Harnessed AI Agents: A Universal Framework for Democratizing Statistical Expertise

Junshui Ma, Ph.D.
AVP & Head of Biometrics Research
Merck & Co., Inc., Rahway, NJ, USA

ASA NJ/PT chapter Spring Symposium
June 26, 2026

In This AI-Era…

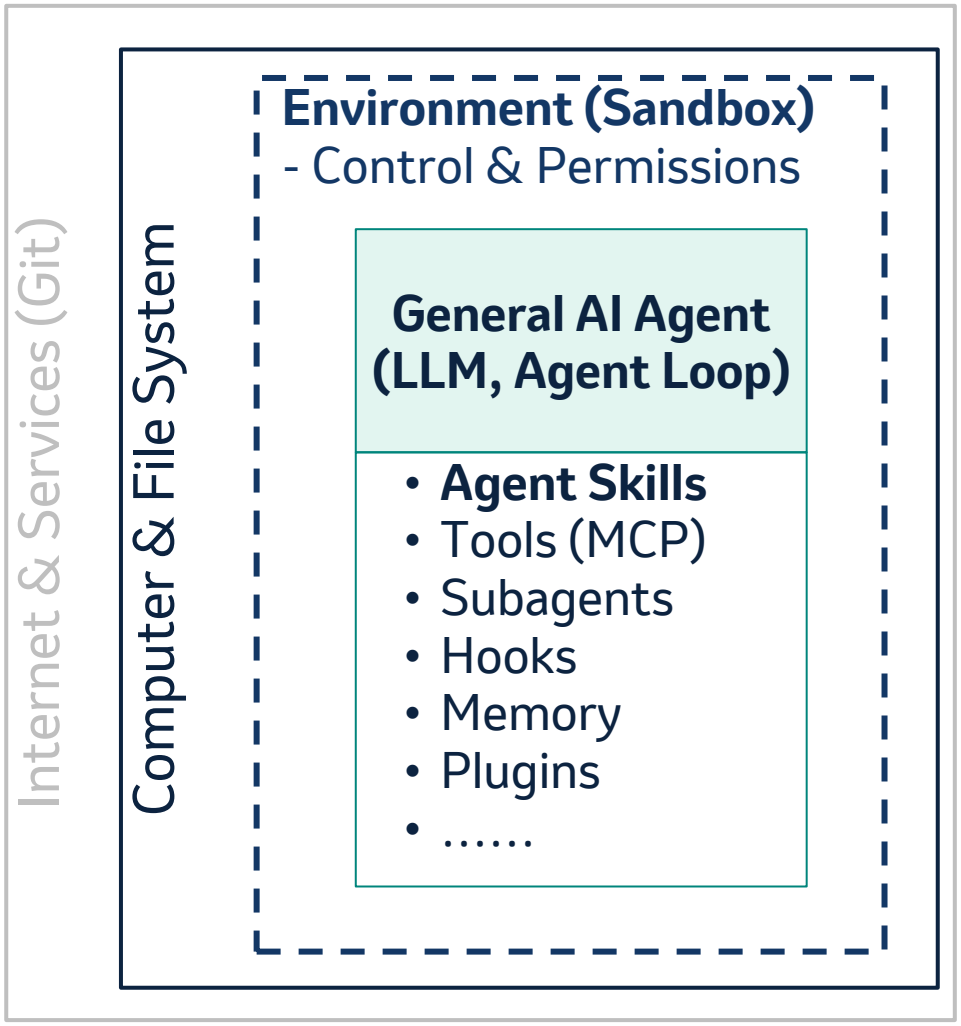
We work with **Harnessed AI Agents**

- **AI Chatbots** just talk, whereas **AI Agents** do things.
- 50+ AI agent platforms are competing, with a few beginning to dominate.
  - Claude Code (Anthropic)
  - Codex (OpenAI)
  - Pi Agent (Open Source)
  - ……

Agent Harness

- The **agent harness** is a system of additional layers applied to general-purpose AI agents to ensure they work effectively for **your needs**.

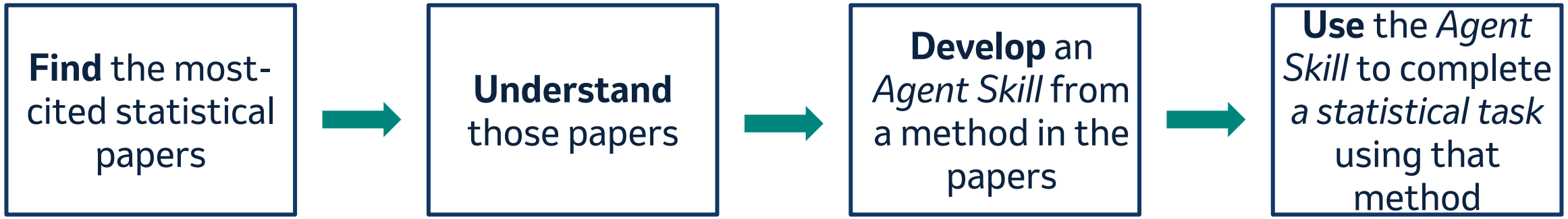
- You control and facilitate your AI agents through your **agent harness.**



MERCK

Our Plan

- The agent harness system is highly **personalized** and **dynamic**.
- This plan **demonstrates** snapshots of several elements within my agent harness system.



These four tasks were completed within 2-3 hours while I attended several meetings, thanks to my agent harness system.

What Are The 10 Most-Cited Papers in *Statistics in Medicine*?

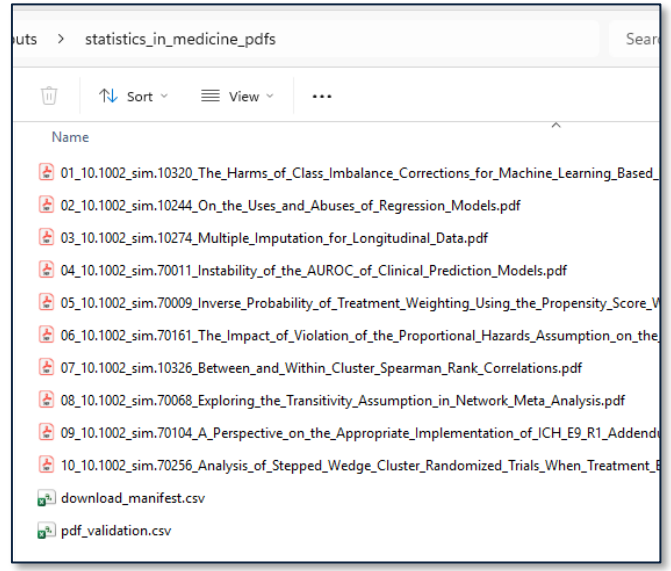
Prompt: List the 10 most-cited papers in *Statistics in Medicine* for **2025** and **2026**

Codex

Rank	Citations	Publication date	Paper	DOI
1	30	2025-01-26	The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study	10.1002/sim.10320
2	18	2025-06-01	On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching	10.1002/sim.10244
3	17	2025-01-23	Multiple Imputation for Longitudinal Data: A Tutorial	10.1002/sim.10274
4	15	2025-02-08	Instability of the AUROC of Clinical Prediction Models	10.1002/sim.70011
5	14	2025-02-06	Inverse Probability of Treatment Weighting Using the Propensity Score With Competing Risks in Survival Analysis	10.1002/sim.70009
6	12	2025-06-01	The Impact of Violation of the Proportional Hazards Assumption on the Calibration of the Cox Proportional Hazards Model	10.1002/sim.70161
7	12	2025-01-24	Between- and Within-Cluster Spearman Rank Correlations	10.1002/sim.10326
8	11	2025-03-30	Exploring the Transitivity Assumption in Network Meta-Analysis: A Novel Approach and Its Implications	10.1002/sim.70068
9	8	2025-05-01	A Perspective on the Appropriate Implementation of ICH E9(R1) Addendum Strategies for Handling Intercurrent Events	10.1002/sim.70104
10	7	2025-09-01	Analysis of Stepped-Wedge Cluster Randomized Trials When Treatment Effects Vary by Exposure Time or Calendar Time	10.1002/sim.70256

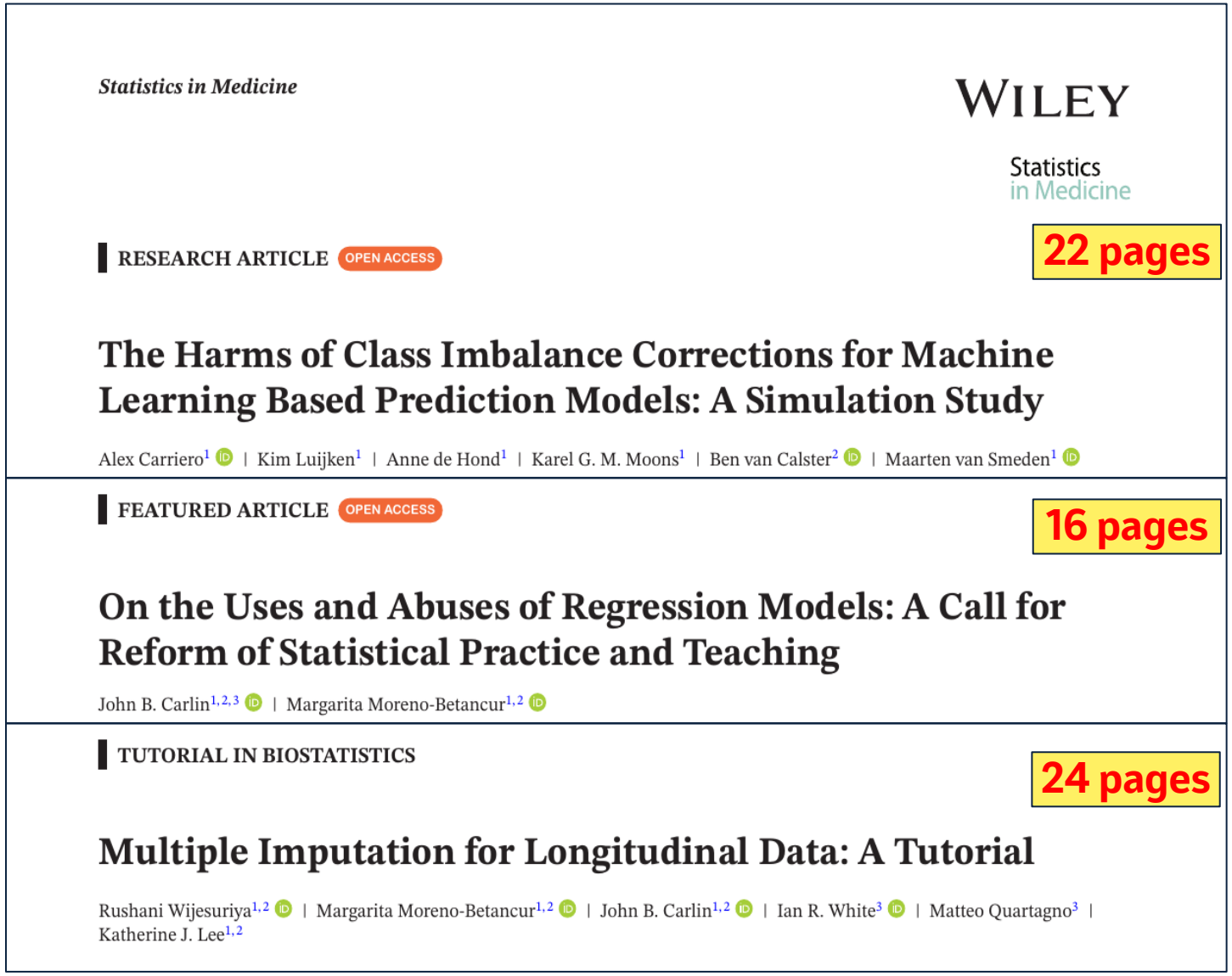
Prompt: Download all the papers as PDFs

Codex



Different Agent systems can have different capabilities, e.g., *Microsoft's Copilot failed to produce meaningful results in this case.*

Understand The Top 3 Papers via Agent-Assisted Reading



Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

22 pages

The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study

Alex Carriero[1] | Kim Luijken[1] | Anne de Hond[1] | Karel G. M. Moons[1] | Ben van Calster[2] | Maarten van Smeden[1]

FEATURED ARTICLE OPEN ACCESS

16 pages

On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching

John B. Carlin[1,2,3] | Margarita Moreno-Betancur[1,2]

TUTORIAL IN BIOSTATISTICS

24 pages

Multiple Imputation for Longitudinal Data: A Tutorial

Rushani Wijesuriya[1,2] | Margarita Moreno-Betancur[1,2] | John B. Carlin[1,2] | Ian R. White[3] | Matteo Quartagno[3] | Katherine J. Lee[1,2]

How can I grasp the points, insights, and key details of these papers quickly?

Agent-assisted Reading:

- For each paper, the *Claude Code agent* was used to generate a ***paper-note***:
  1. Use **"Read-Research-Paper" skill** to generate the ***initial paper-note***.
  2. Use my **"pi-ask" subagent** to critically review it.
  3. Finalize the ***paper-note.***
- I read the ***paper-notes***, referring to the original papers as needed.

MERCK

Agent Skill

- **Agent Skill**, proposed by Anthropic in 2025, injects specific knowledge into AI agents.
- **Agent Skill** is essentially **a folder of files** to teach AI agents specialized capabilities when needed.
- **The "New Employee" Analogy**
  - Imagine you just hired a capable person who lacks company-specific knowledge.
  - **Agentic Skill** resembles providing the new employee with a folder of instructions and examples for completing a specific task in your specific company.

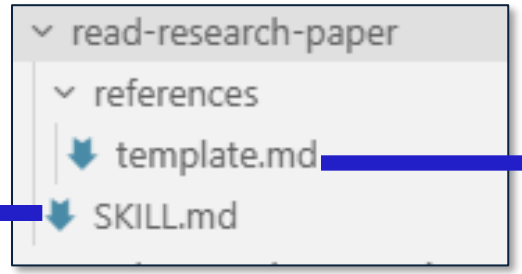
```
Protocol-Writing-Skill/
   ├── SKILL.md          # The required core file
   ├── scripts/          # executable/example codes
   │  └── example_code.R
   ├── references/       # extra on-demand details
   │  └── best_practices.md
   └── assets/           # templates, images, logos
      └── Protocol_template.doc
```

Transferring a Skill = Copying a folder of files

The *Read-Research-Paper* Skill

```
---
name: read-research-paper
description: "Reads a research paper and tells it back as one continuous story — the life of the
paper's core claim, told on a seven-beat spine (protagonist / bind / old road / turn / method /
payoff / kernel): born in a bind measured against a base-rate ruler, crystallized as a bold conjecture,
argued through mechanism and evidence, distilled into a new way of seeing, then walked out of the paper
— life-tested and cashed into falsifiable predictions. Output opens with a scannable TL;DR card
(one-liner / big idea / remember just three things) that compresses the whole story for the time-poor
reader and the six-months-later self, then tells the full story in Markdown. The job is storytelling
that makes the paper land for a smart non-specialist, not academic critique. Use when the user shares
an arXiv link, paper URL, PDF, or asks to read, explain, walk through, or analyze a research paper.
Trigger words: 'read this paper', 'explain this paper', 'walk me through', 'analyze this paper', 'break
down this paper', or when the user shares an academic paper."
---

# read-research-paper: tell a paper as one story

The hard part of reading a paper isn't understanding it — it's being able to *retell* it. Explain it to
a smart person who doesn't know the field, well enough that they can repeat it back, and only then have
you actually read it.

This is a storytelling job. Behind every paper there's a protagonist, a bind, a wall they hit, a turn,
a method, a payoff, and a kernel. Stand that spine up first, then hang the content on it. Skip the
spine and you get ten separate status reports stapled together — the reader drifts off after two pages.

The real protagonist of the story is the paper's *claim* — the bold conjecture the authors are willing
to bet on. The whole piece is the life of that claim: born in a bind (predecessors stuck on a dilemma,
with the water level of the field sitting right there), made to stand up under argument (mechanism and
evidence holding it up), and finally walked out of the paper — past the advisor's scrutiny, down into
lived experience, and out toward a bet on the future. When the paper ends, the claim doesn't leave the
room.

## The overall goal (read this first)

A **smart person who doesn't know this field** should finish your write-up able to retell the paper as
a story — able to say five things:

1. **The bind** — the wall in front of the protagonist (concrete, down to one example), plus the water
level of this problem space (the base rate: where the state of the art sits, how big a typical step is)
2. **The claim** — the bold conjecture the authors bet on, in one sentence
3. **The argument skeleton** — how the mechanism moves, how the evidence holds it up (including the
most counterintuitive side-finding — often the most interesting beat in the story)
4. **The kernel** — the portable new lens: *oh, you can look at it this way*
5. **The test** — where the claim holds and where it breaks in real life; if it's right, what we should
```

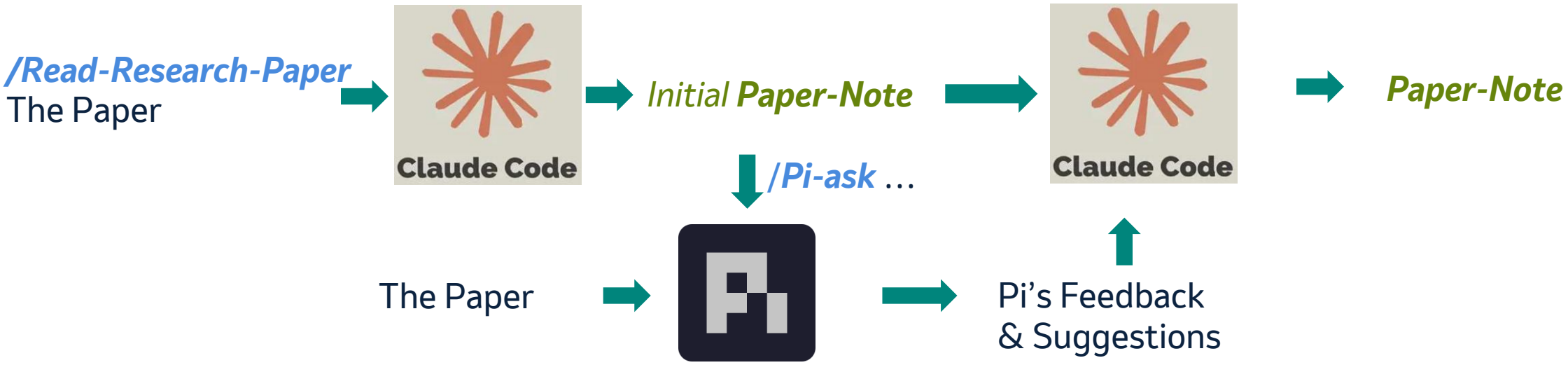
~ 410 lines




```
---
title:    "{3-8 word idea-bone line — the story-kernel line, distilled like an aphorism / chapter
title, no jargon; see SKILL.md 'How to write the title line'}"
subtitle: "{the paper's original title, as published}"
date:     {YYYY-MM-DD}
source:   {URL or source description}
authors:  {author list}
venue:    {publication venue / year}
tags:     [paper]
---

# {title}

## TL;DR

{A card for two kinds of reader: the one who hasn't decided whether to read the whole thing, and the
six-months-later self who just wants the core. Three lines, each finer than the last — the long story
below is just these three lines unfolded beat by beat. Written last, placed first: you can't press it
out without understanding the whole paper.}

### One-liner

{State the whole paper plainly: what bind the protagonist was stuck in, what the authors bet, how it
turned out. Not the soul-line title above (that one is forged into an aphorism and needs a subtitle to
catch it) — this is the line anyone gets in one read: "what did this paper do."}

### Big idea

{If you could carry out only one idea, which is it. Same core as "Insight" — flashed up front here for
direction; "Insight" lands it after the story and gives a second use. Don't reuse the same sentence in
both.}

### Remember just three things

{In six months you'll forget most of it; these three you won't: why the bind is worth solving, what
holds the claim up, what you carry out of the paper. The five things from "the overall goal" pressed
into three — load-bearing points for rebuilding the story, not the bet (that's in "Test"), not the
action list (that's in "Takeaways").}

## Problem

{Set the stage. A concrete example — who the protagonist is, what they run into — that pulls the reader
to the scene of the bind. State it in a sentence or two.}
```

Pi-ask: The Third-party Review Skill and Subagent

- Never blindly trust the output from any AI agent.
- The primary agent's output should be critically reviewed by a third-party reviewer.
- The ***Pi-ask***, a skill-driven subagent built from the **GPT-5.5**-based **Pi agent**, acts as the third-party reviewer.



MERCK

The Generated *Paper-Note*

Problem defines the issue
↓
Translation turns it into a claim
↓
Key Concepts hands you the tools
↓
Insight is what sticks
↓
Advisor/Test/Takeaways applies the claim to the real world

Section	What it does (Using Paper #2 as The Example)
TL;DR	For the reader in a hurry, three zoom levels: (1) one-liner, (2) big idea, (3) three things to remember.
Problem	The bind. lived through one example: a 1995 study whose "parallel curves" were a built-in assumption mistaken for a finding. Sets the ruler: regression misused in ~44% of papers.
Translation	The turn. States the claim (*name your question (describe / predict / cause) before fitting a model)* and shows the same regression meaning three different things across three studies.
Key Concepts	The three tools the argument can't do without: the three question-types, i.e., the estimands. Define them first.
Insight	The kernel you carry out: a model is a verb, not a noun — something you do to answer a question, not a picture of reality.
Advisor Review	A senior statistician's verdict: what's solid, what's assumed (is the trichotomy exhaustive?), how much to trust the 44%. Lands on strong accept.
Test	Takes the claim outside: where it holds and breaks in a real example (café foot-traffic), plus one falsifiable bet about journal reform.
Takeaways	What you can do tomorrow: write your question before your model; stop letting the data pick the question.

The Three One-liners

Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study

Alex Carriero[1] | Kim Luijken[1] | Anne de Hond[1] | Karel G. M. Moons[1] | Ben van Calster[2] | Maarten van Smeden[1]

FEATURED ARTICLE OPEN ACCESS

On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching

John B. Carlin[1,2,3] | Margarita Moreno-Betancur[1,2]

TUTORIAL IN BIOSTATISTICS

Multiple Imputation for Longitudinal Data: A Tutorial

Rushani Wijesuriya[1,2] | Margarita Moreno-Betancur[1,2] | John B. Carlin[1,2] | Ian R. White[3] | Matteo Quartagno[3] | Katherine J. Lee[1,2]

Any common theme?

When predicting rare outcomes, machine-learning practices tell you to "fix" the lopsided data by rebalancing the classes before training. The authors showed that this “fix” consistently made the model overestimate rare cases, which often couldn’t be undone by re-calibration.

Medical researchers often use multivariable regression to best fit the data and draw conclusions from its coefficients. The authors argued that this is backwards: you must first pin down your questions, because the same regression means different things depending on why you ran it. Using regressions the wrong way produces confident nonsense.

When imputing missing values in a longitudinal study, the method must respect the data’s grouping and the planed downstream analysis. This tutorial ensures that the imputations follow these principles and provides guidance on when to use which method.

MERCK

A Skill for Multiple Imputation (MI) for Longitudinal Data

TUTORIAL IN BIOSTATISTICS

Multiple Imputation for Longitudinal Data: A Tutorial

Rushani Wijesuriya[1,2] | Margarita Moreno-Betancur[1,2] | John B. Carlin[1,2] | Ian R. White[3] | Matteo Quartagno[3] | Katherine J. Lee[1,2]

- Paper #3 offers **comprehensive guidelines** on the task of MI for Longitudinal Data.
  - **Rule**: Match the imputation method with the downstream analysis; do not always use the most complex imputation tools.
- Developing an agent skill, ***Longitudinal-MI***, from these guidelines would enable non-experts to perform the MI task properly.

➔ Democratizing "*MI for longitudinal data*"

MERCK

Develop and Refine *Longitudinal-MI* Skill

/STORM-Research
MI for longitudinal data

→ A broad web research using the **STORM 4-phase research method**

↓

/Read-Research-Paper
Paper #3

→ ***Paper note*** of Paper #3

↓

/Skill-Creator
Research+***Note***+**Paper 3**

→ The Initial *Longitudinal-MI* Agent Skill

↓

/Pi-Ask
for fidelity review

→ The Reviewed *Longitudinal-MI* Agent Skill

↓

Refine */Longitudinal-MI*
with simulated tasks

→ The Delivered *Longitudinal-MI* Agent Skill (ver. 1)

```
.claude/skills/longitudinal-mi/
├── SKILL.md                          # the governing principle + decision tree + workflow
├── references/
│   ├── methods.md                    # full method catalog, settings, failure modes
│   ├── decision_guide.md             # lookup tables: which method for which analysis
│   └── verified_simulation_run.txt   # the 200-rep proof (regenerated evidence)
├── scripts/
│   ├── simulate_and_test.R           # Monte Carlo proof of the thesis + a self-test
│   └── impute_longitudinal.R         # runnable end-to-end: the paper's real methods
└── evals/evals.json                  # 3 realistic test prompts + assertions
```

Use the *Longitudinal-MI* Skill

The "MindGrow" Task:

- **Question**: Does having depression early lower later math scores?
- **True Effect**: $B = -0.30$
- **Population**: Same students measured at grades 7, 9, and 11.
- **Missing Data**: 21% of data missing (worse at later grades), in both the outcome and covariates

Use ***/longitudinal-MI*** skill to impute the dataset, and then do the analysis.

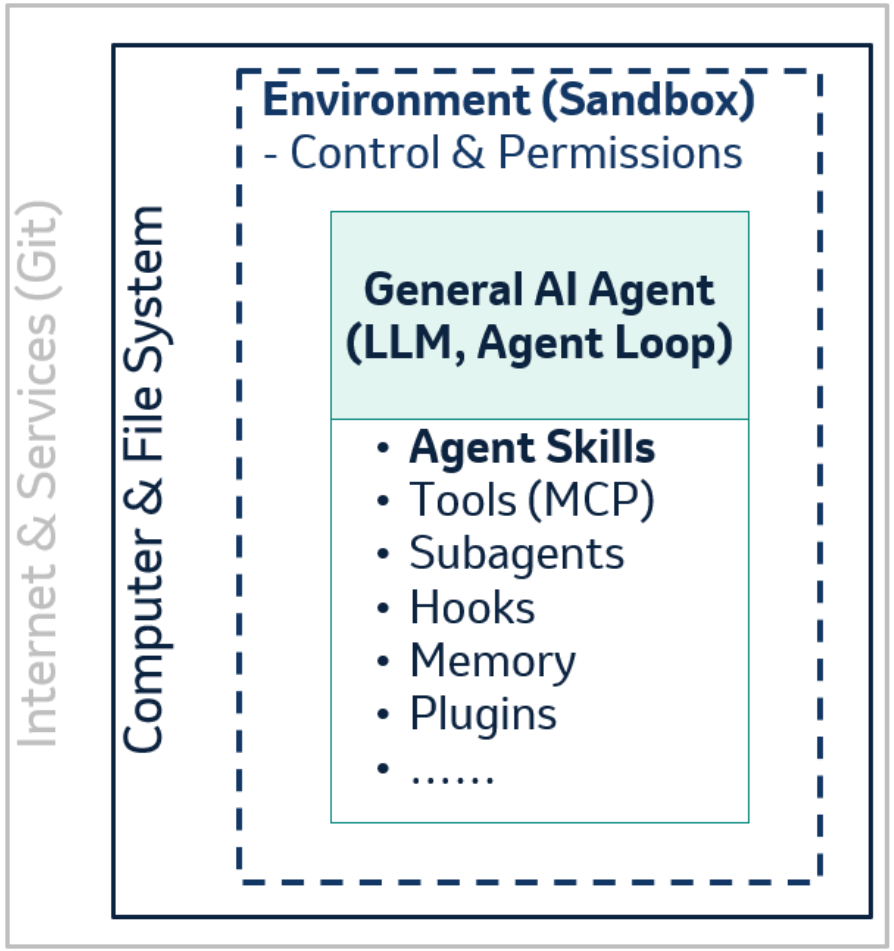
Method	$\widehat{B}$	Between-student Variance
Full Data (Gold Standard)	−0.312	0.824 – Reference
Complete Case Only (No MI)	−0.305	0.814 (99% of Reference)
Naïve MI (Ignore grouping)	-0.295	0.448 (54% of Reference) – **Collapsed**
MI by the Skill (FCS-WIDE*)	-0.314	0.835 (101% of Reference) – **Recovered**

** **FCS-WIDE**: Based on the guidance, the Skill selected the FCS (Fully Conditional Specification) method from the R package `mice` and transformed the data into wide format (i.e., one student per row; different grades as columns).*

MERCK

An Agent-Oriented Permissive and Facilitating Environment

- *"The four tasks were completed within 2–3 hours while I attended several meetings, thanks to my agent harness system."*
- My agent harnessed system was specially set up for agents' autonomous, long-term operation.
  - Running agents within a **Sandbox** that I specially designed for my tasks and infrastructures.
  - Supplying context, knowledge, and tools necessary to complete the tasks efficiently.
  - Granting agents **full permission** within the sandbox environment.
- Therefore, I can let agents automatically
  - finish complex tasks during my meetings, or
  - work on many tasks simultaneously.



The Common Theme

Common Problem: Local methods >> Big picture

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study

Alex Carriero[1] | Kim Luijken[1] | Anne de Hond[1] | Karel G. M. Moons[1] | Ben van Calster[2] | Maarten van Smeden[1]

Rebalance Methods >> Question of Risk Prediction

FEATURED ARTICLE OPEN ACCESS

On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching

John B. Carlin[1,2,3] | Margarita Moreno-Betancur[1,2]

Regression Modeling >> Question to Address

TUTORIAL IN BIOSTATISTICS

Multiple Imputation for Longitudinal Data: A Tutorial

Rushani Wijesuriya[1,2] | Margarita Moreno-Betancur[1,2] | John B. Carlin[1,2] | Ian R. White[3] | Matteo Quartagno[3] | Katherine J. Lee[1,2]

MI Methods >> Downstream Data Analysis Plan

MERCK

Are Statisticians Facing a Crisis in this AI Era?

- In my opinion, statistics was already in crisis before 2022.
- Powerful agents only forces us to seriously reconsider where our value lies.

Summary

- **Agent harness**: Extra layers added to general agents to enhance their effectiveness for you.
- Among various agent harness technologies, **Agent Skill** was specifically introduced.
- A 4-step task (i.e., find, understand, develop and use) was used to demonstrate my **workflow with harnessed agents**.
- Harnessed agents erode statisticians' moats but also present **opportunities**.