LLM-Agent Composition for Document-Grounded Conceptual Design of a Natural-Circulation Cooling Loop

Byeongha Jo [a,b], Yonggyun Yu [c*]
[a] *Korea Atomic Energy Research Institute, 111 Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, Korea*
[b] *University of Science and Technology, 217 Gajeong-ro, Yuseong-gu, Daejeon, Korea*
[c] *National Research Council of Science and Technology, Sejong, Korea*
*Corresponding author: ygyu@nst.re.kr

**Keywords : LLM agents, natural circulation, conceptual design, requirement traceability, agent composition*

1. Introduction

Large-language-model (LLM) agents are beginning to participate in nuclear conceptual design, where work must be organized and evaluated so every claim stays traceable to evidence. We have been developing ReactorStudio, a configuration-management platform compiling requirements from organizational levels (program-level, attributed, chat-derived) into machine-checked constraints and recording every agent solver invocation in a server-side tool-call ledger; agent self-reports are never accepted as evidence — satisfaction verdicts are recomputed on the server from ledger-derived values. A stage-gated review workflow for evidence-bounded claims was reported previously [1]. Representative LLM-agent studies for engineering and science — role-decomposed multi-agent frameworks [2], autonomous scientific-experimentation agents [3], and autonomous-agent surveys [4] — emphasize task completion, not jointly addressing requirement provenance across organizational roles, external physics verification, and auditable evidence over agent self-report. Whether decomposing a single agent into multiple roles yields any benefit under matched constraints remains unclear, motivating this comparison. This paper presents the next benchmark in this program: LLM-agent conceptual design of a planar natural-circulation cooling loop, comparing three agent compositions — single, two-role proposer/verifier, and four-role adding a requirement analyst and executor (Table II). The central finding is that specification completeness, more than composition, determined the observed ranking: an incomplete specification produced composition-dependent but physically implausible solutions, while justified design bounds drove all three compositions to essentially identical designs — at which point additional agent roles raised cost by roughly an order of magnitude with no meaningful quality gain. This paper makes three contributions: (1) a document-grounded, server-rejudged benchmark with an external solver and tool-call ledger — not agent self-report — as the evidentiary source for every verdict; (2) a controlled three-composition comparison under matched model, tools, and tool-call budget; and (3) the specification-completeness finding above.

2. Problem Definition

The design object is a closed planar natural-circulation loop — heat source at bottom (the reactor core, a fixed input), riser, heat sink at top (a heat exchanger boundary), and downcomer — driven by buoyancy alone, no pump. The agent controls two design variables: pipe inner diameter D and thermal-center elevation difference Δz between source and sink. Unlike pure component sizing, the objective is nonlinearly coupled to geometry: buoyancy head grows with Δz and achievable temperature rise while frictional and form losses grow as diameter shrinks, so removable heat duty is an implicit function of (D, Δz). Table I lists the benchmark inputs, roles, and sources, distinguishing values transcribed from the NuScale US460 FSAR chapter [5] from declared modeling bounds. Transcribed values were re-checked in repeated independent passes, correcting two provenance-label errors without changing any value. The natural-circulation flow (730.9 kg/s) is a reference value the judge does not enforce, since loop mass flow is instead solved from the buoyancy-friction momentum balance. The program-level requirement is the heat-removal duty; attributed and chat-derived requirements supply the two judge boundary conditions. For the two design variables, no corresponding design values exist in the reference document (an integral PWR lacks a single loop diameter or a thermal-center elevation scalar), so direct transcription is structurally impossible; the scoring reference for (D, Δz) is therefore a grid reference computed with the same judge physics, undisclosed to the agents.

Table I. Benchmark inputs, their roles, and provenance.

Quantity	Value	Role	Source
Heat-removal duty	250 MWt (±1%)	Judged requirement	FSAR Ch. 5 [5]
Natural-circulation flow	730.9 kg/s	Reference only	FSAR Ch. 5 [5]
Cold-side temperature	482 °F	Boundary condition	FSAR Ch. 5 [5]
Hot-side temperature ceiling	598 °F	Boundary ceiling (T_hot as limit)	FSAR Ch. 5 [5]
RCS design pressure	2,200 psia	Fluid property	FSAR Ch. 5 [5]

Quantity	Value	Role	Source
Pipe inner diameter D	300–1,500 mm	Design variable	Modeling declaration
Thermal-center elevation Δz	1–20 m	Design variable	Modeling declaration

Each variable's declared bounds (Table I) carry a documented rationale: they match the grid-reference domain, and the Δz ceiling is a modeling declaration covering the document-transcribed riser/downcomer axial lengths (2.7–13.9 m) of the reference integral PWR [6], not a document-stated limit. The platform further restricts permissible claims: safety validation, licensing satisfaction, and marine applicability are prohibited claim types; document values are treated as established design points rather than physical optima, and evaluation statements are bounded by the model chain employed.

3. Agentic Workflow and Evaluation Design

Each benchmark run proceeds as simulation-in-the-loop: the agent proposes (D, Δz), calls a deterministic physics-analysis tool through the Model Context Protocol (MCP), reads the returned duty margin, and iterates until the server judges all requirements satisfied or the tool budget is exhausted, with environment-injected boundary conditions the agent cannot alter. Two candidate judges were prepared: a one-dimensional loop momentum/energy network (nonlinear friction closure) and a two-dimensional buoyant CFD solve (OpenFOAM v2512); a judge is adopted only after calibration verifies energy closure, since residual convergence alone can mask a large energy-closure imbalance. The 2-D CFD candidate, run live, showed a self-sustained circulation-flow oscillation and failed the steady-state energy-closure criterion (ratios 0.86 and 0.92 against a 0.03 tolerance) — consistent with the classical oscillatory instability of single-phase differentially heated loops [7] — so the one-dimensional judge was adopted. The adopted judge closes the buoyancy-friction momentum balance $\rho\beta g\Delta T_{max}\Delta z = (fL_{tot}/D + \Sigma K)\dot{m}^2/(2\rho A^2)$, with $L_{tot} = 2\Delta z + 2L_h$ and $\Delta T_{max} = T_{hot,ceiling} - T_{cold}$, using Boussinesq fluid properties (ρ, β, cp, μ) from IAPWS-IF97 at the loop's average state point (2,200 psia, mean of T_{cold} and $T_{hot,ceiling}$). Friction factor f follows the Churchill correlation [8] across the laminar-through-turbulent range at the declared pipe roughness; mass flow $\dot{m}$ (removable duty $Q = \dot{m}\cdot cp\cdot\Delta T_{max}$) is the unique positive root of this monotonic residual, found by bracketed bisection at 1e-9 relative tolerance (deterministic), cross-checked against an independent deep-laminar closed form ($f = 64/Re$) to within 1e-6 relative error. Its limitation is a smooth power-law response that may admit few-call approximate inversion, so iteration-trajectory metrics are a secondary axis. The three compositions are portable agent/skill definitions, invoked through the same product trigger; Table II summarizes each composition's roles and responsibility split.

Table II. The three agent compositions.

Composition	Roles	Responsibility split
Single	one agent	proposes, calls the tool, iterates
Two-role	proposer + verifier	proposer suggests, verifier recommends, orchestrator calls tools
Four-role	analyst + proposer + executor + reviewer	analyst interprets requirements and objective, executor calls one tool per instruction, reviewer recommends

In all three compositions the only decision authority is server re-judgment; no role can declare satisfaction or termination. Identical budgets means the same case, requirement set, MCP tool set, wall-clock timeout, and a combined eight-call solver budget shared across roles, with over-budget runs rejected at server verification against the tool-call ledger. All compositions used the same server-configured model (claude-sonnet-5, default decoding settings); they differ only in delegated-role access. Turn and token budgets were not separately constrained, and per-run model identifiers were not logged. Because the machine-checked requirements are pure feasibility conditions, the harness also gives every composition the same selection objective — prefer the smallest diameter among satisfying candidates — so the bounded campaign measures execution efficiency on a fully specified, smooth monotonic task, while the unbounded round addresses an incomplete specification. Metrics follow three axes — workflow quality, reference proximity, and server-recomputed top-requirement margin — with trajectories derived from the tool-call ledger. The benchmark design was revised after an independent-analysis gate: a first bounded-free round (N=10 per composition) revealed agents exploiting the then-unconstrained elevation variable, reaching satisfying designs at Δz up to 100 m — three to seven times the transcribed piping lengths of the reference design — with quality rankings inverting once filtered. Variable bounds were then declared and enforced inside the analysis tool (out-of-range calls rejected and ledgered), and the full campaign was re-run under identical conditions.

4. Results

In the bounded campaign (N=10 per composition, sequential), all 30 runs satisfied every requirement under server re-judgment, with zero ledger-verification failures. Under the shared min-D objective, the three compositions converged: Δz sat at the declared 20 m ceiling in 30 of 30 runs, and the final diameter spanned only 735.0–737.0 mm (per-composition std 0.2–0.5 mm). A continuous reference $D^* = 731.9$ mm (a scalar root solve of the same judge physics at the 20 m bound) places these diameters 3.1–5.1 mm above the continuum feasibility edge. Table III reports full results;

mean absolute duty error decreased from 0.229 MWt (single) to 0.077 MWt (two-role) to 0.039 MWt (four-role), all remaining an order of magnitude below the ±1% acceptance tolerance (≈2.5 MWt) without altering requirement satisfaction, leaving the three compositions operationally indistinguishable in design quality. Multi-agent compositions nonetheless incurred order-of-magnitude higher inference cost — approximately 9–12× mean cost and 12–14× wall-clock time (Table III) — despite making no more, and on average fewer, tool calls.

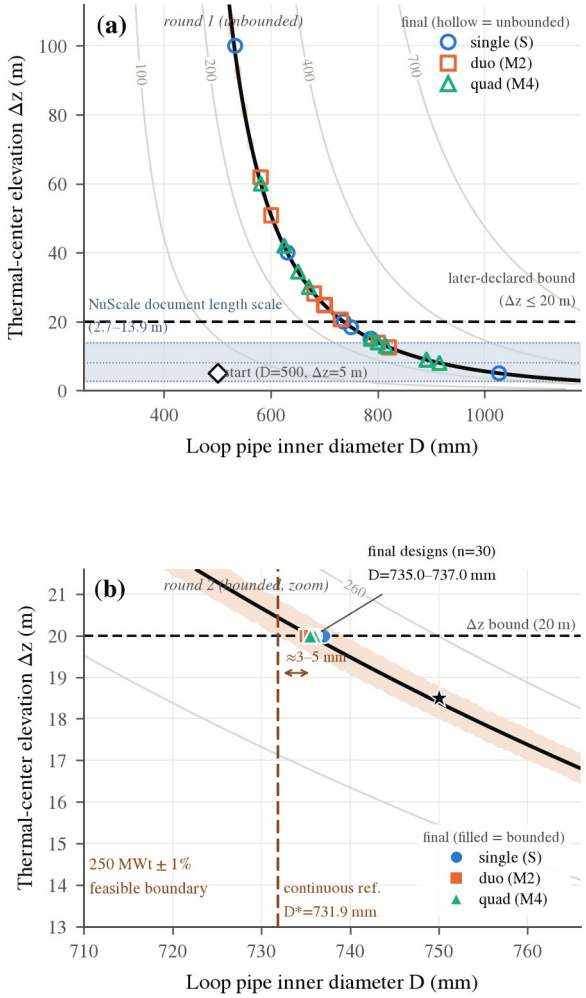


Fig. 1. Two campaigns in the design space. (a) Unbounded-round final designs on the document-derived elevation scale, with the later-declared Δz ≤ 20 m bound. (b) Bounded-round zoom: satisfying band, final designs, grid-resolution reference (star; 750 mm, 18.5 m), and continuous reference D* (dashed).

In the unbounded first round, the two-role composition appeared to achieve the smallest diameters (median 700 mm vs. 750 mm for the single agent), but the independent-analysis gate traced this to the elevation exploitation noted above, and the ranking reversed once filtered (elevation-bounded duo subsample n=2, underscoring the weak support for either ranking). Composition differences thus appeared under an incomplete specification, not on the bounded, fully specified task.

Three limitations apply: runs were sequential, not interleaved; the harness instruments per-run totals only, so per-role LLM-call/token breakdowns are unavailable and the cost increase cannot be decomposed beyond tool-call counts; and reported cost is the top-level session total — controlled delegation runs on the same CLI mechanism confirmed it includes delegated sub-agent usage, though a per-run decomposition of the campaign itself is not retroactively available.

Table III. Bounded-campaign results (N=10 each; mean ± std; all runs satisfied all requirements).

Profile	D (mm)	|ΔQ| (MWt)	Tool calls	Cost (USD)	Time (s)
Single	735.83 ± 0.47	0.229 ± 0.326	6.6 ± 0.8	0.31 ± 0.10	50 ± 12
Two-role	735.41 ± 0.22	0.077 ± 0.144	6.1 ± 1.0	2.81 ± 0.84	603 ± 212
Four-role	735.56 ± 0.16	0.039 ± 0.108	5.6 ± 1.2	3.81 ± 0.84	691 ± 165

5. Summary

We defined a document-grounded planar natural-circulation loop benchmark and compared single, two-role, and four-role LLM agent compositions under server-side re-judgment and ledger-derived metrics. The specification, not the composition, dominated the observed ranking: a bounded, fully specified objective drove all three compositions to the same effective design at sharply different costs, while the unbounded round exposed agents exploiting under-specified requirements. This equal-quality result applies only to a two-variable, monotonic-smooth-physics task with a fully specified objective and hard-bounded search space. The benchmark — its claim-scope restrictions, judge calibration, and specification-completeness lesson — is a reusable template for agentic design in nuclear engineering.

REFERENCES

[1] B. Jo and Y. Yu, A Stage-Gated Workflow for Evidence-Bounded Review of AI-Generated Nuclear Technical Claims: A Reactor Coolant System Pilot, Proc. NAT 2026, submitted, 2026.
[2] S. Hong et al., MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework, Proc. ICLR, 2024.
[3] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, Autonomous Chemical Research with Large Language Models, Nature, Vol. 624, No. 7992, pp. 570-578, 2023.
[4] L. Wang et al., A Survey on Large Language Model Based Autonomous Agents, Frontiers of Computer Science, Vol. 18, 186345, 2024.
[5] NuScale Power, LLC, NuScale US460 SDAA FSAR, Chapter 5, Rev. 1, NRC ADAMS ML23304A338, 2023.
[6] NuScale Power, LLC, NuScale US460 SDAA FSAR, Chapter 4, Rev. 2, NRC ADAMS ML25099A249, 2025.
[7] P. Welander, On the Oscillatory Instability of a Differentially Heated Fluid Loop, J. Fluid Mech., Vol. 29, No. 1, pp. 17-30, 1967.
[8] S. W. Churchill, Friction-Factor Equation Spans All Fluid-Flow Regimes, Chem. Eng., Vol. 84, No. 24, 1977.