBENCH, a benchmark for evaluating the efficacy of *Skills augmentation* in LLM-based agents. Built on the Harbor framework (Merrill et al., 2026; Harbor Framework Team, 2026), each task adopts a containerized structure with an environment that includes agent Skills and related data, a deterministic verification test, and an oracle solution. Following best practices for agentic benchmarks (Zhu et al., 2025; Anthropic, 2026), we ensure strict isolation and deterministic verification. Unlike Terminal-Bench, which evaluates raw model and agent harness capability, SKILLSBENCH introduces a key methodological difference: we evaluate every task under both *vanilla* (no Skills) and *Skills-augmented* conditions, enabling direct measurement of Skills efficacy.

### 2.1. Skills Specification

A **Skill** is an artifact that satisfies four criteria:

- **Procedural content:** Contains how-to guidance (procedures, workflows, SOPs), not factual retrieval
- **Task-class applicability:** Applies to a class of problems, not a single instance
- **Structured components:** Includes a SKILL.md file plus optional resources (scripts, templates, examples)
- **Portability:** Skills are solely based on file systems, so it’s easy to edit, version, share, and be used across different skills-compatible agent harnesses.

This definition explicitly *excludes*: system prompts (lack structure and resources), few-shot examples (Brown et al., 2020) (declarative, not procedural), RAG retrievals (Lewis et al., 2020) (factual, not procedural), and tool documentation (Schick et al., 2023; Qin et al., 2024) (describes capabilities, not procedures). We acknowledge this boundary is not absolute (for example, a StackOverflow answer may blend factual and procedural content), but our criteria provide operational clarity for benchmark construction. We highlight the distinguishing features of Skills compared to other augmentation paradigms in Table 1.

In SKILLSBENCH, each Skill is a modular package located in `environment/skills/` containing:

- **SKILL.md:** Natural language instructions specifying *how* to approach a class of tasks, i.e., workflows, standard operating procedures, or domain conventions.
- **Resources:** Executable scripts, code templates, reference documentation, or worked examples that the agent may invoke or consult.

*Table 1.* Comparison of runtime augmentation paradigms. Skills uniquely combine modular packaging with procedural guidance and optional executable resources, while remaining portable across models and harnesses.

	Prompts	RAG	Tools	Skills
Modular/reusable	×	✓	✓	✓
Procedural guidance	Limited	×	×	✓
Executable resources	×	×	✓	✓
Cross-model portable	✓	✓	✓	✓

## 2.2. Task Specification

Each task in SKILLSBENCH is a self-contained module comprising four components:

- **Instruction.** A human-readable task description specifying the objective, input format, and expected output. We write instructions to be solvable by a knowledgeable human without access to the paired Skills, though the Skills may substantially reduce time-to-solution.
- **Environment.** A Docker container with task-specific data files and a `skills`/ subdirectory containing modular Skills packages. The container ensures reproducibility through isolated dependencies and clean file system state.
- **Solution.** A reference implementation demonstrating the task’s resolvability. This oracle validates that each task has at least one correct solution path.
- **Verifier.** Deterministic test scripts with programmatic assertions, including numeric tolerances where appropriate. This ensures reproducible pass/fail determination without LLM-as-a-judge variance, following execution-based evaluation best practices (Wang et al., 2023b; Brown, 2025).

## 2.3. Dataset Construction

The expressive and flexible nature of Skills and our task specifications enables broad coverage across diverse domains and problem types. To maximize this diversity, we adopted a **community-driven, open-source contribution model**: 105 contributors from both academia and industry submitted 322 candidate tasks. We count submissions that included the full task specification (instruction, environment, solution, and verifier), along with author-assessed difficulty ratings. From this pool, we curated the final SKILLSBENCH dataset.

## 2.4. Contributing Principles

Contributors must satisfy explicit requirements designed to ensure task quality and prevent shortcuts:

**Human-Authored Instructions.** Task instructions must be written by humans, not generated by language models. We enforce this because LLMs generated queries will be

confined by the distributions of LLMs, which are the subject of our evaluations, and LLM-generated queries are most of the time of low quality.

**Skill Generality.** Skills must provide procedural guidance for a *class* of tasks, not solutions to specific instances. Instructions must not reference which Skills to use, meaning agents must discover and apply Skills autonomously. This ensures we measure genuine Skill utilization rather than instruction-following.

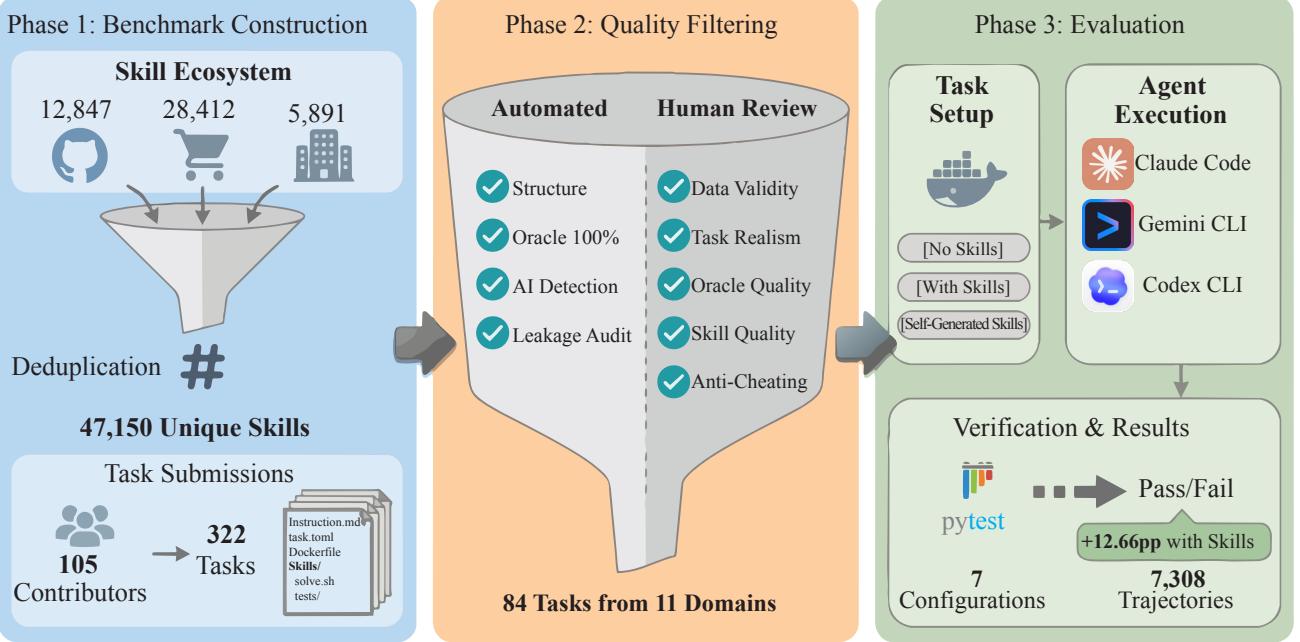
**Deterministic Verification.** All success criteria must be testable through programmatic assertions. We target minimal number of tests needed for verification, avoiding both insufficient coverage and redundant test bloat that leads to artificial low pass rates. Tests must include informative error messages and use parametrization rather than duplication.

**Automated Validation.** Each submission undergoes automated validation before human review:

- **Structural validation:** Required files present (`instruction.md`, `task.toml`, `solve.sh`, `test_outputs.py`), correct directory layout, valid TOML/YAML syntax.
- **Oracle execution:** Reference solution must achieve 100% test pass rate. Tasks with failing oracles are rejected.
- **Instruction quality:** Instruction must be human written (we apply both human review and GPTZero review, and we achieve human label on 100% of our tasks). We also evaluate instructions by six criteria (explicit output paths, structured requirements, success criteria, constraints listed, context-first ordering).

**Human Review.** After automated checks pass, maintainers conduct manual review evaluating five criteria: (1) *data validity*: input data must reflect real-world complexity; synthetic or toy data is rejected unless justified; (2) *task realism*: scenarios must reflect realistic professional workflows without artificial difficulty; (3) *oracle quality*: reference solutions should match how domain experts would solve the task; (4) *Skill quality*: Skills must be error-free, internally consistent, and genuinely useful for similar tasks beyond this benchmark; (5) *anti-cheating*: tasks must prevent shortcut solutions (editing input data, extracting answers from test files, exploiting verifier implementation). Reviewers run benchmark experiments with and without Skills across multiple agents to confirm each task provides meaningful signal about Skill efficacy. Figure 2 provides an end-to-end overview of the three-phase pipeline: benchmark construction, quality filtering, and evaluation.

**Leakage Prevention.** To prevent Skills from encoding task-specific solutions, we enforce explicit authoring guide-



**Figure 2.** SKILLSBENCH pipeline overview. **Phase 1 (Benchmark Construction):** We aggregate Skills from three sources—open-source repositories (12,847), the Claude Code ecosystem (28,412), and corporate partners (5,891)—yielding 47,150 unique Skills after deduplication. In parallel, 322 contributors submit 105 candidate tasks. **Phase 2 (Quality Filtering):** Each task undergoes automated checks (structural validity, AI detection, leakage audit) and human review (data validity, task realism, oracle quality, Skill quality, anti-cheating), producing 84 tasks spanning 11 domains. **Phase 3 (Evaluation):** Tasks are executed under three conditions (no Skills, with curated Skills, self-generated Skills) across three commercial agent harnesses (Claude Code, Gemini CLI, Codex CLI). Deterministic pytest verifiers produce pass/fail outcomes; 7 agent-model configurations yield 7,308 trajectories, with curated Skills providing +12.66pp average improvement.

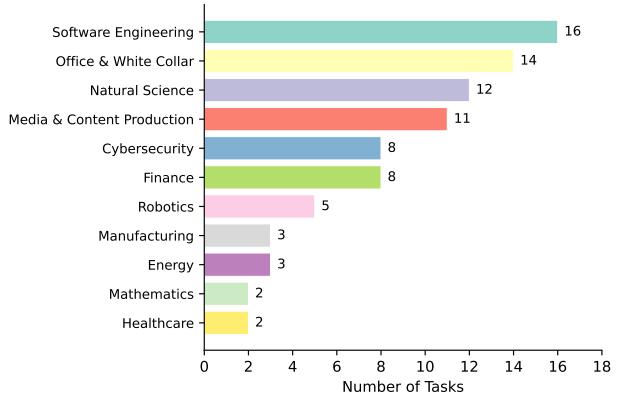
**Table 2.** Task difficulty stratification based on human completion time.

Difficulty	Tasks	Human Time
Core	17 (19.8%)	<60 min
Extended	43 (50.0%)	1–4 hours
Extreme	26 (30.2%)	>4 hours

lines and conduct leakage audits. A Claude Code Agent SDK-based validation agent runs in CI to detect potential Skill-solution leakage; failed tasks are rejected. Skills must NOT contain: task-specific filenames, paths, or identifiers; exact command sequences that solve benchmark tasks; constants, magic numbers, or values from task specifications; references to specific test cases or expected outputs. Skills must be applied to a class of tasks, not a single instance; provide procedural guidance (how to approach), not declarative answers (what to output); and be authored independently of benchmark specifications.

## 2.5. Benchmark Composition

SKILLSBENCH comprises 84 tasks across 11 domains, with category distribution shown in Figure 3. We stratify tasks



**Figure 3.** SKILLSBENCH consists of tasks spanning 11 domains, by difficulty, measured by estimated completion time by individuals whom we consider median specialists for the tasks, without the assistance of AI tools. Original task contributors provided human time estimates, reviewed by an additional set of reviewers from the maintainers who are experts in the same domain.

## 3. Experimental Setup

We evaluate three commercial agent harnesses on SKILLSBENCH across seven frontier models under three Skills con-

ditions, resulting in 7,308 valid trajectories. A trajectory is valid when the agent passes, fails, or times out on a task without infrastructure and runtime errors. Each trajectory is one agent’s attempt at solving a single task under a specific Skills condition.

### 3.1. Agent Harnesses

We evaluate three commercial-line agents: Claude Code ([Anthropic, 2025b](#)), Codex CLI ([OpenAI, 2025](#)), and Gemini CLI ([Google, 2025](#)).

### 3.2. Models

We select seven frontier models: GPT-5.2 ([OpenAI](#)), Claude Opus 4.5, Claude Opus 4.6, Claude Sonnet 4.5, Claude Haiku 4.5 ([Anthropic](#)), Gemini 3 Pro, and Gemini 3 Flash ([Google](#)). All models use temperature 0 for deterministic sampling.

We evaluate each model using its compatible agent harness. Claude Code runs with all four Claude models; Gemini CLI runs with Gemini models; Codex CLI runs with GPT-5.2. This yields 7 model-harness configurations. The full configuration matrix is in [Table 7](#).

### 3.3. Skills Conditions

We evaluate each task under three conditions:

- **No Skills:** Agent receives only `instruction.md`, no Skills present in environment.
- **With Skills:** Complete `environment/skills/` directory with all examples, code snippets, and resources.
- **Self-Generated Skills:** No Skills provided, but the agent is prompted to generate relevant procedural knowledge before solving the task. This isolates the impact of LLMs’ latent domain knowledge.

The self-generated condition is evaluated on 5 of 7 configurations (all Claude Code models and Codex); Gemini CLI does not support this condition.

### 3.4. Evaluation Protocol

We provide Skills as system-level context preceding the task instruction in SKILLSBENCH. We list the injection format and context management details in [Appendix E](#). For each condition, the agent interacts with the containerized environment until task completion, timeout, or round limit. The verifier then executes deterministic assertions to produce a binary pass/fail outcome.

### 3.5. Metrics

**Pass Rate.** The primary metric is pass rate, following the scoring methodology of Terminal-Bench ([Merrill et al.,](#)

2026): for each task, we average the binary reward across 5 trials, then average these task-level scores using a fixed denominator of 84 (the number of evaluated tasks). **Normalized Gain.** Following Hake’s formulation from physics education research ([Hake, 1998](#)), we define normalized gain as:

$$g = \frac{\text{pass}_{\text{skill}} - \text{pass}_{\text{vanilla}}}{1 - \text{pass}_{\text{vanilla}}} \quad (1)$$

**Interpreting Normalized Gain.** Normalized gain has known limitations: a model scoring 90% vanilla and 95% with Skills yields  $g = 0.5$ , identical to a model scoring 10% and 55%. These represent different phenomena (ceiling effects vs. genuine scaffolding). We report both absolute improvement ( $\Delta$ ) and normalized gain ( $g$ ) to enable nuanced interpretation. High  $g$  with low

$$\Delta_{\text{abs}}$$

suggests ceiling effects; high  $g$  with high  $\Delta$  suggests substantial scaffolding. We interpret the claim of “consistent scaffolding efficiency” as similar *proportional* benefit, not identical absolute improvement.

## 4. Results

Our results are twofold: 1. main evaluation three Skills conditions across 7 LLM–agent combinations on 84 tasks, and 2. detailed analysis of Skills design factors including quantity, complexity, and domain effects.

### 4.1. Experiment 1: Skills Efficacy Across LLM–Agent Combinations

We evaluate how curated and self-generated Skills affect agent performance across commercial model-harness configurations. We test each configuration under three conditions on all 84 tasks: no Skills, with curated Skills, and with self-generated Skills (where supported).

#### 4.1.1. MAIN RESULTS

[Table 3](#) presents pass rates across all three conditions for each model-harness combination, ordered by with-Skills performance.

#### Finding 1: Skills provide substantial but variable benefit.

Skills improve performance by +16.2pp on average across 7 model-harness configurations, but with high variance across configurations (range: +13.6pp to +23.3pp). This variability suggests that Skills efficacy depends strongly on the specific agent-model combination, contradicting the assumption of uniform Skills benefits.

#### Finding 2: Gemini CLI + Gemini 3 Flash achieves maximum performance.

The best-performing configura-

**Table 3.** Pass rates (%) and normalized gain ( $g$ ) across three Skills conditions on 84 tasks.  $g$  measures proportional improvement toward perfect performance (Equation 1). Configurations ordered by with-Skills pass rate. “–” indicates condition not evaluated.

Harness	Model	No Skills	Curated Skills		Self-Generated	
			Pass Rate	$g$ (%)	Pass Rate	$g$ (%)
Gemini CLI	Gemini 3 Flash	31.3	48.7	25.3	–	–
Claude Code	Opus 4.5	22.0	45.3	29.9	21.6	-0.5
Codex	GPT-5.2	30.6	44.7	20.3	25.0	-8.1
Claude Code	Opus 4.6	30.6	44.5	20.0	32.0	+2.0
Gemini CLI	Gemini 3 Pro	27.6	41.2	18.8	–	–
Claude Code	Sonnet 4.5	17.3	31.8	17.5	15.2	-2.5
Claude Code	Haiku 4.5	11.0	27.7	18.8	11.0	0.0
<b>Mean</b>		24.3	40.6	21.5	21.0	-1.8

tion is Gemini CLI with Gemini 3 Flash, achieving 48.7% pass rate with Skills. Notably, Claude Code with Opus 4.5 achieves the highest *improvement* (+23.3pp), reflecting Claude Code’s (Anthropic, 2025b) native Skills integration optimized for the Agent Skills specification (Anthropic, 2025a).

**Finding 3: Self-generated Skills provide negligible or negative benefit.** When prompted to generate their own procedural knowledge before solving tasks, models achieve -1.3pp on average compared to the no-Skills baseline. Only Opus 4.6 shows a modest improvement (+1.4pp); Codex + GPT-5.2 degrades substantially (-5.6pp), and remaining models are flat or negative. This contrasts sharply with curated Skills (+16.2pp), demonstrating that effective Skills require human-curated domain expertise that models cannot reliably self-generate. Trajectory analysis reveals two failure modes: (1) models identify *that* domain-specific knowledge is needed but generate imprecise or incomplete procedures (e.g., listing “use pandas for data processing” without specific API patterns), and (2) for high domain-knowledge tasks (manufacturing, financial), models often fail to recognize the need for specialized Skills entirely, attempting solutions with general-purpose approaches.

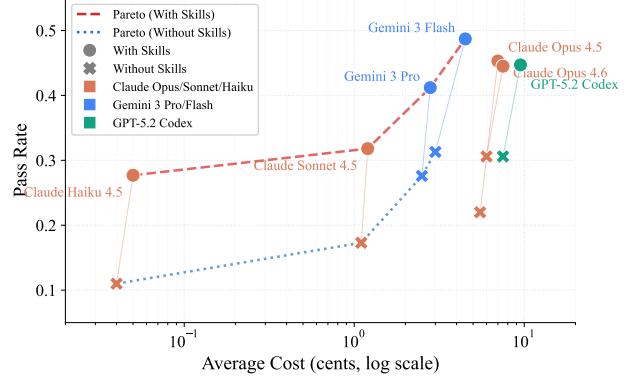
#### 4.1.2. HARNESS-SPECIFIC RELIABILITY

Beyond Skills efficacy, we observe reliability differences across commercial harnesses:

- **Claude Code:** Highest skills utilization rate; improvements range from +13.9pp (Opus 4.6) to +23.3pp (Opus 4.5), with all Claude models consistently benefiting.
- **Gemini CLI:** Highest raw performance (Gemini 3 Flash achieves 48.7% with Skills); improvements range from +13.6pp to +17.4pp.
- **Codex CLI:** Competitive raw performance (44.7% with Skills); frequently *neglects* provided Skills—agents acknowledge Skills content but often implement solutions independently.

#### 4.1.3. DOMAIN-LEVEL ANALYSIS

**Finding 4: Skills benefit varies widely across domains.** Table 4 presents Skill efficacy by domain, revealing substan-



**Figure 4.** Pareto frontier of pass rate vs. cost across model-harness configurations. Filled markers indicate with-Skills conditions; hollow markers indicate without-Skills. Skills shift the Pareto frontier upward, with Gemini 3 Flash and Claude Opus dominating the with-Skills frontier. Cost positions in this figure reflect the evaluation infrastructure’s pricing model. Trajectory analysis reveals that Flash consumes 2.3× more input tokens per task than Pro (1.08M vs. 0.47M), a compensatory strategy where the smaller model substitutes iterative exploration for reasoning depth. At official API pricing (\$0.50 vs. \$2.00 per 1M input tokens), Flash’s 4× lower per-token cost more than offsets this higher volume, making Flash 44% cheaper per task (\$0.55 vs. \$0.98).

**Table 4.** Skills efficacy by domain across 84 evaluated tasks (11 domains). All domains show positive aggregate delta, though individual tasks within each domain may show negative effects.

Domain	With Skills	No Skills	$\Delta_{abs}$
Healthcare	86.1%	34.2%	+51.9
Manufacturing	42.9%	1.0%	+41.9
Cybersecurity	44.0%	20.8%	+23.2
Natural Science	44.9%	23.1%	+21.9
Energy	47.5%	29.5%	+17.9
Office & White Collar	42.5%	24.7%	+17.8
Finance	27.6%	12.5%	+15.1
Media & Content Production	37.6%	23.8%	+13.9
Robotics	27.0%	20.0%	+7.0
Mathematics	47.3%	41.3%	+6.0
Software Engineering	38.9%	34.4%	+4.5

tial heterogeneity. Healthcare (+51.9pp) and Manufacturing (+41.9pp) benefit most, while Mathematics (+6.0pp) and Software Engineering (+4.5pp) show smaller gains. Domains requiring specialized procedural knowledge that is underrepresented in model pretraining (e.g., clinical data harmonization, manufacturing workflows) show the largest improvements, whereas domains with strong pretraining coverage benefit less from external procedural guidance.

#### 4.1.4. TASK-LEVEL ANALYSIS

Analysis of 84 individual tasks reveals high variance in Skills effectiveness:

**Top Skills beneficiaries.** Tasks showing largest improvements: mario-coin-counting (+85.7pp, from 2.9%

to 88.6%), sales-pivot-analysis (+85.7pp), flood-risk-analysis (+77.1pp), sec-financial-report (+74.3pp). These tasks involve specialized procedural knowledge rarely covered in pretraining.

**Skills hurt performance on some tasks.** Despite positive domain-level aggregates, 16 of 84 tasks show negative Skills deltas: taxonomy-tree-merge (-39.3pp), energy-ac-optimal-power-flow (-14.3pp), trend-anomaly-causal-inference (-12.9pp), exoplanet-detection-period (-11.4pp). These failures suggest Skills may introduce conflicting guidance or unnecessary complexity for tasks models already handle well.

## 4.2. Experiment 2: Skills Design Factors

To understand *how* Skills design affects efficacy, we analyze the relationship between Skills quantity, complexity, and performance.

### 4.2.1. SKILLS QUANTITY ANALYSIS

*Table 5.* Pass rates by number of Skills provided. 2–3 Skills shows optimal benefit.

Skills Count	With Skills	No Skills	$\Delta_{abs}$
1 skill	42.2%	24.4%	+17.8
<b>2–3 skills</b>	<b>42.0%</b>	<b>23.4%</b>	<b>+18.6</b>
4+ skills	32.7%	26.9%	+5.9

**Finding 5: 2–3 Skills are optimal; more Skills show diminishing returns.** *Table 5* presents performance stratified by number of Skills provided per task. Tasks with 2–3 Skills show the largest improvement (+18.6pp), while 4+ Skills provide only +5.9pp benefit. This non-monotonic relationship suggests that excessive Skills content creates cognitive overhead or conflicting guidance.

### 4.2.2. SKILLS COMPLEXITY ANALYSIS

*Table 6.* Pass rates by Skills complexity level. Detailed and compact Skills outperform comprehensive ones.

Complexity	Pass Rate	$\Delta_{abs}$	N
Detailed	42.7%	+18.8	1165
Compact	37.6%	+17.1	845
Standard	37.1%	+10.1	773
Comprehensive	39.9%	-2.9	140

**Finding 6: Moderate-length Skills outperform comprehensive ones.** We present the effects of Skills documentation complexity on performance in *Table 6*. Detailed (+18.8pp) and compact (+17.1pp) Skills provide the largest benefit, while comprehensive Skills actually hurt performance (-2.9pp). This suggests that focused procedural

guidance is more effective than exhaustive documentation—agents may struggle to extract relevant information from lengthy Skills content, and overly elaborate Skills can consume context budget without providing actionable guidance.

### 4.2.3. MODEL SCALE EFFECTS

We study the effects of the foundation models’ scale across Claude model family (Opus, Sonnet, Haiku 4.5).

**Finding 7: Smaller model + Skills can exceed larger model without Skills.** Claude Haiku 4.5 with Skills (27.7%) outperforms Haiku without Skills (11.0%) by +16.7pp. Meanwhile, Claude Opus 4.5 without Skills achieves 22.0%. This demonstrates that Skills can partially compensate for model capacity limitations on procedural tasks.

## 5. Discussion

**Skills close procedural gaps.** Skills are most helpful when success depends on concrete procedures and verifier-facing details (steps, constraints, sanity checks), rather than broad conceptual knowledge. We observe large gains on domains with specialized workflows or brittle formats, and smaller or negative effects when models already have strong priors and the Skill adds overhead or conflicts.

**Harnesses mediate Skills use.** Skills efficacy depends not only on Skills quality but also on how the harness implements Skills. Some harnesses reliably retrieve and use Skills, while others frequently acknowledge Skills content but proceed without invoking them. Structured interfaces can also introduce long-trajectory failure modes (e.g., format drift), reducing the influence of early-injected Skills. This motivates evaluating Skills under multiple harnesses rather than treating “with Skills” as a single condition.

**Implications for Skills authoring.** Our analysis suggests that concise, stepwise guidance with at least one working example is often more effective than exhaustive documentation; overly long Skills definitions can increase context burden without improving decisions. Modular Skills also appear to compose better on multi-part tasks, and Skills should explicitly match harness constraints (e.g., repeated format reminders for JSON-only protocols).

### 5.1. Limitations and Future Work

**Coverage and generalization.** SKILLSBENCH focuses on terminal-based, containerized tasks for reproducible evaluation, so results may not directly transfer to GUI agents, multi-agent coordination, or very long-horizon workflows. We also evaluate a limited set of models and harnesses; commercial harness behavior and Skills integration can change over time. A natural extension is to develop multi-modal

skills and protocols for vision-language agents operating in GUI environments.

**Causal attribution and controls.** Skills injection increases context length, so observed gains could partly reflect “more context” rather than procedural structure. Our self-generated Skills condition suggests structure matters—models cannot reliably produce effective procedural guidance despite having the same context budget—but future work requires stronger length-matched baselines (e.g., random/irrelevant text and retrieval-only documentation controls). These baselines also enable studying **automatic Skills synthesis** from demonstrations or documentation and isolating which Skills components (steps, examples, code resources) drive improvements.

**Determinism, contamination, and ecological validity.** Containerization provides state isolation but not perfect determinism or immunity to training-set leakage. We mitigate with multiple runs, a leakage audit (§2.4), and paired (Skills vs. no Skills) comparisons, yet cannot eliminate all nondeterminism or memorization effects. Future work should evaluate **ecosystem-representative settings**, including lower-quality and automatically-selected Skills, and study **Skills composition**—when multiple Skills help or interfere, and whether composite performance can be predicted from atomic Skills effects.

## 6. Related Work

SKILLSBENCH connects to prior work on (1) benchmarking LLM agents, (2) augmenting agents with procedural knowledge and tools, and (3) evaluating improvements across heterogeneous systems.

**Agent benchmarks.** Recent evaluate end-to-end agent capability across realistic environments, including Terminal-Bench (Merrill et al., 2026), SWE-bench and follow-ons (Jimenez et al., 2024; Yang et al., 2024; 2025). Broader environment coverage appears in AgentBench and interactive/web/GUI settings (Liu et al., 2023; Zhou et al., 2024b; Koh et al., 2024; Xie et al., 2024). Other suites emphasize tool-mediated workflows, interactive execution feedback, or domain specialization (Yao et al., 2025; Trivedi et al., 2024; Yang et al., 2023; Chan et al., 2025; Zhang et al., 2024; Zhuo et al., 2025; Austin et al., 2021; Ye et al., 2025). These benchmarks measure how well a fixed agent completes tasks. SKILLSBENCH instead measures *augmentation efficacy* via paired evaluation.

**Procedural augmentation and tool use.** Prior work augments agents with structured reasoning or external knowledge, e.g., CoALA and Voyager (Sumers et al., 2023; Wang et al., 2023a), chain-of-thought and ReAct for multi-step problem solving (Wei et al., 2022; Yao et al., 2023; 2022; Shinn et al., 2023; Madaan et al., 2023; Zhou et al., 2023;

2024a), and retrieval/tool use (Lewis et al., 2020; Zhou et al., 2022; Schick et al., 2023; Qin et al., 2024), and declarative optimization frameworks (Khattab et al., 2023). Skills combine procedural guidance with executable resources (§2.1). Despite many augmentation methods, benchmarks rarely quantify their actual impact.

**Skills ecosystems and evaluation methodology.** Anthropic’s Agent Skills and MCP specifications (Anthropic, 2025a; 2024) formalized skill packages and tool connectivity, while agent CLIs (Claude Code, Gemini CLI, and Codex) provide real-world harnesses (Anthropic, 2025b; Google, 2025; OpenAI, 2025). SKILLSBENCH evaluates both commercial harnesses and a model-agnostic harness based on Terminal-Bench (Merrill et al., 2026) to separate model and harness effects. Finally, broader benchmarking motivates careful reporting and comparability (Mattson et al., 2020; Chiang et al., 2024; Srivastava et al., 2023); we report both absolute gains and normalized gain (Hake, 1998) to compare improvements across different baselines (§3.5).

## 7. Conclusion

We introduced SKILLSBENCH, the first benchmark to systematically evaluate Agent Skills as first-class artifacts. Across 84 tasks, 7 agent-model configurations, and 7,308 trajectories under three conditions (no Skills, curated Skills, self-generated Skills), our evaluation yields four key findings: (1) curated Skills provide substantial but variable benefit (+16.2 percentage points average, with high variance across domains and configurations); (2) self-generated Skills provide negligible or negative benefit (-1.3pp average), demonstrating that effective Skills require human-curated domain expertise; (3) less is more—focused Skills with 2–3 modules outperform comprehensive documentation; and (4) Skills can partially substitute for model scale, enabling smaller models to match larger ones on procedural tasks. These results establish that Skills efficacy is not universal but context-dependent, motivating paired evaluation as standard practice for agent augmentation research. SKILLSBENCH provides both the empirical foundation and open infrastructure for principled Skills design, selection, and deployment.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning through improved evaluation methodology for AI agents. We see no immediate ethical concerns beyond those inherent to agent systems generally. By enabling better evaluation of agent augmentation strategies, our work may contribute to more reliable and capable AI assistants.

## References

- Anthropic. Introducing the model context protocol. <https://www.anthropic.com/news/model-context-protocol>, November 2024. Open standard for connecting AI systems with data sources.
- Anthropic. Equipping agents for the real world with agent skills. <https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills>, October 2025a. Anthropic Engineering Blog.
- Anthropic. Claude code: an agentic coding tool. <https://github.com/anthropics/clause-code>, 2025b.
- Anthropic. Demystifying evals for AI agents. <https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents>, January 2026. Anthropic Engineering Blog.
- Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- Brown, W. Verifiers: Environments for llm reinforcement learning. <https://github.com/PrimeIntell-ect-ai/verifiers>, 2025.
- Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Madry, A., and Weng, L. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=6s5uXNWGIh>.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL <https://dl.acm.org/doi/10.5555/3648699.3648939>.
- Google. Gemini cli: An open-source ai agent that brings the power of gemini directly into your terminal. <https://github.com/google-gemini/gemini-cli>, 2025.
- Hake, R. R. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, 66(1):64–74, 1998.
- Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, January 2026. URL <https://github.com/laude-institute/harbor>.
- Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=VTF8yNQM66>.
- Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.
- Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., and Fried, D. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, 2024.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- Mattson, P., Cheng, C., Diamos, G., Coleman, C., Micikevicius, P., Patterson, D., Tang, H., Wei, G.-Y., Bailis, P., Bittorf, V., et al. Mlperf training benchmark. *Proceedings of Machine Learning and Systems*, 2:336–349, 2020.

- Merrill, M. A., Shaw, A. G., Carlini, N., Li, B., Raj, H., Bercovich, I., Shi, L., Shin, J. Y., Walshe, T., Buchanan, E. K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang, M., Nezhurina, M., Jitsev, J., Lu, D., Mastromichalakis, O. M., Xu, Z., Chen, Z., Liu, Y., Zhang, R., Chen, L. L., Kashyap, A., Uslu, J.-L., Li, J., Wu, J., Yan, M., Bian, S., Sharma, V., Sun, K., Dillmann, S., Anand, A., Lanpouthakoun, A., Koopah, B., Hu, C., Guha, E., Dreiman, G. H. S., Zhu, J., Krauth, K., Zhong, L., Muennighoff, N., Amanfu, R., Tan, S., Pimpalgaonkar, S., Aggarwal, T., Lin, X., Lan, X., Zhao, X., Liang, Y., Wang, Y., Wang, Z., Zhou, C., Heineman, D., Liu, H., Trivedi, H., Yang, J., Lin, J., Shetty, M., Yang, M., Omi, N., Raoof, N., Li, S., Zhuo, T. Y., Lin, W., Dai, Y., Wang, Y., Chai, W., Zhou, S., Wahdany, D., She, Z., Hu, J., Dong, Z., Zhu, Y., Cui, S., Saiyed, A., Kolbeinsson, A., Hu, J., Ryttig, C. M., Marten, R., Wang, Y., Dimakis, A., Konwinski, A., and Schmidt, L. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. URL <https://arxiv.org/abs/2601.11868>.
- OpenAI. Codex cli: Lightweight coding agent that runs in your terminal. <https://github.com/openai/codex>, 2025.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Pan, M. Z., Cemri, M., Agrawal, L. A., Yang, S., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Ramchandran, K., Klein, D., et al. Why do multiagent systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.
- Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., li, d., Liu, Z., and Sun, M. Toolllm: Facilitating large language models to master 16000+ real-world apis. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/file/28e50ee5b72e90b50e7196fde8ea260e-Paper-Conference.pdf](https://proceedings.iclr.cc/paper_files/paper/2024/file/28e50ee5b72e90b50e7196fde8ea260e-Paper-Conference.pdf).
- Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.
- Sumers, T., Yao, S., Narasimhan, K. R., and Griffiths, T. L. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.
- Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Trivedi, H., Khot, T., Hartmann, M., Manku, R., Dong, V., Li, E., Gupta, S., Sabharwal, A., and Balasubramanian, N. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850. URL <https://aclanthology.org/2024.acl-long.850/>.
- Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.
- Wang, Z., Zhou, S., Fried, D., and Neubig, G. Execution-based evaluation for open-domain code generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1271–1290, 2023b.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

- Yang, J., Prabhakar, A., Narasimhan, K., and Yao, S. Intercode: Standardizing and benchmarking interactive coding with execution feedback. *Advances in Neural Information Processing Systems*, 36:23826–23854, 2023.
- Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., and Press, O. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.
- Yang, J., Lieret, K., Jimenez, C. E., Wettig, A., Khandpur, K., Zhang, Y., Hui, B., Press, O., Schmidt, L., and Yang, D. Swe-smith: Scaling data for software engineering agents. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025 D&B Spotlight)*, 2025. URL <https://arxiv.org/abs/2504.21798>. arXiv:2504.21798, accepted at NeurIPS 2025 (Spotlight).
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. R.  $\tau$ -bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=r0NSXZpUDN>.
- Ye, C., Yuan, S., Cooray, S., Dillmann, S., Roque, I. L. V., Baron, D., Frank, P., Martin-Alvarez, S., Koblischke, N., Qu, F. J., Yang, D., Wechsler, R., and Ciucu, I. Replicationbench: Can ai agents replicate astrophysics research papers?, 2025. URL <https://arxiv.org/abs/2510.24591>.
- Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., Jones, E., Hussein, G., Liu, S., Jasper, D., et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*, 2024.
- Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. Language agent tree search unifies reasoning acting and planning in language models. In *International Conference on Machine Learning (ICML)*, 2024a. arXiv:2310.04406.
- Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., and Chi, E. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. arXiv:2205.10625.
- Zhou, S., Alon, U., Xu, F. F., Jiang, Z., and Neubig, G. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations*, 2022.
- Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., and Neubig, G. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=oKn9c6ytLx>.
- Zhu, Y., Jin, T., Pruksachatkun, Y., Zhang, A. K., Liu, S., Cui, S., Kapoor, S., Longpre, S., Meng, K., Weiss, R., Barez, F., Gupta, R., Dhamala, J., Merizian, J., Giulianelli, M., Coppock, H., Ududec, C., Kellermann, A., Sekhon, J. S., Steinhardt, J., Schwettmann, S., Narayanan, A., Zaharia, M., Stoica, I., Liang, P., and Kang, D. Establishing best practices in building rigorous agentic benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL <https://openreview.net/forum?id=E58HNCqoaA>.
- Zhuo, T. Y., Chien, V. M., Chim, J., Hu, H., Yu, W., Widayarsi, R., Yusuf, I. N. B., Zhan, H., He, J., Paul, I., Brunner, S., GONG, C., Hoang, J., Zebaze, A. R., Hong, X., Li, W.-D., Kaddour, J., Xu, M., Zhang, Z., Yadav, P., Jain, N., Gu, A., Cheng, Z., Liu, J., Liu, Q., Wang, Z., Lo, D., Hui, B., Muennighoff, N., Fried, D., Du, X., de Vries, H., and Werra, L. V. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=YrycTj11L0>.

## A. Skill Ecosystem Analysis

To contextualize SKILLSBENCH within the broader landscape of agent augmentation, we analyze the existing ecosystem of publicly available Skills.

### A.1. Data Collection

We collected Skills from three sources:

- Public GitHub repositories tagged with “claude-skills” or “agent-skills” (N=12,847)
- Community marketplaces including Smithery.ai and skillmp.com (N=28,412)
- Corporate partner contributions (N=5,891)

After deduplication (based on SKILL.md content hash), we retained 47,150 unique Skills from 6,323 repositories.

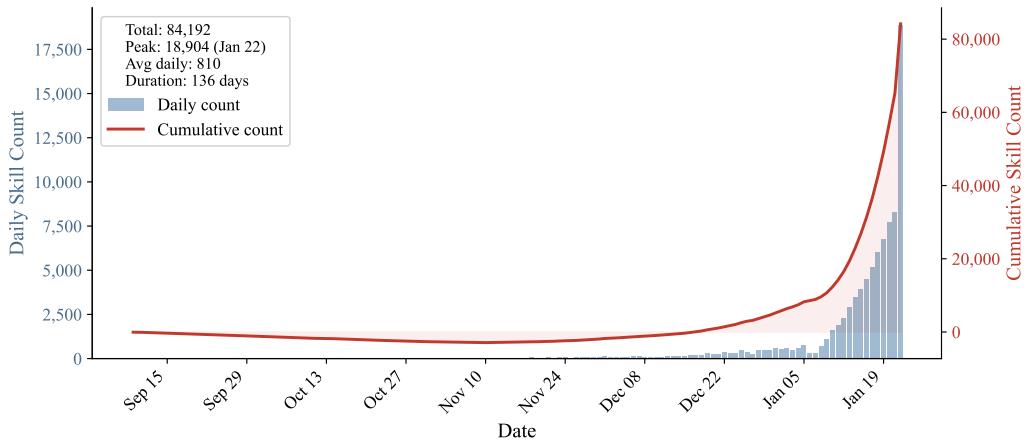


Figure 5. Temporal dynamics of Skill creation over 136 days. Daily additions (bars, left axis) remained modest through late 2025, then surged to a peak of 18,904 in January 2026. The cumulative curve (line, right axis) reflects exponential-like growth, reaching 84,192 total Skills.

### A.2. Skill Characteristics

**Size Distribution.** Skill sizes follow a log-normal distribution with median 2.3 KB (IQR: 0.8–6.1 KB). The largest Skills (top 1%) exceed 50 KB and typically include extensive code resources. Figure 6 shows the SKILL.md token distribution, and Figure 7 shows total Skill size distribution.

**Domain Coverage.** Skills span diverse domains (Figure 8):

- Software Development: 38% (git workflows, code review, testing)
- Data Analysis: 22% (pandas, SQL, visualization)
- DevOps/Infrastructure: 15% (Docker, Kubernetes, CI/CD)
- Writing/Documentation: 12% (technical writing, API docs)
- Other: 13% (scientific computing, finance, etc.)

**Structural Patterns.** Most Skills (78%) follow the standard structure with SKILL.md plus optional resources. Figure 9 shows that most Skills contain very few files (median of one, concentrated below five). Figure 10 confirms the ecosystem is documentation-heavy: markdown files dominate, followed by scripting and configuration code.

### A.3. Quality Indicators

We developed a quality scoring rubric based on:

1. **Completeness:** Presence of required components (0–3 points)

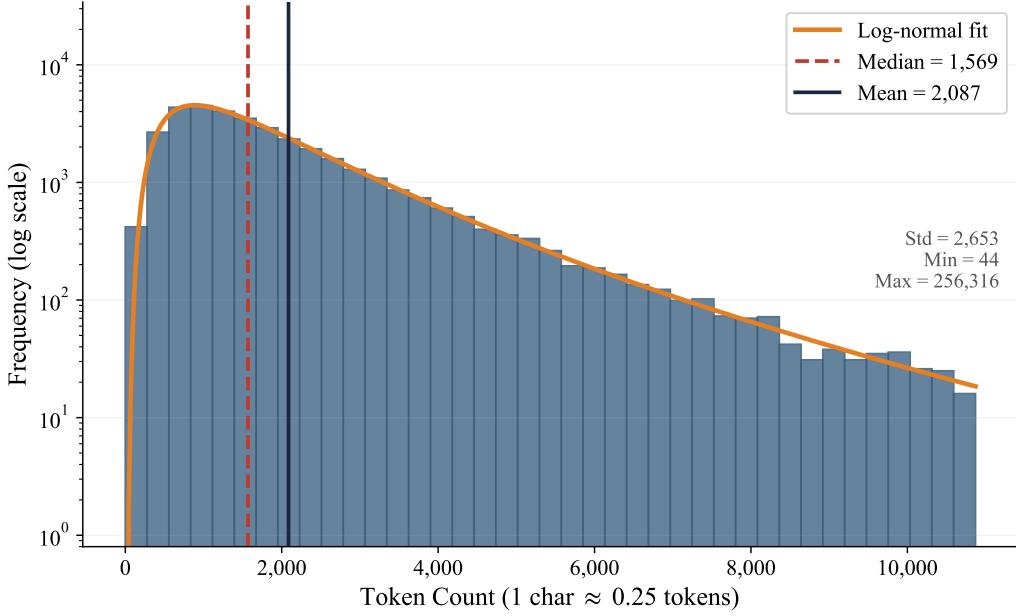


Figure 6. Token distribution of SKILL.md files (n=36,338, 99.5th percentile shown). Most Skills are lightweight with median  $\sim 1.5$ k tokens.

2. **Clarity:** Readability and organization (0–3 points)
3. **Specificity:** Actionable vs. vague guidance (0–3 points)
4. **Examples:** Presence and quality of examples (0–3 points)

Mean quality score across the ecosystem is 6.2/12 ( $SD=2.8$ ), indicating substantial room for improvement in Skill authoring practices.

#### A.4. Implications for Benchmark Design

This ecosystem analysis directly informed SKILLSBENCH construction:

- **Domain selection:** Task categories mirror ecosystem coverage, ensuring Skills exist for evaluation
- **Quality awareness:** Ecosystem mean quality of 6.2/12 motivated our leakage audit and authoring guidelines—low-quality Skills would confound efficacy measurement
- **Skill selection:** We selected benchmark Skills from the top quality quartile (score  $\geq 9/12$ ) to isolate the effect of procedural knowledge from Skill quality variance
- **Size constraints:** Median Skill size ( $\sim 800$  tokens) informed our 8K context budget allocation

**Limitation: Benchmark vs. Ecosystem Gap.** Our 84 evaluated tasks with high-quality Skills represent an optimistic scenario. Real-world Skill usage involves lower-quality Skills (ecosystem mean: 6.2/12 vs. benchmark mean: 10.1/12) and imperfect Skill-task matching. Future work should evaluate with ecosystem-representative Skill samples.

## B. Task Specification and Review Process

This appendix provides full details of the task specification format and quality control process summarized in §2.

### B.1. Task Directory Structure

Each task is a self-contained directory with the following layout:

```
tasks/<task-id>/
  instruction.md      # Task instructions for the agent
```

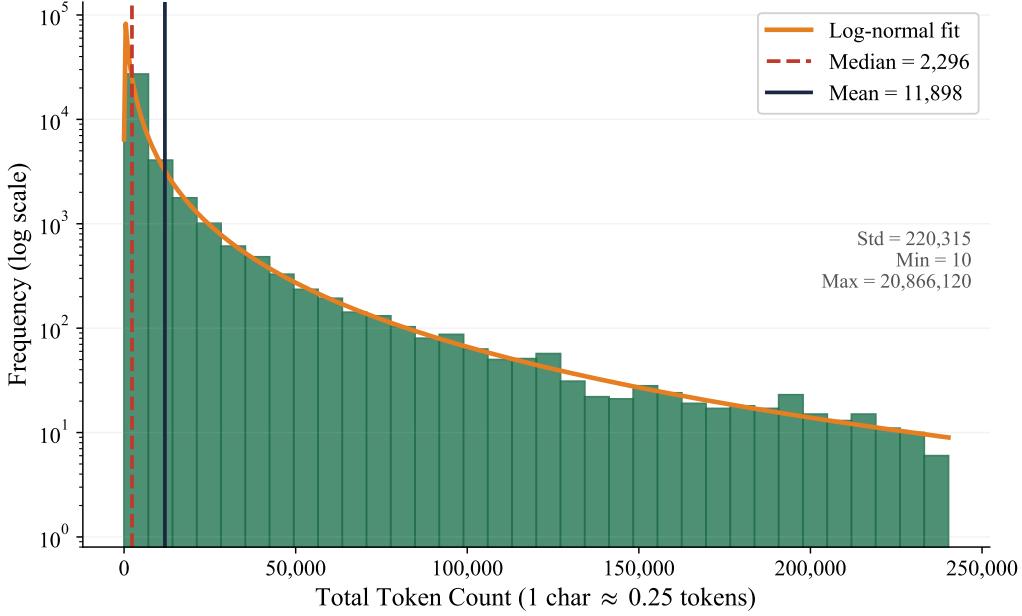


Figure 7. Total Skill size distribution ( $n=37,078$ , 99.5th percentile shown, excluding `metadata.json`). Median total size remains under 2.5k tokens, with distribution highly skewed toward concise artifacts.

```

task.toml          # Metadata and resource configuration
environment/
Dockerfile        # Container setup
skills/            # Curated Skills (absent in no-Skills)
<skill-name>/
  SKILL.md       # Required per skill
  scripts/        # Optional executable code
  references/     # Optional reference documentation
solution/
  solve.sh       # Oracle solution (must pass 100%)
tests/
  test.sh         # Runs pytest inside the container
  test_outputs.py # Programmatic assertions

```

## B.2. Task Configuration (`task.toml`)

Each task specifies metadata and resource limits in TOML format:

```

version = "1.0"

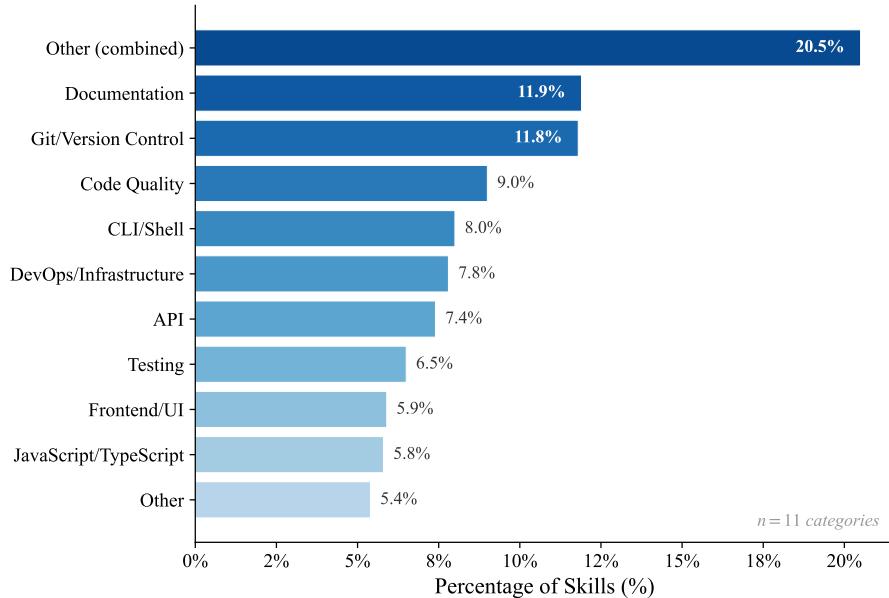
[metadata]
author_name = "Contributor Name"
author_email = "email@example.com"
difficulty = "medium"      # easy | medium | hard
category = "finance"
tags = ["pandas", "data-analysis", "spreadsheet"]

[verifier]
timeout_sec = 900.0        # 600-900s typical

[agent]
timeout_sec = 900.0        # 600-1200s typical

[environment]

```



*Figure 8.* Distribution of Skill categories. The top 10 categories account for 79.6% of all Skills, with Documentation (11.9%), Git/Version Control (11.8%), and Code Quality (9.0%) leading. No single category dominates, reflecting diverse developer needs across documentation, infrastructure, testing, and frontend tasks.

```
build_timeout_sec = 600.0
cpus = 1           # 1-4 cores
memory_mb = 4096   # 2048-10240 MB
storage_mb = 10240 # 10 GB standard
```

### B.3. Verification Infrastructure

Verification uses pytest with the CTRF (Common Test Report Format) output. The `test.sh` script installs dependencies, runs pytest, and writes a binary reward:

```
#!/bin/bash
pip3 install --break-system-packages pytest pytest-ctrf
mkdir -p /logs/verifier
pytest --ctrf /logs/verifier/ctrf.json \
    /tests/test_outputs.py -rA -v
if [ $? -eq 0 ]; then
    echo 1 > /logs/verifier/reward.txt
else
    echo 0 > /logs/verifier/reward.txt
fi
exit 0
```

A task passes if and only if all assertions in `test_outputs.py` succeed (reward = 1). Partial credit is not awarded in the main evaluation.

### B.4. PR Review Process

Each submitted task undergoes a multi-stage review:

1. **Automated CI:** Structural validation (`harbor tasks check`), oracle execution (`harbor run -a oracle`, must pass 100%), and AI-detection screening (GPTZero) on `instruction.md`.
2. **Maintainer review:** Evaluates data validity, task realism, oracle quality, Skill quality, and anti-cheating robustness. Reviewers run benchmark experiments with and without Skills across multiple agents.

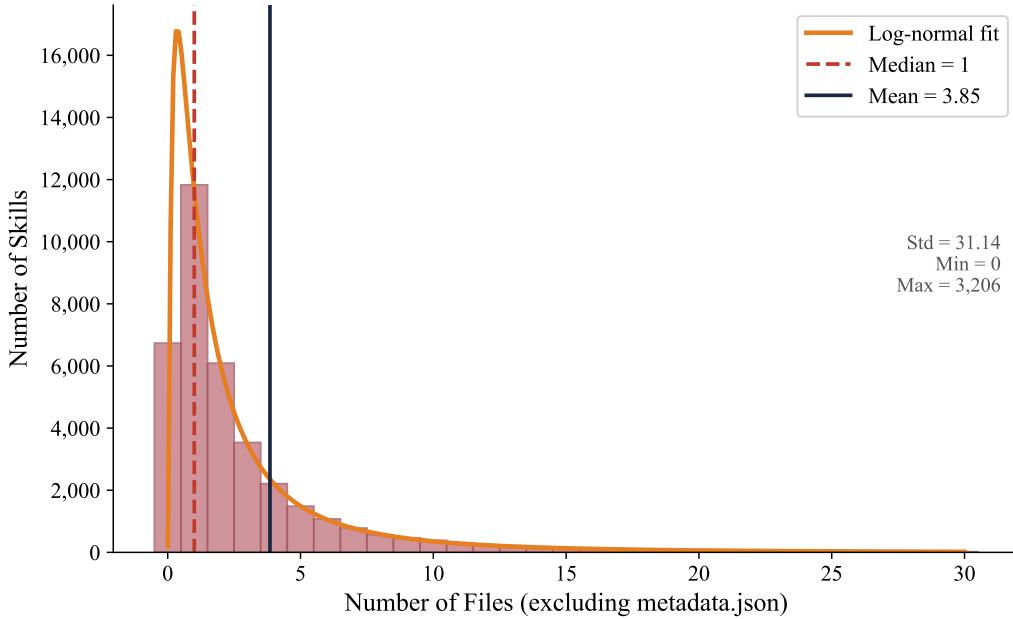


Figure 9. File count distribution per Skill. Most Skills contain 1–5 files.

3. **Benchmark report:** For each task, reviewers produce a structured report documenting oracle results, agent pass rates with and without Skills, failure analysis, and a final verdict (approve, major changes needed, or reject).

Of 322 candidate submissions from 105 contributors, 86 tasks passed all review stages and were included in the final benchmark (26.7% acceptance rate).

## B.5. Contributor Checklist

All task submissions must satisfy the following requirements before review:

- `instruction.md` is human-written (not model-generated); verified via GPTZero and human review
- Skills are error-free, factually correct, and generalizable to similar tasks
- Task is realistic and grounded in professional workflows
- Task requires domain knowledge and is significantly easier with Skills
- Outputs are deterministic and verifiable with programmatic assertions
- Test count is below 10 unless justified; tests cover distinct criteria
- Oracle solution passes 100% (`harbor run -a oracle`)
- Task tested with agent both with and without Skills

## B.6. Maintainer Review Policy

Maintainers evaluate each submission against seven criteria:

1. **AI detection:** Verify `instruction.md` and `task.toml` are manually written using GPTZero and human review. PRs with intentional grammar errors designed to circumvent AI detectors are closed.
2. **Data quality:** Data must be real-world and appropriately complex. AI-generated or toy data is rejected.
3. **Task validity:** Tasks must be grounded in realistic professional scenarios. Artificially inflated complexity is rejected.
4. **Oracle quality:** Simple solutions (e.g., an Excel formula or short script) are preferred over over-engineered oracle implementations.
5. **Author history:** Authors flagged multiple times across PRs are closed automatically.
6. **Test parsimony:** Fewer than 10 test cases unless justified; tests should cover distinct criteria rather than repeat similar checks.

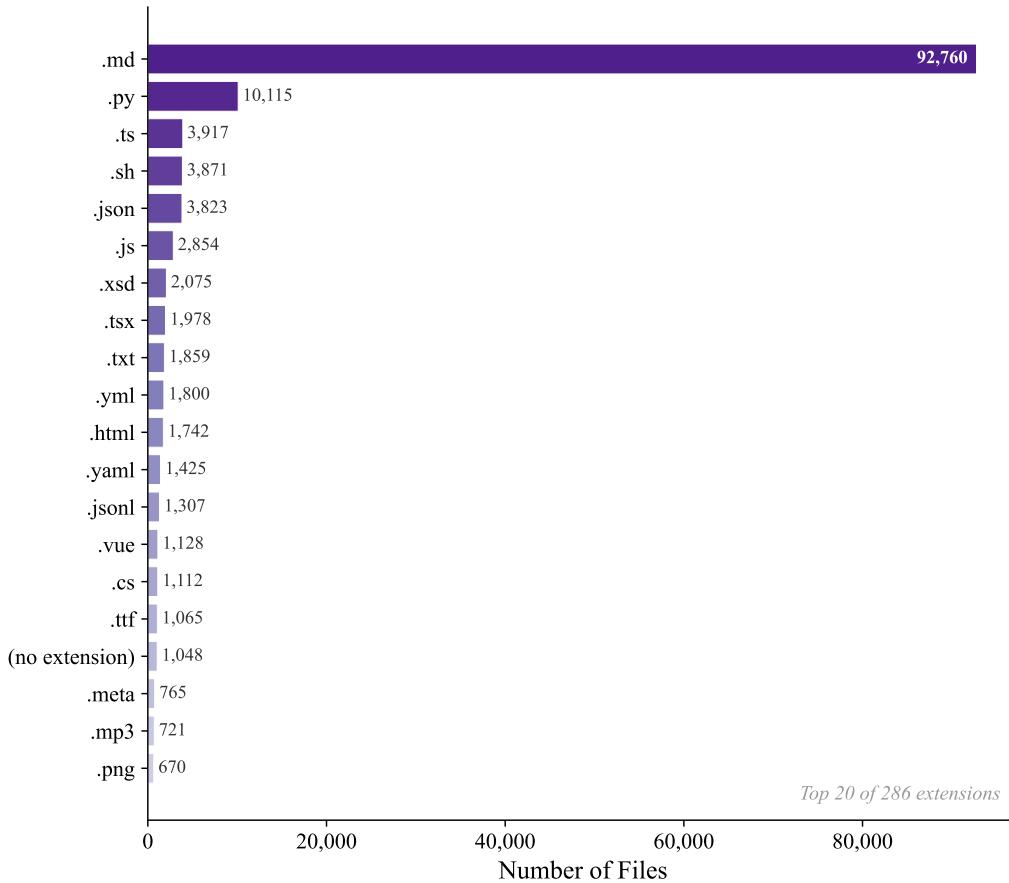


Figure 10. File extension distribution. Markdown files dominate, indicating Skills prioritize natural-language instructions over executable implementations.

7. **Multimodal verification:** For multimodal tasks (audio, PPTX, video, PDF), maintainers personally inspect agent output to verify correctness beyond programmatic assertions.

## B.7. Automated CI Pipeline

The CI pipeline performs the following checks on each PR:

- **Structural validation** (`harbor tasks check`): Verifies required files exist, TOML schema is valid, Dockerfile builds, and test structure is correct.
- **Oracle execution** (`harbor run -a oracle`): Runs the oracle solution end-to-end and requires 100% test pass rate.
- **AI-detection screening**: Runs GPTZero on `instruction.md` to flag potential model-generated content.
- **LLM-backed quality checks**: Automated verification of behavior consistency between instructions and tests, anti-cheating measures, pinned dependency checks, typo detection, and hardcoded solution detection.

## B.8. Benchmark Report Template

For each task, reviewers produce a structured report documenting:

1. **Task metadata:** Name, category, difficulty, tags, description, Skills provided, key requirements.
2. **Oracle results:** Pass/fail status, reward, tests passed, timing.
3. **Agent results:** Pass rates per agent-model combination, with and without Skills, including execution time.

4. **Skills impact:** Quantified comparison of with-Skills vs. without-Skills performance per agent.
5. **Failure analysis:** Per-test breakdown of failures including actual vs. expected output, root cause, and evidence from trajectories.
6. **Recommendation:** One of: APPROVE, APPROVE WITH CAVEATS, MAJOR CHANGES NEEDED, or REJECT.

## B.9. Review Lifecycle

PRs progress through a defined label-based lifecycle:

1. **WIP → Need review:** Author signals readiness for initial review.
2. **Need review → Reviewing:** A maintainer begins active review and benchmark experiments.
3. **Reviewing → Change requested / Major change needed / Critical change needed:** Issues identified; author must address.
4. **Change requested → Take another look:** Author responds after changes.
5. **Ready to merge → Good task:** All reviews passed; task included in benchmark.

Critical changes include unrealistic task scenarios, AI-generated instructions, or synthetic data. Major changes include incorrect tests, unreliable verifiers, or poor Skill quality requiring re-evaluation.

## B.10. Task Quality Criteria

Tasks are evaluated against the following criteria:

- **Realistic:** Grounded in professional workflows that people in that domain actually perform
- **Skill-dependent:** Significantly easier with Skills than without—tasks solvable without any procedural guidance are rejected
- **Verifiable:** Deterministic outputs testable with programmatic assertions; LLM-as-judge is not used
- **Composable:** Tasks should exercise 3–6 Skills together; instructions never reference which Skills to use
- **Test parsimony:** Fewer than 10 test cases unless justified; tests should cover distinct criteria rather than repeat similar checks

## C. Experimental Setup Details

This appendix provides full details of the experimental setup summarized in §3.

### C.1. Model and Harness Configurations

Table 7 presents all 7 agent-model configurations evaluated in the main experiments.

*Table 7.* Agent harnesses and models evaluated. Total: 7,308 valid trajectories across 7 configurations and 3 conditions (no Skills, with Skills, self-generated Skills). Self-generated condition evaluated on Claude Code and Codex configurations only.

	<b>Harness</b>	<b>Model</b>	<b>Provider</b>	<b>Runs</b>
Claude Code	Opus 4.5	Anthropic	1092	
	Opus 4.6	Anthropic	1260	
	Sonnet 4.5	Anthropic	1092	
	Haiku 4.5	Anthropic	1092	
Gemini CLI	Gemini 3 Pro	Google	840	
	Gemini 3 Flash	Google	840	
Codex	GPT-5.2	OpenAI	1092	

### C.2. Harness Descriptions

We evaluate three commercial agent harnesses:

- **Claude Code** ([Anthropic, 2025b](#)): Anthropic’s agent with native Skill integration
- **Gemini CLI** ([Google, 2025](#)): Google’s open-source terminal agent

- **Codex CLI** (OpenAI, 2025): OpenAI’s lightweight coding agent

These tightly couple specific models with proprietary agent logic, representing real-world deployment conditions.

**Model Family Consideration.** Claude models have been trained with awareness of the Agent Skills specification (Anthropic, 2025a), which may confer advantages when processing Skill-formatted instructions.

### C.3. Agent Interface

Agents interact with the environment through a standardized interface:

```
class BaseAgent(ABC):
    @abstractmethod
    def step(self, obs: str) -> str:
        """obs: terminal output -> action"""
        pass
```

### C.4. Skill Injection Mechanism

Skills are injected into each task’s Docker container by copying the `environment/skills/` directory to agent-specific paths. Each Dockerfile includes:

```
# Copy skills to agent-specific locations
COPY skills /root/.claude/skills      # Claude Code
COPY skills /root/.codex/skills        # Codex CLI
COPY skills /root/.gemini/skills       # Gemini CLI
COPY skills /root/.agents/skills      # Portable agents
```

Each agent harness discovers and loads Skills from its respective directory at runtime using its native skill scanning mechanism. Claude Code and Codex read the `SKILL.md` frontmatter (name and description) to determine relevance; Gemini CLI exposes an `activate_skill` tool that agents invoke explicitly. Instructions never reference which Skills to use—agents must discover and apply them autonomously.

### C.5. Skill Structure

Each Skill is a directory containing a required `SKILL.md` file with YAML frontmatter and optional bundled resources:

```
skill-name/
  SKILL.md          # Required: YAML frontmatter + instructions
  scripts/          # Optional: executable code
  references/       # Optional: reference documentation
```

The `SKILL.md` frontmatter specifies the Skill’s name and a one-line description used by agents for skill discovery. The body contains procedural guidance, code examples, and usage patterns.

### C.6. Self-Generated Skills Condition

For the self-generated condition, tasks are identical to the no-Skills baseline except that the following prompt is appended to each `instruction.md`:

#### Important: Generate Skills First

Before attempting to solve this task, please follow these steps:

1. Analyze the task requirements and identify what domain knowledge, APIs, or techniques are needed.
2. Write 1–5 modular skill documents that would help solve this task. Each skill should: focus on a specific tool, library, API, or technique; include installation/setup instructions if applicable; provide code examples and usage patterns; be reusable for similar tasks.
3. Save each skill as a markdown file in the `environment/skills/` directory with a descriptive name.
4. Then solve the task using the skills you created as reference.

The `environment/skills/` directory is empty at the start—agents must populate it before solving the task. No curated Skills are provided. The self-generated condition is evaluated on Claude Code (all four models) and Codex (GPT-5.2) only; Gemini CLI does not support this condition due to its explicit skill-activation interface.

### C.7. Software and Model Versions

Table 8 lists the exact model identifiers and harness versions used in all experiments.

*Table 8.* Model API identifiers and harness versions.

Display Name	API Model ID	Harness Version
Claude Opus 4.5	claude-opus-4-5@20251101	Claude Code 2.1.19
Claude Opus 4.6	claude-opus-4-6	Claude Code 2.1.19
Claude Sonnet 4.5	claude-sonnet-4-5@20250929	Claude Code 2.1.19
Claude Haiku 4.5	claude-haiku-4-5@20251001	Claude Code 2.1.19
GPT-5.2	openai/gpt-5.2-codex	Codex CLI
Gemini 3 Pro	gemini/gemini-3-pro-preview	Gemini CLI
Gemini 3 Flash	gemini/gemini-3-flash-preview	Gemini CLI

### C.8. Container Environment

All tasks run in Docker containers built from an `ubuntu:24.04` base image. Per-task resource allocation is specified in `task.toml`:

- **CPUs:** 1–4 cores (task-dependent)
- **Memory:** 2–10 GB (task-dependent)
- **Storage:** 10 GB (standard across all tasks)
- **GPU:** None (no tasks require GPU)

Containers are deleted after each trial (`delete: true`) to ensure no state leaks between runs.

### C.9. Inference Configuration

- **Temperature:** 0 (deterministic sampling)
- **Max rounds:** Level-dependent (10/30/50 for Core/Extended/Extreme)
- **Context management:** Sliding window with 8K token limit; oldest turns dropped when exceeded
- **Timeout:** Per-task, specified in `task.toml` (range: 600–1200s)

### C.10. Experiment Orchestration

- **Runs per task:** 5 for main conditions (no Skills, with Skills); 3 for self-generated condition
- **Concurrent trials:** 256 (main conditions); 128 (self-generated)
- **Retry policy:** No retries for main conditions; up to 3 retries for self-generated (excluding `VerifierTimeoutError`, `BadRequestError`, `RateLimitError`, `AgentTimeoutError`)
- **Excluded tasks:** `mhc-layer-impl` (GPU requirement) and `fix-visual-stability` (verifier timeout) are excluded from all experiments

The evaluation comprises 7,308 valid trajectories across 7 configurations, 84 tasks, and 3 conditions. We report mean pass rates with 95% bootstrap confidence intervals.

## D. Task-Level Results

Figure 11, Figure 12, and Figure 13 show the task-level pass rates per model across all 84 evaluated tasks. Results are averaged over 5 trials per task-model-condition combination. Tasks (rows) are sorted by average with-Skills pass rate; models (columns) are sorted by aggregate with-Skills score.

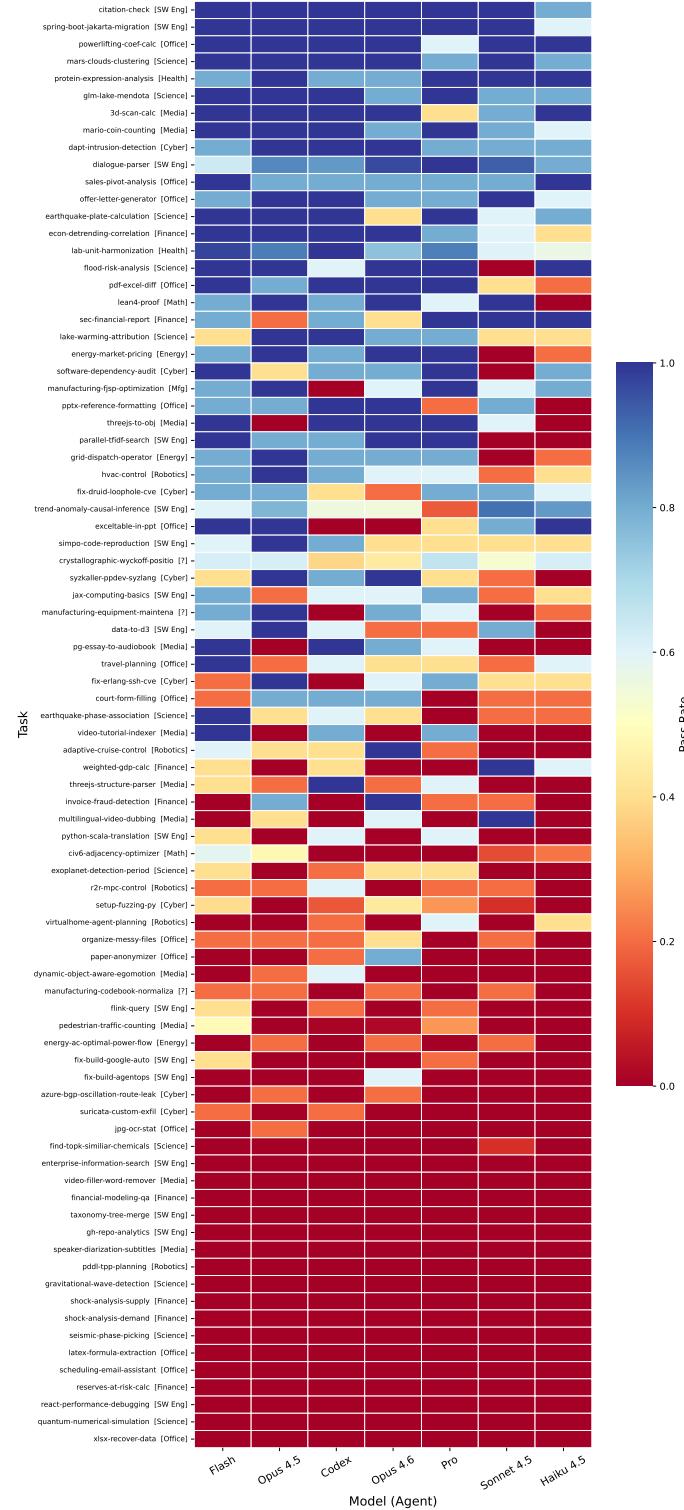


Figure 11. Task pass rate per model with curated Skills. The grid reveals a common set of easy tasks (top, uniformly blue) solved by all models, and hard tasks (bottom, uniformly red) unsolved even with Skills. Tasks along the diagonal transition from solvable to unsolvable as model capability decreases.

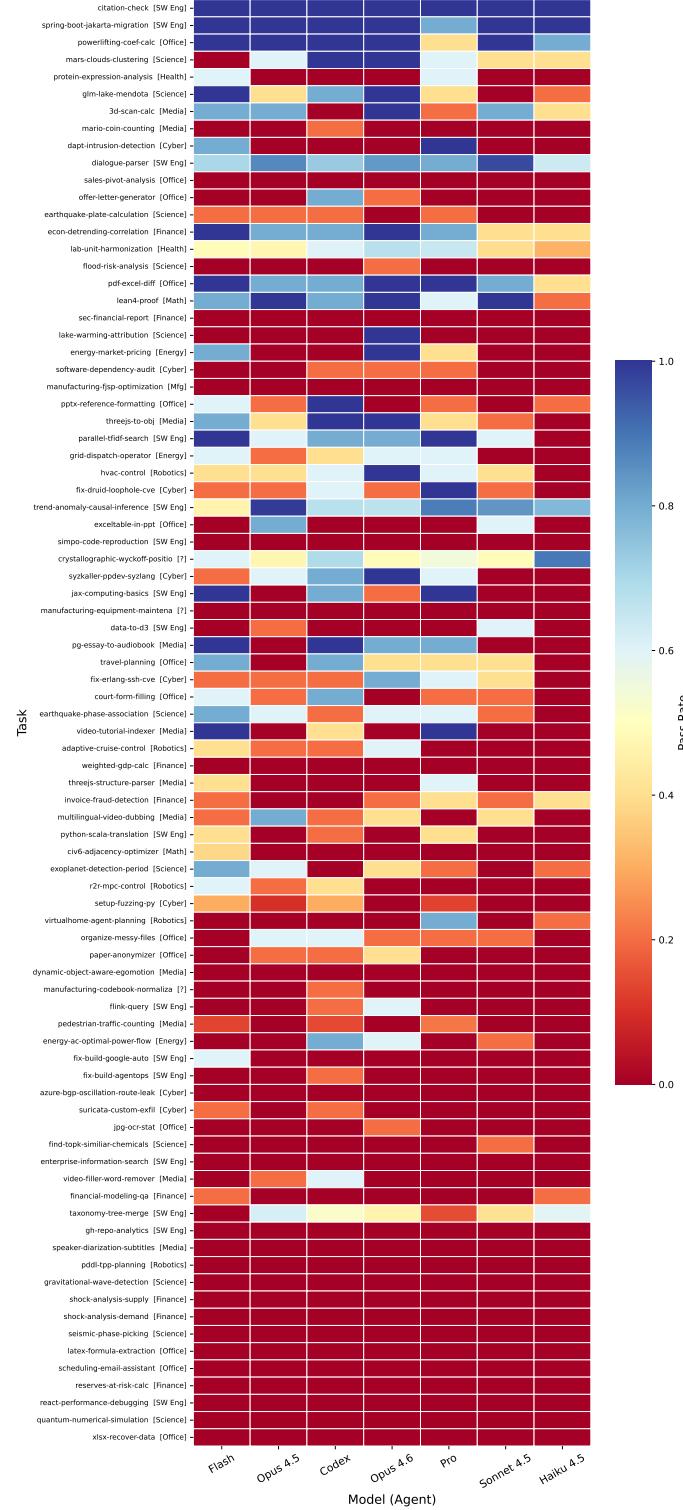


Figure 12. Task pass rate per model without Skills (baseline). Compared to Figure 11, the blue region contracts substantially, confirming that Skills shift many tasks from unsolved to solved.

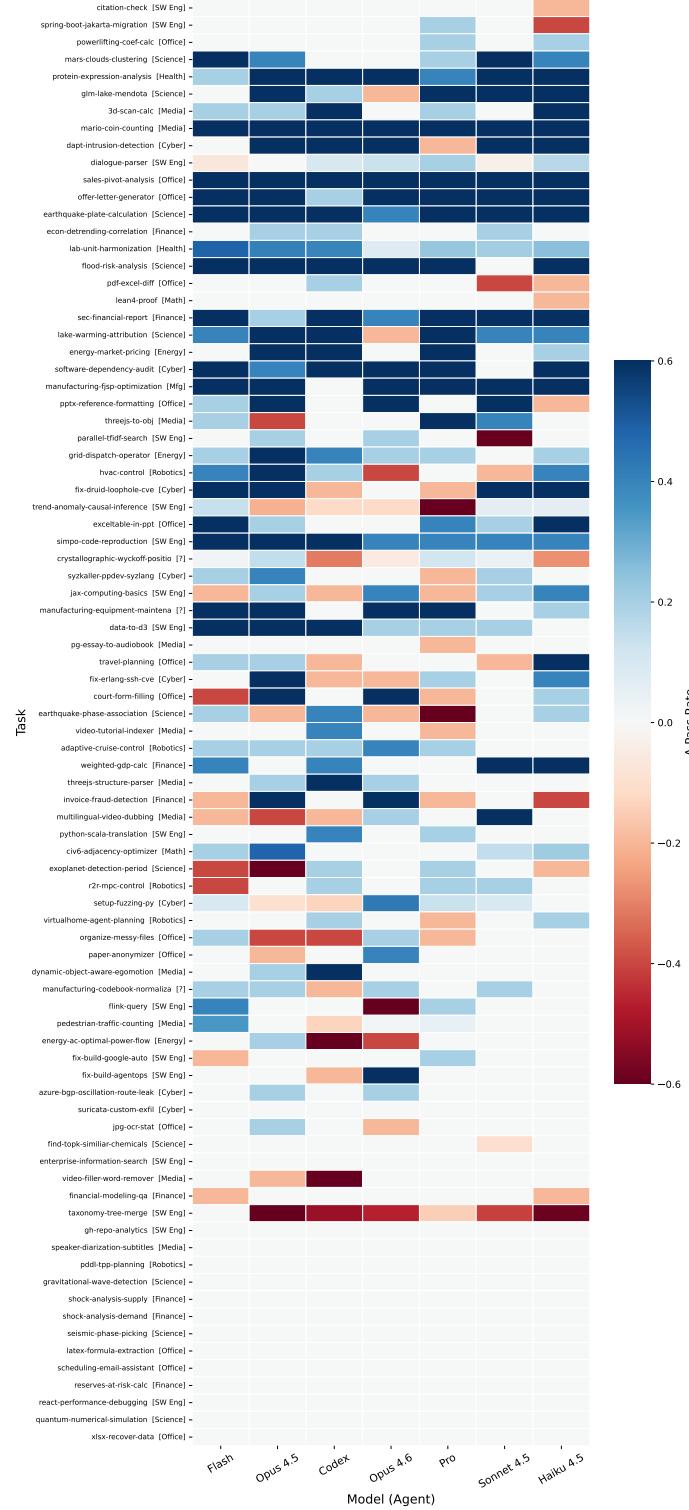


Figure 13. Skills uplift per task (with Skills – without Skills). Blue cells indicate positive uplift; red cells indicate tasks where Skills hurt performance. The majority of cells are blue, confirming broad Skill benefit. A small number of tasks show negative delta for specific models.

## E. Additional Experimental Details

### E.1. Confidence Interval Calculation

We compute 95% confidence intervals using the percentile bootstrap method with 1,000 resamples. For normalized gain, we compute CIs on the gain metric directly rather than on the component pass rates.

### E.2. 10-Run Validation

For a subset of configurations (GPT-5.2 and Claude Opus 4.5, all 26 Extreme tasks, no-Skills and with-Skills conditions), we conducted 10 runs instead of 5. Results show:

- Mean  $\Delta$  within 1.2pp of 5-run estimates
- Standard error reduced by  $\sim 30\%$
- All conclusions remain unchanged

This validates that 5 runs provides sufficient precision for our main findings.

## F. Complete Task List

[Table 9](#) enumerates all 86 tasks in SKILLSBENCH, organized by domain. Two tasks are excluded from evaluation: `mhc-layer-impl` (requires GPU) and `fix-visual-stability` (verifier timeout), leaving 84 evaluated tasks. Difficulty levels are assigned by task authors and validated during review.

*Table 9.* Complete list of SKILLSBENCH tasks (86 total, 84 evaluated). Tasks are grouped by domain and sorted alphabetically within each domain.

Task ID	Domain	Diff.	Description
<code>azure-bgp-oscillation</code>	Cybersecurity	Med	Detect BGP route oscillation and leaks in Azure Virtual WAN topology
<code>dapt-intrusion-detection</code>	Cybersecurity	Hard	Compute network statistics from PCAP file (DAPT2020 traffic)
<code>fix-druid-loophole-cve</code>	Cybersecurity	Hard	Fix Apache Druid 0.20.0 arbitrary code execution vulnerability
<code>fix-erlang-ssh-cve</code>	Cybersecurity	Hard	Fix Erlang/OTP SSH server unauthenticated RCE vulnerability
<code>setup-fuzzing-py</code>	Cybersecurity	Med	Set up continuous fuzzing for 5 Python libraries
<code>software-dependency-audit</code>	Cybersecurity	Med	Security audit of npm dependencies for vulnerabilities
<code>suricata-custom-exfil</code>	Cybersecurity	Med	Write Suricata signatures to detect HTTP data exfiltration
<code>syzkaller-ppdev-syzlang</code>	Cybersecurity	Med	Write syzkaller syzlang support for Linux parallel port driver
<code>energy-ac-optimal-power</code>	Energy	Med	Solve AC optimal power flow for day-ahead market schedule
<code>energy-market-pricing</code>	Energy	Hard	Analyze congestion pricing with counterfactual transmission relaxation
<code>grid-dispatch-operator</code>	Energy	Med	Decide generator dispatches for economically efficient DC power flow
<code>econ-detrending-corr.</code>	Finance	Med	Calculate Pearson correlation between detrended consumption and investment
<code>financial-modeling-qa</code>	Finance	Hard	Analyze large-scale financial data and answer modeling questions
<code>invoice-fraud-detection</code>	Finance	Hard	Detect invoice fraud by cross-referencing PDFs, Excel, and CSV
<code>reserves-at-risk-calc</code>	Finance	Med	Calculate Reserves-at-Risk using IMF gold price and reserves data
<code>sec-financial-report</code>	Finance	Hard	Analyze hedge fund activities comparing Q2 vs Q3 2025 SEC filings
<code>shock-analysis-demand</code>	Finance	Med	Estimate investment shock to Georgia (demand-side macro framework)
<code>shock-analysis-supply</code>	Finance	Hard	Estimate investment shock to Georgia (Cobb-Douglas production)

Continued on next page

Table 9 – continued from previous page

Task ID	Domain	Diff.	Description
weighted-gdp-calc	Finance	Med	Calculate weighted mean of GCC net exports as % of GDP
lab-unit-harmonization	Healthcare	Med	Harmonize clinical lab data units across healthcare systems
protein-expression-analysis	Healthcare	Med	Analyze differential protein expression in cancer cell line data
mfg-codebook-norm.	Manufacturing	Med	Normalize manufacturing defect reason texts to standard codebook
mfg-equipment-maint.	Manufacturing	Med	Answer reflow machine maintenance questions from handbook data
mfg-fjsp-optimization	Manufacturing	Med	Optimize flexible job shop schedule with downtime constraints
civ6-adjacency-optimizer	Mathematics	Hard	Optimize adjacency bonuses in Civilization 6 district placement
lean4-proof	Mathematics	Med	Complete a Lean4 formal proof for a sequence problem
3d-scan-calc	Media & Content	Hard	Calculate mass of 3D printed part from binary STL file
dynamic-obj-egomotion	Media & Content	Med	Analyze camera egomotion and detect dynamic objects in video
mario-coin-counting	Media & Content	Med	Count coins/enemies/turtles in Super Mario video frames
multilingual-video-dub	Media & Content	Med	Perform multilingual video dubbing with TTS alignment
pedestrian-traffic-count	Media & Content	Hard	Count pedestrians in surveillance camera videos
pg-essay-to-audiobook	Media & Content	Med	Convert Paul Graham essays to audiobook via TTS
speaker-diarization-subtitles	Media & Content	Hard	Perform speaker diarization and generate subtitles from video
threejs-structure-parser	Media & Content	Med	Parse Three.js file to extract 3D object part-level structure
threejs-to-obj	Media & Content	Med	Convert Three.js 3D object to OBJ format for Blender
video-filler-word-remover	Media & Content	Med	Detect filler words and their timestamps in interview video
video-tutorial-indexer	Media & Content	Hard	Find chapter start timestamps in a 23-min Blender tutorial
crystallographic-wyckoff	Natural Science	Med	Analyze crystal structure Wyckoff positions from CIF files
earthquake-phase-assoc.	Natural Science	Hard	Perform seismic phase association to identify earthquake events
earthquake-plate-calc	Natural Science	Med	Find earthquake furthest from Pacific plate boundary
exoplanet-detection	Natural Science	Med	Detect exoplanet transit period from TESS lightcurve data
find-topk-chemicals	Natural Science	Med	Find top- $k$ similar chemicals using Morgan fingerprints
flood-risk-analysis	Natural Science	Med	Identify flood stations from USGS streamflow data
glm-lake-mendota	Natural Science	Hard	Run General Lake Model to simulate lake water temperature
gravitational-wave-det.	Natural Science	Med	Detect gravitational wave signal via matched filtering
lake-warming-attribution	Natural Science	Med	Attribution analysis of lake warming from climate data
mars-clouds-clustering	Natural Science	Hard	Optimize DBSCAN to cluster Mars cloud annotations
quantum-numerical-sim	Natural Science	Med	Simulate open Dicke model steady state and Wigner function
seismic-phase-picking	Natural Science	Hard	Pick P and S wave arrival times from earthquake traces
court-form-filling	Office & White Collar	Easy	Fill California Small Claims Court PDF form from case desc.
exceltable-in-ppt	Office & White Collar	Med	Update embedded Excel table in PPTX with exchange rates
jpg-ocr-stat	Office & White Collar	Hard	OCR scanned receipt images and extract data into Excel
latex-formula-extraction	Office & White Collar	Med	Extract all LaTeX formulas from a research paper PDF
offer-letter-generator	Office & White Collar	Easy	Fill Word template with employee data for offer letter
organize-messy-files	Office & White Collar	Med	Organize 100+ PDF/PPTX/DOCX files into 5 subject folders
paper-anonymizer	Office & White Collar	Med	Anonymize research papers by redacting author-identifying info
pdf-excel-diff	Office & White Collar	Med	Identify differences between PDF backup and current Excel
powerlifting-coef-calc	Office & White Collar	Easy	Calculate IPF powerlifting scores in Excel

Continued on next page

Table 9 – continued from previous page

Task ID	Domain	Diff.	Description
pptx-reference-formatting	Office & White Collar	Med	Detect/format dangling paper titles in PowerPoint slides
sales-pivot-analysis	Office & White Collar	Med	Create pivot tables from population PDF and income Excel
scheduling-email-asst	Office & White Collar	Med	Read meeting request emails and propose calendar times
travel-planning	Office & White Collar	Med	Build 7-day travel itinerary with budget/pet constraints
xlsx-recover-data	Office & White Collar	Med	Recover missing values in NASA budget Excel file
adaptive-cruise-control	Robotics	Med	Implement adaptive cruise control simulation via PID
hvac-control	Robotics	Med	Implement temperature controller to maintain 22.0°C
pddl-tpp-planning	Robotics	Med	Solve travelling purchase problem using PDDL
r2r-mpc-control	Robotics	Med	Implement MPC controller for Roll-to-Roll manufacturing line
virtualhome-agent-plan	Robotics	Med	Solve airport ground traffic planning tasks using PDDL
citation-check	Software Eng.	Med	Verify bibliography integrity; identify fake citations in BibTeX
data-to-d3 dialogue-parser	Software Eng.	Med	Visualize stock data using D3.js v6 as single-page web app
	Software Eng.	Easy	Implement dialogue parser converting text to structured JSON
enterprise-info-search	Software Eng.	Hard	Retrieve information from heterogeneous enterprise data
fix-build-agentops	Software Eng.	Easy	Fix build errors in Python (AgentOps) codebase
fix-build-google-auto	Software Eng.	Easy	Fix build errors in Java (google/auto) codebase
flink-query	Software Eng.	Hard	Implement Flink job on Google cluster-usage traces dataset
gh-repo-analytics	Software Eng.	Med	Generate community pulse summary for cli/cli GitHub repo
jax-computing-basics	Software Eng.	Med	Complete programming tasks using JAX language
parallel-tfidf-search	Software Eng.	Med	Parallelize a TF-IDF document search engine
python-scala-translation	Software Eng.	Med	Translate Python tokenizer code to Scala 2.13
react-perf-debugging	Software Eng.	Hard	Debug and fix performance issues in Next.js e-commerce app
simpo-code-reproduction	Software Eng.	Hard	Reproduce SimPO loss function from paper description
spring-boot-jakarta-mig	Software Eng.	Hard	Migrate Java 8/Spring Boot 2.7 to Java 21/Spring Boot 3.2
taxonomy-tree-merge	Software Eng.	Hard	Unify product category taxonomies from Amazon/FB-/Google
trend-anomaly-causal	Software Eng.	Hard	Identify anomalous sales patterns and causal factors
<i>Excluded from evaluation:</i>			
fix-visual-stability*	Software Eng.	Hard	Fix visual instability/layout shifts in Next.js app
mhc-layer-impl*	Software Eng.	Hard	Implement DeepSeek mHC layer (requires GPU)

\*Excluded: mhc-layer-impl requires GPU resources not available in our container environment; fix-visual-stability exhibits consistent verifier timeouts across all configurations.

## G. Comprehensive Results Summary

Table 10 consolidates all aggregate results across 7 model-harness configurations and 3 Skills conditions. Scores are computed using Method D (task-mean with fixed denominator of 84 evaluated tasks), consistent with Terminal-Bench (Merrill et al., 2026) scoring methodology. Confidence intervals are 95% bootstrap CIs with 1,000 resamples.

### Key observations.

- Curated Skills improve performance by +16.2pp on average (range: +13.6 to +23.3pp), corresponding to a normalized gain of 21.5%.
- Claude Code + Opus 4.5 shows the largest absolute improvement (+23.3pp) and normalized gain (29.9%), reflecting Claude Code’s native Skills integration.
- Gemini CLI + Gemini 3 Flash achieves the highest absolute pass rate (48.7%) with Skills, despite a smaller normalized gain (25.3%).

*Table 10.* Comprehensive results across all model-harness configurations and Skills conditions. Pass rates (%) computed using Method D scoring (84-task fixed denominator, 5 trials per task).  $\Delta_S$  = curated Skills improvement;  $g$  = normalized gain =  $\frac{\text{With Skills} - \text{No Skills}}{100 - \text{No Skills}} \times 100$ ; Self-Gen = self-generated Skills;  $\Delta_G$  = self-generated improvement over no Skills. “–” = condition not evaluated. Models ordered by with-Skills pass rate.

Harness	Model	No Skills	95% CI	With Skills	95% CI	$\Delta_{\text{abs}}$ (pp)	$g$ (%)	Self-Gen	$\Delta_G$ (pp)
Gemini CLI	Gemini 3 Flash	31.3	$\pm 3.0$	48.7	$\pm 3.1$	+17.4	25.3	–	–
Claude Code	Opus 4.5	22.0	$\pm 2.8$	45.3	$\pm 2.5$	+23.3	29.9	$21.6 \pm 3.3$	-0.4
Codex	GPT-5.2	30.6	$\pm 3.1$	44.7	$\pm 3.0$	+14.1	20.3	$25.0 \pm 4.0$	-5.6
Claude Code	Opus 4.6	30.6	$\pm 2.6$	44.5	$\pm 3.1$	+13.9	20.0	$32.0 \pm 3.1$	+1.4
Gemini CLI	Gemini 3 Pro	27.6	$\pm 3.0$	41.2	$\pm 3.1$	+13.6	18.8	–	–
Claude Code	Sonnet 4.5	17.3	$\pm 2.5$	31.8	$\pm 2.9$	+14.5	17.5	$15.2 \pm 3.2$	-2.1
Claude Code	Haiku 4.5	11.0	$\pm 2.1$	27.7	$\pm 2.9$	+16.7	18.8	$11.0 \pm 3.2$	0.0
<b>Mean</b>		<b>24.3</b>		<b>40.6</b>		<b>+16.2</b>	<b>21.5</b>	<b>21.0</b>	<b>-1.3</b>

- Self-generated Skills yield -1.3pp on average; only Opus 4.6 shows marginal improvement (+1.4pp).
- Confidence intervals are tighter for no-Skills conditions (lower variance) than for with-Skills conditions, suggesting Skills introduce additional sources of variance.

## H. Token Usage and Cost Efficiency

This section presents token usage statistics and cost analysis across all model-harness configurations. Token data is extracted from trial `result.json` files where available.

### H.1. API Pricing

Table 11 lists the API pricing used for cost calculations, verified as of February 2026.

*Table 11.* API pricing per million tokens (February 2026). Prices reflect standard (non-batch, non-cached) rates.

Model	Provider	Input (\$/MTok)	Output (\$/MTok)
Claude Haiku 4.5	Anthropic	\$1.00	\$5.00
Claude Sonnet 4.5	Anthropic	\$3.00	\$15.00
Claude Opus 4.5	Anthropic	\$5.00	\$25.00
Claude Opus 4.6	Anthropic	\$5.00	\$25.00
GPT-5.2	OpenAI	\$1.75	\$14.00
Gemini 3 Pro	Google	\$2.00	\$12.00
Gemini 3 Flash	Google	\$0.50	\$3.00

Notes: Gemini 3 Pro uses tiered pricing (\$2/\$12 for  $\leq 200K$  context; \$4/\$18 for  $> 200K$ ). GPT-5.2 offers cached input at \$0.175/MTok (90% discount). Anthropic batch API offers 50% off standard rates. Costs reported here use standard non-cached, non-batch rates for consistency.

### H.2. Token Usage by Model and Condition

Token counts are recorded at the harness level. Codex CLI and Gemini CLI record aggregate token usage in each trial’s `result.json` (~80–97% of trials have data); Claude Code does not record aggregate counts in `result.json`. For Claude Code, per-call usage is available in session JSONL logs; however, the with-Skills condition JSONL files use a different format that prevents recovery, and output token counts from JSONL are unreliable (capturing only a subset of API calls). We report Claude Code data only where available and flag the data source.

Table 12 presents mean token usage per trial. Token usage includes both prompt (input) and completion (output) tokens.

Table 12. Mean token usage per trial by model and Skills condition (thousands of tokens). Codex and Gemini data from `result.json`; Claude Code data from session JSONL (input tokens only, output tokens unreliable).  $n$  = trials with available data.

Model	Condition	$n$	Input (K)	Output (K)	Total (K)
GPT-5.2	No Skills	352	961	12.2	974
	With Skills	334	1,087	11.6	1,099
	Self-Gen	206	1,058	13.6	1,072
Gemini 3 Flash	No Skills	405	985	14.2	999
	With Skills	405	1,075	12.1	1,087
Gemini 3 Pro	No Skills	407	495	12.0	507
	With Skills	405	465	10.9	476
Opus 4.6 <sup>†</sup>	No Skills	309	1,947	—	—
	With Skills	308	1,448	—	—

<sup>†</sup>Claude Code input tokens from JSONL session logs (effective input = fresh + cache creation + cache read tokens). Output tokens not reliably recoverable. Other Claude Code models (Opus 4.5, Sonnet 4.5, Haiku 4.5) not shown; with-Skills JSONL data unavailable due to format differences.

### H.3. Estimated Cost per Trial

Table 13 presents estimated cost per trial at standard API pricing (non-cached, non-batch rates from Table 11). Costs are computed from the token usage data in Table 12.

Table 13. Estimated mean cost per trial (\$) at standard API pricing. Costs computed from mean token usage. Claude Code models omitted due to incomplete output token data.

Model	Condition	Est. Cost/Trial	$\Delta$ vs. No Skills
GPT-5.2	No Skills	\$1.85	—
	With Skills	\$2.07	+\$0.22 (+12%)
Gemini 3 Flash	No Skills	\$0.54	—
	With Skills	\$0.57	+\$0.03 (+6%)
Gemini 3 Pro	No Skills	\$1.13	—
	With Skills	\$1.06	-\$0.07 (-6%)

### H.4. Cost-Performance Tradeoff

The Pareto frontier analysis in Figure 4 (main paper) shows that Skills shift the cost-performance frontier upward across all models. Key cost-efficiency findings:

- Gemini 3 Flash consumes  $2.3\times$  more input tokens per task than Gemini 3 Pro (1.08M vs. 0.47M with Skills), a compensatory strategy where the smaller model substitutes iterative exploration for reasoning depth.
- At standard API pricing, Flash’s  $4\times$  lower per-token cost more than offsets higher token volume, making Flash 47% cheaper per task (\$0.57 vs. \$1.06).
- Skills increase input token usage by 6–13% for Codex and Gemini (additional context from Skill documents), but the performance improvement (+16.2pp average) substantially outweighs the marginal cost increase (\$0.03–0.22 per trial).
- Gemini 3 Pro shows a slight *decrease* in token usage with Skills (−6%), suggesting Skills help Pro solve tasks more efficiently, with fewer exploration rounds.

**Cache efficiency.** All models show high cache hit rates: GPT-5.2 at 91–92%, Gemini 3 Pro at 75–76%, and Gemini 3 Flash at 63–67%. Claude Code models (from JSONL data where available) show >99% cache rates, reflecting aggressive prompt caching. In practice, cached pricing reduces actual costs by 50–90% below the standard rates shown in Table 13.

## I. Failure Analysis

This section presents a comprehensive failure analysis across 7,294 evaluated trajectories. We first catalog infrastructure errors (§I.1), then develop a trajectory-level failure taxonomy (§I.2) adapted from the Multi-Agent System Taxonomy

(MAST) (Pan et al., 2025) and the Terminal Agent Taxonomy (Merrill et al., 2026). We classify 5,171 agent failures into five categories based on programmatic analysis of verifier test outputs and error metadata from each trial.

### I.1. Infrastructure Error Summary

**Table 14** summarizes non-agent errors encountered during evaluation. These are infrastructure-level failures unrelated to agent capability.

*Table 14.* Infrastructure errors across 7,414 total trials (including excluded tasks). These errors are treated as reward = 0 in scoring.

Error Type	Count	Root Cause
VerifierTimeoutError	83	Verifier exceeds timeout (2 tasks)
RuntimeError (Docker)	46	Docker build/compose failures
AgentSetupTimeoutError	35	Gemini CLI setup exceeds 360s
RewardFileNotFoundException	2	Verifier didn't produce reward file
AddTestsDirError	1	Couldn't copy tests into container
<b>Total</b>	<b>167</b>	2.3% of all trials

**VerifierTimeoutError (83 trials).** 73 instances on `fix-visual-stability` (excluded from evaluation) and 10 on `react-performance-debugging`. The verifier itself exceeds its timeout budget (300–450s), independent of agent behavior. `fix-visual-stability` was excluded from all experiments due to consistent verifier failures across all configurations.

**RuntimeError (46 trials).** Dominated by a single task: `scheduling-email-assistant` (42 of 46), where the Dockerfile creates symlinks for Google authentication credentials that don't exist in the evaluation environment. The remaining 4 errors are isolated Docker daemon failures (OCI cgroup timeouts, container name conflicts, image race conditions).

**AgentSetupTimeoutError (35 trials).** Exclusively in `withskills-gemini-cli`, affecting 14 tasks. Gemini CLI's agent initialization exceeds the 360-second setup timeout on certain task environments. This does not affect other harnesses.

### I.2. Trajectory-Level Failure Taxonomy

We adapt the Terminal Agent Taxonomy (TAT) from Terminal-Bench (Merrill et al., 2026), which itself derives from MAST (Pan et al., 2025), to classify agent failures on SkillsBench. Where Terminal-Bench uses LLM-as-judge (Docent pipeline with GPT-5) for trajectory classification, we use programmatic analysis of structured test outputs (CTRF reports and pytest logs) from the verifier, as our tasks produce deterministic, machine-verifiable results. We organize failures into five high-level categories:

- **Timeout:** Agent exceeds its allocated execution time (`AgentTimeoutError`). These represent resource exhaustion rather than logical errors—the agent may have been pursuing a viable strategy that simply could not complete within the budget. Corresponds to TAT's “Unaware of Termination Conditions” when the agent fails to prioritize within time constraints.
- **Execution:** Agent produces incorrect or missing output due to implementation errors:
  - *No Output Produced:* Required output files are absent (all verifier tests fail on file existence checks). Often caused by early crashes, environmental setup failures, or the agent never reaching the output-generation stage.
  - *Specification Violation:* Agent contradicts explicit task directives—e.g., producing hardcoded values instead of required formulas, wrong output format, or violating structural constraints. Corresponds to TAT's “Disobey Specification.”
  - *Domain Knowledge Gap:* Agent misinterprets domain-specific concepts—e.g., unit conversion errors, incorrect scientific formulas, or misapplied domain algorithms.
  - *Incorrect Implementation:* Correct approach but flawed logic—e.g., off-by-one errors, incorrect data transformations, or algorithm bugs.
  - *Tool/Environment Failure:* Agent fails to install dependencies, handle permissions, or configure the execution environment.

- **Coherence:** Agent produces a partial but structurally sound solution:
  - *Incomplete Solution:* Some verifier tests pass while others fail on missing components. The agent completes part of the task correctly but leaves key deliverables unfinished. Corresponds to TAT’s “Premature Termination” when the agent declares completion prematurely.
- **Verification:** Agent produces structurally complete output that fails quality thresholds:
  - *Quality Below Threshold:* All required files exist and have correct structure, but computed values, metrics, or outputs fall outside acceptable tolerances. This is the most common failure mode, reflecting the difficulty of SkillsBench tasks even when agents understand the task structure. Corresponds to TAT’s “Weak Verification” when the agent’s self-checks fail to catch quality issues.
- **Unknown:** Verifier output is absent or cannot be classified by our heuristics (4.4% of failures).

### I.3. Overall Failure Distribution

**Table 15** presents the distribution of failure modes across 5,171 agent failures (excluding 110 infrastructure errors). The dominant failure mode is *Quality Below Threshold* (49.8%), indicating that agents typically understand the task structure and produce output, but their solutions are insufficiently accurate. *Agent Timeout* is the second most common (17.8%), followed by *Incomplete Solution* (10.2%) and *No Output Produced* (7.9%).

*Table 15.* Failure mode distribution across 5,171 agent failures (excluding infrastructure errors). Classification is based on programmatic analysis of CTRF test reports and verifier logs.

Category	Failure Mode	Count	%
Execution	No Output Produced	411	7.9
	Domain Knowledge Gap	251	4.9
	Specification Violation	171	3.3
	Incorrect Implementation	68	1.3
	Tool/Environment Failure	16	0.3
<i>Subtotal</i>		917	17.7
Coherence	Incomplete Solution	527	10.2
Verification	Quality Below Threshold	2,577	49.8
Timeout	Agent Timeout	922	17.8
Unknown	Unclassified	228	4.4
<b>Total</b>		<b>5,171</b>	<b>100.0</b>

### I.4. Failure Distribution by Model

**Table 16** shows how failure modes vary across models. Key observations:

- **Opus 4.6 has the highest timeout rate** (29% of its failures) but the fewest Execution failures (10%), suggesting it pursues more ambitious strategies that risk running out of time.
- **Gemini 3 Flash/Pro have the lowest timeout rates** (9–10%) but higher Execution failure rates (23–25%), indicating faster but more error-prone execution.
- **Haiku 4.5 has the highest Execution failure rate** (21%) and highest overall failure rate (85.2% of trials), consistent with its lower capability producing more fundamental errors.
- **Verification failures are uniformly dominant** across all models (48–53%), confirming that quality—not structure—is the primary bottleneck.

### I.5. Success Case Studies

We present representative examples where curated Skills transformed agent outcomes from failure to success, illustrating the mechanisms through which procedural knowledge improves performance.

**Skills bridge domain-specific API gaps: sales-pivot-analysis.** Without Skills, all 7 models scored 0% on this task, which requires creating Excel pivot tables programmatically from population and income data. Agents consistently loaded the data correctly but failed at pivot table creation—Codex attempted manual DataFrame reshaping instead of using

## SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks

---

*Table 16.* Failure category distribution by model (% of each model’s agent failures). Sorted by overall failure rate.

Model	Trials	Fail%	Timeout	Execution	Coherence	Verification	Unknown
Haiku 4.5	1,074	85.2%	13%	21%	9%	53%	4%
Sonnet 4.5	1,072	80.0%	19%	18%	10%	48%	5%
Opus 4.5	1,077	71.9%	19%	17%	11%	49%	5%
Gemini 3 Pro	820	67.1%	9%	23%	13%	52%	3%
Opus 4.6	1,245	67.1%	29%	10%	8%	48%	5%
Codex	1,076	67.8%	22%	14%	12%	49%	4%
Gemini 3 Flash	820	61.8%	10%	25%	9%	51%	5%

`openpyxl`’s pivot table API, producing structurally incorrect output (10/23 tests failed with “list index out of range” on missing pivot objects). With Skills providing step-by-step guidance for the `openpyxl` pivot table workflow, 6 of 7 models achieved  $\geq 80\%$  pass rate. Average improvement: +85.7pp.

**Skills provide critical data processing pipelines: `flood-risk-analysis`.** This task requires identifying flood-risk stations from USGS streamflow data using return period estimation. Without Skills, agents attempted ad-hoc statistical approaches—e.g., simple threshold-based detection or incorrect distribution fitting—achieving only 2.9% pass rate. The curated Skill specified the Log-Pearson Type III distribution, the standard USGS methodology for flood frequency analysis, including the exact `scipy` function calls and parameter interpretation. With Skills, pass rate rose to 80.0% (+77.1pp), with all models correctly applying the USGS-standard methodology.

**Skills encode regulatory knowledge: `sec-financial-report`.** Analyzing hedge fund activities from SEC 13F filings requires understanding specific regulatory formats, CIK lookup procedures, and filing comparison methodology. Without Skills, no model could complete the task (0% pass rate)—agents either failed to locate the correct filings or misinterpreted the tabular data format. The curated Skill documented the SEC EDGAR API endpoints, 13F-HR filing structure, and cross-quarter comparison methodology. With Skills, pass rate reached 75.0% (+75.0pp).

**Skills prevent common implementation pitfalls: `manufacturing-fjsp-optimization`.** The flexible job-shop scheduling problem requires constraint-aware optimization with machine downtime windows. Without Skills, agents produced naive schedules ignoring maintenance constraints (0% pass rate). The curated Skill outlined the constraint propagation approach, objective function formulation, and OR-Tools solver configuration. With Skills, agents successfully formulated and solved the optimization problem (68.6% pass rate, +68.6pp).

### I.6. How Skills Change Failure Patterns

Comparing failure mode distributions between the No Skills and With Skills conditions reveals where Skills have the most impact ([Table 17](#)).

*Table 17.* Failure mode distribution by condition (% of that condition’s agent failures). Skills reduce the overall failure rate from 78.4% to 61.1% while shifting the failure composition.

Condition	Fail Rate	Timeout	Execution	Coherence	Verification
No Skills	78.4%	16.1%	17.1%	10.7%	52.1%
With Skills	61.1%	18.6%	21.1%	8.9%	46.6%
Self-Generated	80.9%	19.9%	13.9%	11.2%	50.4%

**Skills primarily reduce Verification failures.** The absolute number of Quality Below Threshold failures drops from 1,184 (No Skills) to 819 (With Skills), a 30.8% reduction. This accounts for the majority of the improvement: Skills provide domain-specific guidance that helps agents produce higher-quality outputs on tasks they already structurally understand.

**Skills slightly increase the relative share of Timeouts.** While the absolute timeout count decreases from 367 to 328, its share of failures increases from 16.1% to 18.6%. This is because Skills reduce easy failures faster than hard ones—agents that previously produced low-quality output now spend longer pursuing better solutions, sometimes exceeding time limits.

**Skills reduce Coherence failures.** Incomplete Solution failures drop from 243 to 156 (35.8% reduction), indicating that Skills help agents identify and complete all required deliverables rather than leaving components unfinished.

**Self-Generated Skills underperform curated Skills.** The Self-Generated condition shows a higher failure rate (80.9%) than No Skills (78.4%) with more Timeout failures (19.9%), suggesting that agents spend additional time generating and then attempting to use self-authored Skills documents, which may contain errors or misleading guidance.

## I.7. Qualitative Failure Examples

We present representative examples from each major failure mode, drawn from verifier test outputs across failed trajectories.

**Quality Below Threshold: `earthquake-plate-calculation`.** The agent correctly identified the target earthquake event, extracting latitude, longitude, magnitude, and timestamp accurately (7/8 tests passed). However, the computed distance from the nearest plate boundary was 3,562 km instead of the expected 3,878 km—an 8.2% error that exceeded the  $\pm 0.01$  km tolerance. The agent applied the correct Haversine formula but used an incorrect plate boundary coordinate, demonstrating that even when agents understand the computational method, domain-specific data interpretation remains error-prone.

**Agent Timeout: `gravitational-wave-detection`.** The agent was terminated by `AgentTimeoutError` before producing any output file. All 8 verifier tests were skipped after the file existence check failed. This task requires processing LIGO strain data with signal processing pipelines (bandpass filtering, matched filtering, SNR calculation)—a computationally intensive workflow that likely exceeded both the model’s reasoning and the execution time budget.

**Incomplete Solution: `shock-analysis-supply`.** The agent created a structurally correct Excel workbook and passed 6 of 9 tests, correctly setting up sheet structures, formula templates, and some data imports. However, it failed to: (1) populate employment/labor data from the Penn World Tables (PWT), (2) execute the HP filter optimization solver, and (3) compute the depreciation rate. These three missing components represent the most domain-specific and computationally demanding aspects of the task.

**No Output Produced: `gh-repo-analytics`.** All 8 verifier tests errored at the fixture stage with “Missing /app/report.json,” meaning the agent never created the required output file. The task requires interacting with a local Gitea server, cloning repositories, and computing analytics—a multi-step pipeline where failure at any early stage prevents all downstream output.

**Specification Violation: `latex-formula-extraction`.** The agent extracted LaTeX formulas from a PDF but included markdown headers alongside the formulas in the output file, producing 6 entries instead of the required 5. The specification required one formula per line wrapped in `$$` delimiters; the extraneous headers violated this format constraint. This failure illustrates agents’ tendency to over-include content rather than strictly adhering to output specifications.

**Domain Knowledge Gap: `exceltable-in-ppt`.** The agent correctly updated the primary exchange rate cell in the PowerPoint-embedded Excel table (6/8 tests passed) but failed to recompute inverse rates and dependent cells, producing `NaN` propagation through the spreadsheet. The underlying issue was misunderstanding how Excel formula dependencies cascade in embedded workbooks—a domain-specific detail that Skills could address.

**Skills transforming outcomes: `sales-pivot-analysis`.** Without Skills, Codex populated source data correctly but could not create Excel pivot tables (10/23 tests failed with “list index out of range” on missing pivot objects). With Skills, all 23 tests passed—the Skills provided Office-specific guidance for programmatic pivot table creation that the agent could not discover independently. This task exemplifies the  $0\% \rightarrow 85.7\%$  improvement pattern, where Skills bridge a specific capability gap.

## I.8. Unsolved Tasks

16 of 84 evaluated tasks (19%) have a 0% pass rate across all models and conditions, revealing the current frontier of agent capability. These tasks cluster into three patterns:

- **Computationally intractable** (6 tasks): gravitational-wave-detection, quantum-numerical-simulation, seismic-phase-picking, shock-analysis-demand, speaker-diarization-subtitles, taxonomy-tree-merge. These tasks require either domain-specific signal processing pipelines, large-scale numerical optimization, or computationally expensive algorithms that exceed the agent’s time budget.
- **Complex multi-step pipelines** (6 tasks): gh-repo-analytics, pedestrian-traffic-counting, enterprise-information-search, setup-fuzzing-py, reserves-at-risk-calc, shock-analysis-supply. These require chaining multiple tools, APIs, or data sources where failure at any stage prevents downstream progress.
- **Strict specification tasks** (4 tasks): latex-formula-extraction, pddl-tpp-planning, react-performance-debugging, xlsx-recover-data. These have narrow success criteria where even small deviations from the expected output format or values cause all tests to fail.

### I.9. Tasks With Largest Skills Impact

Table 18 lists the 10 tasks where Skills produced the largest improvement in pass rate. These tasks share a common pattern: they require domain-specific procedural knowledge (e.g., Excel pivot table APIs, financial modeling formulas, scientific data processing pipelines) that is well-suited to being encoded in Skill documents. The average improvement for these top-10 tasks is +70.4 percentage points.

*Table 18.* Tasks with largest Skills impact (With Skills pass rate – No Skills pass rate). Only tasks with  $\geq 5$  trials per condition shown.

Task	No Skills	With Skills	$\Delta$
mario-coin-counting	2.9%	88.6%	+85.7pp
sales-pivot-analysis	0.0%	85.7%	+85.7pp
flood-risk-analysis	2.9%	80.0%	+77.1pp
sec-financial-report	0.0%	75.0%	+75.0pp
protein-expression-analysis	17.1%	91.4%	+74.3pp
offer-letter-generator	14.3%	86.5%	+72.2pp
earthquake-plate-calculation	11.4%	82.9%	+71.4pp
dapt-intrusion-detection	25.7%	96.9%	+71.2pp
manufacturing-fjsp-optimization	0.0%	68.6%	+68.6pp
software-dependency-audit	8.6%	69.4%	+60.9pp

### I.10. Agent Timeout Analysis

Agent timeouts (`AgentTimeoutError`) represent 17.8% of agent failures, affecting 922 trials across all configurations. Unlike infrastructure errors, agent timeouts reflect the agent exceeding its allocated execution time.

*Table 19.* Agent timeout rates by model in the with-Skills condition (420 trials per model, 84 tasks  $\times$  5 trials). Effective rate includes retroactive timeout enforcement for Gemini CLI (see §C).

Model	Timeouts	Rate (%)
Opus 4.6	90	21.4
GPT-5.2	78	18.6
Sonnet 4.5	66	15.7
Opus 4.5	64	15.2
Haiku 4.5	59	14.0
Gemini 3 Pro	56	13.3
Gemini 3 Flash	49	11.7

**Timeout concentration.** Timeouts are concentrated on specific tasks rather than spread uniformly. Of 84 tasks, 39 (46%) have at least one timeout in the with-Skills condition. Tasks such as taxonomy-tree-merge, video-filler-word-remover, and gravitational-wave-detection consistently timeout across all models (5/5 trials), suggesting these tasks exceed the computational budget regardless of model capability.

**Gemini CLI timeout enforcement bug.** The `withskills-gemini-cli` configuration recorded 0 `AgentTimeoutErrors` (vs. 13% expected based on other configurations), indicating timeout enforcement was disabled for these runs. We retroactively identified affected trials by comparing execution times to per-task timeout limits derived from all observed `AgentTimeoutErrors` across configurations (see §C). Eight Gemini with-Skills trials that scored reward > 0 were overridden to reward = 0, reducing Flash by 0.7pp and Pro by 0.7pp.

### I.11. Data Cleaning: Gemini CLI Retry Irregularities

The `withskills-gemini-cli` configuration had retry irregularities: 13 tasks had more than the expected 10 trials (5 per model), and 64 launched jobs never completed (missing `result.json`). We cap at 5 trials per task-model (first 5 chronologically) and treat missing results as reward = 0. Six dropped valid trials had reward = 1.0 (see §C for details).