VIBE BUILDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated building design must comply with seismic and wind codes and satisfy structural mechanics constraints, yet most existing agents produce visually plausible models without verification grounded in mechanical analysis and code compliance. We introduce the Vibe Building task and propose **PE-Loop** (**P**hysics-**E**ngine-in-the-**Loop**), an agent in which a deterministic physics engine is the sole source of evaluation signals, mapping code constraints to a physics process reward, while the language model is confined to proposing discrete revisions (section menu, topology, and lateral system). Designs are verified by held-out seismic and wind time-history checks and a constructability gate. On VB-Bench, 3,577 physics-adjudicated building instances across six code families, PE-Loop achieves the highest verified success rate under three of four backbone LLMs, the highest held-out seismic pass rate under all four, and the highest held-out wind pass rate under three. Replacing the physics verdict with a language-model judge, all else fixed, leaves 58.43% of accepted designs noncompliant. These results suggest that reliable structural design rests less on a stronger LLM proposer than on an adjudicator the proposer cannot influence, a division of labor for agents whose outputs must hold up in the physical world.

1 INTRODUCTION

In vibe coding (Karpathy, 2025), a developer states intent in natural language and accepts whatever the model produces, as long as it runs. The criterion is permissive because errors are cheap: a faulty program can be reverted and regenerated within seconds. Buildings afford no such recourse. A structural design must satisfy seismic and wind codes and preserve a safety margin that dead loads erode over decades (Meng et al., 2026); an oversight at the modeling stage compounds silently into irreversible failure (Jiang et al., 2026; Chen et al., 2026). Automating design has long been a goal of digital engineering (Makoond et al., 2024). We ask what vibe coding becomes when "runs" is replaced by "stands," and argue that the answer depends not on how a design is generated but on *who is allowed to accept it*.

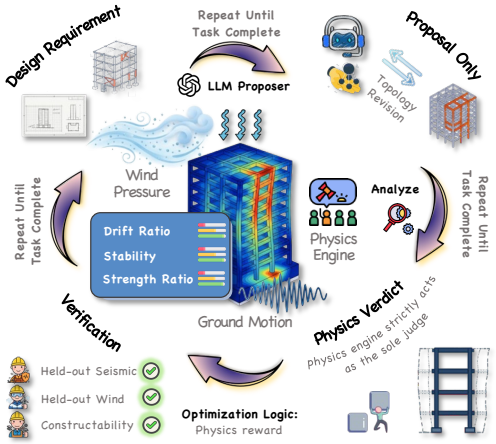


Figure 1: An illustration of PE-Loop.

LLM agents now write and run finite-element code (Qi et al., 2025; Geng et al., 2025) and automate code checking (Liang et al., 2025; Madireddy et al., 2025) (Fig. 1), but the physics engine is usually passive: it executes a committed solution or verifies a known answer. Generative layout methods (Liao et al., 2021; Gu et al., 2024b; Radler et al., 2025) let physics guide the search but cannot read a natural-language brief. Loop- and graph-engineering paradigms (Wang et al., 2026c; He, 2026a; Feng et al., 2026; He, 2026b) supply the orchestration for such multi-stage tasks, yet a language model still commonly renders the verdict, which is neither reliable nor auditable (Zheng et al., 2023; Gu et al., 2024a; Boiko et al., 2023), and without external feedback such a model does not reliably correct its own errors (Huang et al., 2024; Kambhampati et al., 2024). Unit tests

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

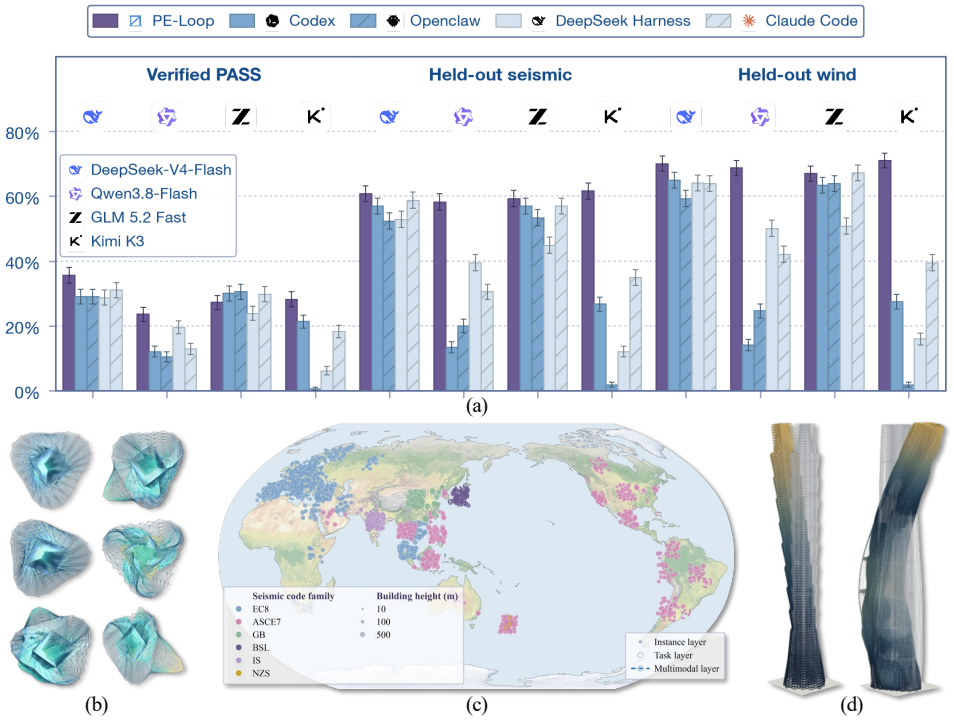


Figure 2: **Swap the backbone, keep the reliability. (a)** Across four LLMs, the held-out pass rate of PE-Loop varies by 3.3 pp (seismic) and 4.1 pp (wind), versus 23.8 and 27.7 pp for the best baseline per backbone; whiskers: ± 1 SE. **(b)** First six plan-view modes of a Shanghai Tower model; torsional mode 5 appears only in plan. **(c)** Instance locations by code family. **(d)** Two VB-Bench instances.

(Jimenez et al., 2024) and verifier-driven search (Novikov et al., 2025; DeepSeek-AI, 2025) move the verdict to a program, but against a few targets or one scalar score, not the coupled constraints a structure must satisfy at once. We therefore hold that only an unpersuadable arbiter should accept a design: a deterministic physics engine that evaluates each candidate under the code-prescribed wind (MOHURD, 2012) and seismic (MOHURD, 2010a) envelope and returns a per-constraint margin that fluent but unfaithful text cannot alter.

We instantiate this position as **PE-Loop** (**P**hysics-**E**ngine-in-the-**Loop**), a generate-verify-reflect loop that terminates only when every strength and serviceability constraint holds. The physics engine holds sole acceptance authority, and its margins form a physics process reward (PPR) that couples one-dimensional section search with topology repair; the language model only proposes discrete revisions. Because the verdict never depends on the proposer, the held-out pass rate of PE-Loop is nearly invariant across four backbones, whereas the best baseline varies by more than 20 pp (Fig. 2a). To evaluate it, we formalize Vibe Building, which keeps the vibe-coding interaction but moves the verdict to a physics engine, and build VB-Bench: 3,577 physics-adjudicated instances across six code families, with held-out dynamic re-checks and a constructability gate. Concurrently, Luo et al. (2026) close a parameter-level repair loop around a finite-element verifier on 44 member-scale cases; we add topology-level repair, which they identify as the limit of such loops, held-out dynamic loading, and a counterfactual that swaps only the verdict holder, on a benchmark two orders of magnitude larger (Appendix A).

Taken together, our work makes the following contributions:

- **Vibe Building.** We formalize the generation of code-compliant building designs from multimodal input, with a protocol combining in-loop compliance, held-out dynamic re-checks, and a constructability gate.
- **VB-Bench.** We construct 3,577 physics-adjudicated instances across six code families, 1,311 gradable tasks, and a structural-code knowledge base in which every limit is traceable to its clause.

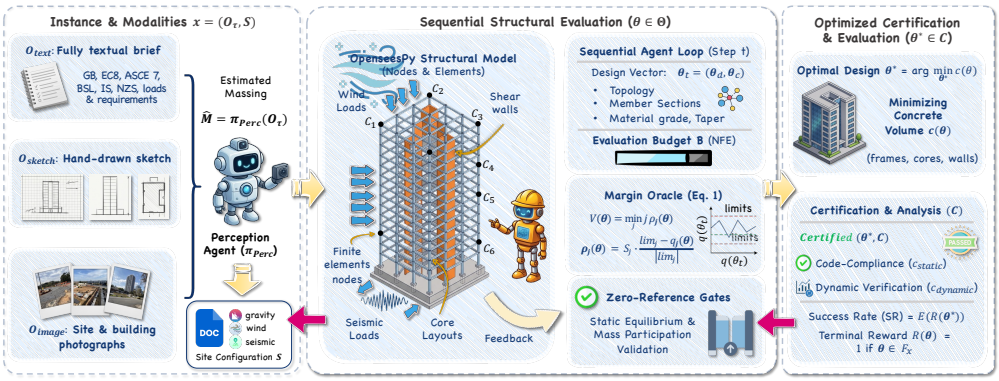


Figure 3: The definition and overview of the Vibe Building task.

- **PE-Loop.** We propose an agent in which a deterministic physics engine holds sole acceptance authority and its constraint-wise margins drive both section search and topology repair.
- **Physics adjudication is decisive.** PE-Loop achieves the best verified success under three of four backbones and the best held-out seismic pass rate under all four; replacing its verdict with an LLM judge turns 58.43% of accepted designs into false positives.

2 Task: Vibe Building

Instances and modalities. An instance $x = (O_\tau, S)$ pairs an *observation* O_τ of an underlying massing model M with a *site* S that fixes the code family (GB, EC8, ASCE 7, BSL, IS, or NZS; MOHURD, 2010a; CEN, 2004; ASCE, 2022; MLIT, 1950; BIS, 2016; Standards New Zealand, 2004) and its gravity, wind, and seismic loading provisions. The observation takes one of three modalities, $\tau \in \{\text{text}, \text{sketch}, \text{image}\}$: a numeric text brief, a sketch with text, or photographs alone (Fig. 3).

Constraints and physical margins. Each instance carries a constraint set $G_x = \{g_1, \ldots, g_m\}$, where each g_j bounds a demand quantity $q_j(\theta)$ of a candidate realization $\theta \in \Theta$ against a code-prescribed limit lim_j; the space $\Theta = \Theta_d \times \Theta_b$ factorizes into discrete choices (lateral system, core and wall counts, core plan geometry, and a section triple of beam scale, concrete grade, and taper on a finite menu) and the square-column side b on a 50-mm grid. Demand quantities are evaluated under the code-prescribed envelope of gravity, wind, and seismic load cases. Two zero-reference gates (static equilibrium; effective mass participation) reject invalid models before any code limit is consulted. Each constraint contributes a signed, normalized margin

$$\rho_j(\theta) = s_j \, \frac{\text{lim}_j - q_j(\theta)}{|\text{lim}_j|}, \qquad V(\theta) = \min_j \rho_j(\theta), \tag{1}$$

where $s_j \in \{\pm 1\}$ encodes the inequality direction; $\rho_j(\theta) \geq 0$ iff g_j is satisfied, and $V(\theta)$ is the worst-case margin. We write $\boldsymbol{\rho}(\theta) = (\rho_1(\theta), \ldots, \rho_m(\theta))$ for the full margin vector and $\boldsymbol{q}(\theta) = (q_1(\theta), \ldots, q_m(\theta))$ for the demand vector, both returned to the agent at every evaluation. Each instance additionally carries a detailing set G_x^{build} of code clauses that bound member proportions of a realization, such as the beam clear-span-to-depth ratio; their margins follow Eq. 1 with q_j a geometric ratio of θ and cost no simulator call.

Episode protocol and objective. An episode on instance x unfolds under the margin oracle of Eq. 1. It is a sequential interaction bounded by an evaluation budget B counted in simulator calls (number of function evaluations, NFE): at step t the agent selects $\theta_t \in \Theta$ and observes $\big(\boldsymbol{q}(\theta_t), \boldsymbol{\rho}(\theta_t)\big)$; the episode ends when the agent returns a final design $\theta^\star$, when the budget is exhausted before a design is returned, or when the agent declares ESCALATED, a certifiable admission that the available design levers are insufficient, which is itself a scored outcome. Writing $F_x = \{\theta \in \Theta : \bigwedge_{j=1}^m g_j(\theta)\}$ for the feasible set, the terminal reward is sparse, $R(\theta) = \mathbb{1}[\theta \in F_x]$,

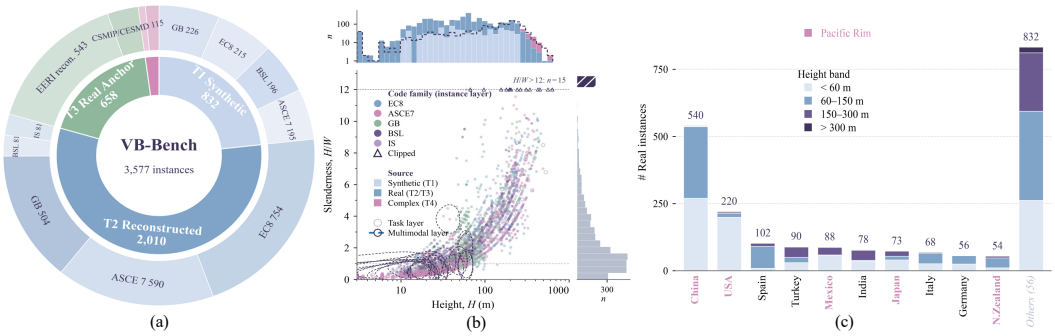


Figure 4: **VB-Bench overview.** (a) Instance composition by route (inner) and code family (outer). (b) Height–slenderness: instance (filled), task (open circles), multimodal (dashed). (c) Geographic coverage (top-10 countries, Pacific Rim in pink).

and the ideal return minimizes material among feasible realizations:

$$\theta^\star = \arg\min_{\theta \in F_x} c(\theta), \tag{2}$$

with $c(\theta)$ the concrete volume of the realized structure, accounting for frames, cores, and walls alike.

Certification and multimodal grounding. The return of Eq. 2 is certified before it counts: the agent returns $(\theta^\star, \mathcal{C})$, where the certificate $\mathcal{C} = (\boldsymbol{\rho}, \boldsymbol{\rho}^{\mathrm{L4}}, \boldsymbol{\rho}^{\mathrm{build}})$ records, respectively, the code-compliance margins, a held-out dynamic verification under recorded ground motions and field-recorded or spectrum-consistent wind series that the agent never optimizes against, and compliance with the detailing set G_x^{build}. For $\tau \in \{\mathrm{sketch}, \mathrm{image}\}$, the agent first applies a perception operator π_{perc} to its observation, yielding a massing estimate $\hat{M} = \pi_{\mathrm{perc}}(O_\tau)$, but adjudication is always performed on the ground-truth M, so design errors are never confounded with perception errors. The primary metric aggregates these certified returns over an evaluation set $\mathcal{X}$ of instances into the one-shot success rate

$$\mathrm{SR} = \mathbb{E}_{x \sim \mathcal{X}} \left[R^{\mathrm{cert}}(\theta_x^\star) \right], \tag{3}$$

where $\theta_x^\star$ is the design returned on instance x and R^{cert}, defined in Eq. 9, is the conjunction of the three certificate components.

3 Benchmark: VB-Bench

VB-Bench has two layers (Fig. 4; details in Appendix B). The instance layer includes 3,577 physics-adjudicated buildings from four construction routes, forming a difficulty gradient: synthetic (T1, 832), literature-grade reconstructions (T2, 2,010), real anchors (T3, 658), and complex-massing towers (T4, 77). The task layer includes 1,311 gradable tasks stratified by solvability and question type (intent L1, design L2, modeling L3), with 1,197 paired variants and a 372-task quarantine set against contamination. Every sample passes expert review, physical adjudication under envelope gravity, wind, and seismic load cases with height-tiered limits, and time-history verification. Every limit, including that from detailing clause $l_n/h_b \geq 4$ of G_x^{build}, carries a clause number traceable to a structural-code knowledge base compiled from 86 codes across six code families.

4 Method: Physics-Engine-in-the-Loop

PE-Loop rests on one invariant: the only authority that advances a design is the deterministic margin oracle of Section 2, never the self-assessment of a language model (Zheng et al., 2023; Shinn et al., 2023); agent outputs change the next candidate θ_t, never the verdict. Constraints analytic in θ are enforced by construction, and only those needing the simulator are adjudicated. Figure 5 gives the overview; pseudocode (Algorithm 1), proofs, and implementation details are in Appendix D.

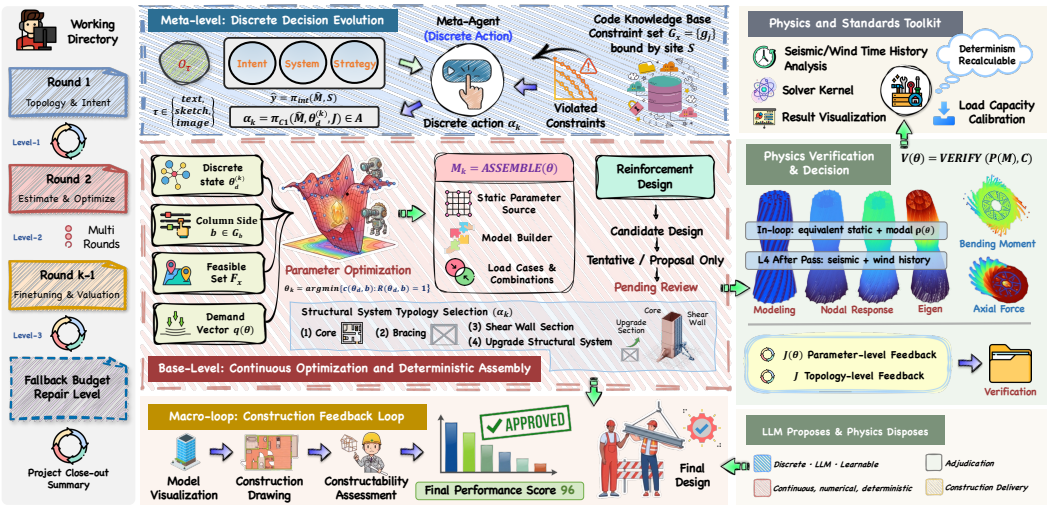


Figure 5: Overview of PE-Loop.

Physics process signal. Because $\rho_j(\theta) \geq 0$ iff g_j holds (Eq. 1), feasibility and reward are functions of the margins alone, and the violated set directs repair:

$$R(\theta) = \mathbb{1}[V(\theta) \geq 0] = \prod_{j=1}^{m} \mathbb{1}[\rho_j(\theta) \geq 0], \tag{4}$$

$$\mathcal{J}(\theta) = \{\, j : \ \rho_j(\theta) < 0 \,\}. \tag{5}$$

We call $\boldsymbol{\rho}$ the *physics process reward* (PPR) and R the verifiable outcome reward (Uesato et al., 2022; Lightman et al., 2024; Shao et al., 2024); $R(\theta_t)$ is observable at every step. The loop consumes only the sign of $\boldsymbol{\rho}$; magnitudes enter only the certificate (Eq. 9).

Design space and intent. The agent searches one coordinate on a grid, the square-column side b, and treats everything else as discrete:

$$\theta = (\theta_d, b), \quad \theta_d = (n_{\text{core}}, n_{\text{wall}}, \kappa, \varsigma), \quad b \in G_b = \{0.30 + 0.05\, i \text{ m} : \ i = 0, \ldots, 44\}, \tag{6}$$

with core and wall counts $n_{\text{core}}, n_{\text{wall}}$, core plan geometry κ, and section triple ς (beam scale, concrete grade, taper) on a finite menu (Appendix D.2). For sketch and image inputs one vision-language call gives $\hat{M} = \pi_{\text{perc}}(O_\tau)$; an intent call $\pi_{\text{int}}(\hat{M}, S)$ then fixes the initial core and wall counts (Eq. 10, Appendix D.2). Neither call is charged to B, and adjudication stays on the true M.

Column-axis search (C2). For fixed θ_d write $\tilde{\theta}(b) = \Pi_{\text{build}}(\theta_d, b)$ for the projected candidate. We adopt Assumption 1 (monotone margin along the projected column axis): $b \leq b' \Rightarrow V(\tilde{\theta}(b)) \leq V(\tilde{\theta}(b'))$. The feasible sides then form an up-set of G_b, and since concrete volume is strictly increasing in b (Lemmas 1–2, Appendix D.5),

$$b^\star(\theta_d) = \min\{b \in G_b : \ R(\theta_d, b) = 1\} = \arg\min\{c(\theta_d, b) : \ b \in G_b, \ R(\theta_d, b) = 1\}, \tag{7}$$

which exists iff $R(\theta_d, b_{\max}) = 1$. C2 finds it by an upper-bound probe, $\lceil \log_2 45 \rceil = 6$ bisection steps, and a confirmation, which together take at most 8 evaluations; with one re-verification of the returned design, one round costs at most 9. Throughout C2, ς stays frozen, up to the beam-scale clip of Π_{build} in the projected variant.

Discrete decision (C1) and outer loop. Round k runs C2 and returns PASS if $b^\star$ exists, or ESCALATED if $k = K$ or NFE $\geq B$. Otherwise, the language model picks one action from $\mathcal{A}$ at the section, topology, or system level (adjust ς; add a core or wall; adjust κ; change the lateral system):

$$\theta_d^{(k+1)} = a_k\big(\theta_d^{(k)}\big), \qquad a_k = \pi_{\text{C1}}\big(\hat{M}, \ \theta_d^{(k)}, \ \mathcal{J}_k\big) \in \mathcal{A}, \tag{8}$$

Table 1: Five systems × four LLMs on the topology-decision instances of the main split; effort `high`. Per block, **best** and second; tied values share a mark; lower is better for NFE@p, Exh., Wall, T/O, and Rank; ESC is unranked. ±: Agresti–Coull SE (SEM for NFE@p). Wall in min, P/h in passes per hour, others in %. PASS, Seis., Wind, Build, and Verif. divide by the 390 topology-decision answers per arm, ESC, Exh., T/O, and Valid by all 720 answers, and pass^3 by the 130 instances. Definitions and denominators in Appendix E; THA: time-history analysis.

	Compliance		Efficiency				Reliability	Integrity	Held-out THA		Constr.	Composite		
Arm	**PASS**	**ESC**	**NFE@p**	**Exh.**	**Wall**	**T/O**	**pass^3**	**Valid**	**Seis.**	**Wind**	**Build**	**Verif.**	**P/h**	**Rank**
DeepSeek-V4-Flash														
Codex	82.05±1.95	7.78±1.01	84.90±1.40	59.72±1.82	6.3	6.94±0.96	66.15±4.10	91.39±1.06	56.92±2.50	64.87±2.41	50.77±2.52	28.97±2.29	7.8	3.04
Openclaw	73.85±2.22	12.78±1.25	86.79±1.43	58.19±1.83	11.0	7.78±1.01	48.46±4.32	93.75±0.92	52.31±2.52	59.23±2.48	47.18±2.52	28.97±2.29	4.0	4.29
DeepSeek Harness	82.82±1.91	6.81±0.95	98.08±0.60	91.81±1.03	8.6	7.22±0.98	67.69±4.06	94.72±0.85	52.82±2.52	64.10±2.42	**54.10**±2.51	28.72±2.28	5.8	3.50
Claude Code	83.85±1.87	5.56±0.87	93.99±1.12	83.06±1.40	6.9	6.53±0.93	66.92±4.08	94.03±0.90	58.72±2.48	63.85±2.42	52.05±2.52	31.03±2.33	7.3	2.83
PE-Loop (ours)	**91.28**±1.45	11.39±1.19	**62.51**±0.38	**2.09**±0.31	**1.6**	**1.67**±0.51	**68.46**±4.03	**98.89**±0.43	**60.77**±2.46	**70.00**±2.31	46.41±2.51	**35.64**±2.42	**34.2**	**1.33**
Qwen3.8-Flash														
Codex	14.10±1.77	34.31±1.77	**28.24**±5.08	13.47±1.28	13.4	18.89±1.46	2.31±1.64	56.39±1.84	13.33±1.73	14.10±1.77	12.56±1.69	12.05±1.66	0.6	3.83
Openclaw	32.56±2.36	1.11±0.43	98.72±0.50	42.92±1.84	30.0	56.11±1.84	7.69±2.47	54.31±1.85	20.00±2.03	24.62±2.18	20.77±2.05	10.51±1.57	0.7	4.50
DeepSeek Harness	65.90±2.39	3.61±0.72	83.88±1.47	30.42±1.71	23.2	27.08±1.65	40.00±4.24	77.50±1.56	39.49±2.46	50.00±2.52	**42.56**±2.49	19.49±2.01	1.7	2.67
Claude Code	56.15±2.50	12.50±1.24	80.25±1.57	17.08±1.40	17.1	16.39±1.38	31.54±4.03	68.19±1.73	30.51±2.32	42.05±2.49	37.44±2.44	12.82±1.71	2.0	2.75
PE-Loop (ours)	**77.44**±2.12	11.67±1.20	62.74±0.38	**5.27**±0.19	**4.4**	**1.53**±0.49	**63.85**±4.16	**98.89**±0.43	**58.21**±2.49	**68.72**±2.34	34.87±2.40	**23.59**±2.15	**10.7**	**1.25**
GLM 5.2 Fast														
Codex	83.33±1.89	2.50±0.61	96.86±0.78	80.56±1.48	11.0	10.56±1.15	64.62±4.14	90.00±1.13	56.92±2.50	63.33±2.43	52.05±2.52	30.00±2.31	4.6	2.92
Openclaw	77.95±2.10	2.78±0.64	98.95±0.31	79.44±1.51	18.4	15.69±1.36	56.15±4.29	89.72±1.14	53.33±2.51	63.85±2.42	50.77±2.52	**30.51**±2.32	2.5	3.58
DeepSeek Harness	69.49±2.32	0.56±0.34	85.31±1.32	35.00±1.77	25.0	26.39±1.64	38.46±4.21	86.94±1.26	44.87±2.51	50.77±2.52	50.00±2.52	23.85±2.16	1.7	4.42
Claude Code	**86.92**±1.72	1.53±0.49	90.85±0.93	45.42±1.85	8.5	5.97±0.90	**80.77**±3.47	92.78±0.98	56.92±2.50	**67.18**±2.37	**56.41**±2.50	29.74±2.31	6.1	**2.00**
PE-Loop (ours)	76.92±2.13	10.42±1.15	**63.31**±0.39	**6.41**±0.24	**2.5**	**1.11**±0.43	63.85±4.16	**96.25**±0.73	**59.23**±2.48	66.92±2.37	40.00±2.47	27.18±2.25	**18.3**	2.08
Kimi K3														
Codex	30.51±2.32	20.42±1.50	**56.71**±4.01	13.61±1.28	8.1	4.44±0.79	11.54±2.88	86.39±1.28	26.67±2.23	27.44±2.25	24.62±2.18	21.28±2.07	2.2	2.50
Openclaw	5.13±1.16	0.42±0.31	95.00±2.57	2.50±0.61	30.0	92.92±0.97	0.00±1.05	8.19±1.03	1.79±0.75	1.79±0.75	5.13±1.16	0.51±0.51	0.1	4.71
DeepSeek Harness	22.56±2.12	0.00±0.20	76.17±2.76	4.58±0.80	30.0	73.61±1.64	1.54±1.47	32.08±1.74	12.05±1.66	15.90±1.86	17.18±1.91	6.15±1.25	0.5	3.88
Claude Code	54.10±2.51	0.69±0.36	82.36±1.56	17.92±1.43	21.2	29.31±1.69	24.62±3.76	61.11±1.81	34.87±2.40	39.49±2.46	38.21±2.45	18.21±1.96	1.5	2.83
PE-Loop (ours)	**83.08**±1.90	7.64±1.00	63.07±0.33	**0.00**±0.19	**4.5**	**2.50**±0.61	**74.62**±3.80	**97.64**±0.59	**61.54**±2.45	**71.03**±2.29	**43.33**±2.50	**28.21**±2.27	**11.0**	**1.08**

where $\mathcal{J}_k$ is the violated set of the last evaluation of round k: that of $(\theta_d^{(k)}, b_{\max})$ when the probe fails, or of the failed confirmation when the budget truncates the bisection; the split follows comparative advantage: grid search beats the model on the column axis at equal budget, while the model supplies cross-instance experience for discrete repairs. We use $K = 3$ (one decision per lever level) and $B = 15$, shared across all compared systems. In the `ultra` tier, a material loop C3 further lowers concrete volume along ς, accepting only feasible, strictly cheaper candidates (Eq. 12, Appendix D.7); the four tiers form one anytime family (Appendix D.8).

Constructability by construction. The detailing clause $l_n/h_b \geq 4$ is analytic in θ, so before every evaluation, a projection Π_{build} clips the beam-depth scale and caps $b_{\max}$ such that a compliant beam always exists, at no oracle cost (Eq. 11, Appendix D.3). Assumption 1 is stated on the projected path, so Lemma 1 applies to the capped grid verbatim, and Lemma 4 guarantees feasibility of every returned design even where the assumption fails. We evaluate Π_{build} as an ablation under one backbone (Section 5.2); the "PE-Loop" rows of Table 1 run without it, so that their Build column measures the detailing gap that search alone leaves open.

Final verification (L4) and certificate. The returned design undergoes one seismic response-history analysis under recorded ground motions and one wind time-history analysis, both elastic single-direction analyses (Appendix D), neither of which ever enters the loop, yielding held-out margins $\boldsymbol{\rho}^{\text{L4}}$ (drift ratio; peak acceleration above 150 m). With the detailing margins $\boldsymbol{\rho}^{\text{build}}$, the certificate and certified success are

$$\mathcal{C} = \big(\boldsymbol{\rho},\ \boldsymbol{\rho}^{\text{L4}},\ \boldsymbol{\rho}^{\text{build}}\big)(\theta^\star), \qquad R^{\text{cert}} = \mathbb{1}\big[V \geq 0 \ \wedge\ \min \boldsymbol{\rho}^{\text{L4}} \geq 0 \ \wedge\ \min \boldsymbol{\rho}^{\text{build}} \geq 0\big](\theta^\star), \quad (9)$$

and R^{cert} enters Eq. 3. Swapping the in-loop verdict R for an LLM judge, all else fixed, quantifies the cost of misplaced acceptance authority (Section 5.2).

5 Experiments

Setup. We evaluate five systems, namely PE-Loop, Codex, Openclaw, DeepSeek Harness, and Claude Code, each driven by four LLMs (DeepSeek-V4-Flash, Qwen3.8-Flash, GLM 5.2 Fast, and Kimi K3), on the 260-instance main split of VB-Bench (the selection procedure is described in Appendix B). All rates are computed on the topology-decision instances of this split (all 130 for every system on every seed, Appendix E) at effort level `high` (B=15, K=3) and averaged over

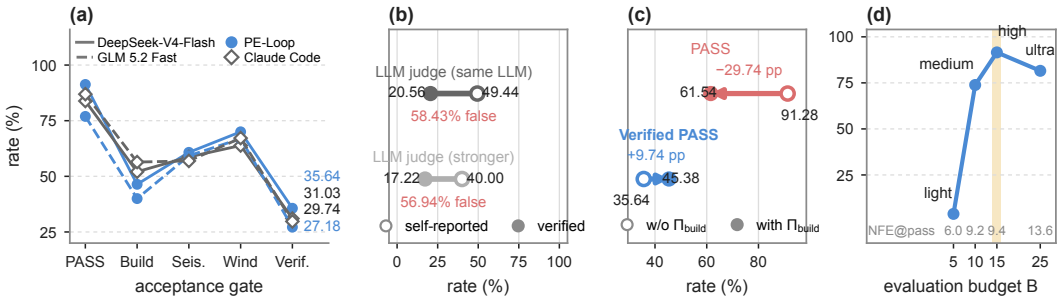


Figure 6: Where certified success is won, and three interventions on PE-Loop. (a) Gate survival of PE-Loop and the baseline leading raw PASS in both blocks, Claude Code, under DeepSeek-V4-Flash (solid) and GLM 5.2 Fast (dashed); all blocks in Fig. 7. (b) LLM judge in place of the margin oracle: self-reported (open) vs. physics-verified (filled) PASS (Table 5). (c) Rates without (open) and with (filled) Π_{build} (Table 6). (d) PASS against the evaluation budget (seed 0), with NFE@pass below.

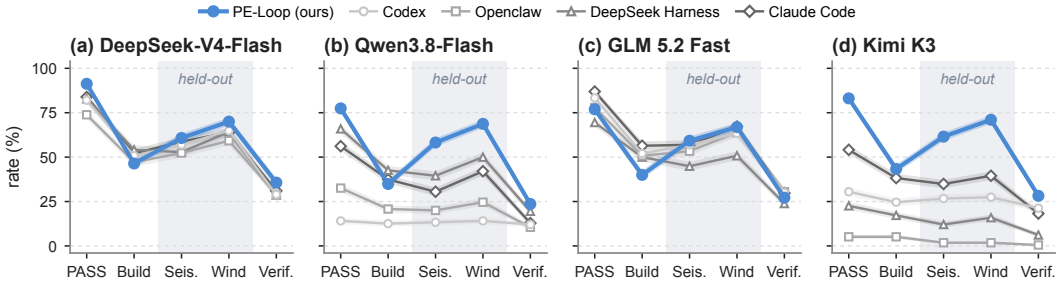


Figure 7: Survival through the acceptance gates on all four LLM blocks (same construction as Fig. 6a; values from Table 1; shaded band: the held-out re-checks).

three seeds. Only this stratum separates systems: section-only controls are passed by 93.6–97.9% of answers under every system and boundary instances by none (Table 4). Within each LLM block, the instances, the margin oracle (Eq. 1), and the budget are identical, so the driving model is the only free variable. Verified PASS, defined as the conjunction of static compliance, the held-out time-history verification (Eq. 9), and the constructability gate, serves as the empirical counterpart of R^{cert}; column definitions and denominators are given in Appendix E. Comparisons against LLM-free search use a separate search-sufficient reference set (GB-59, 59 instances $\times$ 3 seeds), which is never pooled with the main split. Of the 130 scored instances, 109 are text briefs, and 21 are sketch or image inputs; for the latter, the agent perceives $\hat{M}$, and adjudication runs on the true M. All rates pool the two groups.

5.1 Main Results

Compliance tracks the verdict, not the model. PE-Loop holds 76.92–91.28% PASS and 62.51–63.31% NFE@pass across the four LLMs; only wall-clock time moves (1.6 to 4.5 min). A baseline is its own judge and inherits the model's variance: Codex falls from 83.33% under GLM 5.2 Fast to 14.10% under Qwen3.8-Flash. Section 5.2 isolates the cause by swapping the verdict alone. The range across arms on raw PASS exceeds the 15 pp separation we pre-registered in every block (17.43 pp under DeepSeek-V4-Flash and GLM 5.2 Fast, the narrowest); the claims that follow nonetheless rest on the held-out and verified columns and on the four-backbone invariance, not on raw compliance alone.

Certified success is not raw compliance. Under GLM 5.2 Fast, the 10.00 pp raw lead of Claude Code (86.92 vs. 76.92) shrinks to 2.56 pp certified (29.74 vs. 27.18); under DeepSeek-V4-Flash, our 7.43 pp raw lead (91.28 vs. 83.85) shrinks to 4.61 pp certified (35.64 vs. 31.03), and the loss is concentrated at the constructability gate: 50.84% of our statically compliant designs clear the

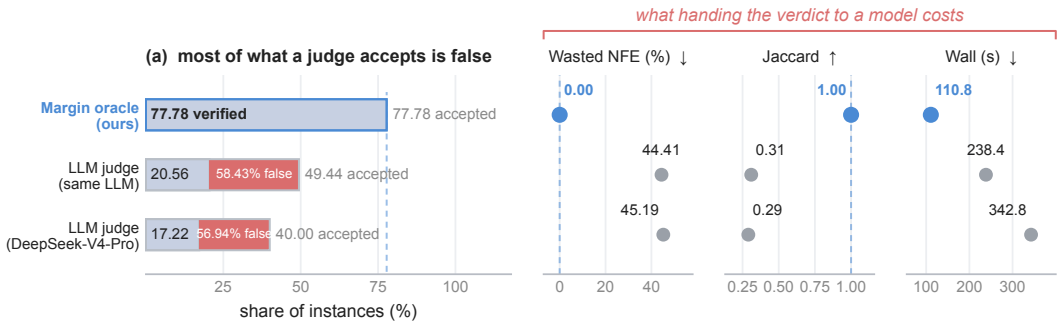


Figure 8: Who holds the verdict (Table 5). (a) What each gate accepts, split into physics-confirmed (light) and false (coral); dashed line: what the oracle delivers. (b–d) The price. Details in Appendix E.

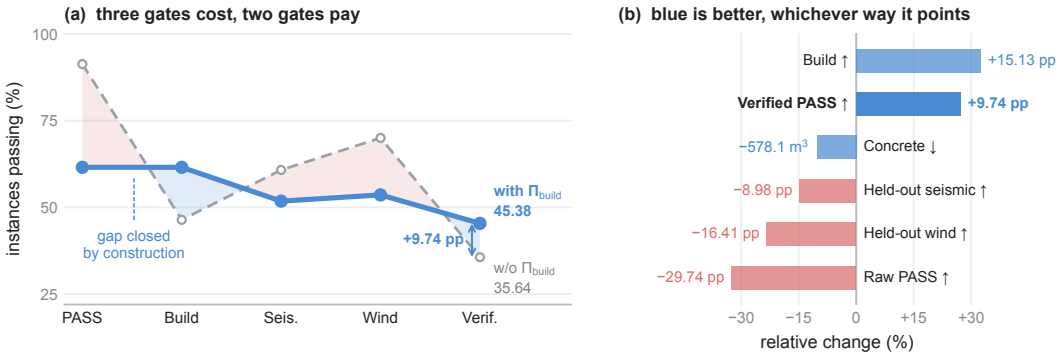


Figure 9: Π_{build} trades raw for certified compliance (Table 6). (a) Between the curves, coral is what it costs and blue what it pays; (b) relative change, blue is the better direction. Details in Appendix E.

detailing clause against 62.08% for Claude Code, whereas at the held-out checks the two arms are within 4 pp of each other conditional on compliance (Fig. 6a; all blocks in Fig. 7; Appendix E). We hold the best Verified PASS in three of four blocks, the best seismic in all four, and the best wind in three; both exceptions, under GLM 5.2 Fast, lie within about one standard error of the difference (30.51 for Openclaw vs. 27.18 for PE-Loop on Verified PASS, $z \approx 1.03$; 67.18 for Claude Code vs. 66.92 for PE-Loop on wind, $z \approx 0.08$). Under DeepSeek-V4-Flash, our weakest gate is constructability (46.41% vs. 47.18–54.10%), diagnosed in Section 5.2; under Kimi K3, we already lead that gate.

5.2 Ablations

Who holds the verdict (Fig. 8; Table 5). When R (Eq. 4) is replaced by an LLM judge that sees the brief, the code limits, and the design parameters but no simulator output, with all else fixed, the same LLM claims 49.44% PASS but delivers 20.56%: 58.43% of its acceptances are false, and 44.41% of its evaluations are spent on misdirected repairs. A stronger, disjoint judge is no better (56.94% false positives) and roughly triples wall-clock time, while the margin oracle claims exactly what it delivers (77.78%; the 60 pre-registered instances hold 33 of the 34 failures of the main-table run, whose PE-Loop passes 81.67% of them). Who holds the verdict drives compliance, not a correlate, consistent with the unreliability of unaided self-verification (Huang et al., 2024; Kambhampati et al., 2024) and with the optimization pressure a learned judge invites (Gao et al., 2023).

Constructability (Fig. 9; Table 6). Our constructability deficit comes from lever order: the outer loop enlarges the column before changing topology, and thick columns trip $l_n/h_b \geq 4$ (Appendix E). The gap is not a quirk of our loop: on the reference set, every system returns unbuildable designs, at 42.75–72.41% of its feasible answers (Fig. 15 in Appendix D.3). Moreover, 123 of our 126 violations are repairable by clipping one section scale, and Π_{build} does exactly that inside the loop:

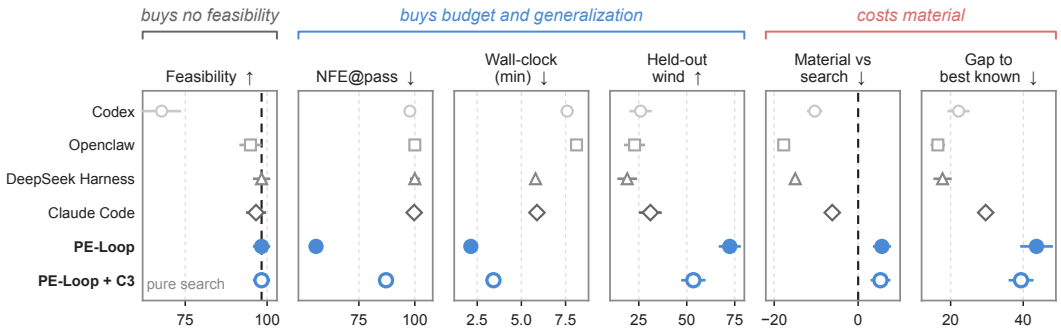


Figure 10: What the language model buys on the search-sufficient reference set (59 GB-family instances $\times$ 3 seeds, DeepSeek-V4-Flash; means $\pm$ SE). Dashed line: exhaustive LLM-free section search. Details in Appendix E.

raw PASS falls from 91.28 to 61.54% as thick-column designs that passed only by violating the clause are removed, but Verified PASS rises from 35.64 to 45.38%, and concrete falls from 5,610.1 to 5,032.0 m^3.

5.3 Budget and what the language model buys

Budget discipline is structural. The escalation action makes the budget a property of the algorithm: because one outer round costs at most 9 oracle calls (Lemma 1), the exhaustion rate of PE-Loop never exceeds 6.41% under any of the four LLMs (Table 1). Under DeepSeek-V4-Flash, the baselines, which have no budget-aware escalation action, exhaust it on 58.19–91.81% of their answers, and PE-Loop answers at 62.51% NFE@pass compared with 84.90–98.08%. The budget is also a dial, in the spirit of test-time compute scaling (Snell et al., 2025): on the identical 130 instances (seed 0), success runs 3.85 / 73.85 / 91.54% from `light` to `high` (Fig. 6d), and `ultra` trades 10.00 pp of it for one more outer round (Appendix E).

What the language model buys (Fig. 10). Against LLM-free pure search on the reference set, the feasibility of PE-Loop ties at 98.31% with no discordant instances (n_{01}=0, n_{10}=0, N=59; McNemar's test, McNemar, 1947): the model buys no feasibility. It buys budget (NFE@pass 55.86% vs. 97.78–100.00%) and held-out generalization (72.41% wind vs. 18.97–31.03%). The price is material: 5.66% more concrete than pure search, where every baseline spends less (-17.74 to -6.17%), because $b^\star$ (Eq. 7) is minimal along a single coordinate; C3 recovers part of it, and closing the rest is the clearest open direction. Among the baselines, only Codex differs significantly from search on feasibility (Appendix E).

6 Conclusion and Limitations

The results locate reliability in who holds the verdict, not in the proposer's strength. Once a deterministic engine alone accepts designs, held-out pass rates and oracle cost become properties of the loop: swapping the backbone barely moves them, whereas swapping only the verdict holder turns most acceptances into false positives, even with a stronger judge, because capability cannot substitute for access to the physics. Three implications follow. First, the language model's role narrows: on instances search alone can solve, it adds no feasibility; its value lies in choosing among discrete structural alternatives and spending a fixed budget well. Second, raw compliance is the wrong yardstick for physical-world agents: certified success reorders systems, and every system we tested, ours included, returns unbuildable designs that surface only under out-of-loop checks, so benchmarks should score certificates rather than claims. Third, a useful verifier is external, deterministic, and constraint-wise: its per-constraint margin both resists persuasion and tells the proposer what to repair. Limits remain: the held-out checks share the loop's physical model, certified success stays below one half, and single-coordinate search costs 5.66% more concrete than exhaustive search; independent-solver validation, broader clause coverage, and multi-coordinate search come next. Wherever an accepted output is costly to reverse, the verdict should rest with an arbiter the proposer cannot persuade.

AI USE STATEMENT

We have not used generative AI tools to generate synthetic data, to develop the theoretical framework, to formulate or prove the claims of Lemmas 1–4, to implement PE-Loop or the evaluation harness, to design the experiments, or to propose hypotheses; translation, dataset cleaning and reformatting, and qualitative data analysis do not apply to this work. We used generative AI tools to help interpret results: to cross-check the numbers reported in the text against the frozen result tables, to compute derived statistics (standard errors and z-scores of differences), and to draft the wording of result interpretations. We also used them to edit and condense text for readability, to create scientific figures and the scripts that generate them, and to write and edit verification code. We reviewed all AI-assisted work: we read back every plotted value from the figure objects and asserted it, the authors recomputed every reported number and derived statistic from the per-episode records, and the authors checked and approved all claims and interpretations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The designs returned by any system evaluated here are schematic and have not been reviewed by a licensed engineer; nothing in this paper substitutes for that review, and the benchmark certificate concerns compliance with the sourced clauses under the stated analyses, not a building's safety. The instances are derived from published building descriptions and public design codes and contain no personal data. The benchmark is intended to measure and improve the verifiability of automated design tools, which we expect to reduce, not increase, the risk that an unverified machine-generated design reaches construction.

REPRODUCIBILITY STATEMENT

We will release VB-Bench (instances, code-family limits, screening and difficulty labels, the 260-instance main split, and the 59-instance search-sufficient reference set), the margin oracle with the held-out seismic and wind analyses, PE-Loop, and the harness that drives all five systems upon publication. Every number in Section 5 and Appendix E is copied from frozen result tables produced by a single campaign; the release will include those tables, the per-episode records, the seeds, the model identifiers and effort tiers of Appendix D.8, the denominators disclosed in Appendix E, and the scripts that regenerate the figures and re-assert every number in the text against the tables. The oracle is deterministic; the residual variance comes from the language models, which we access through commercial APIs and therefore cannot reproduce bit-for-bit, so we report each rate over three seeds with its standard error.

REFERENCES

Alan Agresti and Brent A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998. doi: 10.1080/00031305.1998.10480550.

Timothy D. Ancheta, Robert B. Darragh, Jonathan P. Stewart, Emel Seyhan, Walter J. Silva, Brian S.-J. Chiou, Katie E. Wooddell, Robert W. Graves, Albert R. Kottke, David M. Boore, Tadahiro Kishida, and Jennifer L. Donahue. NGA-West2 database. *Earthquake Spectra*, 30(3):989–1005, 2014. doi: 10.1193/070913EQS197M.

ASCE. ASCE/SEI 7-22: Minimum design loads and associated criteria for buildings and other structures. Technical report, American Society of Civil Engineers (ASCE), Reston, VA, 2022.

BIS. IS 1893 (Part 1):2016: Criteria for earthquake resistant design of structures. Part 1: General provisions and buildings. Technical report, Bureau of Indian Standards (BIS), New Delhi, 2016.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

CEN. EN 1998-1:2004: Eurocode 8: Design of structures for earthquake resistance. Part 1: General rules, seismic actions and rules for buildings. Technical report, European Committee for Standardization (CEN), Brussels, 2004.

Jinxin Chen and Yi Bao. Multi-agent large language model framework for code-compliant automated design of reinforced concrete structures. *Automation in Construction*, 177:106331, 2025. doi: 10.1016/j.autcon.2025.106331.

Weiwei Chen, Jianze Wang, Shitong Xu, Yongqing Jiang, Kaoshan Dai, Mohammad Al Hamaydeh, and Qingyu Chen. Post-earthquake recovery assessment of indoor unanchored non-structural systems using computer vision techniques. *Safety Science*, 204:107407, 2026. doi: 10.1016/j.ssci.2026.107407.

Xiaoqian Chen, Zeyu Zhang, Yu Li, Wen Yao, and Weien Zhou. Research on structure topology optimization design empowered by deep learning method. *Advances in Mechanics*, 54(2):213–258, 2024. doi: 10.6052/1000-0992-23-052. In Chinese.

Kichang Choi, Minwoo Jeong, Taegeon Kim, Seokhwan Kim, Seungwon Baek, and Hongjo Kim. RAMASC: A retrieval-augmented multi-agent framework for automated structural calculation. *Advanced Engineering Informatics*, 74:104698, 2026.

A. G. Davenport. The spectrum of horizontal gustiness near the ground in high winds. *Quarterly Journal of the Royal Meteorological Society*, 87(372):194–211, 1961. doi: 10.1002/qj.49708737208.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Paula Doubrawa, Alex Rybchuk, and Stefano Letizia. Constraints for rotor inflow benchmarks: Unstable cases. Zenodo dataset, IEA Wind Task 57, 2025. URL https://doi.org/10.5281/zenodo.15747594. CC BY 4.0.

Changyu Du, Sebastian Esser, Stavros Nousias, and André Borrmann. Text2BIM: Generating building models using a large language model-based multiagent framework. *Journal of Computing in Civil Engineering*, 40(2):04025142, 2026.

FEMA. Quantification of building seismic performance factors. Technical Report FEMA P-695, Federal Emergency Management Agency, prepared by the Applied Technology Council, Washington, DC, 2009.

Yuyuan Feng, Zhishang Xiang, Chaobin Yang, Qichao Ma, Zerui Chen, Yujing Zhang, Ke Huang, Chuanjie Wu, Zhaoxu Liu, Yili Wang, Xin He, Jiapu Wang, Zijin Hong, Hao Chen, Yuanchen Bei, Kun Wang, Shengyuan Chen, Ningyu Zhang, Enyan Dai, Linhao Luo, Qingyi Pan, Qi Wang, Wenqi Fan, Guangjing Wang, Na Zou, Yangqiu Song, Xin Wang, Zechao Li, Xia Hu, Qing Li, Xiao Huang, Zhihong Zhang, Jinsong Su, Qinggang Zhang, and Yi Chang. Graph engineering in the era of LLM agents: From individual intelligence to system intelligence. *arXiv preprint arXiv:2608.21156*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. RLEF: Grounding code LLMs in execution feedback with reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/gehring25a.html.

Z. Geng, J. Liu, R. Cao, L. Cheng, H. Wang, and M. Cheng. A lightweight large language model-based multi-agent system for 2D frame structural analysis. *arXiv preprint arXiv:2510.05414*, 2025.

Ziheng Geng, Ian Franklin, Santiago Martinez, Jiachen Liu, Yunhe Zhao, and Minghui Cheng. Agentic large language models for automated structural analysis of 3D frame systems. *arXiv preprint arXiv:2606.06525*, 2026.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024a.

Yi Gu, Yuli Huang, Wenjie Liao, and Xinzheng Lu. Intelligent design of shear wall layout based on diffusion models. *Computer-Aided Civil and Infrastructure Engineering*, 39(23):3610–3625, 2024b.

Xingang Guo, Yaxin Li, Xiangyi Kong, Yilan Jiang, Xiayu Zhao, Zhihua Gong, Yufan Zhang, Daixuan Li, Tianle Sang, Beixiao Zhu, et al. Toward engineering AGI: Benchmarking the engineering design capabilities of LLMs. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025.

Chaoyue He. Awesome Loop Engineering. https://github.com/ChaoYue0307/awesome-loop-engineering, 2026a. Field guide and implementation kit for Loop Engineering, version 0.9.0.

Chaoyue He. Awesome Graph Engineering: A field guide, dataset, and interactive atlas for programmable AI-agent organizations. https://github.com/ChaoYue0307/awesome-graph-engineering/releases/tag/v1.3.0, 2026b.

Seokjae Heo. Enhancing reliability and automation of LLM-based structural analysis using a hybrid multi-agent pipeline. *Scientific Reports*, 16:19458, 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Donald E. Hudson, A. Gerald Brady, Mihailo D. Trifunac, and A. Vijayaraghavan. Strong-motion earthquake accelerograms: Digitized and plotted data. Volume II, Part A: Corrected accelerograms and integrated ground velocity and displacement curves. Technical Report EERL 71-50, Earthquake Engineering Research Laboratory, California Institute of Technology, Pasadena, CA, 1971.

Yongqing Jiang, Jianze Wang, Zhiqi Shen, Zhenghong Lin, Jiayuan Wang, Yijian Yang, Kaoshan Dai, and Haoran Luo. Rethinking scientific modeling: Toward physically consistent and simulation-executable programmatic generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11189–11200, 2026. doi: 10.1145/3770855.3818987.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

J. C. Kaimal, J. C. Wyngaard, Y. Izumi, and O. R. Coté. Spectral characteristics of surface-layer turbulence. *Quarterly Journal of the Royal Meteorological Society*, 98(417):563–589, 1972. doi: 10.1002/qj.49709841707.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Andrej Karpathy. There's a new kind of coding I call "vibe coding". Post on X (formerly Twitter), February 2025. URL https://x.com/karpathy/status/1886192184808149383.

Jeonghye Kim, Sojeong Rhee, Minbeom Kim, Dohyung Kim, Sangmook Lee, Youngchul Sung, and Kyomin Jung. ReflAct: World-grounded decision making in LLM agents via goal-state reflection. *arXiv preprint arXiv:2505.15182*, 2025.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

Chen Liang, Zhaoqi Huang, Haofen Wang, Fu Chai, Chunying Yu, Huanhuan Wei, Zhengjie Liu, Yanpeng Li, Hongjun Wang, Ruifeng Luo, and Xianzhong Zhao. AECBench: A hierarchical benchmark for knowledge evaluation of large language models in the AEC field. *Advanced Engineering Informatics*, 71:104314, 2026.

Haoran Liang, Yufa Zhou, Mohammad Talebi-Kalaleh, and Qipei Mei. Automating structural engineering workflows with large language model agents. *arXiv preprint arXiv:2510.11004*, 2025.

Wenjie Liao, Xinzheng Lu, Yuli Huang, Zhe Zheng, and Yuanqing Lin. Automated structural design of shear wall residential buildings using generative adversarial networks. *Automation in Construction*, 132:103931, 2021.

Wenjie Liao, Xinzheng Lu, Yifan Fei, Yi Gu, and Yuli Huang. Generative AI design for building structures. *Automation in Construction*, 157:105187, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Jiachen Liu, Ziheng Geng, Ran Cao, Lu Cheng, Paolo Bocchini, and Minghui Cheng. A large language model-empowered agent for reliable and robust structural analysis. *Structure and Infrastructure Engineering*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Xinzheng Lu, Wenjie Liao, Yu Zhang, and Yuli Huang. Intelligent structural design of shear wall residence using physics-enhanced generative adversarial networks. *Earthquake Engineering & Structural Dynamics*, 51(7):1657–1676, 2022.

Jianbin Luo, Weibin Lin, Yiran Lin, Qing Wei, and Wei Guo. Verification-driven closed-loop multi-agent large language model framework for code-compliant structural design. *arXiv preprint arXiv:2608.07978*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Soumya Madireddy, Lu Gao, Zia Din, Kinam Kim, Ahmed Senouci, Zhe Han, and Yunpeng Zhang. Large language model-driven code compliance checking in building information modeling. *arXiv preprint arXiv:2506.20551*, 2025.

Nirvan Makoond, Andri Setiawan, Manuel Buitrago, and Jose M Adam. Arresting failure propagation in buildings through collapse isolation. *Nature*, 629(8012):592–596, 2024. doi: 10.1038/s41586-024-07268-5.

Frank McKenna. OpenSees: A framework for earthquake engineering simulation. *Computing in Science & Engineering*, 13(4):58–66, 2011. doi: 10.1109/MCSE.2011.66.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Shiqiao Meng, Ying Zhou, Bingxu Liao, Mushi Chang, Tianshu Zhang, Abouzar Jafari, and Abderrahim Djerrad. A physics-informed foundation model for rapid high-fidelity structural response prediction. *Nature Communications*, 17:8884, 2026. doi: 10.1038/s41467-026-75508-5.

MLIT. Building Standard Law of Japan (Act No. 201 of 1950) and the Building Standard Law Enforcement Order, with related notifications. Ministry of Land, Infrastructure, Transport and Tourism (MLIT), Tokyo, 1950. As amended.

MOHURD. GB 50011-2010: Code for seismic design of buildings (2016 edition). Technical report, Ministry of Housing and Urban-Rural Development of the People's Republic of China (MOHURD) and General Administration of Quality Supervision, Inspection and Quarantine (AQSIQ); China Architecture & Building Press, Beijing, 2010a. National standard, issued 2010-05-31, partially revised 2016.

MOHURD. JGJ 3-2010: Technical specification for concrete structures of tall building. Technical report, Ministry of Housing and Urban-Rural Development of the People's Republic of China (MOHURD); China Architecture & Building Press, Beijing, 2010b. Industry standard, effective 2011-10-01.

MOHURD. GB 50009-2012: Load code for the design of building structures. Technical report, Ministry of Housing and Urban-Rural Development of the People's Republic of China (MOHURD) and General Administration of Quality Supervision, Inspection and Quarantine (AQSIQ); China Architecture & Building Press, Beijing, 2012. National standard, effective 2012-10-01.

Nathan M. Newmark. A method of computation for structural dynamics. *Journal of the Engineering Mechanics Division*, 85(3):67–94, 1959.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pp. 278–287. Morgan Kaufmann, 1999.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent Q: Advanced reasoning and learning for autonomous AI agents. *arXiv preprint arXiv:2408.07199*, 2024.

Y. Qi, R. Xu, and X. Chu. FeaGPT: An end-to-end agentic-AI for finite element analysis. *arXiv preprint arXiv:2510.21993*, 2025.

Sizhong Qin, Yi Gu, Yao Jiang, Ao Cai, Changjian Zhou, Shaoxuan Shuai, Jiachang Wang, Tianhao Shen, Yueqiang Li, Xinhao Li, Li Zeng, Yueshi Chen, Dachen Gao, Genrong Xu, Wenjie Liao, and Xinzheng Lu. StructureClaw: Traceable LLM agents and an executable benchmark for structural engineering workflows. *arXiv preprint arXiv:2607.14896*, 2026.

A. Radler, E. Volkmann, J. Brandstetter, and A. Berzins. Diverse topology optimization using modulated neural fields. *arXiv preprint arXiv:2502.13174*, 2025.

Matthew Renze and Erhan Guven. Self-reflection in LLM agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Alex Rybchuk and Stefano Letizia. Constraints for rotor inflow benchmarks: Stable cases. Zenodo dataset, IEA Wind Task 57, 2025. URL https://doi.org/10.5281/zenodo.15747684. CC BY 4.0.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jack Wei Lun Shi, Minghao Dang, Wawan Solihin, Leong Hien Poh, and Justin K. W. Yeoh. Reinforcement learning to improve large language model-based automated code compliance systems. *arXiv preprint arXiv:2606.22402*, 2026.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Masanobu Shinozuka and C.-M. Jan. Digital simulation of random processes and its applications. *Journal of Sound and Vibration*, 25(1):111–128, 1972. doi: 10.1016/0022-460X(72)90600-1.

Ole Sigmund and Kurt Maute. Topology optimization approaches: A comparative review. *Structural and Multidisciplinary Optimization*, 48(6):1031–1055, 2013.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*, 2025.

Standards New Zealand. NZS 1170.5:2004: Structural design actions. Part 5: Earthquake actions – New Zealand. Technical report, Standards New Zealand, Wellington, 2004. Incorporating Amendment 1 (2016).

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024a. URL https://arxiv.org/abs/2305.16291.

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. Reinforcement learning for self-improving agent with skill library. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1529–1550, 2026a. URL https://aclanthology.org/2026.acl-long.69/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024b.

Quankai Wang, Yulin Xie, Tongfei Yang, Minghui Cheng, and Ran Cao. Human-enhanced loop modeling (HELM): Agent-based finite element modeling of concrete bridge barriers. *arXiv preprint arXiv:2606.12025*, 2026b.

Yi Wang, Haopeng Zhang, Chengxiang Huang, Rui Dai, Kaikui Liu, Piotr Koniusz, and Xiangxiang Chu. LoopArena: Benchmarking models as runtime controllers for loop engineering. *arXiv preprint arXiv:2608.28281*, 2026c.

Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and Botian Shi. EvolveR: Self-evolving LLM agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*, 2025.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Yajie Yang, Wenxiang Chen, Zhihao Zhang, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. AgentPRM: Process reward models for LLM agents via step-wise promise and progress. *arXiv preprint arXiv:2511.08325*, 2025.

Jun Yang, Kentaro Yaji, and Shintaro Yamasaki. Data-driven topology design based on principal component analysis for 3D structural design problems. *Structural and Multidisciplinary Optimization*, 68(5):103, 2025. doi: 10.1007/s00158-025-04036-5.

Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, Xin Li, Bo Zhang, and Liang He. AutoSkill: Experience-driven lifelong learning via skill self-evolution. *arXiv preprint arXiv:2603.01145*, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025.

Sompote Youwai, David Phim, Vianne Gayl Murcia, and Rianne Clair Onas. Investigating the potential of large language model-based router multi-agent architectures for foundation design automation: A task classification and expert selection study. *arXiv preprint arXiv:2506.13811*, 2025.

Tao Zhang, Zhenhai Liu, Yong Xin, and Yongjun Jiao. MooseAgent: A LLM based multi-agent framework for automating MOOSE simulation. *arXiv preprint arXiv:2504.08621*, 2025.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yu-Jie Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Minjie Zhu, Frank McKenna, and Michael H. Scott. OpenSeesPy: Python library for the OpenSees finite element framework. *SoftwareX*, 7:6–11, 2018. doi: 10.1016/j.softx.2017.10.009.

A Related Work

Language model agents for structural analysis and design. Research on applying language models to structural analysis and design spans two levels. At the generation level, LLM agents translate natural language into executable finite-element code (Qi et al., 2025), evolving toward lightweight, router-based multi-agent architectures for frame structures (Geng et al., 2025; 2026), specialized member modeling (Wang et al., 2026b), code-compliant reinforced concrete design (Chen & Bao, 2025), foundation type selection (Youwai et al., 2025), multiphysics solver generation (Zhang et al., 2025), retrieval-augmented structural calculation (Choi et al., 2026), reliability-oriented structural analysis (Liu et al., 2026), and editable building models (Du et al., 2026). At the verification level, multi-agent systems automate design review (Liang et al., 2025) and code compliance checking of building information models, increasingly refined through reinforcement learning (Madireddy et al., 2025; Shi et al., 2026), and benchmarks now grade domain knowledge (Liang et al., 2026), simulation-validated engineering designs (Guo et al., 2025), and the executable evidence chain from request to report (Qin et al., 2026). In most of these cases, the solver merely executes a task whose output is already known or verifiable. Hence, evaluation reduces to reproducing or checking a known answer rather than judging an undetermined design. Separately, data-driven

methods generate shear-wall layouts with adversarial and diffusion models (Liao et al., 2021; Lu et al., 2022; Gu et al., 2024b; Liao et al., 2024) or combine generative neural fields, principal component analysis, and topology optimization (Sigmund & Maute, 2013) to produce diverse layouts from numerical objectives (Radler et al., 2025; Chen et al., 2024; Yang et al., 2025), without reading a natural-language brief. Concurrently, two systems close a repair loop around a deterministic structural solver (Heo, 2026; Luo et al., 2026). Luo et al. (2026) let the LLM adjust member sections from the verifier's exceedance report on 44 member-scale cases (beams, trusses, and low-rise planar frames), report compliance rising from 56.8% to 98.6%, and attribute it to the verifier because it does not change detectably across two backbones; they note that the remaining failures stem from the initial topology, which a parameter-level loop cannot repair. We corroborate their attribution and extend it with topology-level repair by the language model while a deterministic search sets member sections, a benchmark of 3,577 building-scale instances across six code families, held-out seismic and wind re-checks that the loop never optimizes against, a counterfactual that replaces only the verdict holder (Table 5), and a budget-bounded search with correctness guarantees (Lemmas 1–4) compared against four coding harnesses under four backbones.

LLM agents grounded in physics and simulator feedback. Agents coupled with tools, physical models, or simulators interleave reasoning with action (Yao et al., 2023b) and explicitly plan and correct candidate decision paths using environment feedback or tree search (Yao et al., 2023a), turning one-shot generation into multi-step search and verification (Kim et al., 2025; Zhou et al., 2024; Putta et al., 2024). Related work achieves policy improvement without parameter updates through structured self-reflection (Shinn et al., 2023), distilling long-term interaction experience into reusable policy principles (Wang et al., 2024a; Zhao et al., 2024) or structured skill libraries that evolve continually via reinforcement learning signals (Wu et al., 2025; Yang et al., 2026; Wang et al., 2026a). Other work trains self-correction directly through multi-turn online reinforcement learning (Renze & Guven, 2024; Kumar et al., 2024). Who verifies matters: language-model judges carry systematic biases (Zheng et al., 2023; Wang et al., 2024b; Gu et al., 2024a), models do not reliably correct their own reasoning without external feedback (Huang et al., 2024), refinement helps mainly when the critique is grounded in a tool (Madaan et al., 2023; Gou et al., 2024), and policies optimized against a learned proxy overfit it (Gao et al., 2023), which motivates letting the model generate while sound external critics accept (Kambhampati et al., 2024). Regarding feedback granularity, process supervision outperforms outcome supervision (Uesato et al., 2022; Lightman et al., 2024), with simulator outputs, execution feedback, or process reward model returns used to score intermediate decisions and code-level proposals step by step (Gehring et al., 2025; Xi et al., 2025); reinforcement learning from verifiable rewards trains against programmatic checkers (Shao et al., 2024; DeepSeek-AI, 2025), test-time computation can substitute for model scale when a reliable verifier exists (Snell et al., 2025), evaluator-in-the-loop search discovers new constructions and algorithms (Romera-Paredes et al., 2024; Novikov et al., 2025; Lu et al., 2024), and executable benchmarks grade agents by running their outputs, from unit tests (Jimenez et al., 2024) to repeated-trial reliability measured by pass^k (Yao et al., 2025). A more demanding setting remains open, one in which an accepted proposal must satisfy several externally specified performance levels simultaneously, rather than depending on a single one-shot reward signal. This departs from building-design practice, where multiple codes and performance levels are checked in parallel, and motivates the closed-loop mechanism with multidimensional verification developed in this paper.

B VB-BENCH DETAILS

Instance and task layers. The instance layer provides 3,577 physically adjudicated buildings (Fig. 4), grouped by construction route into synthetic instances (T1, 832), literature-grade reconstructions (T2, 2,010), real anchors (T3, 658), and complex-massing towers (T4, 77), which together supply both the source of design problems and a gradient of difficulty. The task layer contains 1,311 independently collected, gradable tasks, stratified by solvability and question type: intent (L1), design (L2), and modeling (L3). L2 spans the full set, while L1 (338) and L3 (384) are the subsets for which we could source intent and model ground truth. Contamination-resistant components comprise 1,197 paired variants and a quarantine set of 372 tasks.

From the instance layer to the experimental sets. The experiments do not run on the whole instance layer. Static screening (missing ground truth, too few stories, non-reinforced-concrete sys-

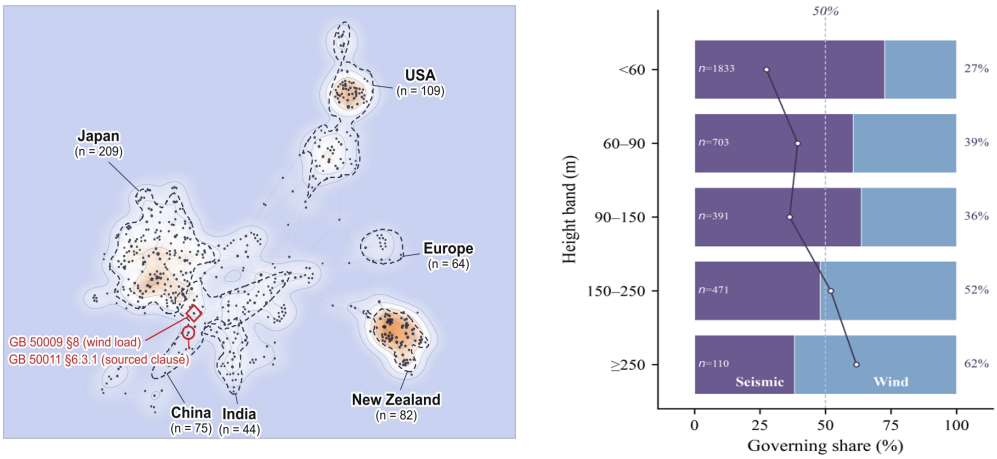


Figure 11: **VB-Bench sources and governing loads.** (a) The structural-code knowledge base as one cross-linked graph. Each dot is a wiki entry (a code source document, an institution, or a clause concept) placed by a force-directed layout of the 1,847 wikilinks between entries; the shading is entry density weighted by link degree, from pale blue (sparse) through white to orange (hub pages). Dashed outlines enclose the densest region of each jurisdiction's entries, labeled with the number drawn; Europe's entries are the most dispersed, so its outline marks only its densest cluster. Two landmark pages are marked: GB 50011-2010 Clause 6.3.1, which sources the detailing rule of Table 6, and the wind-load chapter of GB 50009-2012. 583 of the 625 entries (121 sources, 79 institutions, 425 concepts, distilled from 86 original codes) form the connected component shown; the remaining 42 unlinked or small-component pages are omitted. (b) Wind vs. seismic governing share by height band; wind governs the majority of instances above 150 m.

tems, infeasible slenderness) and a dynamic solvability calibration leave a difficulty map of 3,216 instances: 1,836 that a section change alone solves, 345 that require a topology decision, and 1,035 that no system solves (Table 3); 32 further image-only instances are held out as a multimodal quarantine set and never used here, leaving a pool of 3,184. The 260-instance main split is a seeded, height-stratified sample of that pool: 130 topology-decision instances, 110 section-only controls, and 20 capability-boundary instances above 250 m, with per-band quotas set by supply: only five topology-decision instances exist between 120 and 250 m, so that band's quota was cut from 40 to 20. The freed slots went to the 24–60 m band, where topology decisions are densest. Running five systems on four backbones and three seeds over the full pool was outside the campaign's budget, and only the topology-decision stratum separates systems (Appendix E). The smaller sets of Section 5 derive from the same seed and nest inside one another: the 60-instance set of Table 5 was frozen together with its criteria before any run, and the 59-instance search-sufficient reference set is reused from an earlier campaign, whose exhaustive and grid-search reference lines make it the only set with an LLM-free baseline. PE-Loop and every baseline answered all 260 instances on every seed and backbone (n=390 topology-decision answers and n=720 answers in total per arm and backbone, the 20 boundary instances excluded from every comparative statistic; non-deliveries in Appendix E). Table 1 reports scores only for the 130 topology-decision instances; the section-only and boundary strata are reported in Table 4.

Acceptance. Every sample passes expert review, physical adjudication, and time-history verification. Acceptance uses the envelope of three load cases: gravity (self-weight) is always included, while seismic and wind are evaluated separately, and the most unfavorable value of each response quantity is taken without load-combination factors, since the envelope already provides a conservative upper bound. Acceptance limits scale with building height according to a tiered scale (Fig. 11b).

Ground-motion and wind record libraries. Both held-out re-checks draw from fixed, public record libraries released with the benchmark (Fig. 12). The seismic library holds 62 recorded horizontal accelerograms in units of g: the 44 records of the FEMA P-695 far-field set (FEMA, 2009), 22 shallow-crustal earthquakes of magnitude 6.5–7.6 recorded more than 10 km from the rupture, two horizontal components each, from San Fernando 1971 and Friuli 1976 to Chi-

Chi 1999 and Düzce 1999, with peak accelerations of 0.21–0.82 g, taken from the PEER NGA-West2 database (Ancheta et al., 2014);[1] 16 further PEER records from six events (San Fernando 1971, Imperial Valley 1979, Manjil 1990, Düzce 1999, San Simeon 2003, and Chuetsu-oki 2007); and the two horizontal components of El Centro 1940 (Appendix C). Sampling intervals range from 0.0025 to 0.02 s and durations from 20 to 100 s; during ingestion, we checked every record against the point count and time step in its header and against its published peak acceleration, within 10%. Only the P-695 subset is eligible for the held-out selection (Appendix D); we keep the rest for illustration. The wind library holds sixteen ten-minute wind-speed series at 5 Hz, stored as normalized fluctuations $v(t)/\bar{v}$ together with their turbulence intensity I_u. Eight are field records, resampled from the 20 Hz sonic-anemometer measurements at 110.5 m and 179.5 m on the tall meteorological tower of the IEA Wind Task 57 rotor-inflow benchmarks, four from stable and four from unstable atmospheric cases, with I_u of 0.055–0.105 (Doubrawa et al., 2025; Rybchuk & Letizia, 2025);[2] eight are synthesized by spectral representation (Shinozuka & Jan, 1972) from the Kaimal and Davenport spectra (Kaimal et al., 1972; Davenport, 1961) at target I_u of 0.12, 0.18, and 0.25, the three terrain categories the benchmark distinguishes, with fixed seeds.

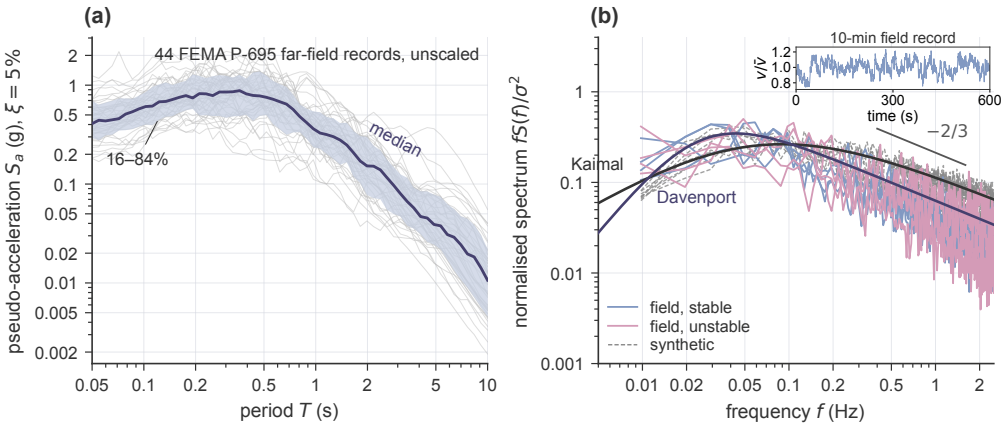


Figure 12: **The held-out record libraries.** (a) Unscaled 5%-damped pseudo-acceleration response spectra of the 44 FEMA P-695 far-field records (gray), with their median and 16–84% band; only these records enter the held-out selection. (b) Normalized power spectra $fS(f)/\sigma^2$, with σ^2 the variance resolved between 0.01 and 2.5 Hz, of the eight field-recorded wind series (color, by atmospheric stability) and the eight spectrum-consistent synthetic series (gray, dashed), against the Kaimal and Davenport targets used for synthesis; the short line marks the $-2/3$ inertial-subrange slope; inset, one field record over its ten minutes. Sources and processing in the text.

Structural-code knowledge base. To support clause-level verification of design limits, we build a parallel knowledge base. It collects 86 original structural design codes from six countries and regions, namely China (GB 50011-2010, JGJ 3-2010, and GB 50009-2012; MOHURD, 2010a;b; 2012), Europe (Eurocode 8; CEN, 2004), the United States (ASCE/SEI 7; ASCE, 2022), India (IS 1893; BIS, 2016), Japan (the Building Standard Law, BSL; MLIT, 1950), and New Zealand (NZS 1170.5; Standards New Zealand, 2004), and organizes them into 625 cross-linked entries comprising 121 code source documents, 79 institution and committee entities, and 425 clause concepts (Fig. 11a). It covers seismic and wind design provisions, load values (dead, live, wind, and seismic), and concrete structural design rules, and tracks each code's revision history. Every limit in the task set carries a clause number traceable to this knowledge base.

Detailing set. Every code-compliance instance carries a detailing set G_x^{build} drawn from its code family with clause provenance in the knowledge base. The current release sources one clause family, the beam clear-span-to-depth limit $l_n/h_b \geq 4$ (GB 50011-2010 Clause 6.3.1, MOHURD, 2010a; JGJ 3-2010 Clause 6.3.1, MOHURD, 2010b), evaluated geometrically from the returned member

[1] https://ngawest2.berkeley.edu/.

[2] Documentation: https://rotor-inflow-benchmark.readthedocs.io/. Data, CC BY 4.0: https://zenodo.org/records/15747594, https://zenodo.org/records/15747684.

sections at no simulator cost; the two further candidate clauses and the reasons for their exclusion are given in Appendix D.3.

C ONE MASSING, FOUR PROJECTORS

Figures 13 and 14 show six buildings, each read from one massing by four tools, to make visible what the margin oracle of Section 4 analyzes. The four columns are the same in both figures. They are headed by the tool that produced them: the massing in Rhino[3], a Blender[4] rendering of it, the frame idealization the physics engine adjudicates in OpenSeesPy (McKenna, 2011; Zhu et al., 2018), and that frame's response to a recorded ground motion. The first two columns are what a designer sees; the last two are what the physics engine sees. Table 2 lists the six frame models.

Rhino (column 1). For the three classics, the panel is a still from a four-viewport screen recording of the interactive Rhino session in which the massing was modeled; it is the first frame in which the model is complete. The benchmark's LLM-free search produced the three towers, so no such session exists, and our headless Rhino export writes box massings only; their panel therefore draws the massing the frame generator starts from, one lofted perimeter ring per floor, in a Rhino-style four-viewport layout. That drawing is ours, not a Rhino screen.

Blender (column 2). For the classics, the panel shows the Blender scene of the same massing, re-rendered headless with Cycles on a light world. For the towers, we loft the column 1 massing, mesh it, and render it with the same world, camera rule, and Cycles settings, with a glazed facade and a story band per floor so the form reads as a building.

OpenSeesPy model (column 3). The panel shows the frame idealization that the benchmark generates and the physics engine analyzes, as four orthographic views (plan, north and east elevations, isometric). Every structural member is drawn; in the tower rows, facade members are hidden so that columns, cores, and belts are visible.

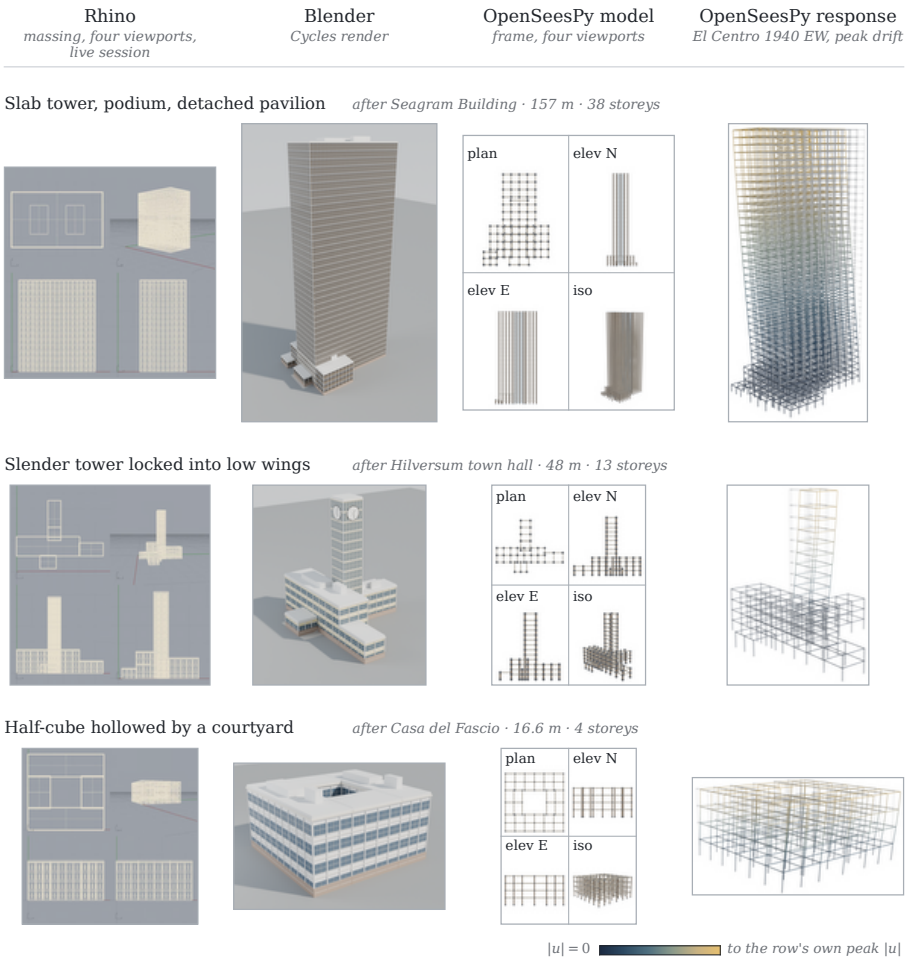
OpenSeesPy response (column 4). The panel shows the deformed shape of the column-3 model at the instant of peak roof drift under a linear time-history analysis run for these figures: the east–west (S90W) component of the 1940 Imperial Valley earthquake recorded at the El Centro terminal substation, USGS station 117, in the Caltech corrected digitization (Hudson et al., 1971) as redistributed by vibrationdata.com[5] (PGA 0.214 g, 2,674 points at $\Delta t = 0.02$ s, 53.5 s), uniform base excitation along global X, Newmark average acceleration (Newmark, 1959) with the effective stiffness factorized once, Rayleigh damping of 5% on the first two modes, every node recorded at every step. Members are colored by $|u|$ from zero to that row's own peak, against the undeformed ghost. This analysis is illustrative: it is not the held-out verification of Eq. 9, which uses the benchmark's ground-motion set and is scored, not drawn. Before any step was run, each rebuilt model was asserted to reproduce the node count, element count, and first period stored with the building; Table 2 gives those counts, T_1, the number of structurally connected pieces the frame generator leaves, the peak roof drift, and the peak $|u|$.

What these figures do not claim. The classics are massings modeled in an interactive Rhino session before this work and turned into frames by the same generator; the towers come from the benchmark's LLM-free search (target derivation, grid search over sections, verification). None of the six is an episode of Section 5. The three mega-frames are certified on drift only, since the mega-frame model does not evaluate top acceleration, period ratio, or the shear-weight and stiffness-weight ratios; the classics carry every check. The response model is screening-level and elastic: linear members, no gravity in the seismic case, no material nonlinearity, no soil, one horizontal component of one record. We did not compare against any measured or published response of these buildings, and the figures make no claim about the analysis's accuracy. They show that one massing consistently drives four tools and what the last two contain.

[3] Rhinoceros 3D, Robert McNeel & Associates, https://www.rhino3d.com.

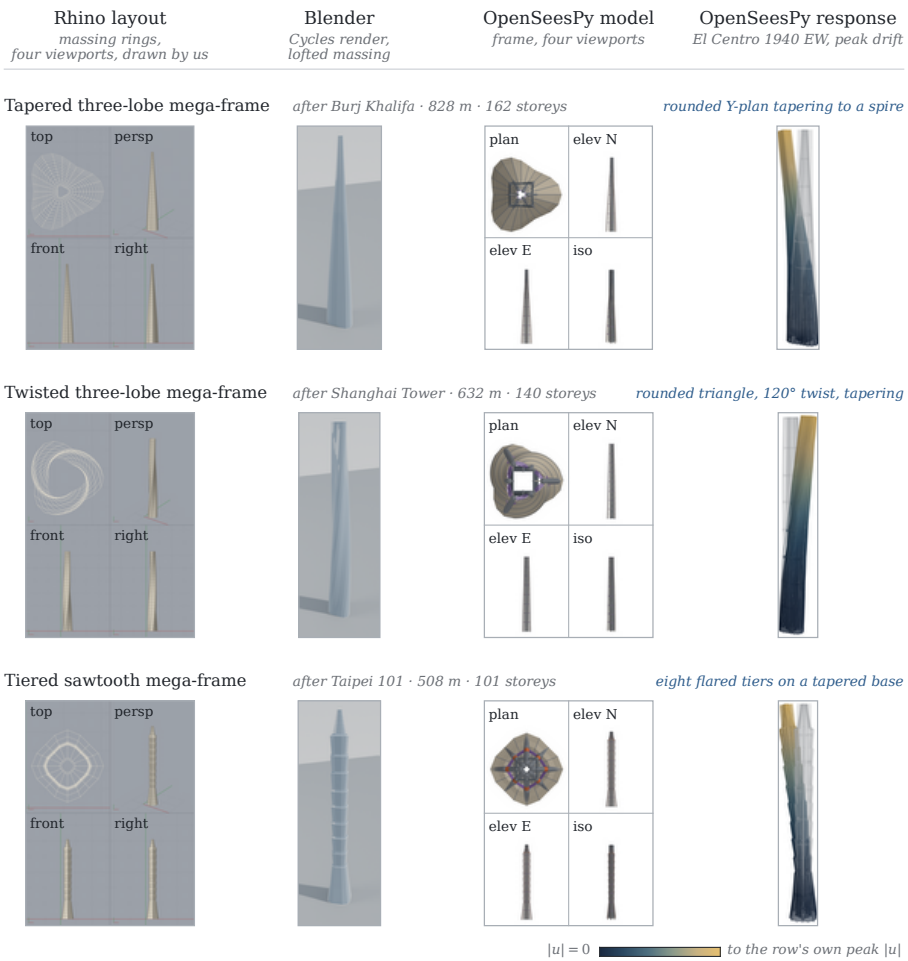
[4] Blender, Blender Foundation, https://www.blender.org.

[5] https://www.vibrationdata.com/elcentro_EW.dat, two-column file of time and acceleration in g.



building	nodes	elements	first T_1 (s)	pieces	peak roof drift	peak $\lvert u \rvert$ (mm)
after Seagram Building	2746	7469	2.5486	2	H/483	325
after Hilversum town hall	243	454	0.9197	4	H/503	96
after Casa del Fascio	280	576	0.4852	1	H/456	42

Figure 13: **Classic buildings, four projectors.** Three massings modeled in an interactive Rhino session before this work: a slab tower on a podium with a detached pavilion (after the Seagram Building, 157 m, 38 stories), a slender tower locked into low wings (after the Hilversum town hall, 48 m, 13 stories), and a courtyard block (after the Casa del Fascio, 16.6 m, 4 stories). Column 1 is a still of that session, column 2 its Blender scene, columns 3 and 4 the frame the physics engine analyzes and its peak response, displacements drawn $\times 50$. The town hall's four structurally separate pieces move independently in column 4; the frame generator leaves this, not a display fault. Columns and analysis as described in Appendix C; model sizes and peak responses in Table 2.



building	nodes	elements	first T_1 (s)	pieces	peak roof drift	peak $\lvert u \rvert$ (mm)
after Burj Khalifa	6194	20736	8.4376	1	H/697	1189
after Shanghai Tower	5358	17964	8.5692	1	H/828	762
after Taipei 101	3264	10096	4.3478	1	H/807	630

Figure 14: **Supertall towers, four projectors.** Three mega-frames from the benchmark's LLM-free search: a rounded Y-plan tapering to a spire (after Burj Khalifa, 828 m, 162 stories), a rounded triangle with a 120° twist (after Shanghai Tower, 632 m, 140 stories), and eight flared tiers on a tapered base (after Taipei 101, 508 m, 101 stories). Column 1 is the lofted massing we drew, column 2 is our Cycles render, and columns 3 and 4 show the frame the physics engine analyzes and its peak response, with displacements drawn $\times 100$. Seen from above, the ring beams of a tapering or twisting tower sit at a different radius on every floor, so the plan viewports of columns 1 and 3 both show a stacked disc. All three are drift-certified only. Columns and analysis are as described in Appendix C; model sizes and peak responses are in Table 2.

Table 2: The six frame models of Figs. 13 and 14: size of the idealization that the physics engine analyzes, first period T_1, structurally connected pieces left by the frame generator, and peak response under El Centro 1940 EW (settings in Appendix C).

Massing (after)	H (m)	Stories	Nodes	Elements	T_1 (s)	Pieces	Peak roof drift	Peak $\lvert u\rvert$ (mm)
Classics (Fig. 13), displacements drawn ×50								
Seagram Building	157	38	2,746	7,469	2.55	2	$H/483$	325
Hilversum town hall	48	13	243	454	0.92	4	$H/503$	96
Casa del Fascio	16.6	4	280	576	0.49	1	$H/456$	42
Supertall towers (Fig. 14), displacements drawn ×100								
Burj Khalifa	828	162	6,194	20,736	8.44	1	$H/697$	1,189
Shanghai Tower	632	140	5,358	17,964	8.57	1	$H/828$	762
Taipei 101	508	101	3,264	10,096	4.35	1	$H/807$	630

D METHOD DETAILS AND DERIVATIONS

D.1 IN-LOOP CONSTRAINT SET AND THE ROLE OF THE PPR

The in-loop constraint set G_x is fixed by the code family and the task type of the instance (Section 2) and is evaluated by equivalent static and modal analysis. Code-compliance instances use the checks of the bound code family for which limits could be sourced, typically upper bounds on the inter-story drift ratio, the torsional displacement ratio, the minimum shear-to-weight ratio, the story lateral-stiffness ratio, the torsional-to-translational period ratio, and the stiffness-to-weight ratio (global stability index); performance-target instances additionally carry a lower bound on the fundamental frequency and an upper bound on the static top displacement. For buildings taller than 150 m, an upper bound on wind-induced peak acceleration also enters G_x and is checked at every evaluation by a modal-superposition wind time history. Response-history analysis under recorded ground motions never enters G_x; candidates rejected by the zero-reference gates of Section 2 are treated as $R = 0$ with every constraint violated.

The PPR is formally close to potential-based reward shaping (Ng et al., 1999). However, it is neither a hand-designed potential nor does it enter the return in the difference form $\gamma\phi(s') - \phi(s)$; it is used only as an observation and a verdict (violated constraints guide repair before feasibility, concrete volume ranks candidates after feasibility). It therefore inherits no policy-invariance guarantee, and such theoretical properties are outside the scope of this work. Because the design loop consumes only the sign of $\boldsymbol{\rho}$, its behavior does not depend on any heuristic weighting of margin magnitudes, and there is no tunable intermediate quantity between guidance and verdict. Magnitudes are returned to the agent and recorded in the certificate but not consumed by the search; using them, e.g., to order C1 actions by the most violated constraint, is a natural extension we leave open.

D.2 INITIAL TOPOLOGY AND THE COLUMN GRID

In $\theta_d = (n_{\text{core}}, n_{\text{wall}}, \kappa, \varsigma)$, the core plan geometry κ consists of the core-to-floor area ratio, the plan aspect ratio, and the plan offset; the section triple ς consists of the beam-section scale, the concrete-grade scale, and the vertical taper, each on a small finite menu. The intent $\hat{y} = (\text{system}, \text{governing load}, \text{strategy}) = \pi_{\text{int}}(\hat{M}, S)$ fixes the initial topology

$$\begin{aligned} n_{\text{core}}^{(0)} &= \mathbb{1}\big[\texttt{add_core} \in \text{strategy} \ \vee \ \text{system} = \text{frame-core}\big], \\ n_{\text{wall}}^{(0)} &= \mathbb{1}\big[\texttt{add_wall} \in \text{strategy} \ \vee \ \text{system} = \text{frame-shear-wall}\big]; \end{aligned} \tag{10}$$

$\kappa^{(0)}$ and $\varsigma^{(0)}$ take the defaults (area ratio 0.15, aspect ratio 1, zero offset; beam-section and concrete-grade scales 1, taper 0). Equation 10 writes the semantic decision explicitly into $\theta_d^{(0)}$, which also lets Section 5 test whether the final design is consistent with the declared intent.

The column grid G_b of Eq. 6 is a convention rather than a derivation, and we state its provenance. The 50 mm step is the construction modulus on which reinforced-concrete column dimensions are specified, so every $b \in G_b$ is buildable. The lower end, 0.30 m, is the smallest section the parametric frame template instantiates. The upper end, 2.50 m, is the clamp on column sections shared by the search spaces of all compared systems; an earlier, agent-specific ceiling of 1.6 m made the agent

escalate before the shared search space was exhausted, so we unified it to the shared clamp. The resulting size $|G_b| = 45$ is what fixes the bisection cost in Lemma 1.

D.3 Detailing set and projection

The frame template sizes members from θ alone: the column is square with side b; the beam depth and width are $h_b = 0.8\,b\,\varsigma_{\text{beam}}$ and $b_b = \max(0.4\,b,\ h_b/2.5)$ with $\varsigma_{\text{beam}} \in [0.5, 2.0]$; the bay width in each plan direction is $W/\max(1, \lfloor W/6\,\text{m} \rceil)$, with $\lfloor\cdot\rceil$ rounding to the nearest integer (and likewise for D), so the beam clear span is $l_n = \text{bay} - b$, taken as the smaller of the two directions. The detailing set G_x^{build} currently sources one clause family, the clear-span-to-depth limit $l_n/h_b \geq 4$ (GB 50011-2010 Clause 6.3.1; JGJ 3-2010 Clause 6.3.1), whose margin follows Eq. 1 with $q = h_b/l_n$, $\text{lim} = 1/4$, and $s = +1$. Because h_b grows with b while l_n shrinks with b, the constraint binds for thick columns: with $\text{bay} = 6$ m and $\varsigma_{\text{beam}} = 1$ it is violated for $b > 6/4.2 = 1.43$ m, i.e., from $b = 1.45$ m on the grid.

Before every evaluation the agent applies the projection

$$\Pi_{\text{build}}: \quad \varsigma_{\text{beam}} \leftarrow \min\{\varsigma_{\text{beam}},\ l_n/(3.2\,b)\}, \qquad b_{\max} \leftarrow \max\{\, b \in G_b :\ b \leq \text{bay}/2.6 \,\}, \quad (11)$$

which follows directly from the clause. Solving $l_n/(0.8\,b\,\varsigma_{\text{beam}}) \geq 4$ for the beam scale gives $\varsigma_{\text{beam}} \leq l_n/(3.2\,b)$; requiring this bound to admit the smallest scale on the menu, $\varsigma_{\text{beam}} \geq 0.5$, gives $\text{bay} - b \geq 1.6\,b$, i.e., $b \leq \text{bay}/2.6$, which caps the grid at $b_{\max}$. Both operations are analytic in θ and cost no oracle calls; because Assumption 1 is stated on the projected path, Lemma 1 applies to the capped grid verbatim (Appendix D.5). Applied post hoc to designs returned without the projection, the same clipping repairs every violation except those with $b > \text{bay}/2.6$, which no beam scale can rescue.

Figure 15 is the measurement that motivated placing the projection inside the loop rather than applying it after the fact. For each compared system, it reports the share of physically feasible returned designs on the 59-instance search-sufficient reference set, pooled over three seeds, that violate the clear-span-to-depth limit. The gap grows with column size and is largest for the agent without the projection, because its bisection settles on the thickest feasible column. All but three of the violations are repairable by clipping the beam scale alone; the three exceptions have $b > \text{bay}/2.6$.

We analyzed two additional clause families that are not sourced. The joint rule that the beam must not be wider than the column (GB 50011-2010 Clause 6.3.2 otherwise requires flat-beam provisions) can bind only in a corner of the menu: $b_b/b \leq \max(0.4,\ 0.32\,\varsigma_{\text{beam}})$ while the tapered top-column side is $(1 - \text{taper})\,b \geq 0.6\,b$. Hence, a violation requires $0.32\,\varsigma_{\text{beam}} > 1 - \text{taper}$, i.e., $\varsigma_{\text{beam}} > 1.875$ at the maximum taper of 0.40. No returned design reaches this corner (Appendix E), so the rule is monitored but not sourced as a constraint. The seismic-grade-dependent minimum column side (300 mm or 400 mm, GB 50011-2010 Clause 6.3.5; JGJ 3-2010 Clause 6.4.1) requires the seismic grade of GB 50011-2010 Table 6.1.2, which is not defined uniformly across the code families of the benchmark; the axial-compression-ratio limit requires member forces that the harness does not expose. Both are reported as limitations.

D.4 Pseudocode

Algorithm 1 gives the episode-level control flow of PE-Loop as implemented, in the notation of Sections 2 and 4; the color convention matches the hatching of the method figure: blue marks the three language-model calls, which are not charged to the budget, and red marks the margin-oracle calls, each of which costs one evaluation, so a round's evaluation count can be read off the red lines directly. In the Π_{build} variant, the projection of Eq. 11 is applied to every candidate before evaluation; the harness-side re-verification of the returned design on the true massing M (Section 2) and the offline final verification L4 (Appendix D.10) are not part of the agent and are omitted.

D.5 Correctness of C2 and its evaluation bound

Assumption 1 (Monotone margin along the projected column axis)**.** *For fixed θ_d and $b \leq b'$ in G_b, $V(\Pi_{\text{build}}(\theta_d, b)) \leq V(\Pi_{\text{build}}(\theta_d, b'))$.*

Assumption 1 is stated on the projected path because Π_{build} may lower the beam scale as b grows. Two facts confine its role. First, the beam clip is inactive on the whole grid whenever $\varsigma_{\text{beam}} \leq$

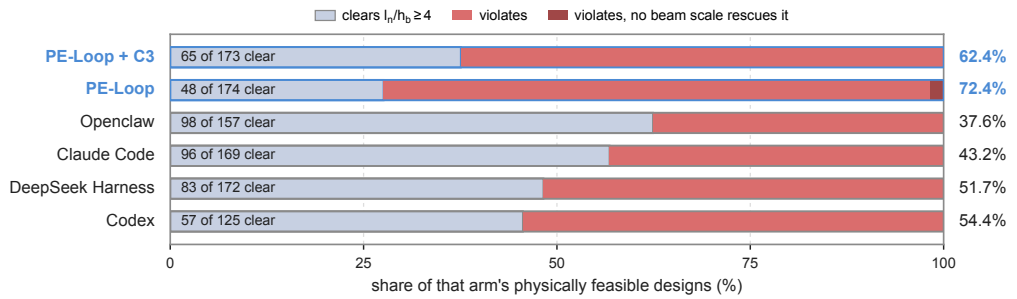


Figure 15: The detailing gap is not ours alone. Each bar splits one arm's feasible designs in the 59-instance reference set by the clause $l_n/h_b \geq 4$; the right rate is the violating share. Every system returns unbuildable designs (42.75–72.41%). PE-Loop without Π_{build} is the worst, but 123 of its 126 violations are repairable by clipping one section scale, which Π_{build} does by construction (Table 6). Coverage in Table 7.

$l_n(b_{\max})/(3.2\,b_{\max})$, since $(\text{bay} - b)/(3.2\,b)$ is decreasing in b; in that case ς is literally frozen and the assumption reduces to monotonicity of the unprojected margin (Lemma 3). Second, whether or not it holds, the returned design is feasible (Lemma 4); only minimality along b can be lost.

Lemma 1. *Under Assumption 1: (i) the feasible section set $U(\theta_d) = \{b \in G_b : R(\Pi_{\text{build}}(\theta_d, b)) = 1\}$ is an up-set of G_b; (ii) $b^\star(\theta_d) = \min U(\theta_d)$ exists if and only if $R(\theta_d, b_{\max}) = 1$, where $b_{\max}$ is the upper end of the grid after the cap of Eq. 11 (2.50 m when the cap is inactive); (iii) bisection locates $b^\star$ with at most $\lceil \log_2 |G_b| \rceil = 6$ evaluations, so one C2 call uses at most $2 + \lceil \log_2 |G_b| \rceil = 8$ evaluations (upper-bound probe, bisection, final confirmation; Algorithm 1, lines 5–12), and one outer round, which re-verifies the returned design once more, uses at most $\lceil \log_2 45 \rceil + 3 = 9$.*

Proof. (i) If $b \in U$ and $b' \geq b$, then by Assumption 1 $V(\theta_d, b') \geq V(\theta_d, b) \geq 0$, hence $b' \in U$. (ii) By (i), $U \neq \varnothing$ if and only if $b_{\max} \in U$; G_b is finite, so a minimum exists whenever U is non-empty. (iii) Bisect the index interval $[0, h_0]$ with $b_{h_0} = b_{\max}$: if the midpoint is feasible, move the right endpoint to the midpoint, otherwise move the left endpoint to the midpoint plus one. By (i) the invariant "the right endpoint is feasible and every index before the left endpoint is infeasible" is preserved, the interval length at least halves at every step, and $h_0 + 1 \leq 45 \leq 2^6$ gives convergence to a single index within 6 steps. Adding the probe, the confirmation, and the final re-verification gives the stated bounds. □

Lemma 2. *If the concrete volume $c(\theta_d, b)$ is strictly increasing in b, then under Assumption 1 $b^\star(\theta_d) = \arg\min\{c(\theta_d, b) : b \in U(\theta_d)\}$, which is the second equality in Eq. 7.*

Proof. U is an up-set with minimum $b^\star$, and a strictly increasing function attains its minimum over U at the minimum element. □

Lemma 3 (Inactive clip)**.** *If $\varsigma_{\text{beam}} \leq l_n(b_{\max})/(3.2\,b_{\max})$, then $\Pi_{\text{build}}(\theta_d, b) = (\theta_d, b)$ for every $b \in G_b$ with $b \leq b_{\max}$.*

Proof. $l_n(b)/(3.2\,b)$ is decreasing in b, so its minimum over the capped grid is attained at $b_{\max}$; below that minimum the clip of Eq. 11 is the identity at every $b \leq b_{\max}$, and the cap is already applied. □

Lemma 4 (Feasibility without Assumption 1)**.** *Any design that C2 returns with status* PASS *satisfies $R = 1$, whether or not Assumption 1 holds; in the Π_{build} variant, it additionally satisfies $\min \boldsymbol{\rho}^{\text{build}} \geq 0$.*

Proof. C2 returns PASS only after the confirmation evaluation (Algorithm 1, line 12) reports $R = 1$ for the very design returned; in the Π_{build} variant, $\min \boldsymbol{\rho}^{\text{build}} \geq 0$ holds by construction of the projection. □

Algorithm 1 PE-Loop episode. Blue: language-model calls, not charged to B. Red: margin-oracle calls (Eq. 1), each charged one evaluation. In the Π_{build} variant, every oracle call is preceded by the projection of Eq. 11. The harness re-verification on M and the held-out L4 analyses are not part of the agent.

Require: instance $x = (O_\tau, S)$; grid $G_b = \{b_0 < \cdots < b_{44}\}$ (Eq. 6); action set $\mathcal{A}$; budget B; rounds K; tier $\in \{\texttt{light}, \ldots, \texttt{ultra}\}$
Ensure: status $o \in \{\text{PASS}, \text{ESCALATED}\}$; design $\theta^\star$; margins $\boldsymbol{\rho}(\theta^\star)$; evaluations NFE

// B0–B2 perception and intent (Eq. 10)
$\hat{M} \leftarrow \pi_{\text{perc}}(O_\tau)$ if $\tau \neq$ text, else parse the brief
$\hat{y} \leftarrow \pi_{\text{int}}(\hat{M}, S)$; $\theta_d^{(0)} \leftarrow$ initial state of Eq. 10; NFE $\leftarrow 0$
for $k = 0, 1, \ldots, K$ **do**
// C2 minimal feasible column side under $\theta_d^{(k)}$ (Eq. 7; Lemmas 1–2)
  $b_{\max} \leftarrow \max\{b \in G_b : b \le \text{bay}/2.6\}$ if Π_{build}, else b_{44} ▷ Π_{build} caps the grid
  EVAL$(\theta_d^{(k)}, b_{\max})$; NFE += 1 ▷ upper-bound probe
  if $R(\theta_d^{(k)}, b_{\max}) = 1$ **then**
    $\ell \leftarrow 0$; $h \leftarrow \text{index}(b_{\max})$
    while $\ell < h$ **and** NFE $< B - 1$ **do** ▷ $\le \lceil \log_2 |G_b| \rceil = 6$ steps
      $m \leftarrow \lfloor (\ell + h)/2 \rfloor$; EVAL$(\theta_d^{(k)}, b_m)$; NFE += 1
      if $R(\theta_d^{(k)}, b_m) = 1$ **then** $h \leftarrow m$ **else** $\ell \leftarrow m + 1$
    end while
    EVAL$(\theta_d^{(k)}, b_\ell)$; NFE += 1 ▷ confirmation; catches budget truncation
    if $R(\theta_d^{(k)}, b_\ell) = 1$ **then**
      $\theta^\star \leftarrow (\theta_d^{(k)}, b_\ell)$; $o \leftarrow$ PASS; **break** ▷ PASS before exhaustion
    end if
  end if
  $\mathcal{J} \leftarrow \{ j : \rho_j < 0 \}$ of the last evaluation ▷ Eq. 5; the only signal C1 receives
// C1 one discrete repair, proposed by the language model (Eq. 8)
  if $k = K$ **or** NFE $\ge B$ **then**
    $\theta^\star \leftarrow (\theta_d^{(k)}, b_{\max})$; $o \leftarrow$ ESCALATED; **break**
  end if
  $a_k \leftarrow \pi_{\text{C1}}(\hat{M}, \theta_d^{(k)}, \mathcal{J}) \in \mathcal{A}$; $\theta_d^{(k+1)} \leftarrow a_k(\theta_d^{(k)})$ ▷ clipped to admissible ranges
end for
// C3 material loop along ς, `ultra` tier only (Eq. 12)
if $o =$ PASS **and** tier $=$ `ultra` **then**
  for all θ' in the candidate menu, while NFE $< B$ **do**
    EVAL(θ'); NFE += 1; **if** $R(\theta') = 1 \ \wedge \ c(\theta') < c(\theta^\star)$ **then** $\theta^\star \leftarrow \theta'$
  end for
end if
if NFE $< B$ **then**
  EVAL$(\theta^\star)$; NFE += 1 ▷ final measurement; L4 and $\mathcal{C}$ offline (Eq. 9)
end if
return o, $\theta^\star$, $\boldsymbol{\rho}(\theta^\star)$, NFE

C2 proceeds in three steps (Algorithm 1, lines 5–12). It first spends one evaluation on $(\theta_d, b_{\max})$; if $R = 0$, then by Lemma 1(ii) $b^\star$ does not exist and the violated-constraint set $\mathcal{J}(\theta_d, b_{\max})$ is handed to C1. Otherwise, it bisects as in Lemma 1(iii), and finally spends one evaluation to confirm $b^\star$. The remaining budget hard-truncates the procedure; if truncation occurs and the returned section fails the confirmation, it is treated as "$b^\star$ does not exist". The section triple ς is frozen throughout the bisection (up to the beam-scale clip of Π_{build}): enlarging b monotonically stiffens the structure, whereas adding the beam section, concrete grade, or taper to the search would break monotonicity and multiply the evaluation count by the number of dimensions, so these three are left to the discrete decision in C1. If Assumption 1 fails on an individual instance (e.g., non-monotone behavior from higher-mode coupling), the final confirmation still guarantees that the returned $b^\star$ is feasible (Lemma 4); only minimality is lost, and this risk is covered by the final verification and by the acceptance protocol of Section 5.

D.6 C1 ACTION SET AND THE BUDGETED OUTER LOOP

C1 presents the language model with a geometry summary, the current θ_d, and the violated-constraint set $\mathcal{J}$, and asks for one action from $\mathcal{A}$. The action set contains five action types on three levels of leverage: at the section level, adjusting the section triple ς; at the topology level, adding a core, adding a shear wall, and adjusting the core plan geometry κ; at the system level, revising the declared lateral system in place and resetting the initial topology by Eq. 10, i.e., an intent-level fallback. Actions only answer "what to change"; out-of-range numerical parameters are clipped to their admissible ranges rather than rejected (core-to-floor area ratio $[0.01, 0.30]$, plan aspect ratio $[0.2, 5.0]$, plan offset $[-0.4, 0.4]$ in each direction, beam-section scale $[0.5, 2.0]$, concrete-grade scale $[1.0, 1.30]$, taper $[0, 0.40]$), so the output of C1 is always a valid point of Θ_d. Adding a core or a wall increments the corresponding count; changing the lateral system resets the counts as in Eq. 10; an unparsable action falls back to adding a core.

The outer loop runs as in Eq. 8: round k first executes C2 and passes if $b^\star$ is found; otherwise, if $k = K$ or the cumulative evaluation count satisfies $\mathrm{NFE}_k \geq B$, the episode terminates as ESCALATED; otherwise, C1 is executed, and the next round begins. Fallback is thus layered: failures with the section below its upper bound are absorbed inside C2 by enlarging b; failures that persist at the upper bound are escalated to C1, where the language model chooses among section-, topology-, and system-level levers conditioned on $\mathcal{J}$ (for instance, torsion-governed failures favor system-level actions, but the model makes this choice and is not a hard-coded rule); ESCALATED is declared only when the budget B or the round limit K is exhausted. This division confines the language model to the discrete semantic space where it has a comparative advantage, while the oracle still adjudicates the consequence of every decision.

D.7 MATERIAL LOOP C3

C3 is enabled only in the `ultra` tier. Starting from the first passing design, which by Eq. 7 already solves Eq. 2 along b, it accepts a candidate θ' iff

$$\theta^\star \leftarrow \theta' \quad \Longleftrightarrow \quad R(\theta') = 1 \ \wedge \ c(\theta') < c(\theta^\star). \tag{12}$$

To reduce c along ς, it enumerates five section-triple candidates (raising the concrete grade, introducing a vertical taper, enlarging the beam section, and their combinations), ordered by the material efficiency measured in a pilot study so that the cheapest lever is tried first, runs a coarse-to-fine column-side search around $b^\star$ for each candidate (steps of 12, 6, 3, and 1 grid cells, stopping at the first feasible reduction), re-checks every candidate with the oracle, and accepts strictly cheaper and only feasible candidates, following Eq. 12. Candidates are evaluated independently from the pre-C3 pair $(\varsigma, b^\star)$ rather than cumulatively, and the cheapest feasible one is adopted (Algorithm 1). C3 has its own cap of 4 evaluations and is bounded by B; the candidate menu is deliberately small because every candidate consumes evaluations.

D.8 COMPUTE TIERS

Structural design has no single correct compute budget: schematic design needs comparable options within minutes, whereas construction-document design needs a design that is right the first time. The architecture therefore does not fix the budget; it instantiates the same B0–C3 pipeline at four tiers that differ only in the budget B, the number of outer rounds K, and two lever switches: `light` $(B, K) = (5, 1)$ for schematic comparison, `medium` $(10, 2)$, `high` $(15, 3)$ as the default and maintable setting, and `ultra` $(25, 4)$ for construction documents. The C3 material loop runs only at `ultra`; the held-out L4 verification runs only at `high` and `ultra`, so we report the two lower tiers on PASS and evaluation count alone. The evaluation count includes only oracle calls; language-model and vision-model calls and the L4 final verification are not charged to B.

The tier values are operating points on one anytime family rather than derived quantities. The `high` tier is the main table setting: $B = 15$ is the budget shared by every compared system, and $K = 3$ allows one topology decision per lever level. The `light` tier sits below the cost of one complete round (9 evaluations, Lemma 1) and can therefore only screen feasibility; at $B = 5$ the confirmation may be the last charged call, so the final measurement is skipped, and the confirmation margins are reported. `medium` affords one complete round plus a fallback; `ultra` adds one more round and

the C3 budget. Section 5 evaluates the family by a success rate that increases from `light` to the main-table `high` setting.

D.9 TERMINATION SEMANTICS

An episode ends in one of four terminal states: PASS (the returned design satisfies all in-loop constraints), ESCALATED (honest termination after the levers or the budget are exhausted), FAIL (execution crash), and UNEVALUATED (not assessable). ESCALATED and FAIL are strictly separated: the former is the legitimate, scored terminal state of Section 2, in which the agent declares that no feasible design was found within the given action set and budget, rather than a feigned convergence; the latter is a system failure. UNEVALUATED is further attributed to the responsible side: non-delivery, timeout, or crash on the side of the system under test is counted as infeasible and enters the denominator, whereas cases in which the harness cannot synthesize the geometry or has no sourced constraint are removed from the denominator. This attribution rule prevents the pass rates in Section 5 from being systematically biased against any system by harness defects.

D.10 FINAL VERIFICATION L4

Dynamic response is the real loading condition, but response-history analysis under recorded ground motions is one to two orders of magnitude more expensive than elastic static analysis, and placing it inside the iteration loop would raise the cost per instance by an order of magnitude; seismic response-history analysis is therefore used only as a final check. After the equivalent-static loop converges, the returned design $\theta^\star$ undergoes one seismic response-history analysis under recorded ground motions and one full wind time-history analysis, yielding the held-out margin vector of Eq. 9. Its components are the inter-story drift ratio under response history (same limit as the in-loop constraint) and the wind-induced peak acceleration (limited by occupancy: 0.15 m/s^2 for residential and apartment buildings and 0.25 m/s^2 for office and hotel buildings, following Clause 3.7.6 of JGJ 3 (MOHURD, 2010b), applied above its 150 m applicability height). Whether seismic response-history analysis is mandatory is prescribed by height, seismic intensity, and site class (Clause 5.1.2 of GB 50011 (MOHURD, 2010a): above 100 m for intensity 7 and for intensity 8 on site classes I–II, above 80 m for intensity 8 on site classes III–IV, above 60 m for intensity 9); our held-out verification runs the seismic response history on every returned design without applying this trigger, so that the held-out set is not systematically biased toward instances that are unlikely to fail. Because the recorded-ground-motion condition never enters G_x, L4 is a held-out test on a loading condition unused inside the loop; it shares the physical model with the loop and therefore cannot detect systematic errors in the model form itself (Section 6). The third certificate component $\boldsymbol{\rho}^{\text{build}}$ is the detailing margin vector in Appendix D.3; for the projected agent, it is non-negative by construction, and for non-projected systems, it records the detailing gap.

Held-out records and scaling. The seismic re-check draws one record per design from the FEMA P-695 far-field subset of the record library of Appendix B: candidates are ranked by spectral-shape misfit to the site design spectrum over $[0.2\,T_1,\ 1.5\,T_1]$, the best match is taken, and it is amplitude-scaled by the geometric-mean log ratio of code to record spectral acceleration over that band, the convention of GB 50011 Clause 5.1.2, ASCE 7 Chapter 16, and EN 1998-1 Clause 3.2.3.1 (MOHURD, 2010a; ASCE, 2022; CEN, 2004). The analysis is an elastic modal-superposition response history in one horizontal direction with 5% Rayleigh damping anchored at T_1 and T_3, restricted to the 5–95% Arias-intensity window; the peak inter-story drift ratio under that record is checked against the instance's drift limit. The re-check runs on every returned design, whether or not Clause 5.1.2 of GB 50011 would require a time-history analysis at that height. The wind re-check draws three of the sixteen ten-minute series of the wind library of Appendix B, their fluctuation amplitude rescaled to the turbulence intensity of the site's terrain category, field records taken first, and applied as quasi-steady buffeting loads $F_i(t) = \bar{F}_i\,[v(t)/\bar{v}]^2$ at the code basic wind pressure. The same modal machinery with 2% damping gives the drift envelope, checked against the static limit unless the first natural frequency exceeds 1 Hz, in which case the code static wind load governs drift, and the top-floor acceleration, converted from the fifty-year to the ten-year return pressure of GB 50009 Table E.5 (MOHURD, 2012) by linear rescaling and checked against the Clause 3.7.6 limits of JGJ 3 (MOHURD, 2010b); for code families whose ten-year pressure is not tabulated, the acceleration check is reported as unevaluated rather than approximated.

Table 3: Difficulty map of the screened corpus and the main-split quotas. Every one of the 3,216 screened instances was labeled by an exhaustive LLM-free search that tries section resizing, then up to four cores, then up to four walls, with at most 45 evaluations per stage: section-only instances are solved by resizing alone, topology-decision instances need a core or wall, and boundary instances are not solved within the action space. The last three columns give how many of each the 260-instance main split draws per height band.

	Screened corpus				Main split		
Height band	n	Section-only	Topology-decision	Boundary	T-dec.	Controls	Boundary
<24 m	1,239	828	49	362	30	30	0
24–60 m	968	580	193	195	65	35	0
60–120 m	615	357	98	160	30	30	0
120–250 m	371	71	5	295	5	15	0
>250 m	23	0	0	23	0	0	20
All	3,216	1,836	345	1,035	130	110	20

D.11 Architectural placement

From a graph-engineering perspective, the pipeline is a task graph whose nodes are deterministic physics tools and language-model calls (perception, intent parsing, topology decision, parameter search, material optimization, final verification) and whose edges represent data dependencies. The graph runs as an iterative loop rather than a one-shot orchestration. In each round, the agent decides the next action (repair or escalate) based on physics feedback rather than being forced along the graph structure. In the layered stack of graph engineering (Feng et al., 2026; He, 2026b), the system sits at the "heavy harness, single loop" position: the acceptance authority is built into the deterministic physics tools of the harness layer rather than delegated to a supervising model of the same origin, which is precisely the structural remedy for the verification-credibility deficit discussed in Appendix A.

E Additional experimental results

Protocol. The 260-instance main split holds 130 topology-decision instances, 110 section-only controls, and 20 capability-boundary instances; the last are excluded from every comparative statistic and listed in Table 4 only to document that no arm solves them. Within an LLM block, every arm faces the same instances on every seed, the same margin oracle (Eq. 1), and the same budget; each rate divides by its own denominator (below). The search-sufficient reference set (GB-59: 59 instances $\times$ 3 seeds) carries the comparisons against the LLM-free reference lines and is never pooled or compared with the main split (caveat (ii) of Appendix E.4).

Representativeness of the main split. The 130 scored instances are a stratified 37.7% sample of the 345 topology-decision instances in the pool, the only stratum that separates systems: section-only instances are passed by 93.6–97.9% of answers under every system through column resizing, and boundary instances by none (Table 4). Sampling is deliberately non-uniform by height, drawing 4.8% of the 1,239 instances below 24 m but 87.0% of the 23 above 250 m, which is what makes the twenty boundary instances available; the split thus reflects the discriminative stratum, not the pool's height distribution. At n=390 answers per cell, the Agresti–Coull standard error (Agresti & Coull, 1998) is about 2–3 pp, several times smaller than the reported effects, so enlarging the split would narrow already narrow intervals. We release the full corpus, including the 1,035 boundary instances; Tables 3 and 4 give the map, the quotas, and the unscored strata.

Column definitions for Table 1. Columns: PASS = static pass rate on the topology-decision instances (↑); ESC = escalation rate (diagnostic, no preferred direction); NFE@p = NFE@pass, mean evaluations of passing answers as a fraction of the budget (↓); Exh. = NFE exhausted, share of answers that run the budget to zero (↓); Wall = median wall-clock minutes, LLM gateway queueing included (↓); T/O = timeout rate (↓); pass^3 = share passing on all three seeds, the pass^k reliability metric of Yao et al. (2025) at k=3 (↑); Valid = validity rate (↑); Seis. / Wind = held-out seismic / wind time-history re-check (↑); Build = constructability gate (↑); Verif. = Verified PASS, the four-

Table 4: Static PASS% on the three strata of the main split under DeepSeek-V4-Flash (topology-decision column as in Table 1; the other two strata are tabulated from seed 0 for the baselines and from all seeds for PE-Loop). Section-only controls are passed by 93.6–97.9% of answers under every system and boundary instances by none, so only the topology-decision stratum separates systems.

Arm	Topology-decision (130)	Section-only controls (110)	Boundary (20)
Codex	82.05±1.95	93.64	0.00
Openclaw	73.85±2.22	94.55	0.00
DeepSeek Harness	82.82±1.91	95.45	0.00
Claude Code	83.85±1.87	97.27	0.00
PE-Loop (ours)	**91.28**±1.45	**97.88**	0.00

way conjunction of PASS, Seis., Wind, and Build (↑); P/h = passes per hour (↑); Rank = mean, over the twelve ranked columns, of the arm's rank within its block (1 = best; ties are broken by table order; ESC is excluded) (↓). All columns are percentages except Wall (minutes), P/h (count per hour), and Rank.

Systems and run conditions. Five arms are compared on every backbone: PE-Loop; Codex, OpenAI's `codex exec` command-line agent; Openclaw, an open-source agent harness; DeepSeek Harness, the model vendor's own scaffold (version 0.1.0-rc.7); and Claude Code (version 2.1.237), driven through an Anthropic-protocol bridge to the same backbone with two known asymmetries that disadvantage it and that we did not remove: its system prompt names a different model than the one serving it, and it truncates context to 200k tokens although the backbone offers more. The four baselines receive the identical brief, shell access, and the same deterministic margin oracle as a callable tool under the same harness-enforced budget (15 at effort `high`); they decide when to stop and what to submit, and the returned design is read from a file they write. Every episode is capped at 1,800 s, and timeouts count as failures; all arms reach one API endpoint through a shared model-level rate-limiting gateway; sampling parameters are provider defaults, since the command-line agents expose no temperature or seed control, and seeds 0–2 are replicate indices. Three cells deserve a note: under GLM 5.2 Fast, 62 completed DeepSeek Harness episodes accumulated at least 300 s of gateway queueing, inflating that cell's Wall, and seeds 1–2 of Openclaw and Claude Code ran through a relay path rather than the central gateway, with different completion rates; under Qwen3.8-Flash, the PASS of Claude Code differs by 14.62 pp between seeds 0 and 2 (2.4 SE), below the 3-SE threshold of our data gate. The campaign comprised about 15,600 agent episodes between 3 and 10 September 2026; the held-out analyses took about 430 CPU core-hours. The 372-task quarantine set and the 1,197 paired variants are reserved for contamination audits and were not used here.

Reading notes for Table 1. Every arm answers the full set on all three seeds under every backbone (n=390 topology-decision, n=720 all, n=130 for pass^3), and each rate divides by the answers the arm was run on. Under GLM 5.2 Fast, Openclaw, DeepSeek Harness, and Claude Code lost 5, 14, and 12 topology-decision answers to the relay path and gateway queueing; these non-deliveries count as failures in every denominator, which disfavors these baselines and is disclosed rather than imputed. PE-Loop's exhaustion stays within 0.00–6.41% because it escalates after the fixed probing phase rather than running the budget down; the baselines have no budget-aware escalation action, so an answer ending without a feasible submission before the budget runs out counts as ESC, and most run to the budget. Seis., Wind, and Build share the PASS denominator, so a design that is not statically PASS fails them too; Verified PASS is their conjunction with PASS. Because every arm answers the same instances on every seed, Table 1 compares arms on identical instance sets within a block; we report unpaired rates with their standard errors.

Why the constructability gate binds for PE-Loop. The cause is lever order (Appendix D.6): a failing candidate is first repaired inside C2 by enlarging b, and only a failure at the top of the column grid escalates to a topology change in C1, so the delivered column is as thick as the section lever alone required. Since $h_b = 0.8\, b\, \varsigma_{\mathrm{beam}}$ grows with b while l_n shrinks (Appendix D.3), thick-column designs violate $l_n/h_b \geq 4$ first, precisely the failure mode Π_{build} removes.

Table 5: Grounding ablation under DeepSeek-V4-Flash: same loop, same budget, same 60 pre-registered topology-decision instances $\times$ 3 seeds; only the gate implementation varies. The ablation is a separate run from Table 1; on the same 180 answers, the main-table run of PE-Loop passes 81.67% (147/180), within sampling variance of the oracle arm here ($z \approx 0.92$). These instances are harder than the full stratum: they hold 33 of the 34 failures of that run. The judge receives the brief, the code limits, and the design parameters, never a simulator quantity. Verified PASS is the harness's deterministic physics check on the submitted design, identical for every arm; for the oracle arm it coincides with the in-loop gate by construction and serves as the lossless reference. Whether the in-loop constraint set is itself sufficient is a separate question, answered by the held-out columns of Table 1; the judge arms were not re-checked under held-out loading.

Gate implementation	Self-rep. PASS%	Verified PASS% ↑	False-pos.% ↓	Wasted NFE% ↓	Jaccard ↑	Wall s
Margin oracle (ours)	77.78%	**77.78%**	**0.00%**	**0.00%**	**1.00**	**110.8**
LLM-judge (same LLM)	49.44%	20.56%	58.43%	44.41%	0.31	238.4
LLM-judge (DeepSeek-V4-Pro)	40.00%	17.22%	56.94%	45.19%	0.29	342.8

Table 6: Constructability projection ablation (main split, topology-decision instances, DeepSeek-V4-Flash). Raw PASS% drops by 29.74 pp; certified success rises by 9.74 pp. With Π_{build}, every statically compliant design satisfies the detailing clause by construction, so Build% equals PASS%, and the detailing gap is zero.

Arm	PASS%	Concrete m^3	Build%	Detailing gap (pp)	THA seis.%	THA wind%	Verified PASS%
PE-Loop	91.28%	5,610.1	46.41%	+44.87	60.77%	70.00%	35.64%
PE-Loop + Π_{build}	61.54%	5,032.0	61.54%	+0.00	51.79%	53.59%	**45.38%**

Where the certified margin moves. Verified PASS is a conjunction, so marginal gate rates do not add; we compare survival conditional on static compliance. Under DeepSeek-V4-Flash, 66.6% of our compliant designs pass the held-out seismic check and 76.7% the wind check (70.0% and 76.1% for Claude Code), while 50.8% clear the detailing clause (62.1%). The certified lead of 4.61 pp is therefore what survives of a 7.43 pp raw lead after a constructability deficit that Π_{build} removes (Table 6). Figure 7 shows the same construction for every block and system.

Detailing gap by column size. The gap grows with column size and is largest for PE-Loop without Π_{build}, whose outer loop pulls the section lever first (median column side 1.275 m against 0.950–1.200 m for the baselines). Of its 126 violations, 123 (97.6%) are repairable by clipping the beam-depth scale alone; the other 3 (2.4%) are topology-level cases with $b > \mathrm{bay}/2.6$. Conditioned on static compliance (Build% over PASS% in Table 1, DeepSeek-V4-Flash), the gate is harshest on us: 50.84% of our compliant designs clear the clause, against 61.88–65.33% for the four baselines; with Π_{build} in the loop the rate is 100% by construction (Table 6).

Scope of the projection ablation. Three limits apply to Table 6: this is the main split, not the 59-instance reference set; one LLM only (DeepSeek-V4-Flash); and it is scored on the 130 topology-decision instances, while the unprojected arm answers the full 260, so it is not a full-coverage control.

Caliber statements. On the search-sufficient reference set, PE-Loop returns 175 answers (174 feasible) and Openclaw, DeepSeek Harness, and Claude Code 176 each, all on 59 instances; Codex returns 149 answers on 51 instances (131 feasible). Cross-system claims use the 51-instance intersection, and paired tests against search use each arm's own coverage (Table 7). Checks whose code limit could not be sourced, or whose quantity the agent never reported, are UNEVALUATED and leave the denominator rather than counting as failures. Two criteria carry no discriminative signal and are not discussed in Section 5: intent-consistency saturates at 98.66–100.00% (range 1.34 pp), and the beam-width rule binds only for $\varsigma_{\mathrm{beam}} > 1.875$ with taper $> 1 - 0.32\,\varsigma_{\mathrm{beam}}$, which no returned design reaches.

Table 7: Coverage on the search-sufficient reference set (59 instances $\times$ 3 seeds; 177 answers maximum). The last column is the count and within-arm share of feasible answers that violate the clear-span-to-depth clause (Fig. 15).

Arm	Instances	Answers	Feasible	Violating $l_n/h_b \geq 4$
Codex	51	149	131	56 (42.75%)
Openclaw	59	176	170	85 (50.00%)
DeepSeek Harness	59	176	169	86 (50.89%)
Claude Code	59	176	171	76 (44.44%)
PE-Loop w/o Π_{build}	59	175	174	126 (72.41%)

E.1 The search-sufficient reference set: 59 instances

Paired significance and failure modes. Against LLM-free exhaustive search (exact McNemar test on feasibility with Holm correction (Holm, 1979), N=59), only Codex differs significantly: 18 instances that search solves and Codex does not, and none the other way ($p_{\text{Holm}} = 1.14 \times 10^{-4}$). Openclaw, PE-Loop, DeepSeek Harness, and Claude Code have 2, 0, 0, and 1 discordant instances, respectively ($p_{\text{Holm}} = 1$); $p = 1$ with no discordant pair is agreement instance by instance, not evidence of equivalence, because the paired test's denominator is the discordant set. Per criterion, with UNEVALUATED checks excluded from the denominator, the failures of Codex concentrate on story drift (12.08%, 18/149), stability (10.74%, 16/149), and the base-shear floor (8.05%, 12/149), with the story-stiffness ratio at 1.34%; every other system fails only the story-stiffness ratio (1.69–1.74%, and never Openclaw) and, at most once each, the period ratio; torsion and wind acceleration are never failed by any system.

E.2 Effort tiers and the anytime contract

All numbers here are seed-0 values on the 130 topology-decision instances (Table 1 pools three seeds, hence its different Wall of 1.6 min). Scoring the four tiers of Appendix D.8 on these identical instances removes the set-composition confound; Fig. 6d plots PASS against B with NFE@pass under each point. Wall-clock is not monotone across tiers (175 / 183 / 104 / 258 s, seed-0 median) because it includes LLM gateway queueing, and neither is success: `ultra` (B=25) returns 81.54% against 91.54% (119 of 130) at `high` (B=15), because the extra rounds let C1 revise the topology of instances the section lever had already settled, not always for the better.

E.3 Denominator disclosure

Every rate divides by the answers each arm was run on: 720 per arm and backbone ($260 - 20$ instances $\times$ 3 seeds) and 180 per arm of the grounding ablation. The 31 GLM-block non-deliveries count as failures and stay in the denominator; otherwise the effective denominator equals the answer count in every row. Agent-side UNEVALUATED answers (at most eleven, for Codex; four for Openclaw, one each for DeepSeek Harness and Claude Code, and three for the stronger judge) also stay in the denominator, as do all timeouts. Rate-limit events were retried rather than dropped, so they do not reduce the denominator, but they are uneven: 0 for PE-Loop, 34 for Openclaw, 90 for Codex, 134 for Claude Code, and 297 for DeepSeek Harness; three of the four baselines thus spent a substantial share of their runs waiting on a gateway, which the Wall column of Table 1 includes.

E.4 Threats to validity of these numbers

Three caveats apply; none can be resolved with more seeds. (i) For most instances, the plan footprint behind the column grid is back-derived from floor area at a fixed aspect ratio of 1.3 (`geom_source` = `from_area`), biasing the bay width and, with it, the detailing margin ρ in Fig. 15 for multi-block massings. (ii) The main split and the reference set differ in instances, seeds, and run directories, so their numbers are never comparable. (iii) In every block, the cross-arm PASS% range exceeds the pre-registered 15 pp separation (narrowest: 17.43 pp, DeepSeek-V4-Flash and GLM 5.2 Fast), but our strongest evidence is the four-LLM grid, held-out columns, and paired tests.