

# Evaluations and Audit Trails for Trustworthy AI Agents in Pharmacometrics



Ari Pritchard-Bell

Amgen, Clinical Pharmacology, Modeling & Simulation

# Elements of Trust

**01**

## **Human skills**

From copilots to engineered multi-agent workflows

**02**

## **Agent skills**

Unit operations and audit trails

**03**

## **Benchmarks and beyond**

Evaluation standards and optimization

# Elements of Trust

**01**

## **Human skills**

From copilots to engineered multi-agent workflows

**02**

## **Agent skills**

Unit operations and audit trails

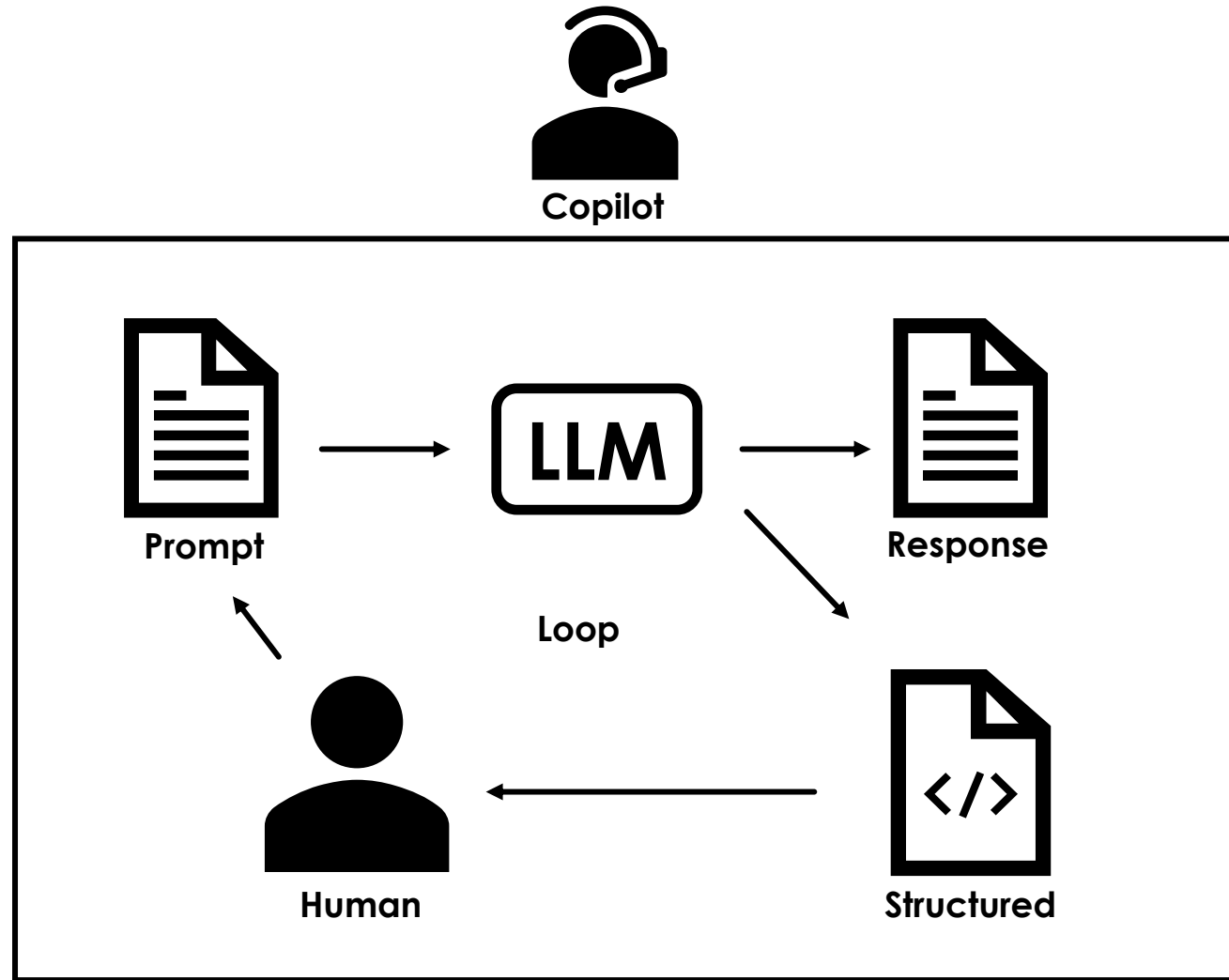
**03**

## **Benchmarks and beyond**

Evaluation standards and optimization

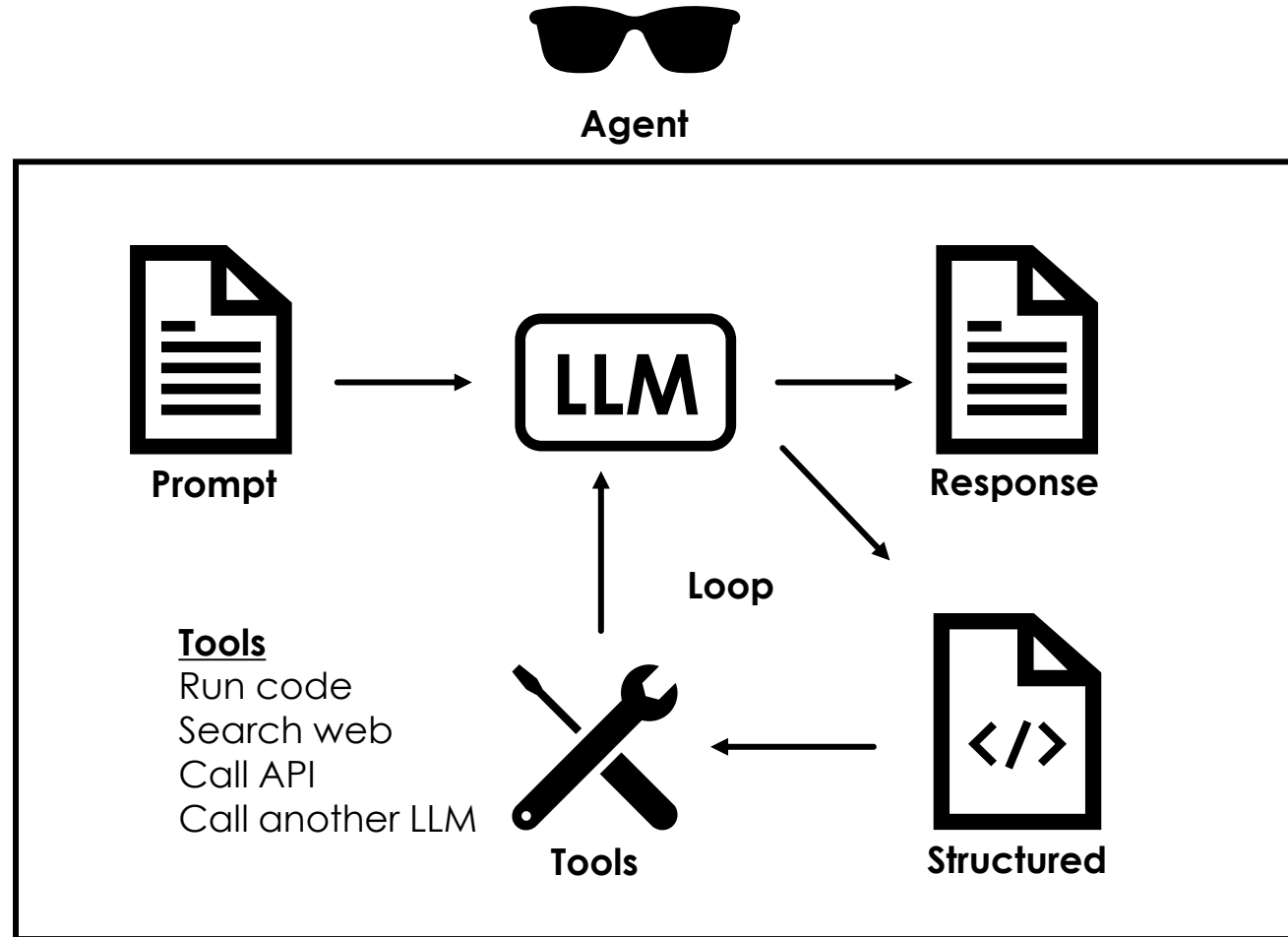
# Agentic Workflows

## Human in the loop

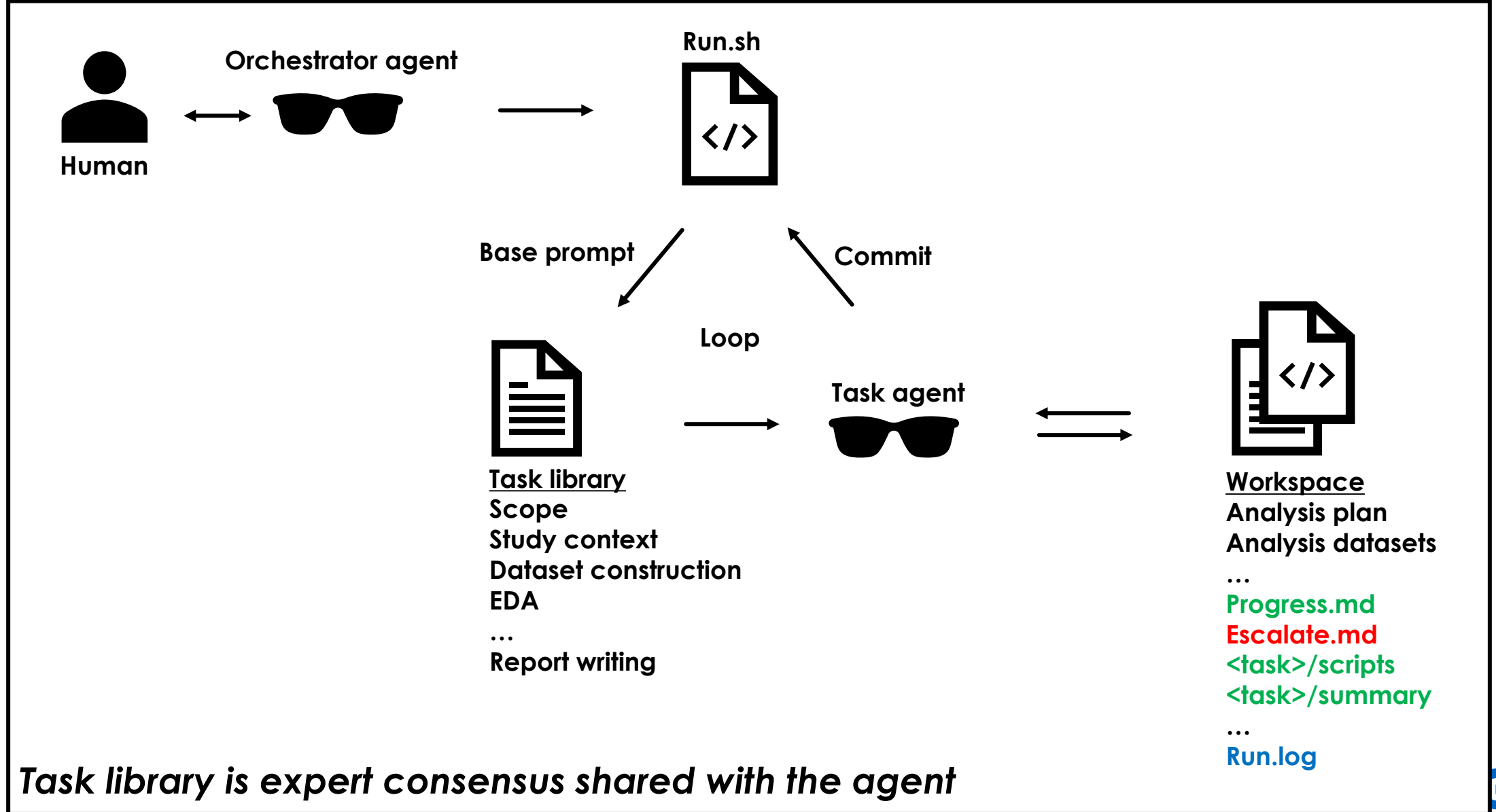


# Agentic Workflows

Agent as a callable function



# Context Engineering



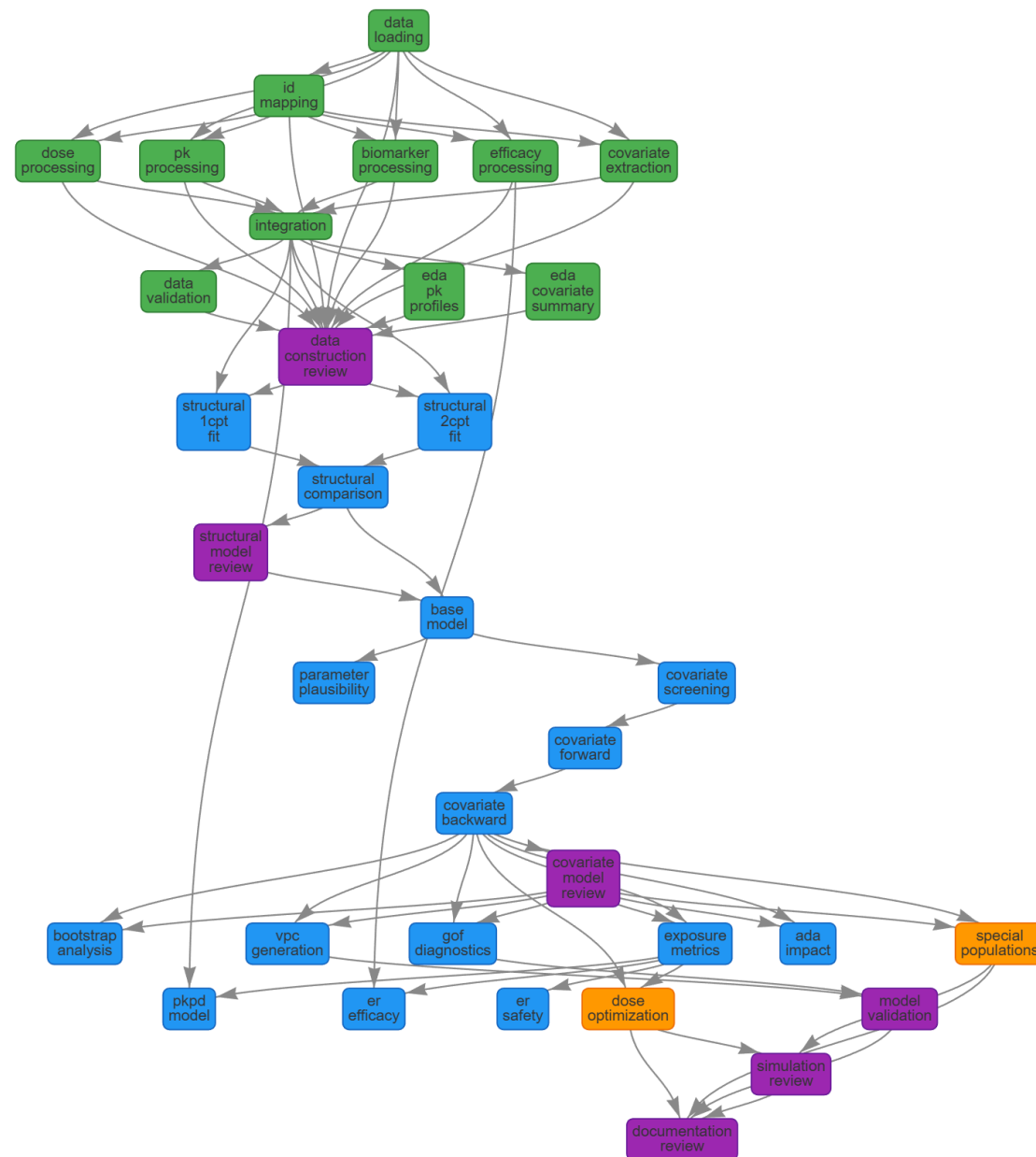
# Task Library

## Expert consensus shared with the agent

```
{
  "task_id": "covariate_analysis",
  "category": "modeling",
  "description": "Assess whether the candidate covariates explain inter-individual variability in the disposition parameter of interest",
  "depends_on": [
    "structural_model",
    "study_context"
  ],
  "rules": [
    "Read study_context.md for the drug and the pre-specified candidate covariate set; use the structural model fit",
    "Test each pre-specified candidate against a stated statistical criterion, and estimate any relationship you retain rather than fixing it",
    "Retain a candidate only if it adds explanatory power beyond what already-supported covariates account for, and the effect is scientifically defensible for this drug",
    "Judge a retained effect by its magnitude, not only its significance: a statistically supported effect whose estimated size is implausible against the drug's mechanism or established physiological expectation is a signal of an inadequate or unconverged underlying fit, to be reconciled before it is reported as final",
    "Bound every fit (capped iterations + timeout); screen candidates cheaply before committing to full estimation, and compute uncertainty only for models you carry forward",
    "Report every candidate's decision and reason; report as supported only the covariates the analysis defends"
  ],
  "produces": "covariate_analysis_results.csv, covariate_analysis_summary.md",
  "verify": "each pre-specified candidate is tested with a reported effect size and statistical basis; every candidate carries a supported/rejected decision with a one-line reason; a retained covariate is defensible beyond a univariate association, and its estimated effect magnitude is checked for plausibility, not reported on significance alone"
}
```

tasks.json covariate\_analysis entry

7



Full published task library



# Elements of Trust

01

## Human skills

From copilots to engineered multi-agent workflows

02

## Agent skills

Unit operations and audit trails

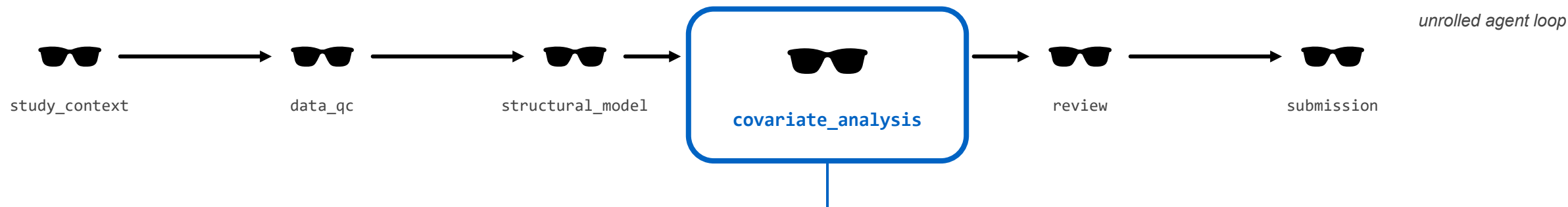
03

## Benchmarks and beyond

Evaluation standards and optimization

# Agent Skills

## Unit operations for pharmacometrics



### Inputs



#### General prompt

the same base prompt for every task



#### Task entry = SKILL.md

rules, examples, produces, verify  
specific context for covariate analysis



#### Workspace: previous task outputs

`study_context.md`, structural model fit, QC'd dataset



**`covariate_analysis`**

fresh agent, one task

### Outputs (`covariate_analysis/`)



`fit_covariate.R`



`covariate_analysis_results.csv`



`covariate_analysis_summary.md`



interim/ fit objects, code logs, etc.

### Between iterations

Persistent files (state) shared by agents



#### `tasks.json`

`covariate_analysis`: passes = true



#### `progress.md`

appended progress log



#### `workspace/`

cumulative outputs, organized by task

# Reproducibility

## Results are reproducible without the agent

### Workspace

tasks.json  
progress.md  
run\_20260921\_135110.log  
scope/  
study\_context/  
data\_qc/  
structural\_model/  
**covariate\_analysis/**  
review/  
submission/  
    submission.yaml  
    scorecard.yaml  
ANALYSIS\_COMPLETE.txt

### covariate\_analysis/



**fit\_covariate.R**  
the script; re-run it, no agent needed



**covariate\_analysis\_results.csv**  
candidate parameter tests, with decisions/justification



**covariate\_analysis\_summary.md**  
method, decisions/justification, final values



**interim/**  
fit objects, forward\_steps.csv, fit\_log.txt

*covariate\_analysis\_summary.md written by agent*

```
# covariate_analysis
## Result
Of the three pre-specified candidates (WT, ALB,
CRCL) tested on both disposition parameters with
IIV (CL, Vc), **only body weight is supported**,
acting on **both CL and Vc** as an allometric
power relationship. ALB and CRCL are rejected.
The final covariate model improves fit by
**dOFV -73.4 over the structural base**.

## Decisions & rationale
- **Retained WT on CL and Vc**: supported
statistically (dOFV -73 total), incrementally
(each clears forward criterion), and
mechanistically (canonical allometric exponents).
- **Rejected ALB**: univariate signal is a WT
artifact; vanishes conditionally.
- **Rejected CRCL**: no signal + mechanistically
implausible (non-renal mAb).
- Uncertainty computed only for the final model.
```

Reviewable

Executable

Traceable

# Elements of Trust

01

## Human skills

From copilots to engineered multi-agent workflows

02

## Agent skills

Unit operations and audit trails

03

## Benchmarks and beyond

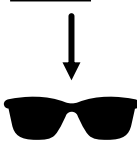
Evaluation standards and optimization

# Benchmark Scenario



## The packet

protocol.md, sap.md, dataset.csv,  
submission.template.yaml



## Agent

fresh directory per run  
claude -p, non-interactive mode



## submission.yaml

one answer per item

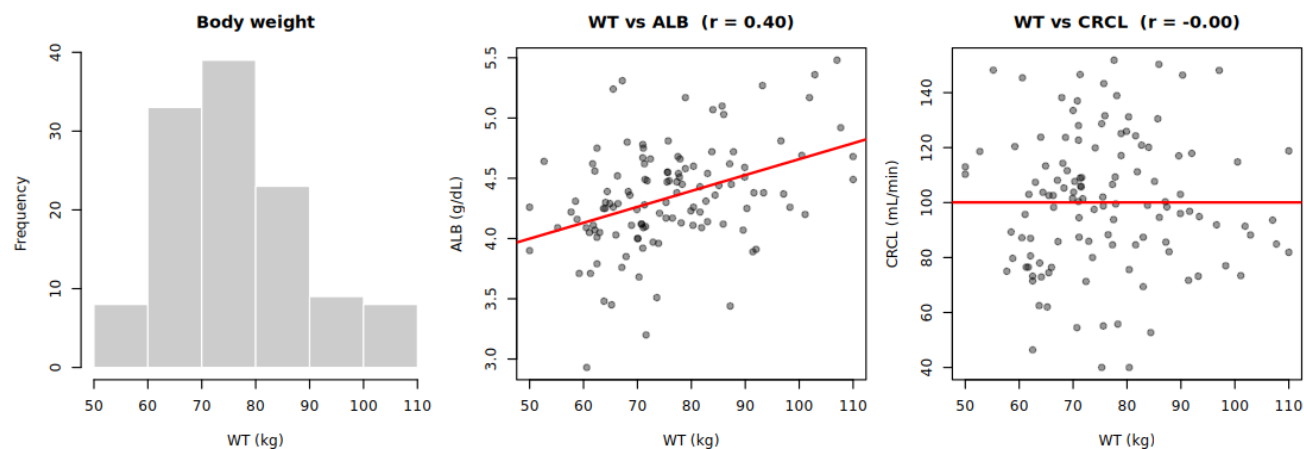
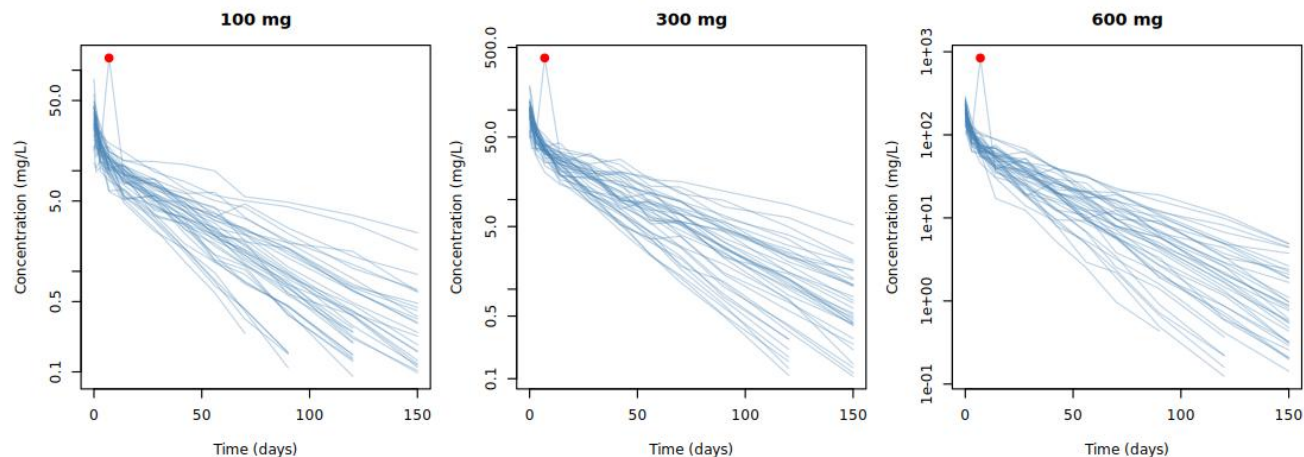


## Scorer: truth.yaml + score.R

0 to 1 per item

First-in-human mAb PopPK, 120 subjects, single 1-h IV infusion,  
100 / 300 / 600 mg, serum PK to day 150, LLOQ 0.1 mg/L

Concentration-time by dose (log scale); red = injected day-7 outliers

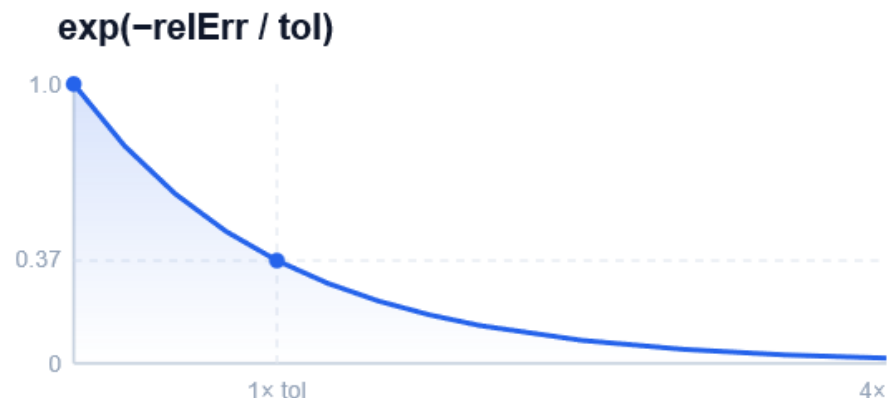


Albumin tracks weight ( $r = 0.40$ ): a correlated decoy; creatinine clearance: an independent decoy

# Evaluation Standards

How do I know the results are “right”?

| ITEM               | AREA               | SCORER      | GROUND TRUTH                                                                                  |
|--------------------|--------------------|-------------|-----------------------------------------------------------------------------------------------|
| structural_ncmt    | structural         | categorical | 2 compartments (exact match)                                                                  |
| disposition_params | estimation         | map         | CL 0.20 · Vc 3.0 · Q 0.6 · Vp 3.0<br>tol 20% per param; aliases resolved                      |
| error_model        | estimation         | map         | proportional 0.15 (tol 25%)                                                                   |
| cov_effects        | covariate-analysis | map         | WT on CL <sup>0.75</sup> & Vc <sup>1.0</sup> (tol 20%)<br>ALB, CrCL = decoys → score 0 (trap) |
| outlier_records    | data-quality       | set         | rows 62 / 692 / 1322 (right set)<br>subjects 5/50/95, day-7 DV ×10 slip                       |



**numeric** score =  $\exp(-\text{relErr} / \text{tol})$

**categorical** exact match → 1, else 0

**set** right set of picks (F1); misses & extras both hurt

precision (P) = correct / submitted

recall (R) = correct / expected

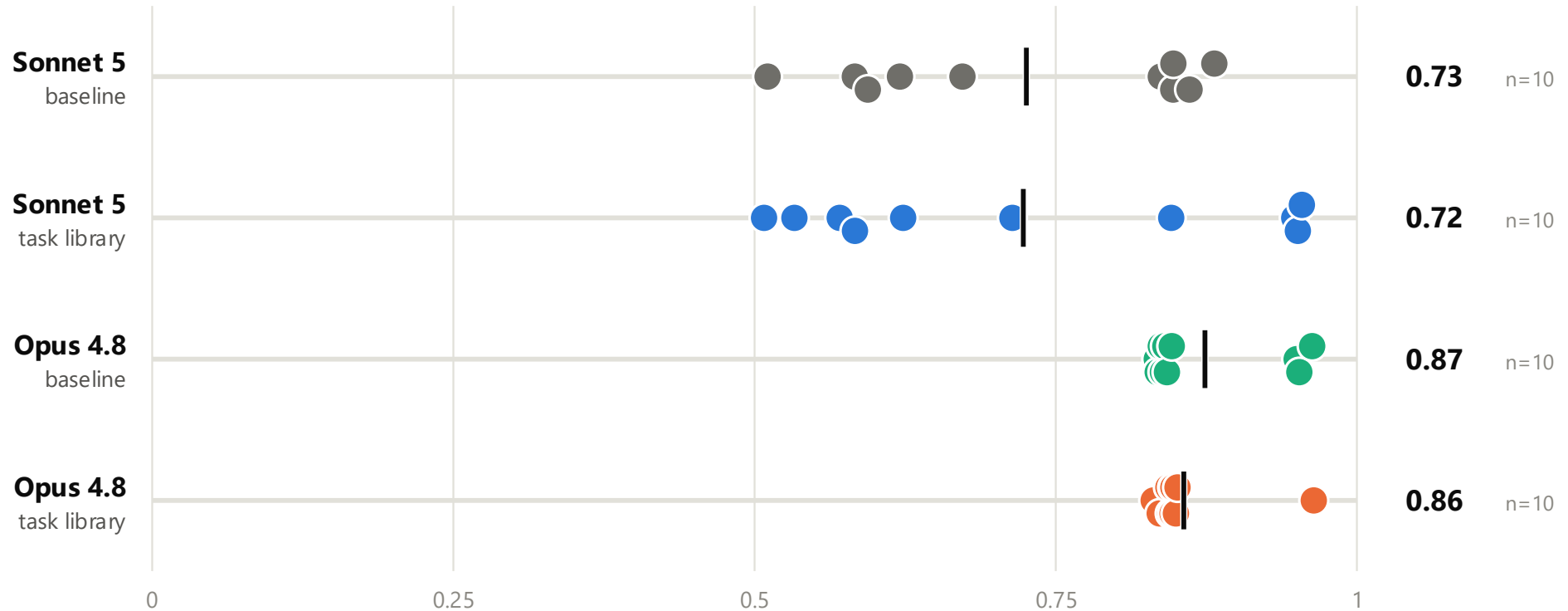
$F1 = 2 \cdot P \cdot R / (P + R)$

**map** like set, but each match graded by closeness

$\text{map} = \sum \exp(-\text{relErr}/\text{tol}) / (\# \text{ distinct names})$

# Evaluation of Model and Harness

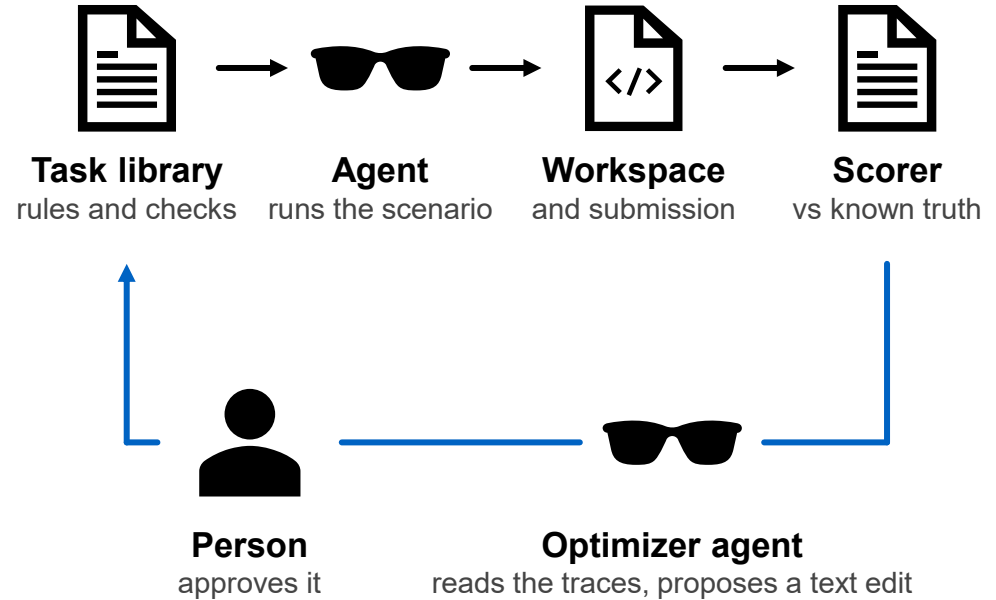
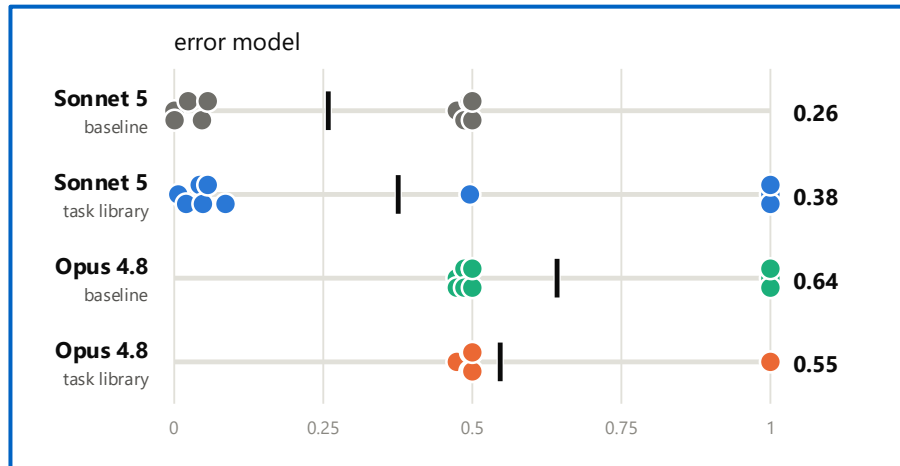
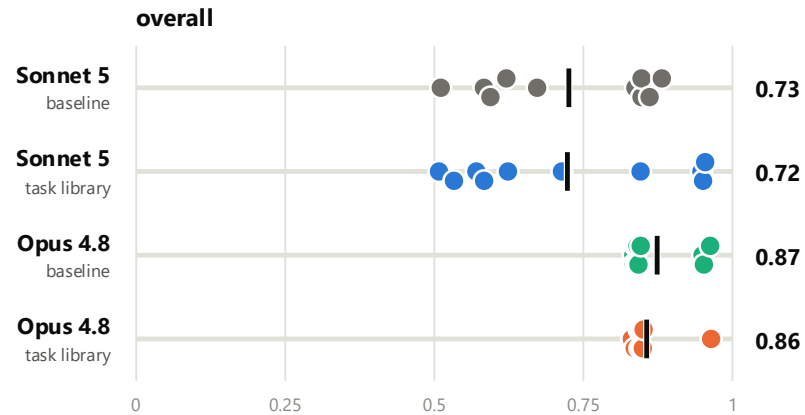
Differentiates model but not harness



PMbench overall score, 40 blinded runs of the same mAb pop-PK study.  
One circle per run (10 per arm); tick = mean.

# Measure then Optimize

Task context iteratively improved against evaluation (trust) metrics



## Error model task/skill improvement

33 of 40 runs kept an additive term the simulation does not have  
Add rule text:

“Keep an additive error term only if the low-concentration data support it.”

PMbench overall score and the three items that vary, 40 blinded runs of the same mAb pop-PK Structural model and outlier handling scored 1.0 in all 40 runs and are not shown.

# Elements of Trust

**01**

## Human skills

From copilots to engineered multi-agent workflows

**The human must know how and when to use an agent**

**02**

## Agent skills

Unit operations and audit trails

**Modular tasks build trust incrementally**

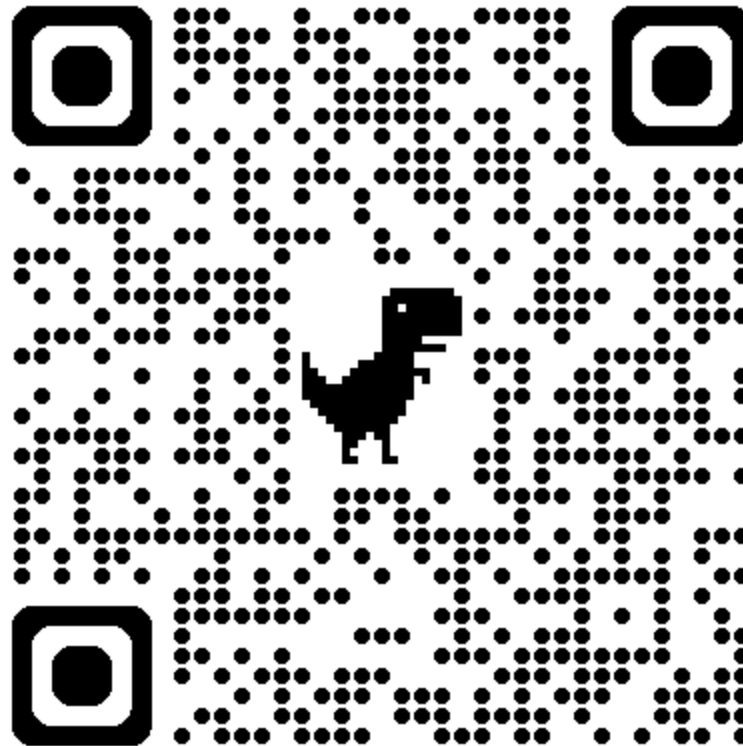
**03**

## Benchmarks and beyond

Evaluation standards and optimization

**Community defined metrics unlock greater autonomy**

# Thank You



<https://github.com/AIML-SIG/Agentic-workflows>