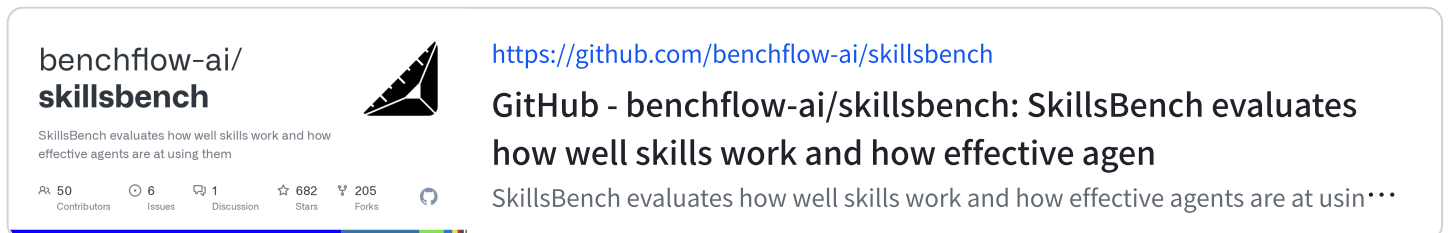


SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks

<https://arxiv.org/pdf/2602.12670>



The screenshot shows the GitHub repository for SkillsBench. The repository name is 'benchflow-ai/skillsbench'. The description reads: 'SkillsBench evaluates how well skills work and how effective agents are at using them'. The repository statistics are: 50 Contributors, 6 Issues, 1 Discussion, 682 Stars, and 205 Forks. The repository URL is <https://github.com/benchflow-ai/skillsbench>. The repository title is 'GitHub - benchflow-ai/skillsbench: SkillsBench evaluates how well skills work and how effective agen'. The repository description is 'SkillsBench evaluates how well skills work and how effective agents are at using...'. The repository is a public repository.

- 首个系统测评Agent Skills框架，第一个领域专家构建综合Agentic Benchmark
- 涵盖86个任务、11个领域、7种agent配置
- 揭示Agent Skills的insights

Skills介绍:

Problem: Agents need procedural knowledge

Claude Code, Codex提供了强大的通用能力，但是:

- Domain Specific Tasks需要特定的流程知识
- Fine tuning牺牲了通用性
- Token开销

What Skills look like

A Skill directory:

```
anthropic_brand/  
├ SKILL.md  
├ docs.md  
├ slide-decks.md  
└ apply_template.py
```

anthropic/brand_style/SKILL.md

YAML Frontmatter

```
---  
name: Anthropic Brand Style Guidelines  
description: Applies Anthropic's official brand colors and  
typography to PowerPoint presentations  
---
```

Metadata

Overview

Markdown

This skill provides Anthropic's official brand identity resources for PowerPoint presentations. It includes a pre-branded template and tools to apply Anthropic styling to existing presentations.

Freeform Content

Colors

Main Colors:

- Dark: `#141413` - Primary text and dark backgrounds
- Light: `#faf9f5` - Light backgrounds and text on dark
- Light Gray: `#e8e6dc` - Subtle backgrounds

Typography

- *Headings*: Poppins (with Arial fallback)
- *Body Text*: Lora (with Georgia fallback)

Progressive Disclosure

Initially, Claude sees just the metadata (from the YAML frontmatter of SKILL.md) from the available Skills.

Name: **Excel Files**

description: Comprehensive Microsoft Excel (.xlsx) document creation, editing, and analysis with support for formulas, formatting, analysis, and visualization. Use when:

- (1) Creating new spreadsheets with formulas & formatting
- (2) Reading and analyzing data,
- (3) Modifying spreadsheets while preserving formulas
- (4) Data analysis and visualization
- (5) Recalculating formulas

Name: **PowerPoint Files**

description: Microsoft PowerPoint (.pptx) presentation creation, editing, and analysis.

1. Creating or modifying presentations
2. Reading and analyzing existing content
3. Analyzing comments or speaker notes

name: **Anthropic Brand Style Guidelines**

description: Applies Anthropic's official brand colors & typography to PowerPoint presentations and other professional docs.

name: **Anthropic BigQuery Schemas**

description: Strategic data analysis using Google BigQuery for the Finance & Strategy team. Provides access to revenue metrics, product usage analytics, sales/CRM data, and analytics via MCP integration.

Progressive Disclosure

If Claude thinks the Skill is relevant to the current task, it loads the full SKILL.md file into context.

anthropic/brand_style/SKILL.md

YAML Fontmatter

```
---
name: Anthropic Brand Style Guidelines
description: Applies Anthropic's official brand colors and typography to PowerPoint presentations
---
```

Overview

Markdown

This skill provides Anthropic's official brand identity resources for PowerPoint presentations. It includes a pre-branded template and tools to apply Anthropic styling to existing presentations.

Colors

Main Colors:

- Dark: `#141413` - Primary text and dark backgrounds
- Light: `#faf9f5` - Light backgrounds and text on dark
- Light Gray: `#e8e6dc` - Subtle backgrounds

Typography

- *Headings*: Poppins (with Arial fallback)
- *Body Text*: Lora (with Georgia fallback)

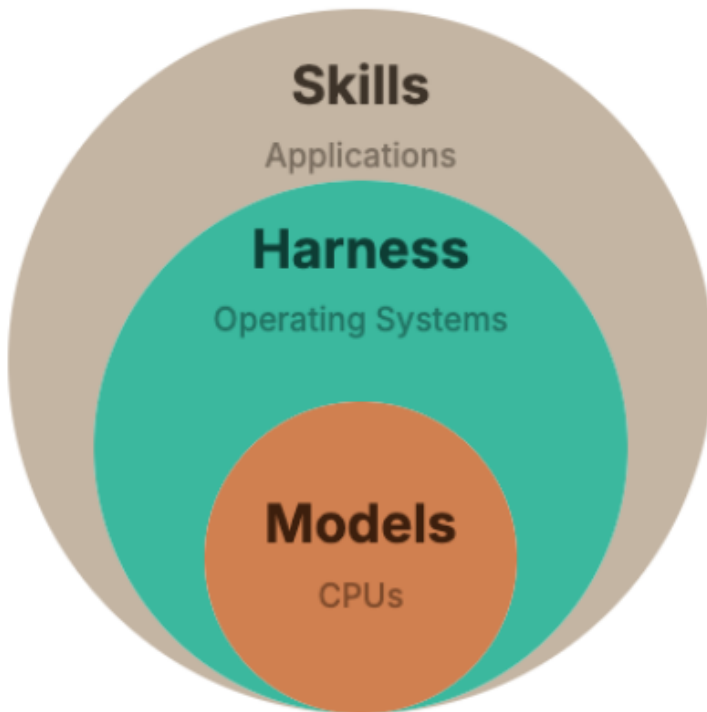
Level	File	Context Window	# Tokens
1	SKILL.md Metadata (YAML)	Always loaded	~100
2	SKILL.md Body (Markdown)	Loaded when Skill triggers	<5k
3+	Bundled files (text files, scripts, data)	Loaded as-needed by Claude	unlimited*

Approach	Token Cost	Performance
Manual instructions	5,000-10,000 tokens/request	Variable quality
Skills (metadata only)	Minimal (just name/description)	Expert-level
Skills (full load)	~5,000 tokens when skill is used	Expert-level

Skills对比其他runtime augmentation范式:

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

Skills的逻辑关系:



- **Skills Layer**

Domain-specific capabilities and workflows that extend agent functionality. Like applications on an OS, skills provide specialized knowledge and tools for particular tasks.

- **Agent Harness Layer**

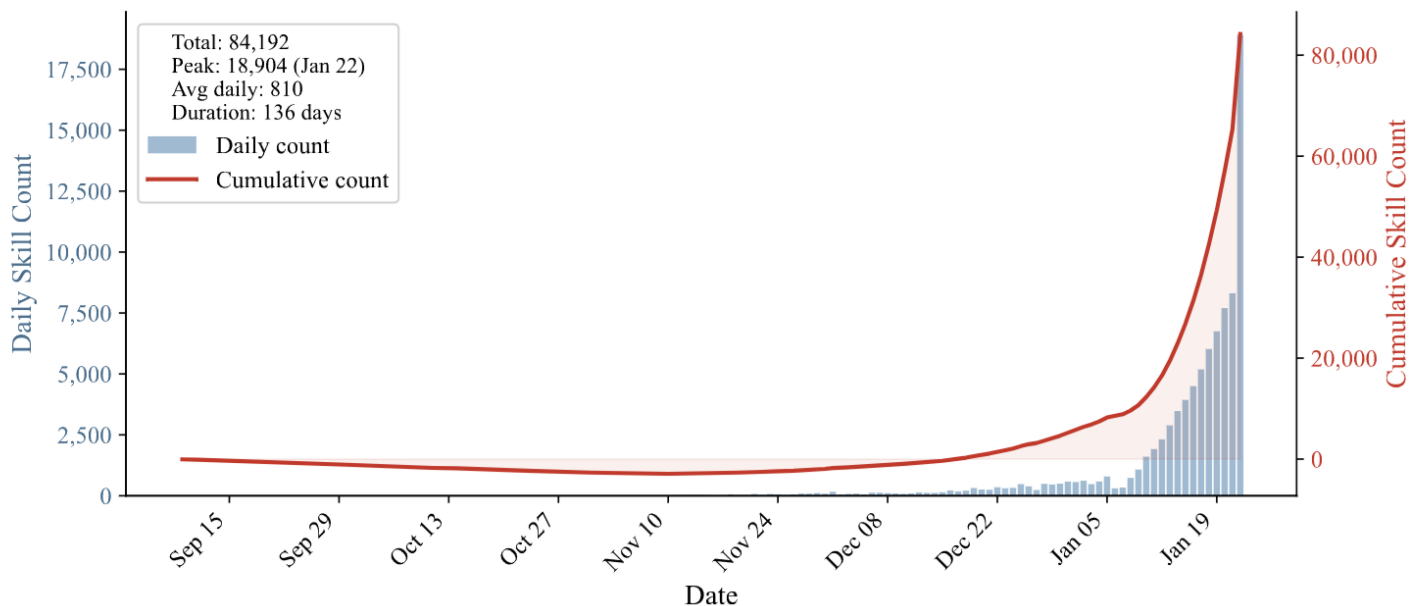
The execution environment that orchestrates agents, manages tool access, and handles I/O. Analogous to an operating system that mediates between applications and hardware.

- **Models Layer**

The foundational AI models that power reasoning and generation. Like CPUs, they provide the raw computational capability that upper layers build upon.

Benchmark概览:

开源社区提供越来越多的skills，但是缺乏对Skills如何/何时提高Agent性能的评价



86 EXPERT-LEVEL TASKS ACROSS 11 DOMAINS

Office & White Collar

Excel, Word, PowerPoint, PDF, Gmail

Natural Science

Physics, Astronomy, Chemistry, Earth Sciences

Finance

Macrofinance, Economics, Portfolio Mgmt

Healthcare

Clinical Lab, Cancer Proteomics

Manufacturing

Codebook, Equipment, Job-Shop Scheduling

Cybersecurity

CVE Patching, CTF, Network Security

Energy

Power Grid, Optimal Power Flow

Mathematics

Formal Proof, Game Optimization

Robotics

PDDL Planning, Control Systems

Media & Content

Video Editing, Audio, Dubbing, TTS

Software Eng.

ML Repro, DevOps, Bug Fix, Migration

DIFFICULTY

Core 20%

Extended 50%

Extreme 30%

■ Core — 17 tasks · <60 min
 ■ Extended — 43 tasks · 1-4 hours
 ■ Extreme — 26 tasks · >4 hours

COMMUNITY MOMENTUM

800+

Community Members

180+

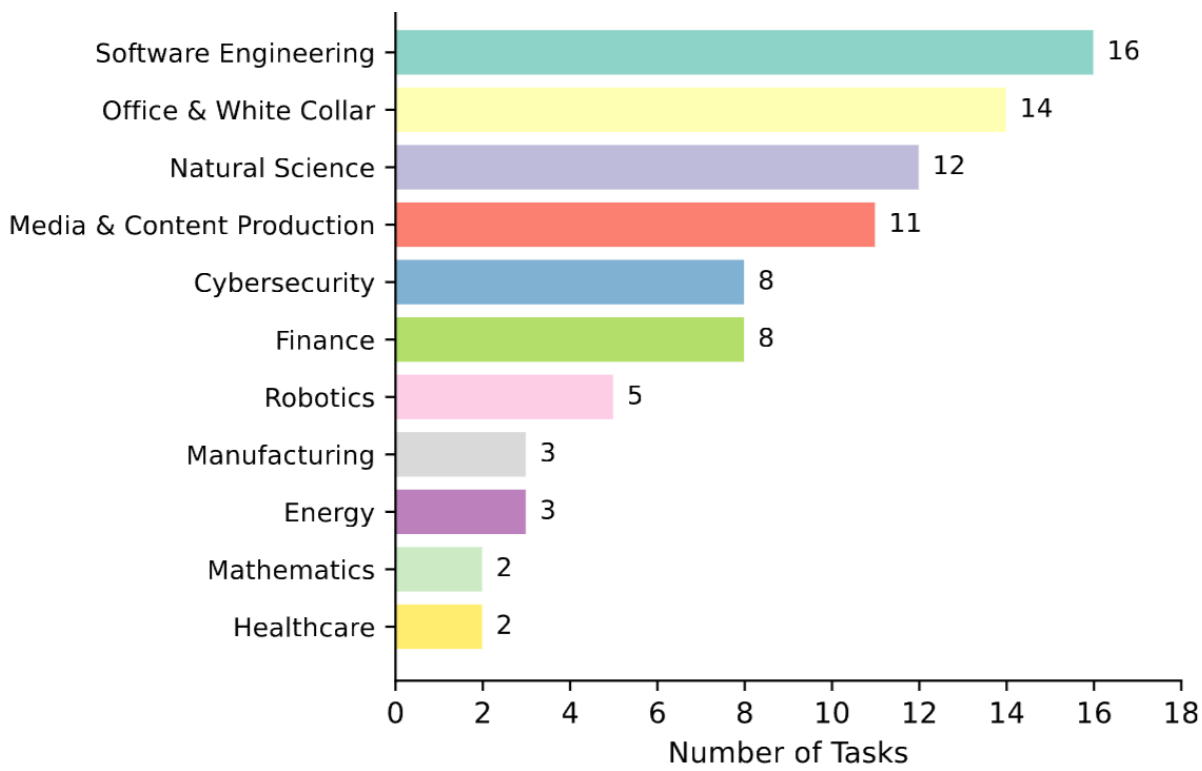
Contributors

80%+

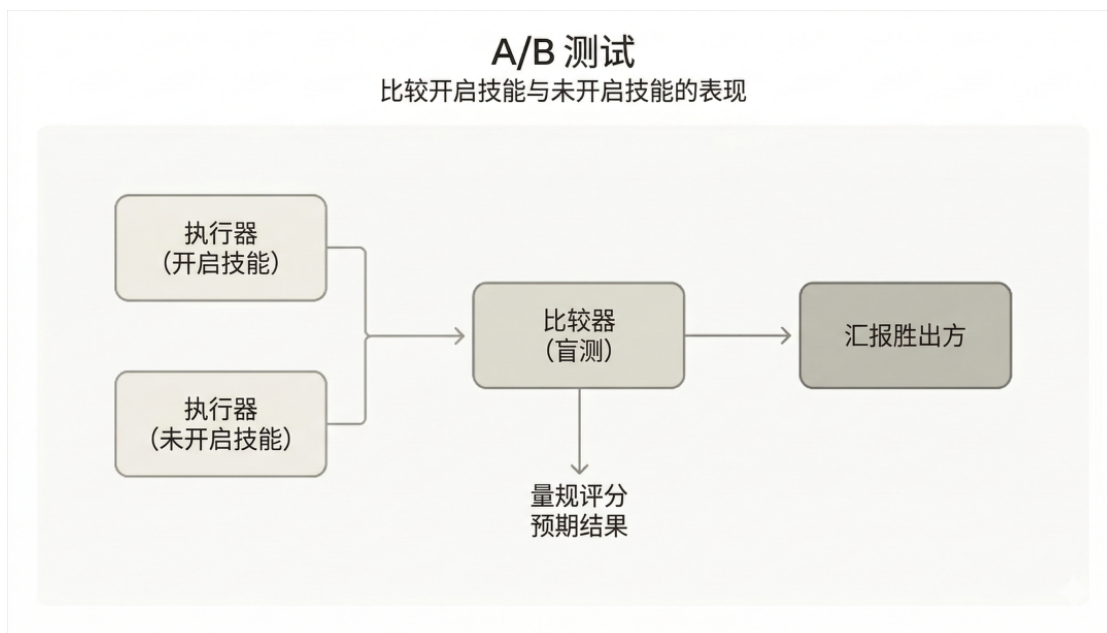
PhDs or Senior Professionals

Configuration划分:

Difficulty	Tasks	Human Time
Core	17 (20%)	< 60 min
Extended	42 (49%)	1-4 hours
Extreme	26 (31%)	> 4 hours



- With Skills
- Without Skills
- Self-generated Skills: 排除LLM内部已有的知识



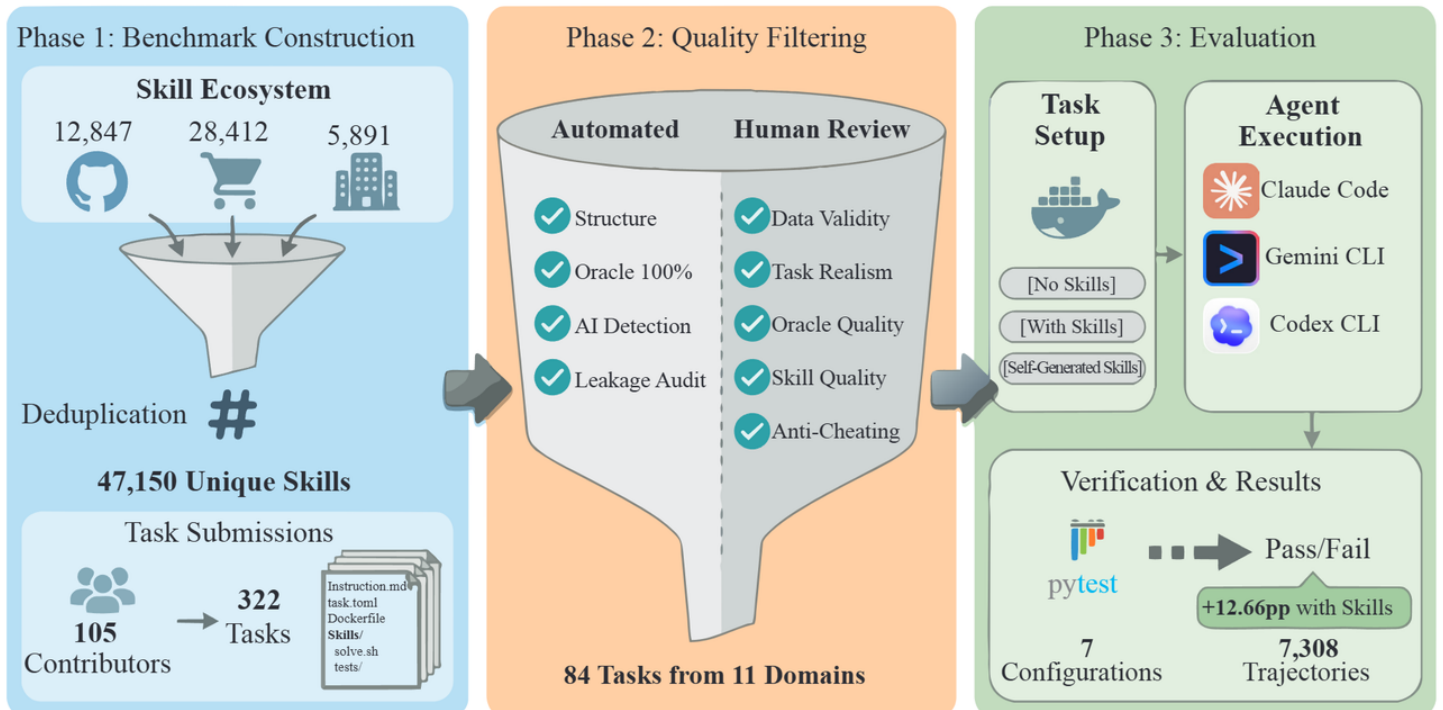
指标:

Pass Rate: the binary reward across 5 trials

Normalized Gain:

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

Benchmark构建:



- Instructions written by humans (verified via human review and GPTZero)
- Deterministic verifiers with programmatic assertions - no LLM-as-judge variance
- Leakage prevention via CI-based audits ensuring Skills provide guidance, not solutions
- Oracle solutions demonstrating resolvability

Key Findings:

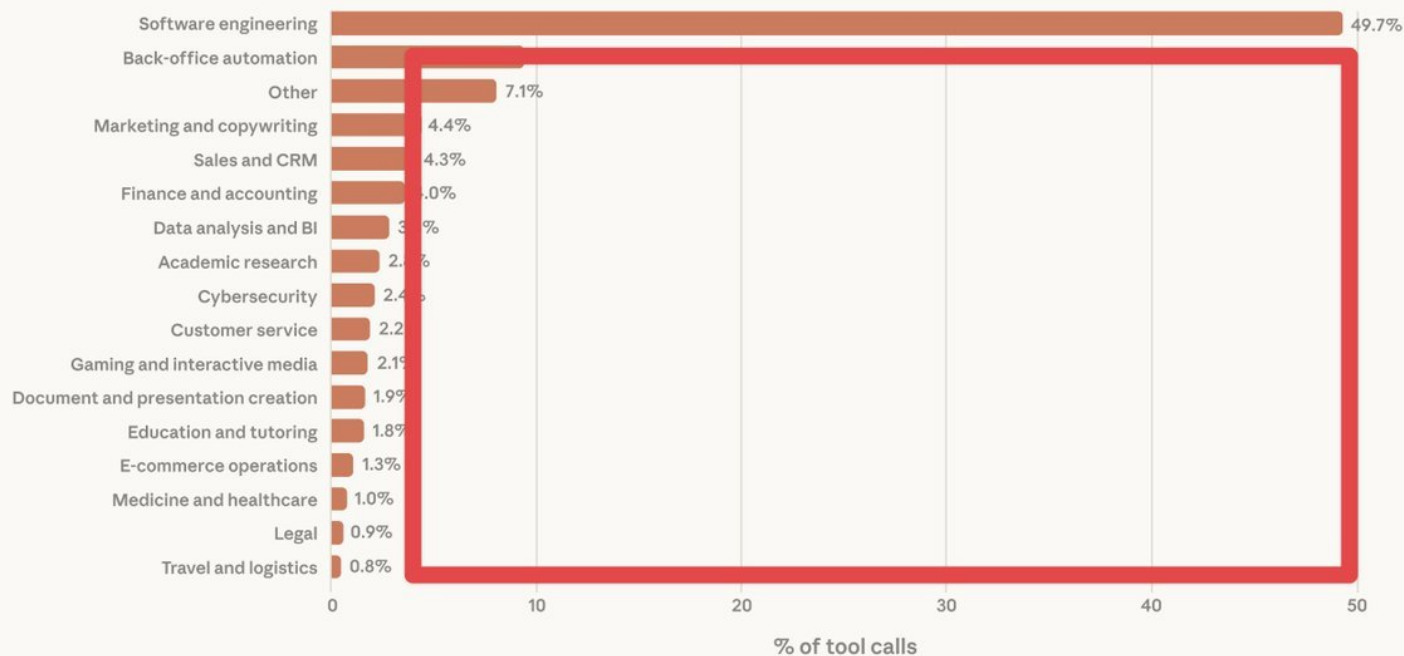
Skills提供了显著的增益，Claude Code的收益最大:

Agent + Model	No Skills	With Skills	Uplift
Gemini CLI (Gemini 3 Flash)	31.3	48.7	+17.4
Claude Code (Opus 4.5)	22.0	45.3	+23.3
Codex (GPT-5.2)	30.6	44.7	+14.1
Claude Code (Opus 4.6)	30.6	44.5	+13.9
Gemini CLI (Gemini 3 Pro)	27.6	41.2	+13.6
Claude Code (Sonnet 4.5)	17.3	31.8	+14.5
Claude Code (Haiku 4.5)	11.0	27.7	+16.7

不同的domain下，skills带来的收益差别非常明显：

Domain	With Skills	No Skills	Δ_{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

In what domains are agents deployed?



Skills的数量与复杂度对性能的影响:

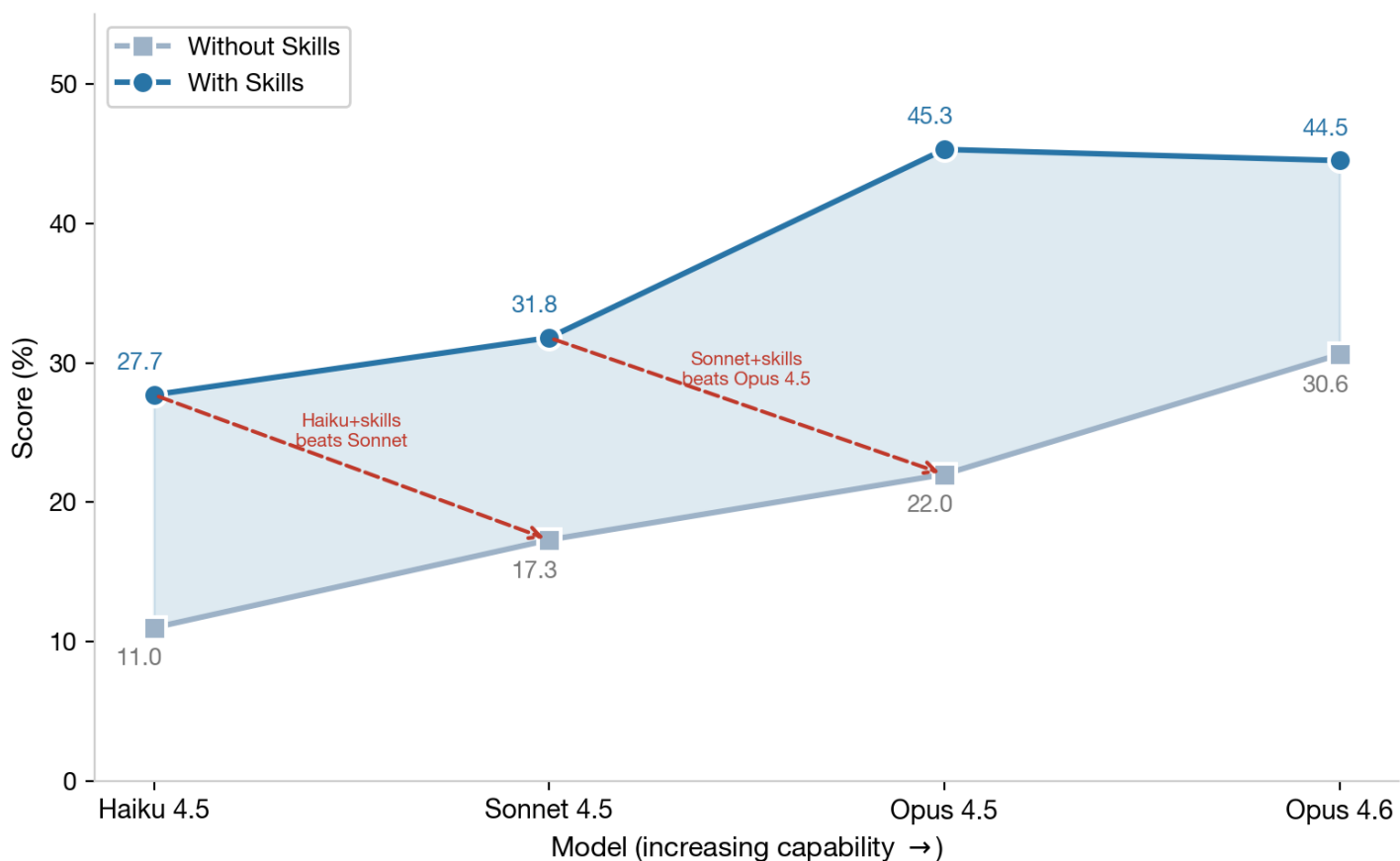
Skills Count	With Skills	No Skills	Δ_{abs}
1 skill	42.2%	24.4%	+17.8
2–3 skills	42.0%	23.4%	+18.6
4+ skills	32.7%	26.9%	+5.9

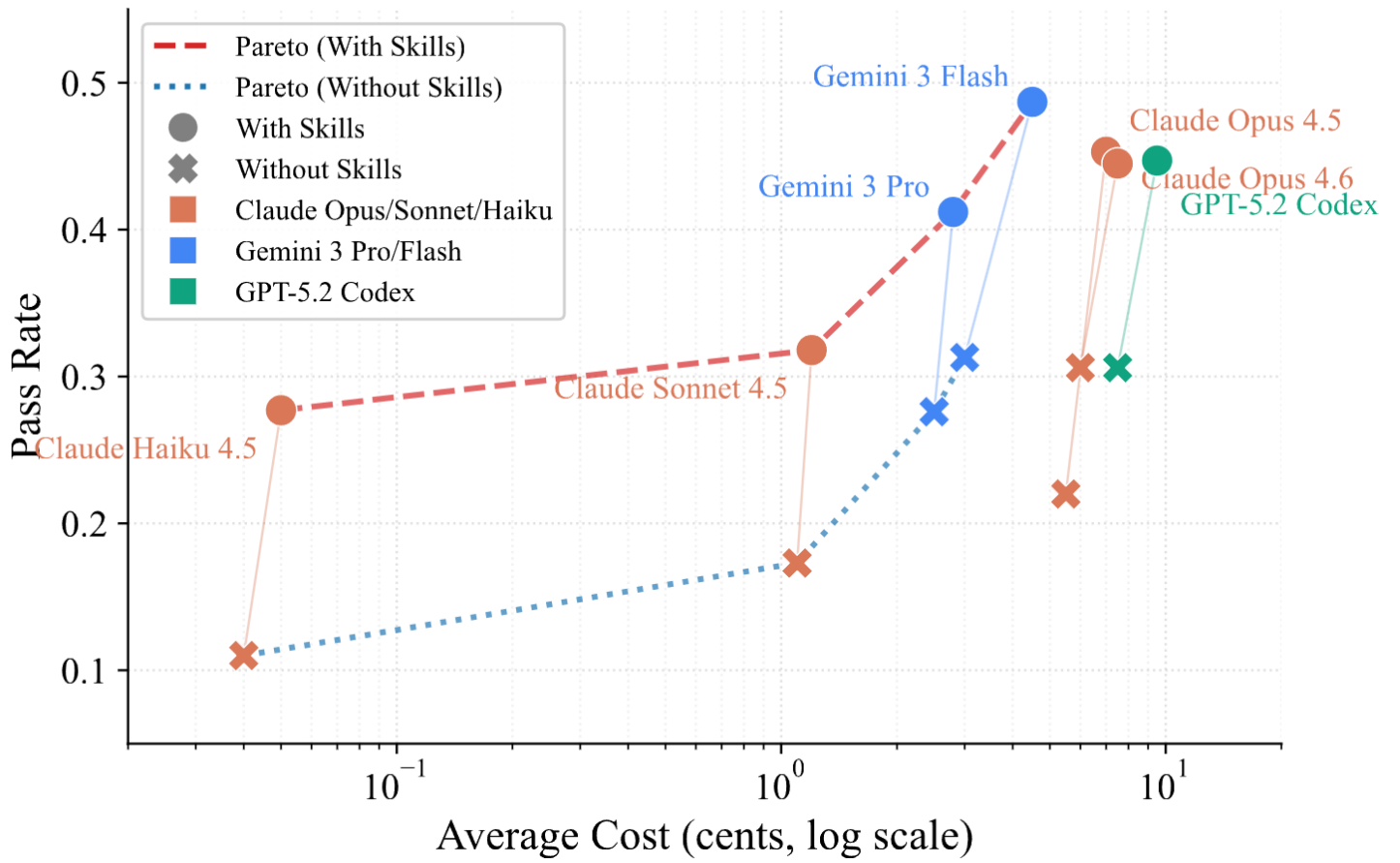
Self-generated Skills基本没有作用:

Case study分析出两种失败模式: procedure不准确; 缺乏领域知识

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
Mean		24.3	40.6	21.5	21.0	-1.8

Skills可以弥补model scale的不足：





Future Work:

- Self-generated Skills: 外部反馈 & Self-evolve, 例如AutoSkill, SkillRL, SkillsCreator(<https://mp.weixin.qq.com/s/vjMG8i7DwQ7R2B1C4AVQdA>)等
- Domain Skills