TrialDesignBench

Benchmarking AI agents on clinical trial statistical design

Eric Zhang · Eikon Therapeutics · 10/09/2026

Agent Intelligence in Practice

⚙ Harness — right context, right time



Context

memory · knowledge (RAG) · workflow (skills)



LLM

the brain



Tools

e.g. MCP

↺ repeat: call tool → update context → decide next step, until task complete

Example: Expanding Context with a Skill

Group Sequential Design skill — given a trial scenario, it designs a group sequential survival-endpoint trial: boundaries, sample size, multiplicity strategy, and a simulation-verified report, through a guided Q&A.

Skill structure

```
group-sequential-design/
├── SKILL.md              # Q&A workflow, design steps
├── reference.md          # design rules, spending fns, alpha recycling
├── examples.md           # R code by design pattern
├── post_design.md        # timing checks, verification
└── scripts/
    ├── compute_event_prob.R
    └── gsd_report_template.py
```

Loaded progressively — only pulled into context when the workflow reaches that step

Output structure

```
output/gsd_{disease}_{endpoi
├── gsd_design.R
├── gsd_results.json
├── multiplicity_diagram.png
├── gsd_verification.R
├── gsd_verification_log.md
├── gsd_report.py
└── gsd_report.docx
```

Demo video

Benchmark Run: Evaluating the Skill

Any technical addition to an agent — skill, tool, workflow — should be benchmarked with-vs-without, not assumed to help. 21 design scenarios, skill vs. no-skill — overall scorecard (best-with-skill run per scenario):

	With Skill	Without Skill
Skill wins	16	—
Ties	4 (#3, #21, #60, #113)	—
No-skill wins	1 (#139)	—
Avg score (excl. #60 trivial)	**90%**	**62%**

1 · Does it add value?

2 · How good with skill?

90% avg vs. 62% no-skill, at each scenario's best run

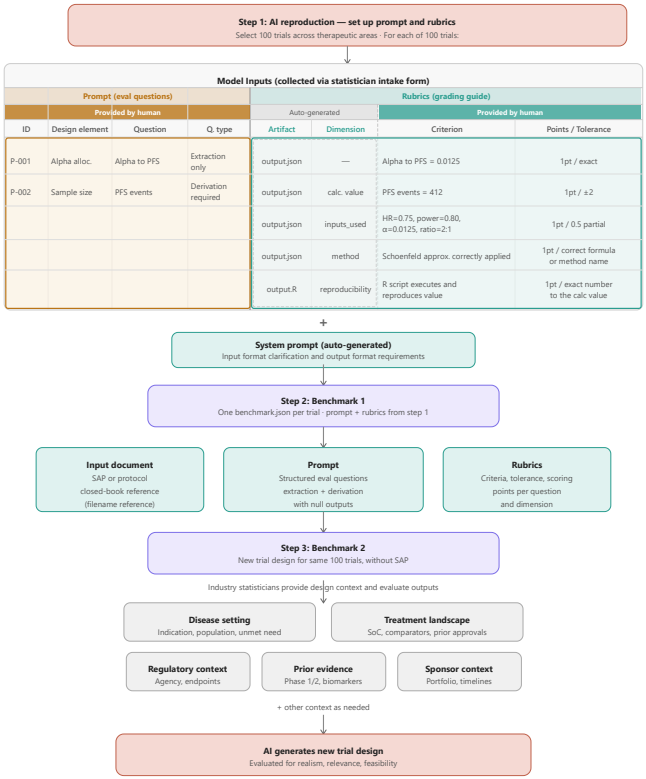
3 · Failure patterns

- **P0** four-arm, contribution-of-components design

TrialDesignBench

- Evaluates whether AI agents can **reproduce** or **design** clinical trial statistical designs
- **Benchmark 1** — reproduce, closed-book with SAP/protocol
- **Benchmark 2** — design, no SAP, only context (disease, treatment, regulatory, sponsor)

🔎 Full workflow spec



Building the Benchmark

1 · Intake form

Statisticians write design questions, and rubrics for a trial

HuggingFace Space ↗

2 · Benchmark review

QA pass on submitted cases before they enter the benchmark

HuggingFace Space ↗

3 · Evaluation framework

Runs models against the benchmark, scores outputs against rubrics

Built on Harbor ↗

Example: Question → Rubric → Grade

KN189 (KEYNOTE-189), P-004 — derivation question:

"What is the total required number of PFS events at its final analysis based on Lachin and Foulkes' method for the variance estimation?"

Inputs used

- Target HR = 0.7 ✓ +1
- Randomization ratio 2:1 ✓ +1

Calculated value

- PFS events at IA2 ≈468 ±3 ✗ 0

Model derived 472 — out of tolerance

Method

- Fixed design via Lachin-Foulkes ✓ +2
- L-F uses event fractions, not allocation fractions ✗ 0

Model scored **13 / 18** on this prompt

Where We Are



26 / 100 trials

toward the target benchmark size

We'd love your help. Write a test case — question + rubrics — from a protocol/SAP you know well or find interesting, and help stress-test AI on real trial design.