SkillLens

From raw experience to skill consumption

A systematic study of model-generated agent skills

Zisu Huang · Fudan University / Microsoft Research

01

Experience

What should a skill learn from?

02

Extraction

What makes a useful skill?

03

Consumption

Who can benefit from it?





Fudan University · Microsoft Research · Shanghai Jiao Tong University

A reusable skill is not yet a useful skill.

Experience can become a reusable procedure—without retraining.

Past trajectories → 

→ Unseen tasks

But generating plausible advice does not establish that an agent can use it.

RQ1 **Do generated skills reliably help?**
Across extractors, consumers and domains.

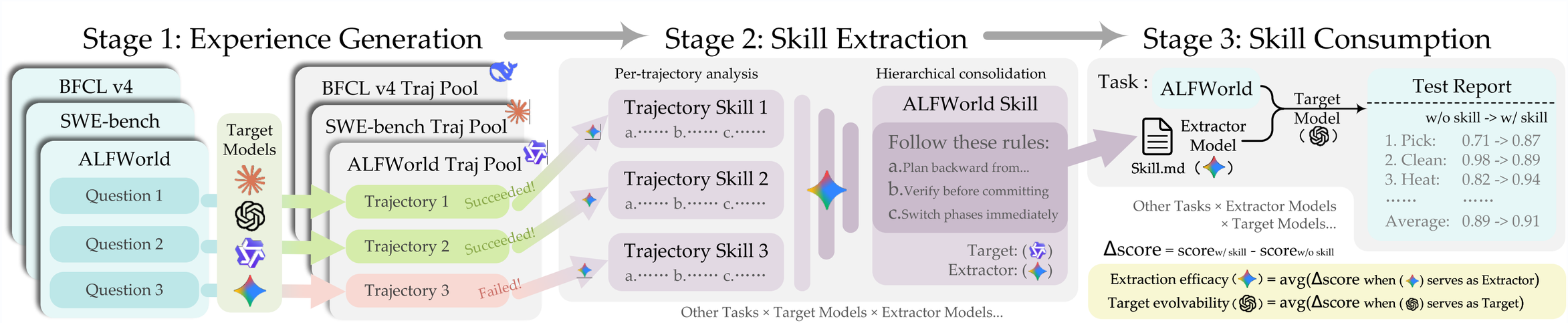
RQ2 **What drives their utility?**
Dissect the full skill lifecycle.

RQ3 **Can diagnosis improve extraction?**
Turn validated observations into guidance.

Evaluate skills by what agents can do with them.

STUDY DESIGN

One lifecycle. A controlled comparison.



5 domains × **5** extractors × **6** target models = **150** configurations

ALFWorld · SpreadsheetBench · SWE-bench-Verified · SEAL-0 · BFCL-v4

Δ = performance with skill − no-skill baseline. Extract from training experience; evaluate on unseen tasks.

Skills help on average. Negative transfer remains common.

112/150 settings improve

37/150 settings regress

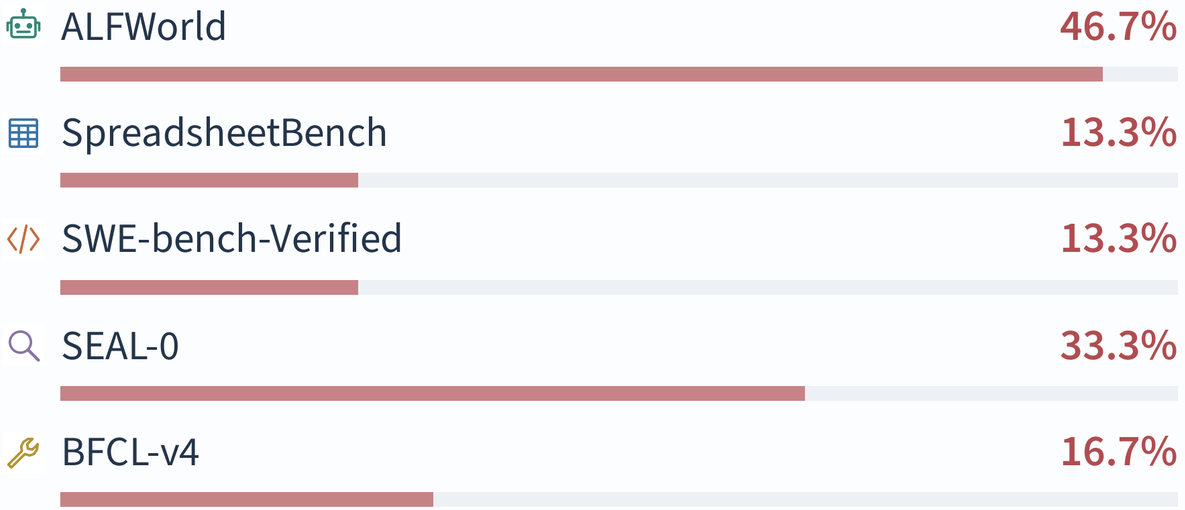


Each dot is one extractor–target–domain configuration. One setting ties.

An average gain does not guarantee a useful skill for a particular agent.

Negative transfer by domain

Share of 30 configurations

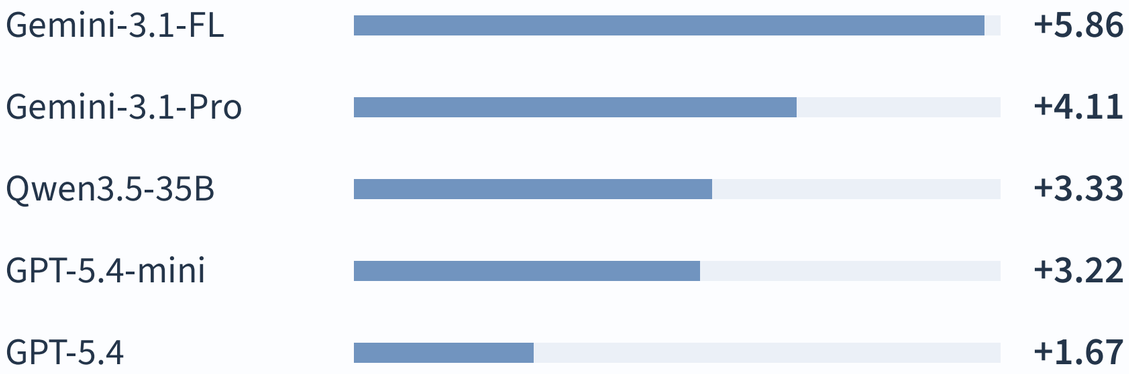


Paper §4.2 · Table 1 · Δ < 0 denotes negative transfer

Execution, extraction and consumption are different capabilities.

Who extracts useful skills?

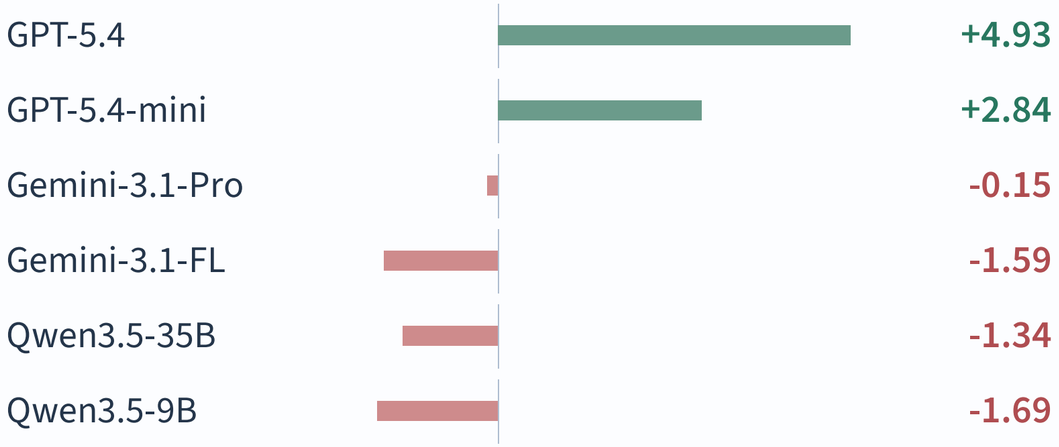
SpreadsheetBench · EE (pp)



GPT-5.4 has the strongest task baseline here, yet the lowest extraction efficacy.

Who benefits from consuming them?

ALFWorld · TE (pp)

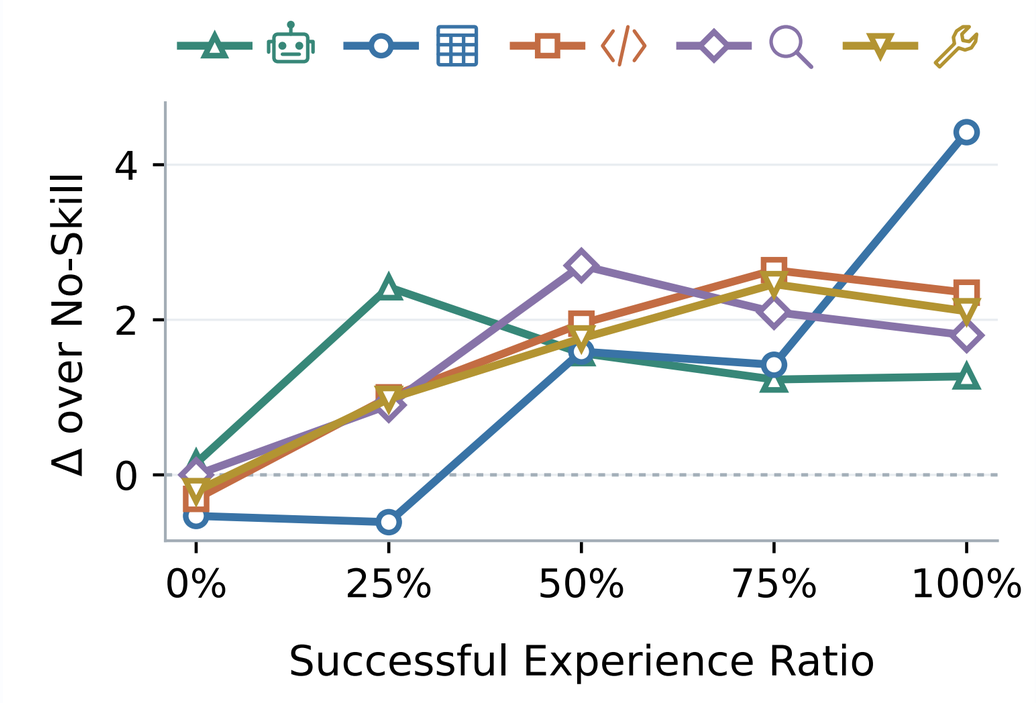


The same pool of extractors helps some targets while others regress on average.

Choosing a skill pipeline is an extractor–consumer–domain compatibility problem.

Paper §4.2 · Table 1 · EE: fixed extractor; TE: fixed target

The best experience mix depends on the domain.



01 **Different domains need different ratios.**

SpreadsheetBench favors all-success experience. ALFWorld benefits from failure-heavy pools.

02 **Success is the foundation.**

All-failure pools are generally weak. Successes provide procedures, not only warnings.

Experience composition is a design choice, not a neutral input.

Paper §5.1 · Figure 2(a) · GPT-5.4-mini extractor

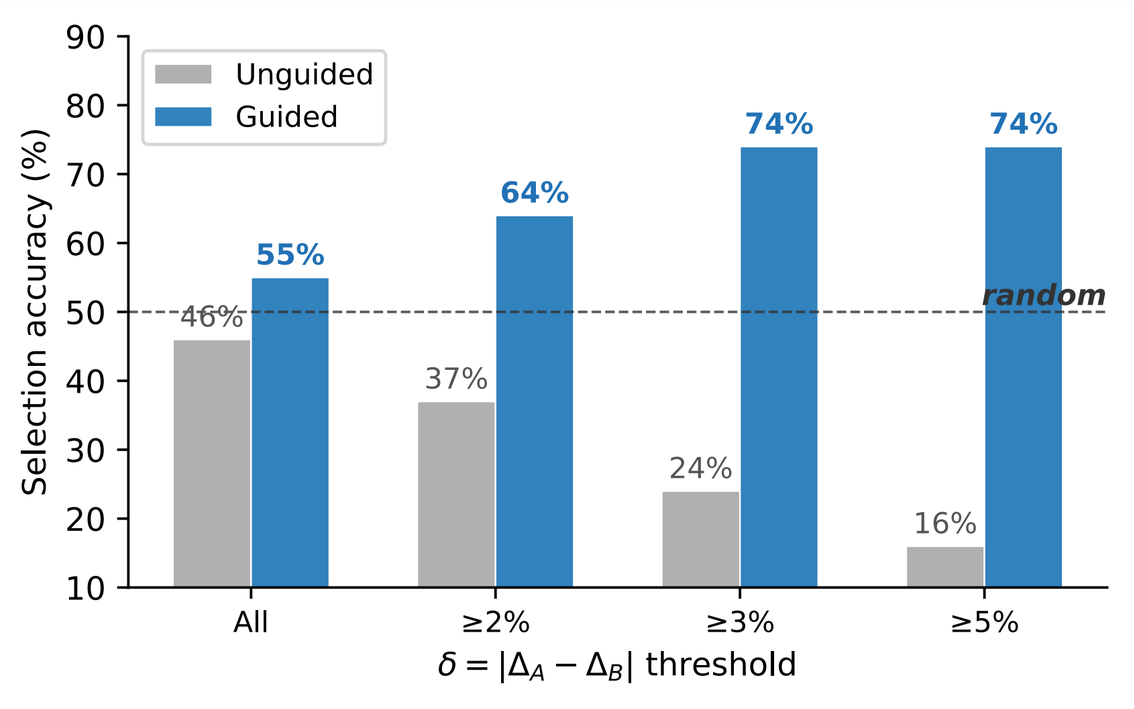
Surface form is a poor proxy for utility.

01 **Changing format does not reliably change utility.**

Across the four tested configurations, the format effect is non-significant: **all p > 0.34.**

02 **Plausible text can mislead a judge.**

Judging skills from their text alone can favor advice that performs worse downstream.



Exploratory pairwise judging; separate from the held-out intervention test.

What the skill says matters more than how it looks.

Paper §5.2 · Appendix C, Table 8 · Appendix E, Figure 5

Hold the skill fixed. The consumer still changes the gain.

CONTROLLED COMPARISON

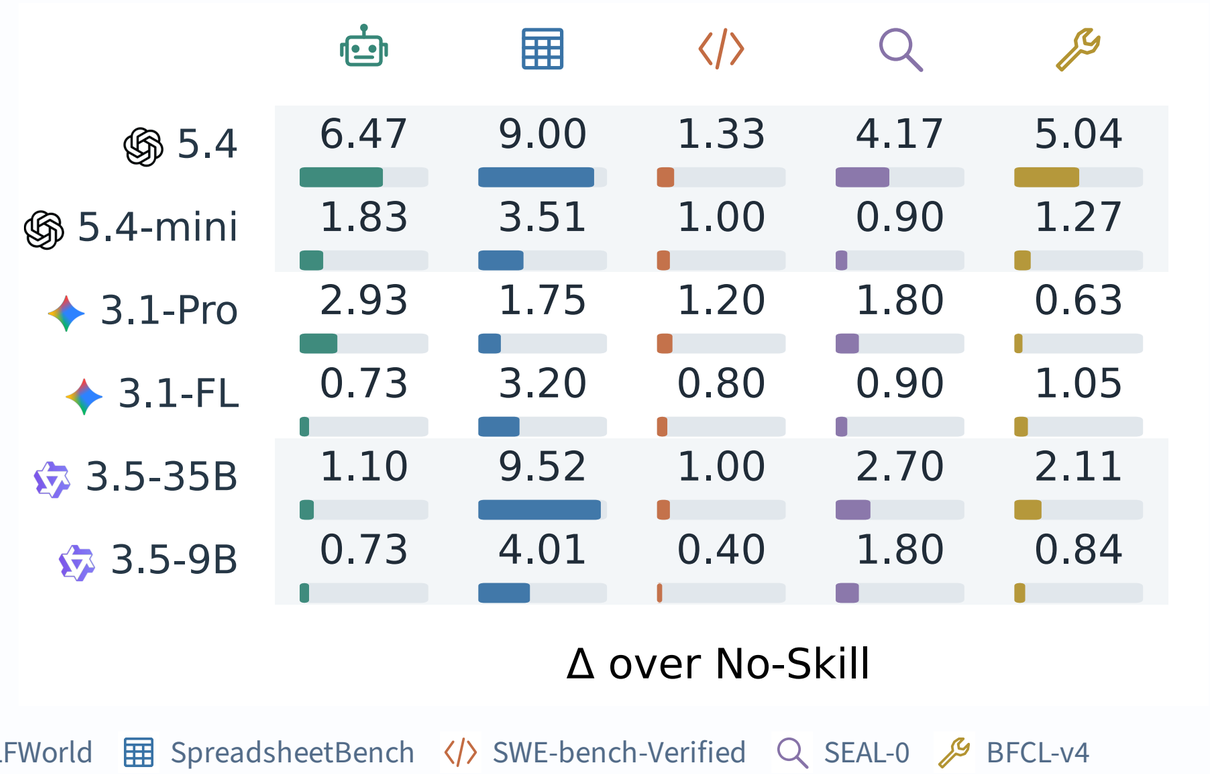
One skill per domain.

SKILL → ○○○ ○○○

Apply its text unchanged to all six targets.

Consumption ability is a target property.
The same advice can translate into very different gains.

Skills steer the model's existing strategy; execution robustness limits the benefit.



Useful skill content and the ability to execute it must be considered together.

Paper §5.3 · Figure 2(b) · Same skill within each domain

Turn utility-grounded diagnosis into an extraction prior.

01 **Contrast**

Compare high- and low-utility skill pairs to discover candidate quality dimensions.

02 **Validate**

Keep dimensions that align with downstream utility on development tasks.

03 **Guide**

Insert the resulting meta-skill into the extractor’s system prompt.

meta-skill — EXTRACTOR SYSTEM PROMPT

01 **Failure mechanism encoding**

Explain what fails and why it fails.

02 **Actionable specificity**

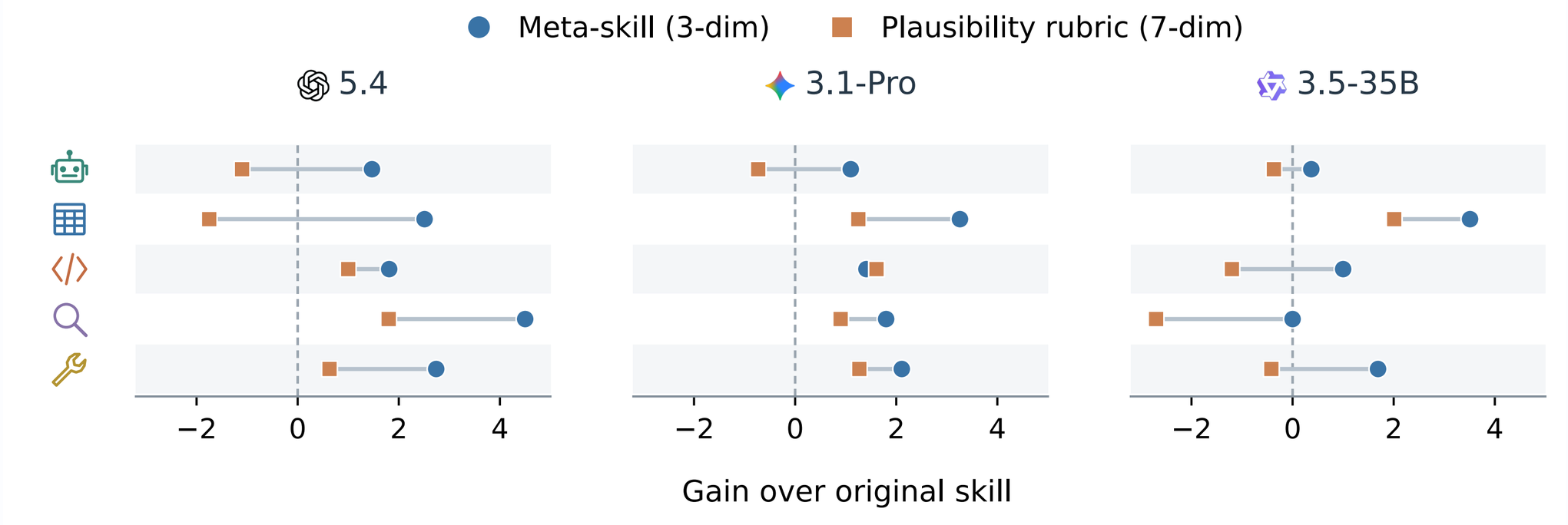
Turn the diagnosis into executable fixes.

03 **High-risk action blacklist**

State the risky actions to avoid.

The intervention changes how a skill is extracted, before any held-out evaluation.

Validated criteria produce more useful skills.



ALFWorld · SpreadsheetBench · SWE-bench-Verified · SEAL-0 · BFCL-v4

+1.95 pp
mean gain over original skills

14 / 15
held-out settings improve; one ties

+0.15 pp
mean gain from a plausibility rubric

Paper §6 · Figure 3 · Disjoint training / development / held-out splits

TAKEAWAYS

Skill quality is a lifecycle property.

01	**Measure utility.**	Generated skills help on average, but negative transfer makes downstream evaluation essential.
02	**Match the lifecycle.**	Experience composition, extraction quality and consumer ability jointly shape the outcome.
03	**Close the loop.**	Validated observations can become generation-time guidance that improves extracted skills.

Study the whole path from experience to downstream behavior.

SkillLens

From experience to useful skills.

Thank you.

Questions & discussion

microsoft.github.io/SkillLens

Paper

Code

Paper, code and project page