WORLD LAUREATES FORUM

2026

International Conference
on AI for Engineering

World Laureates Foundation

INSTITUTE OF AI FOR ENGINEERING

Execution Feedback and Review in LLM-Assisted Engineering Design

A Controlled Study of Power and Water Networks

Yuchen Xie Peiran Ren*

Tongji University * Corresponding author

01 Research Contributions

Existing LLMs for constrained engineering design.

1 Fixed constraints and action checks
Fix assets, permissions and limits. Verify actions.

2 Feedback and additional review
Share drafts, selected tests and one revision.

3 Frozen scoring and local reversion
Hold out final tests. Revert fields to trace effects.

02 Model APIs and CPU Scoring

23 networks screened; **16** feasible references.
13 power + 3 water; 3 model settings × 2 runs.
96 drafts × 7 routes = **672** correlated records.

LLM draft — Fixed scope → **CPU test** — 4/8 dev. tests → **LLM review** — Same / cross → **LLM revise** — One revision

CPU scoring after freeze: 8 dev. + 8 held-out tests

Feedback varies by route. Held-out results stay hidden.

04 Action and Execution Checks

Separate legacy water batch: 6 / 24 labels changed.
4 changed from fail to pass; 2 from pass to fail.

LV rural 2: 16 → 0 → 16
Initial → revised → tap schedule reverted

Net1: 3 → 0; revert threshold: 3, pump speed: 0

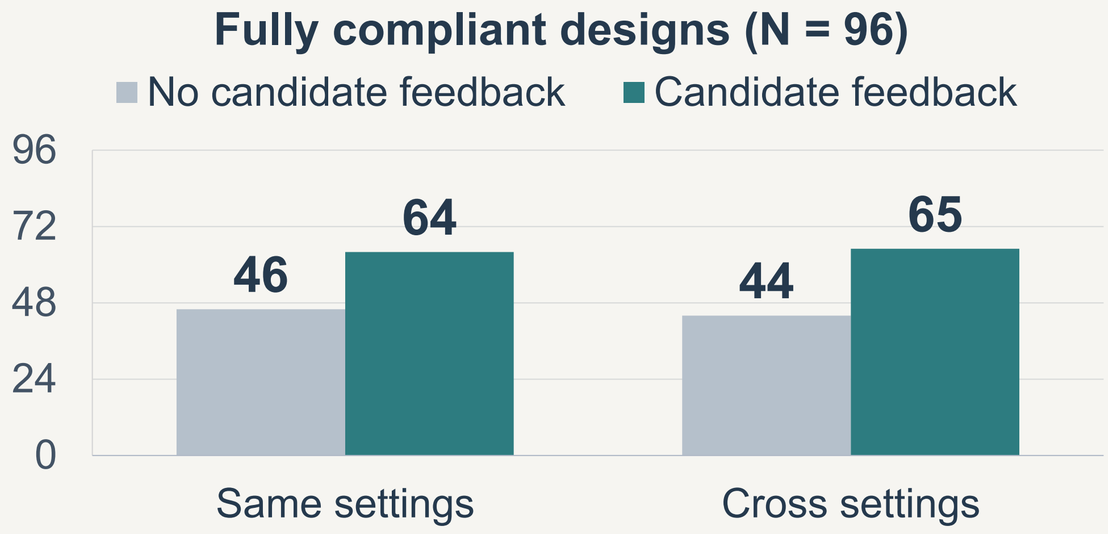
Counts: failed tests out of 16 per design.
2 of 4 post hoc cases, not general causal evidence.

03 Feedback Gains and Limits

Passing all 8 held-out tests (N = 96 each)

Same: 51 → 68 **Cross: 46 → 70**

Chart below: all 16 tests and structural checks passed.



2 × 2 comparison with shared drafts; N = 96 per route.

Paired changes per draft: gains, losses and net gain

Comparison	Gain	Loss	Net
Feedback: same	20	2	18
Feedback: cross	22	1	21
Cross-setting review[1]	5	4	1

Limited additional benefit from review

Full and masked review both pass 64 / 96.
Each wins on 4 drafts. Stable added benefit remains unestablished; review may still help.

[1] Cross vs. same settings, with feedback.
Water ceiling and related sources limit generalization.
672 routes share 96 drafts; they are not independent.
CPU budgets and permissions differ; no fair ranking.

05 Future Work

Broader tasks
Add independent networks and harder water tasks. Test transfer across networks.

Matched resources
Match candidate, test and token budgets plus resets. Add stronger optimization.

Feedback components
Separately mask diagnostics, numeric results and advice. Measure each contribution.

Cross-validation
Prespecify interventions. Use independent solvers or measurements for validation.

Reuse: frozen constraints, scores and CPU replay
github.com/Sumimi24/opentrust-research
examples/engineering-replay/

4 frozen records, 64 test runs: replay matched
Isolated environment on the same host; no model calls.
Refs: Self-Refine (NeurIPS 2023); pandapower docs