NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

M4Bench

Evaluating Procedural Specification for Clinical EHR Derivation Agents

Hannes Ill[1], Rafi Al Attrach[1], Rajna Fani[1], Ahram Han[1,2], Ryohei Yamamoto[1], Marlene Louisa Moerig[1,3], Dukyong Yoon[1,4], Leo Anthony Celi[1,5]

[1]Laboratory for Computational Physiology, MIT · [2]Dept. of Surgery, Seoul National University Hospital · [3]Charité, University Medicine Berlin · [4]Dept. of Biomedical Systems Informatics, Yonsei University College of Medicine · [5]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center

Paper
neurips.cc

Code + data
github.com/hannesill/m4

+0.136 reward with clinician-reviewed skills
Codex, 26/28 tasks improve · Claude Code replication +0.120

+0.162 from a complete operational spec alone
The skill adds +0.024 on top, CI includes 0 (8-task control subset)

25/280 with-skill runs pass every strict check
Key F1 moves +0.003: gains are partial agreement, not solved tasks

1 Derived variables aren't columns

SOFA, KDIGO AKI, MELD, Sepsis-3: EHR studies rebuild them from raw tables, and each implementation choice changes cohorts and conclusions.

`chartevents` `labevents` `inputevents` `outputevents`

↓

PROCEDURE THAT LIVES OUTSIDE THE SCHEMA

item codes · time windows · unit conversions · missingness rules · output grain

↓

SOFA score `per ICU stay`

2 Skills: written procedures

A **skill** is a markdown procedure for one task, adapted from MIT-LCP MIMIC-Code and eICU concept code. No SQL, no reference outputs.

`sofa-24h / skill.md`

```
Scoring rules   6 organ systems, 0–4 points each
Item IDs, tables  vasopressors, PaO₂/FiO₂
                  pairing, ventilation
Time window     first 24 h after ICU admission
Missingness     missing component scores 0
Output grain    one row per ICU stay
```

reviewed by 2 clinicians

3 M4Bench at a glance

28 tasks	**15** task families	**11+17** standard + raw tasks
26+2 MIMIC-IV + eICU	**2** clinician reviewers per task	**1,850** paper-facing runs, traces released

4 One run, end to end

Task + database
Task card names the derivation, keys, and value columns.
- DuckDB clone of MIMIC-IV / eICU, **reference targets removed**
- Standard: intermediate tables kept · Raw: rebuild from source tables

↓

Condition
- no-skill: task + database only
- **with-skill: + task-specific skill.md**
- controls: decoy skill · raw SQL · operational spec · schema probes

↓

Coding agent in Docker
– Codex: GPT 5.4 Mini, GPT 5.5
– Claude Code: Opus 4.7, Sonnet 4.6
– isolated workdir, API-only network
– canary + contamination lint

→ `output.csv`

↓

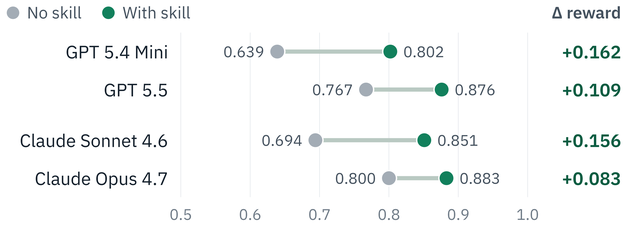
Score vs hidden reference
– **Reward**: mean match rate over value columns
– **Strict pass**: every check passes
– **Key F1**: right rows (output grain)

Two clinicians review every task and skill against source definitions; provenance is recorded per task.

5 Controls isolate the mechanism

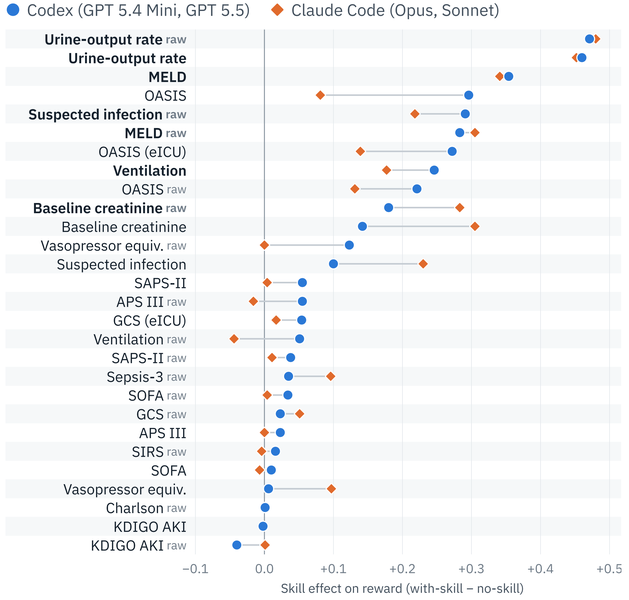
CONDITION	AGENT ALSO GETS	TESTS
No skill	nothing	baseline
With skill	task skill.md	targeted procedure
Decoy skill	another task's skill	extra context?
Raw ref. SQL	reference SQL text	code structure
Operational spec	full rules, no item IDs	completeness vs format
Schema probes	renamed / restructured DB	schema familiarity

6 Skills raise agreement



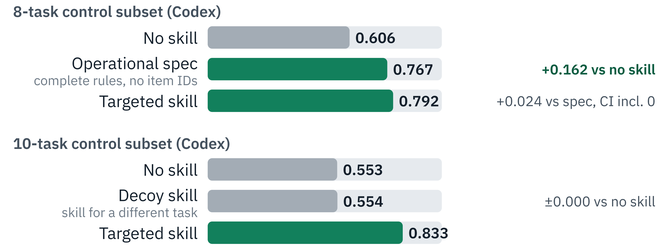
Reward averaged over 28 native-schema tasks, 5 trials per cell. Claude Code replication is descriptive (harness differs by condition).

Per task: gains where implementation decides



- Largest: urine-output windows, MELD, suspected-infection timing, baseline creatinine, ventilation
- Flat: KDIGO, Charlson, SOFA, APS III; six of the eight flattest score ≥ 0.94 without skills

7 Completeness drives the gain



- Decoy skill = no skill: not generic context. Raw SQL ≈ skill: the reference procedure carries the signal

Subsets fixed in advance; they over-represent large-effect tasks.

8 Robust to schema changes

86% of the gain kept with **obfuscated** names

70% kept with **restructured** tables

GPT 5.4 Mini, 26 MIMIC-IV tasks. Descriptive; not a bound on memorization.

9 Higher agreement ≠ solved

13 of 280 no-skill runs pass every strict check

25 of 280 with-skill runs

Key F1 +0.003 468 of 522 failing runs return the right rows but miss a value threshold

Takeaways

- Write the full procedure down: rules, item IDs, windows, grain
- Completeness of the spec matters more than its format
- Treat output as a draft: check keys, duplicates, value ranges; get clinical review

Scope. Scores measure agreement with public-code reference derivations, not clinical correctness. 26 of 28 tasks use MIMIC-IV.

github.com/hannesill/m4 · huggingface.co/datasets/hannesill/m4bench-evidence · illh@mit.edu