



Benchmarking LLM Agents for Simulation-Grounded Process Engineering Reasoning

Jingkang Liang, Gürkan Sin

Process and System Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU), 2800, Kgs. Lyngby, Denmark



Kgs. Lyngby, Denmark,
August 17-18, 2026

Can LLM agents reliably inspect process simulation gather the right evidence, and produce defensible HAZOP-relevant answer?

1. PROBLEM & MOTIVATION

LLMs can interact with process simulation through MCP[1].

Can an LLM reason from a process simulation without overstating what the available evidence supports?

Correctness

Does the answer match the engineering state and process topology?

Evidence

Does the agent retrieve the right variables, metadata, endpoints, and paths?

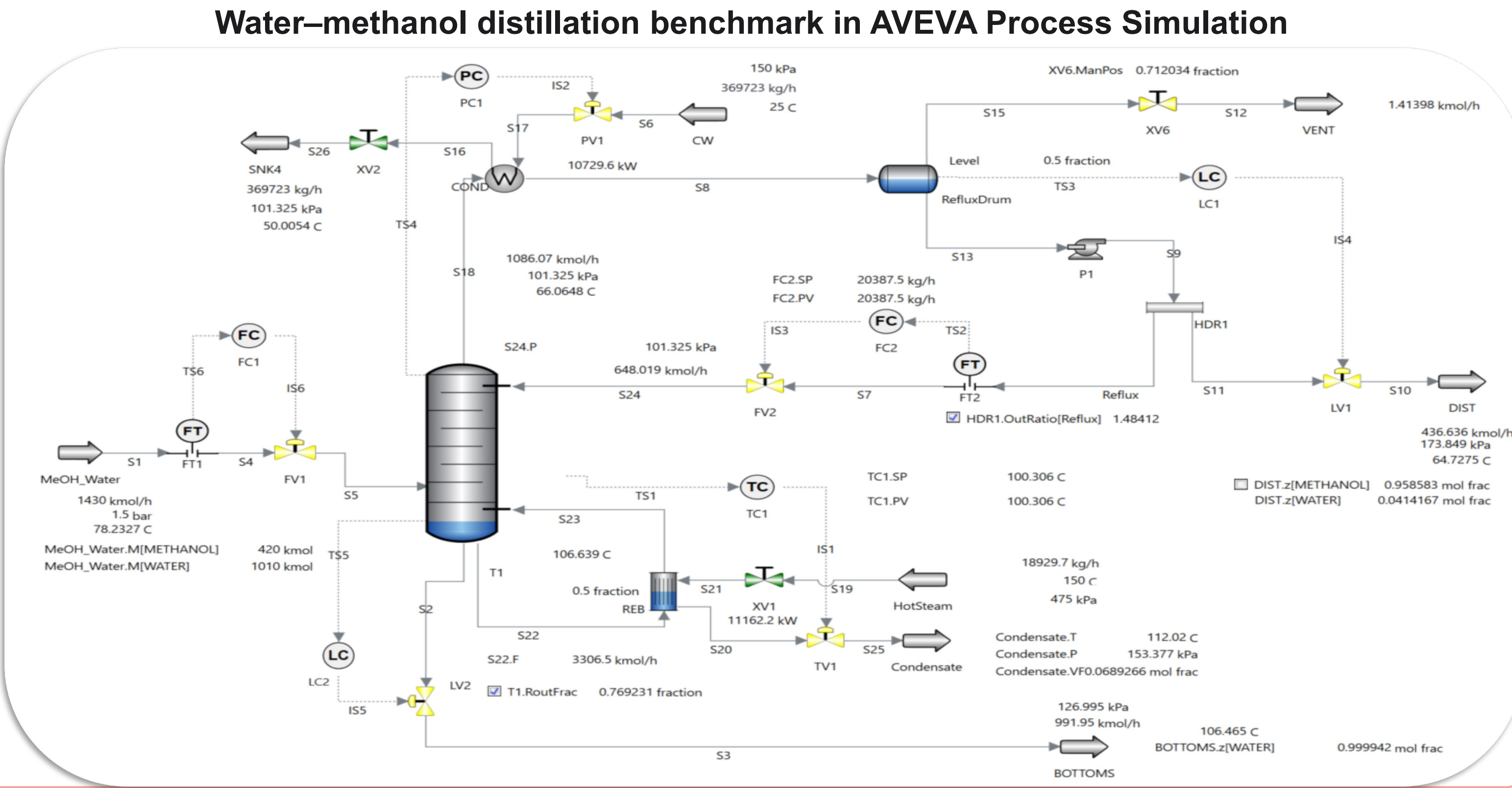
Boundaries

Does it stop when alarms, relief activation, or dynamics are not represented?

2. SIMULATION ENVIRONMENT

FIXED ENVIRONMENT

- Solved steady-state flowsheet
- Feed, column, condenser, reflux drum, reboiler, controllers and products
- Exact model variables and endpoints available through read-only MCP tools
- 13 fixed engineering / HAZOP questions per session



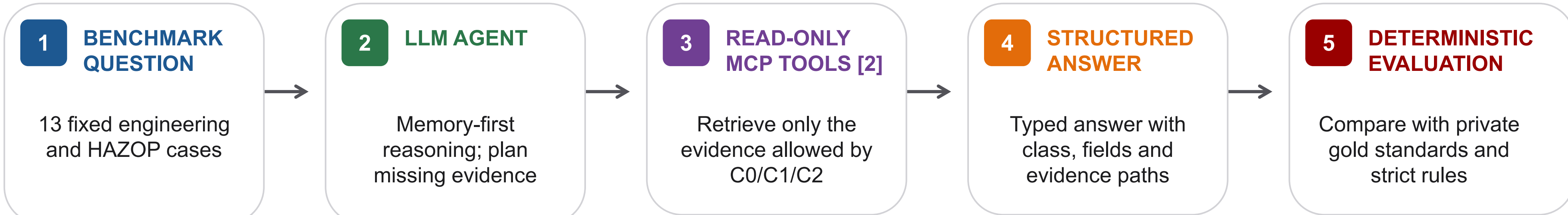
3. BENCHMARK DESIGN

INFORMATION CONDITIONS

C0	C1	C2
SCREENSHOT ONLY	MCP ONLY	SCREENSHOT + MCP
Visible topology and labels No internal variables or metadata	8 read-only AVEVA tools Models, connectors, variables, parameters, metadata...	Same screenshot plus the same read-only MCP tool set

Increasing information availability

Cumulative memory across questions, tools, responses and answers



PUBLIC TO LLM

Questions • condition-specific tool descriptions • structured-answer schema

RECORDED PER RUN

Tool calls and arguments • responses • latency • tokens • cost • failures

PRIVATE FROM LLM

Gold answers • evaluation rules • alternatives • forbidden claims

4. EVALUATION RESULTS

Luna ranks first in every condition; C1 is best for Luna and MiMo, while C2 does not improve on C1.

45 cumulative sessions

585 evaluated turns

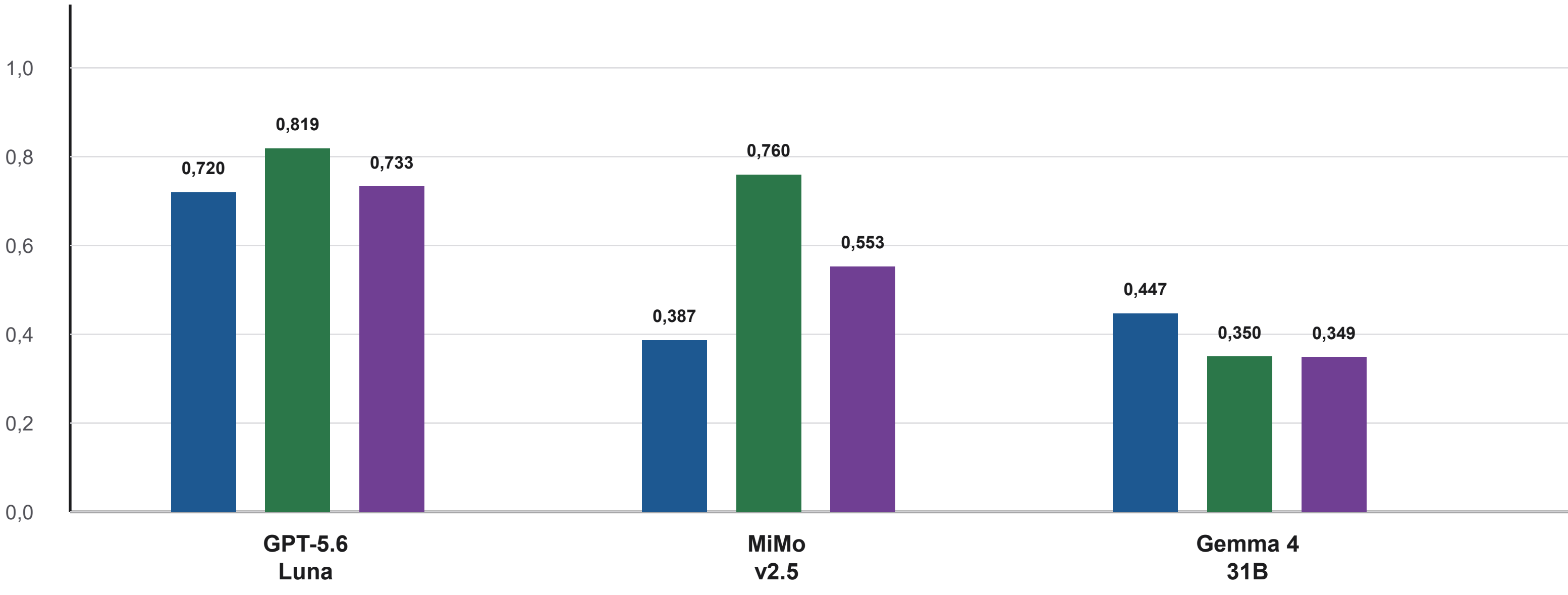
9,879,028 tokens

USD 0.469754 total cost

345 MCP calls

Overall benchmark score by condition

C0 C1 C2



Mean ± SD: Luna 0.720±0.052 / 0.819±0.294 / 0.733±0.284 • MiMo 0.387±0.227 / 0.760±0.274 / 0.553±0.256 • Gemma 0.447±0.044 / 0.350±0.069 / 0.349±0.050

C1/C2 tool-enabled resource use and dominant MCP behaviors

Model	C1 calls	C1 dominant tools	C2 calls	C2 dominant tools	Tokens/run C1	Tokens/run C2	Cost/run C1/C2	Overall
GPT-5.6 Luna [3]	72	metadata 21; connectors 17	71	connectors 20; all vars 18	345,495	358,164	USD 0.00915 / 0.00903	0.757
MiMo v2.5 [4]	43	all vars 12; metadata 11	65	metadata 16; all vars 14	252,887	400,616	USD 0.00584 / 0.00946	0.567
Gemma 4 31B [5]	62	connectors 36	32	connectors 22	267,632	141,683	USD 0.02919 / 0.01442	0.382

Output reliability — all 195 turns per model

Typed Direct Missing



- Luna's missing turns reflect 3 recursion events followed by 23 cascading API errors.
- MiMo's primary failure is schema/format noncompliance (53 direct replies).
- Gemma remained schema-compliant when it answered, but exact field and evidence accuracy was low; provider timeouts and recursion caused another 24 missing turns;

CONCLUSIONS

- Luna ranks first in C0, C1 and C2.
- C1 is best for Luna and MiMo.
- C2 does not improve on C1: extra multimodal context adds complexity after MCP evidence is available.
- Reliability changes the ranking story: MiMo fails mainly on format compliance; Gemma on timeouts and recursion.

TAKE-HOME



Simulation tools can improve grounded reasoning, but only when agents retrieve selectively, preserve structured outputs, and keep cumulative sessions healthy.

References

- Liang, J., Groll, N., & Sin, G. (2026). Large language model agent for user-friendly chemical process simulations. Digital Chemical Engineering, 19, 100312. <https://doi.org/10.1016/j.dche.2026.100312>
- What is the Model Context Protocol (MCP)? Model Context Protoc n.d. <https://modelcontextprotocol.io/docs/getting-started/intro> (December 3, 2025).
- OpenAI. GPT-5.6 Luna. OpenAI <https://developers.openai.com/api/docs/models/gpt-5.6-luna>
- Xiaomi. MiMo-V2.5. Xiaomi MiMo n.d. <https://mimo.mi.com/models/mimo-v2.5> (
- Google. Gemma 4 31B. Hugging Face n.d. <https://huggingface.co/google/gemma-4-31b>

ProSafe

Funded by
the European Union

HORIZON-MSCA-
2022-DN-01 call,
project no.
101119358