

BuildArena: A Physics-Aligned Interactive Benchmark of LLMs for Engineering Construction

Tian Xia^{1,2}, Tianrun Gao¹, Wenhao Deng¹, Long Wei¹, Xiaowei Qian¹, Chenglei Yu¹, Tailin Wu^{1,2†}



ICML
International Conference
On Machine Learning

1 Motivation: From Language To Construction

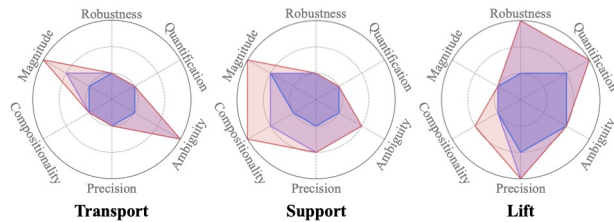


- **Text:** Documents, Literatures, Legals
- **Code:** Sufficient Data, Executable, Verifiable
- **Construction:** Spatial, Connectivity, Physics
- LLMs' construction ability **remain largely unevaluated**.

3 Construction Results: Diverse Solutions

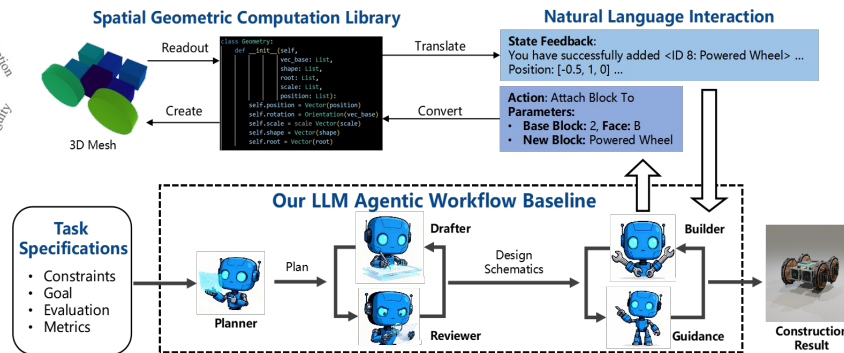


2 Benchmark Tasks: Goal-Driven Construction



- **Transport:** horizontal movement
- **Support:** structural physics, long-horizon
- **Lift:** vertical movement, precision
- **Goal-Driven:** zero-shot prompt

4 Key Feature: Step-wise Verification & Feedback



5 Main Results: Much to Improve

Model	Transport	Support	Lift
GPT-5	42.7%	52.1%	41.1%
Grok-4	11.5%	20.8%	21.8%
Claude-4	12.5%	3.1%	4.2%
Seed-1.6	6.2%	19.3%	2.1%
GPT-4o	6.3%	13.5%	3.6%
Kimi-K2	4.7%	11.5%	3.1%
Qwen-3	5.7%	5.7%	1.0%
DeepSeek-3.1	2.6%	8.3%	3.6%
Gemini-2.0	1.6%	7.8%	0.0%

5 More: Towards Engineering Automation



BuildArena
Paper Link



BuildArena
Homepage



BuildArena 2.0
Construction Challenge