

STARTUP SUMMER EXPERIENCE '26 · FINAL PRESENTATION

AI Skill Evaluator

A skill to evaluate skills — a 5-question framework that scores any internal AI automation on reliability, time saved, and risk, and returns a verdict: keep, optimize, or kill. Built in five weeks inside Coverflex.

Meet the team.



Teresa Pimentel

Economics · NOVA SBE

Framework design and validation
— drafted the scoring model, ran
expert and creator interviews, built
Legal Minion.



Sarp Özveren

Management Engineering · ITU

Scoring engine and automation —
built the scorecard pipeline, Slack
integration, and explain-skill.

Coverflex, in one slide.



One card, every benefit



Flexible compensation for 13,000+ companies and 230,000 employees — meal, insurance, childcare, budgets, flexpay.

Increasingly powered by AI skills



A growing library of internal automations — decks, legal review, inbox triage — each with a real per-run cost.



Martinho Aragão

EIR · Coverflex — our host and advisor inside the company

THE PROBLEM

Powerful automations. No proof they pay off.



Costly to run

Every automation spends real money on every run — API calls, compute, maintenance.



No repeatable decision mechanism

"Should we keep this one?" was answered by gut feel — differently every time.



Invisible value

Time saved never shows up anywhere — nobody sees a number for what a skill returns.

Our mission: Save the value from "spirit world" and make it visible.

LEARNING BY BUILDING

First, we built the things we'd judge.

CHALLENGE 1 · TOGETHER

**inbox-scanner**

Scans the inbox for product-capability questions and drafts replies for human review.

CHALLENGE 2 · TERESA

**Legal Minion**

First-pass contract review for the legal team — yes, that's the official name.

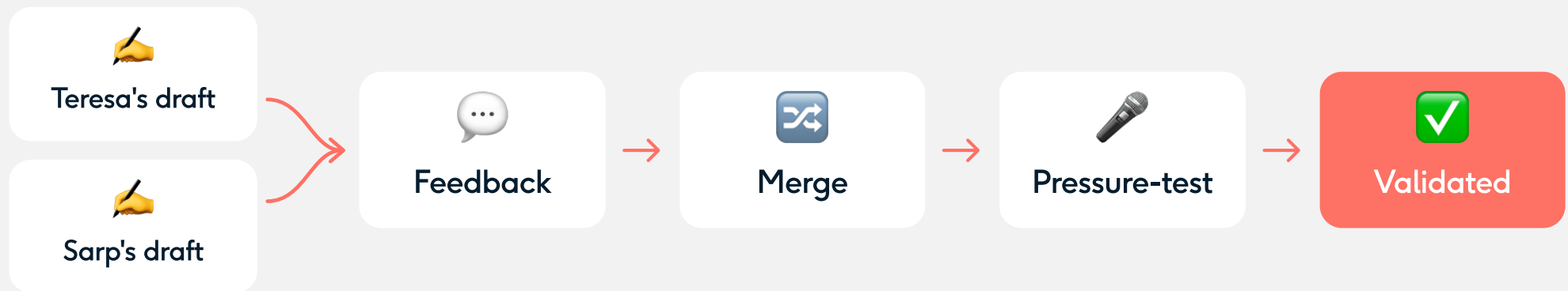
CHALLENGE 2 · SARP

**explain-skill**

Turns any skill's documentation into a visual workflow diagram anyone can read.

HOW THE FRAMEWORK WAS BORN

Two drafts. One framework. Many iterations.



Independent drafts, merged with Martinho's feedback, pressure-tested with Fintech House experts and creators across six departments — then locked.

GETTING SKILLS SCORED

Every skill gets scored. Two ways in.



Weekly Slack nudge

Skills in the library that aren't evaluated yet get an automatic weekly message.



Auto-publish

The scorecard is generated and published automatically with the skill.



SKILL SCORECARD · COVERFLEX-DECK

Manual time per run (minutes)

180

Accuracy (1–3)

1

2

3

Consistency (1–3)

1

2

3

Fabrication (1–3)

1

2

3

Run scorecard ↓

Either way in, the owner answers the same five plain-language questions — no technical knowledge needed.

THE FRAMEWORK IN ACTION

Same questions. Three verdicts.

coverflex-deck

Keep ✓

8.2/10

Strong ROI, trust holds

Reliability 3.83 / 5



Time saving 5 / 5



Risk 3.33 / 5



COST / RUN	ROI / RUN	MIN SAVED
€0.24	+239.2×	114 min

explain-skill

Optimize

5.8/10

One weak spot to fix

Reliability 2.5 / 5



Time saving 3.33 / 5



Risk 3.33 / 5



COST / RUN	ROI / RUN	MIN SAVED
€0.1316	+61.0×	49 min

inbox-scanner

Kill ✕

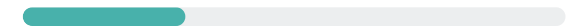
4.0/10

Effort better spent elsewhere

Reliability 3.5 / 5



Time saving 1.5 / 5



Risk 4.0 / 5



COST / RUN	ROI / RUN	MIN SAVED
€0.04	+0.4×	5 min

HOW WE MEASURE THE WORK

What we learned. What we hit.



A solid AI foundation

Coverflex's tooling let us ship real automations in days, not weeks.



Users beat assumptions

Every interview changed the framework — it got good when other people touched it.



Startup pace is real

Five weeks felt like five days — scope ruthlessly, ship weekly.

KPI	TARGET	ACTUAL
Departments covered	≥ 3	6 teams 
Usable by non-technical staff	Yes	5 questions 
Challenges completed	3	3 / 3 

WHERE THIS GOES FROM HERE

Next steps for Coverflex.



Recalibrate with real data

Feed completed scorecards back into the model to tune weights and verdict thresholds.



Integrate Claude's New Cost Tracking Tool

Replace estimated per-run cost with measured spend for sharper ROI numbers.

FIVE WEEKS AGO, THIS QUESTION HAD NO ANSWER

"Is this skill worth it?"

Now, value lives with us in the physical world!

RIP inbox-scanner 

Outcome: a validated 5-question scorecard, adopted across 6 departments, with automated Slack intake — and the first data-backed kill decision at Coverflex.