Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00392-5

Original Article

Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task

Roshan Aryal [1, ✉]

1. *Artificial Intelligence and Data Science, Independent Researcher, Auckland, NZL*

Received: September 02, 2026 | Review began: September 04, 2026 | Review ended: November 24, 2026 | Published: November 25, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background: Closed benchmarks with a hidden answer key dominate large language model (LLM) evaluation, yet many real tasks have no single correct answer. We treat one such task, designing a 24/7 autonomous AI development-and-research workstation for a 32 GB Apple M6 Mac mini, as a case study in evaluation: an open-ended design with no ground-truth solution but many individually falsifiable sub-claims, such as whether a model fits in 32 GB, whether a named tool exists, and whether a response contradicts itself.

Methods: We gave one identical, evidence-demanding prompt to eleven frontier LLM systems and collected thirteen verbatim responses, extracted a 39-axis decision matrix, scored every response on a nine-dimension 0/1/2 rubric with four raters scoring independently (one author-derived baseline and three frontier LLMs on an identical packet), web-verified every factual sub-claim against sources fixed to the capture window, and checked memory feasibility with a standard-library estimator. Sensitivity analyses remove the author-derived rater, remove weakly reliable dimensions, and vary the memory assumptions; agreement statistics carry bootstrap intervals.

Results: Four results stand out in this single-task case study. First, the systems agree far more on the shape of the machine (one large model resident at a time, a coordinator/worker topology, durable SQLite state, private-network remote access) than on the products that realise each choice: median modal agreement 10 versus 5 of 10 systems. Agreement is not evidence that the shared choices are correct. Second, the study-internal rater, itself an LLM, flagged 19 tools and models as hallucinated; on web verification none was confirmed nonexistent: all 19 are real projects, many released or renamed after that rater's training cutoff. The rater's error was in its existence verdicts: independent verification did find genuine factual errors, including wrong attributes for two flagged names. Third, the response with the largest reference apparatus (99 URLs) mostly cited pages that support its claims, but only 13 of the 86 references traceable to the capture day were primary sources and 20 were forks or copies of the projects they describe; reference count says little about source quality. Fourth, in the one system captured in three chat modes, the modes differed on six of the nine rubric dimensions with the prompt held fixed. Rubric totals are rater-dependent: only a few responses keep their third of the ranking when the rater set changes.

Conclusions: On this currency-sensitive task, an LLM asked to grade another LLM's factual grounding penalised the newest, best-grounded answers, because the grader's own knowledge was stale. The study shows that this failure occurs; it does not estimate how often it occurs across tasks. Evaluation pipelines that place an LLM in the grader's seat for currency-sensitive questions should put an external verification step in front of it. We release the corpus, the rubric, four rater score sets, a web-verified register, and standard-library analysis scripts.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Categories: Evaluation Methods, Machine Learning (ML), Natural Language Processing (NLP)

Keywords: large language model evaluation, language model as judge, training cutoff bias, reference hallucination, inter-rater agreement, prompt sensitivity, systems architecture, reproducible benchmark corpus

Introduction

Closed benchmarks dominate LLM evaluation: a task with a hidden answer key, a scalar score, and a leaderboard. Systems architecture is not that. Given a hardware envelope and a requirements list there is no single correct design, but there are many claims inside any proposed design that are individually checkable: does this model fit in 32 GB, does this tool exist, does this version number resolve, does the response contradict itself.

This paper treats "design a hard system under a fixed consumer-hardware constraint" as an LLM task and measures it. The task is deliberately stress-inducing along several dimensions. The hardware is new: the Apple M6 Mac mini was announced on 25 August 2026, six days before the study. The memory ceiling of 32 GB unified memory is tight enough that a single wrong model choice breaks the design. The prompt explicitly asks for sources and for a list of things not to build.

The contributions are a reusable protocol for evaluating open-ended architecture responses (verbatim capture, structured extraction into a 39-axis decision matrix, a nine-dimension 0/1/2 rubric, multi-rater scoring with reported agreement, and automatable falsifiable-sub-claim checks); an empirical map of where eleven systems agree and disagree; a demonstration that, on this task, apparent fabrication was mostly a failure of the LLM rater; constraint-reasoning and recency results, including a memory analysis that reports how much non-model memory each design leaves; a citation-quality analysis and a reasoning-mode sensitivity result; sensitivity analyses for the raters, the rubric dimensions and the memory assumptions; and a synthesised reference architecture together with a frank limitations section that includes a methodological error we made and its fix. The study is a single-task case study: with thirteen captures on one prompt and one hardware specification it shows patterns and failure modes on this instrument, not population-level laws, and we are explicit throughout about which claims are corpus observations and which are candidates for generalisation.

Agentic benchmarks such as SWE-bench and its variants [1], WebArena [2], and others including Terminal-Bench, OSWorld, Tau-Bench and AgentBench all score against an execution oracle: the task has a hidden pass or fail, and the scope has widened from function-level generation to environment-grounded interaction, but the answer key is always present. A survey of the agent-systems landscape [3] and benchmarks of frontier LLMs on real-world assistant tasks [4] describe the same oracle-based setting. Constrained design has no such oracle. We take the stance of scoring only falsifiable sub-claims, and measuring consensus rather than correctness, from systematic-review methodology: PRISMA-S for reporting a reproducible search [5], PRESS for peer-reviewing a search strategy [6], and the National Academies framing of reproducibility and replicability [7].

Using an LLM to score another LLM's output is now standard and its failure modes are catalogued. A survey of LLM-as-a-judge practice [8] documents position bias, verbosity bias, and self-enhancement; a large-scale study across twenty NLP evaluation tasks [9] finds that LLM judges vary substantially by task and must be validated against human annotations before use; and a controlled analysis [10] shows that a judge's ranking can be reversed simply by changing response order, and proposes position calibration as a fix. The standard fixes are randomised ordering, repeated runs, and rubric-based scoring that constrains the judge to factual criteria, which is exactly the instrument we use. Our case study adds an example of a bias not on those lists: a training-cutoff recency bias in which, when the judged content is more current than the judge's own knowledge, the judge can emit false-positive fabrication flags.

The hallucination literature supplies our fabrication vocabulary. A survey of hallucination in large language models [11] separates factual contradiction from factual fabrication, meaning generated content whose facts cannot be verified against any source; that fabrication category is the one our verification register applies to unresolvable tool and model names. Empirically, a controlled study of ChatGPT-generated bibliographies [12] found 55 percent of GPT-3.5 citations and 18 percent of GPT-4 citations fabricated outright, with 24 to 43 percent of the non-fabricated citations still carrying

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

substantive errors; a comparative analysis for systematic reviews [13] reported reference-hallucination rates between 28 and 91 percent and reference precision below 14 percent for every model tested. Recently published content is a recognised knowledge-boundary cause of this behaviour. Our contribution moves that same cutoff mechanism from the author to the judge: it is not only that a model with an old cutoff invents post-cutoff sources, it is that a rater with an old cutoff can flag real post-cutoff artefacts as invented.

Line-level and repository-level code generation is heavily studied, with SWE-agent, AutoCodeRover, Agentless, and OpenHands exploring different issue-resolution agent designs. Work on LLMs at the architecture level is newer and smaller. Dhar and colleagues asked LLMs to write the decision section of architecture decision records from the recorded context and found that GPT-4, without examples, produced relevant and accurate decisions that still fell short of human architects [14]. A multivocal review of generative AI for software architecture found architectural decision support and architecture reconstruction to be the most common uses, and reported that rigorous testing of the generated outputs was usually missing and that sound evaluation frameworks are lacking [15]. Both lines of work judge a design decision against a human reference or a human reader. Whole-system design under a fixed resource envelope, where no reference design exists, is less studied; the nearest analogue is human-expert capacity planning. We treat "produce a deployable architecture from a hardware spec plus a requirements list" as an evaluable LLM task and evaluate it through its falsifiable parts.

Materials And Methods

System	Slug	Base model (free tier)	Notes
Anthropic	claude-sonnet-5	Claude Sonnet 5 (in an AI coding tool, browsing on)	anchor, not blind; excluded from consensus counts and rankings
Mistral	mistral-large-3	Mistral Large 3 (675B/41B)	—
OpenAI	gpt-5	GPT-5.6 Luna	—
Perplexity	perplexity	Sonar (auto)	—
Moonshot	kimi-instant	Kimi K3-class, Instant mode	—
Google	gemini-3.1-pro	Gemini 3.x Pro	—
Alibaba	qwen-3.7-plus	Qwen ~3.7	—
Meta	meta-llama-4	Llama 4 (variant undisclosed)	—
DeepSeek	deepseek-expert	DeepSeek, version not shown (presumed V4-Pro), deep-reasoning mode	canonical DeepSeek answer
DeepSeek	deepseek-instant	DeepSeek, version not shown (presumed V4-Pro), fast mode	mode variant (see Table 6)

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

DeepSeek	deepseek-instant-deepthink	DeepSeek, version not disclosed (presumed V4-Pro), Instant+DeepThink	mode variant (see Table 6)
xAI	grok-4	Grok 4	—
Zhipu	z-ai	GLM-4.6/4.7-class (undisclosed)	—

TABLE 1: The response corpus: 13 captures from 11 systems

We follow published guidance on reporting empirical software-engineering studies that involve LLMs [16]: this section declares the LLM usage and role, reports model, version, interface and browsing state per capture, releases the prompts verbatim, records session provenance in each capture's metadata, and includes an open-weight system in the corpus.

The task instrument

The frozen prompt has a responder-context header that fixes the date at 31 August 2026, asks the model to declare its browsing state and version, and requests a Sources list. The task body asks for a 24/7 autonomous development-and-research workstation on a 32 GB Apple M6 Mac mini (12-core CPU, dual 16-core Neural Engine, 170 GB/s memory bandwidth, 512 GB internal plus 1 TB external SSD), coordinating "100+ logical agents without running 100 LLMs", remotely controllable, and securely sandboxed. It asks for deliverables A to K including an exact install plan and a list of what not to install. Two controlled variants are also frozen: one is a full paraphrase using request-for-comments framing with reshuffled sections and no shared sentences, and one is the base prompt with its anti-anchoring steer removed. Both were re-run on a five-system subset; those results appear under prompt sensitivity. Every other matrix and ranking in this paper is for the base-prompt captures.

The response corpus

Thirteen captures were collected on 31 August 2026 from eleven systems (Table *1*). One system, DeepSeek, contributes three captures in three chat modes of one product, presumed but not confirmed to share a base model. Consensus analysis counts one canonical column per system, giving 10 non-anchor systems; the fabrication and internal-consistency analyses use all 13 captures, because the three DeepSeek modes are a within-product comparison no other vendor provided.

The anchor capture, from Anthropic's Claude Sonnet 5, ran inside the same tool used to build this study's repository, with live browsing and partial authorship of the requested output format. It is retained as the consensus anchor and the only sourced baseline. It is excluded from the consensus counts, the response rankings and the rank-based thirds; it is included, under its own name, in the per-capture fabrication and consistency checks and in the rater-agreement statistics, which compare raters on the same thirteen responses rather than compare systems. The unit of analysis is the system capture, meaning a product interface plus its routing, model, system prompt, browsing state and mode, not an isolated model.

Extraction and rubric scoring

For each capture we retain a verbatim raw response, which is never edited after merge except for a dated correction note, the model's own cited-sources list, and structured reviewer notes, plus one short phrase per decision axis in a 39-axis matrix. Scoring uses a nine-dimension 0/1/2 rubric: hardware-constraint adherence (D1), recency (D2), tool factuality (D3), model factuality (D4), benchmark factuality (D5), citation quality (D6), actionability (D7), security model (D8), and internal consistency (D9). Axis coding used predefined inclusion criteria; for example, "Apple-Silicon-native inference" counts MLX-family runtimes and llama.cpp with Metal, not CPU-only inference. The 39 axes were defined by the author, which introduces researcher degrees of freedom; the coding guide and the per-response phrases are released so the

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

coding can be audited. The rubric mixes two layers, and we do not claim to establish an objective ground truth for overall architectural quality: some dimensions are objective, such as model existence, version existence, memory feasibility under stated assumptions, and self-contradiction, while others are rubric-based judgements of design quality, such as actionability and security adequacy.

The two instruments were written before data collection finished: both are in the repository's first commit on 31 August 2026, which also holds the first four of the thirteen captures. Each axis is a decision every response has to make to answer the prompt, such as which inference engine to run, how many models to keep resident, and which task queue to use. Each rubric dimension targets one of the study's questions (fabrication, constraint reasoning, recency, citation quality, internal consistency) or a deliverable the prompt asks for (an install plan, a security model), and each of the three levels has a written anchor. The instruments were revised twice, and both revisions are versioned. On 1 September 2026, after web verification, the zero level of D3 and D4 was reworded to require a web-verified nonexistent entity. After the agreement analysis, sharper anchors were written for D1, D7, D8 and D9; they are released as a second version and were not applied to any score in this paper. No expert-panel review, pilot study or psychometric validation of either instrument was carried out. The evidence we can offer is limited to three kinds: the factual dimensions D1, D3 and D4 are tied to external checks (the memory estimator and the verification register) rather than to rater judgement alone; agreement is reported per dimension with uncertainty; and the coding guide and every per-response phrase are released for audit. We therefore use the rubric as a structured reading aid for this case study, not as a validated measure of architectural quality, and we rest no conclusion on a fine-grained ranking.

Rater procedure and inter-rater agreement

The nine-dimension rubric was applied independently to all thirteen captures by four raters: one author-derived baseline and three frontier LLMs (GPT-5.6 Sol, Grok 4, and a DeepSeek chat run) given the identical scoring packet. "Independent" here means each scored the same packet without seeing another rater's scores; it does not mean design-independent, because the author-derived baseline also wrote the rubric and the reviewer notes. We therefore treat that rater as an author-derived baseline rather than a neutral fourth opinion, and use GPT-5.6 Sol as the canonical comparison rater. No independent human expert rated the responses; the Discussion sets out which results this affects. Agreement is reported as exact and within-one-point agreement, Cohen's kappa unweighted and quadratic-weighted [17], Gwet's AC1 unweighted and quadratic-weighted [18], Krippendorff's ordinal alpha [19], and, for the per-response totals, Spearman's rho and Kendall's tau-b between every rater pair. The estimator is a released standard-library script.

The order of work is recorded in the repository history. The rubric and axes were committed on 31 August 2026. The web-verification register that overturned the initial fabrication flags was committed at 09:55 on 1 September. The author-derived baseline scores, which were derived from each capture's reviewer notes after those notes had been corrected against the register, were committed together with the three LLM raters' scores at 12:05, and the first manuscript draft at 12:14. Scoring therefore preceded the writing of conclusions, but the author-derived baseline was not blind to the verification results. For that reason every ranking result is also reported without it (Sensitivity analyses). The 37 disagreements in the canonical pair were adjudicated by a fixed, documented rule: the rubric text decides, and for the factual dimensions the verification register decides; each decision and its reason is listed in the released adjudication report. Because adjudication decides cell by cell, an adjudicated total can fall below every rater's total. This happens for three DeepSeek captures, where adjudication applied recommend-then-forbid contradictions and unresolved model names that the raters had scored more leniently (Figure *1*).

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

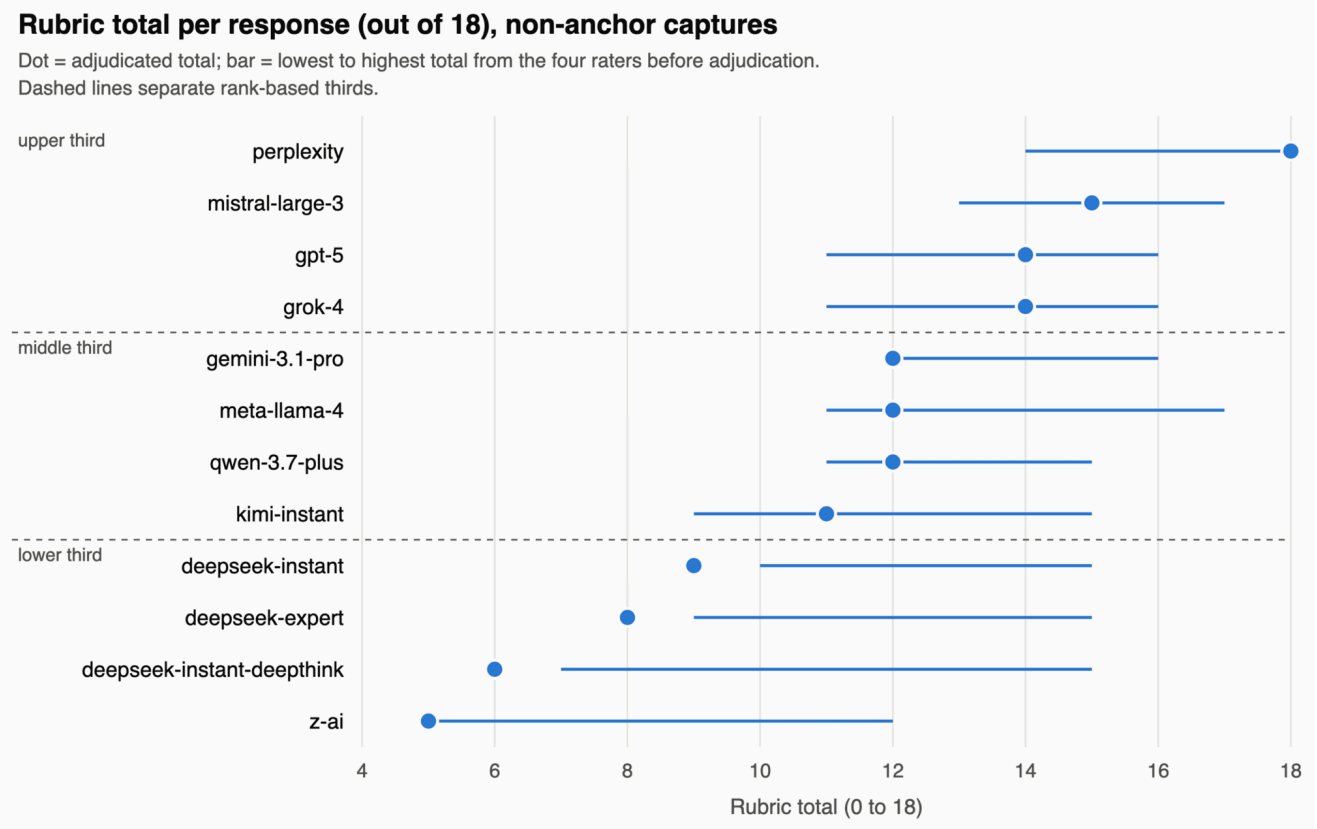

FIGURE 1: Rubric total per response

Dot: adjudicated total out of 18. Bar: lowest to highest total from the four raters before adjudication. Dashed lines separate the rank-based thirds. Non-anchor captures only.

Two scoring passes exist and are not interchangeable. In the full four-rater pass an intermediate version of the packet listed twelve real post-cutoff tools in one rule while instructing raters to treat them as real; all three second-rater runs saw it. That pass supplies the agreement statistics and, with this caveat, the rater ranges and the sensitivity analyses; it does not supply the canonical D3 and D4 levels. A clean re-run of dimensions D3 and D4 used an uncontaminated packet that names nothing, forbids reading the verification register, mandates web verification, and scores an unresolved entity as 1. That clean pass is canonical for the D3 and D4 levels. One of the two fresh raters used for it, Perplexity, was discarded because it scored tool factuality 0 for seven responses by treating its own search misses as evidence of non-existence.

The discarded Perplexity run participated only in the clean D3 and D4 re-run; it was never part of the four-rater pass and enters no agreement statistic. Every inter-rater figure reported below is computed over the author-derived baseline, GPT-5.6 Sol, Grok 4, and the DeepSeek chat run, and is therefore unaffected by that exclusion. The canonical D3 and D4 levels used for the adjudicated totals come from the surviving clean-packet rater, GPT-5.6 Sol, except where the verification register overrides a cell; each such override is recorded with its reason in the adjudicated score file. The released scripts regenerate every figure from the scored packets, so the effect of including or excluding any single rater can be checked directly.

Fact verification

Every falsifiable sub-claim is recorded verbatim and checked. Tool and version existence is checked against a registry or repository fixed to the capture date, and the per-entity verdicts (entity exists, identity correct, attribute correct, claim supported, and a final category) are released as a register. URLs in a response's citation apparatus are classified on three axes: whether the URL resolves, whether it supports the claim it is attached to (fully, partly, or not), and whether the source is primary, secondary, tertiary, or irrelevant. "Non-supporting or unresolvable" is the union of does-not-resolve and

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

does-not-support. The operational decision rules for each axis, with examples, are released with the data. A URL resolves only if it returns the document the reference names, judged as of the capture window from an archived copy where the live page has changed. A reference supports a claim fully only if the page states the claim, including any number, version or attribute the claim gives. A commit, pull request or skill file cited for a performance or capability claim does not support it unless the file itself contains the evidence. The Meta capture's 99 references were coded against these rules from evidence of their state on the capture day: a Wayback Machine capture on or before 31 August 2026 where one exists (23 references) and, for GitHub references, the object's own dated record from the GitHub API (a commit, an issue or pull request, or a README or file at the last commit before the cutoff). A reference that is gone today and has no archived copy is recorded as unknown. Every verdict, its reason and its evidence are released.

Memory-budget estimation

A standard-library estimator adds the weights of a response's co-resident model set, meaning what the response says stays loaded, and a key-value cache estimate, and compares the sum plus a non-model reserve with 32 GB. Weight sizes are taken from the response's own figures where it gives them and otherwise from published 4-bit artefact sizes; each preset in the released script documents its source. Where a response states a size that the verification register contradicts, the arithmetic uses the stated size, because it tests the response's own budget, and the text reports the verified size separately. The key-value cache is the context length the response states multiplied by a per-1,000-token rate set by the largest resident model. The estimator's default reserve is macOS and core services at 7 GB, the Python supervisor and workers at 1.5 GB, SQLite with a vector index at 0.5 GB, and a headless browser at 2.5 GB, 11.5 GB in total; a design fits with headroom when at least 2 GB remains free.

Every one of these figures is an assumption, so instead of a single verdict we report, for each design, the break-even reserve: 32 GB minus weights minus key-value cache, which is the largest non-model allowance under which the design still fits. We compute it with the key-value rate halved, as set, and doubled, and compare it with the 11.5 GB default and with a 7 GB bound for a lean headless host (Figure *2*). Three presets encode an alternative reading of a response rather than its stated design (for example, both of a response's large models loaded when the response itself says one at a time); they are marked and are not used for any verdict on a response. A design that swaps models gets one preset per phase, including any helper model the response shows running alongside the heavy one.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

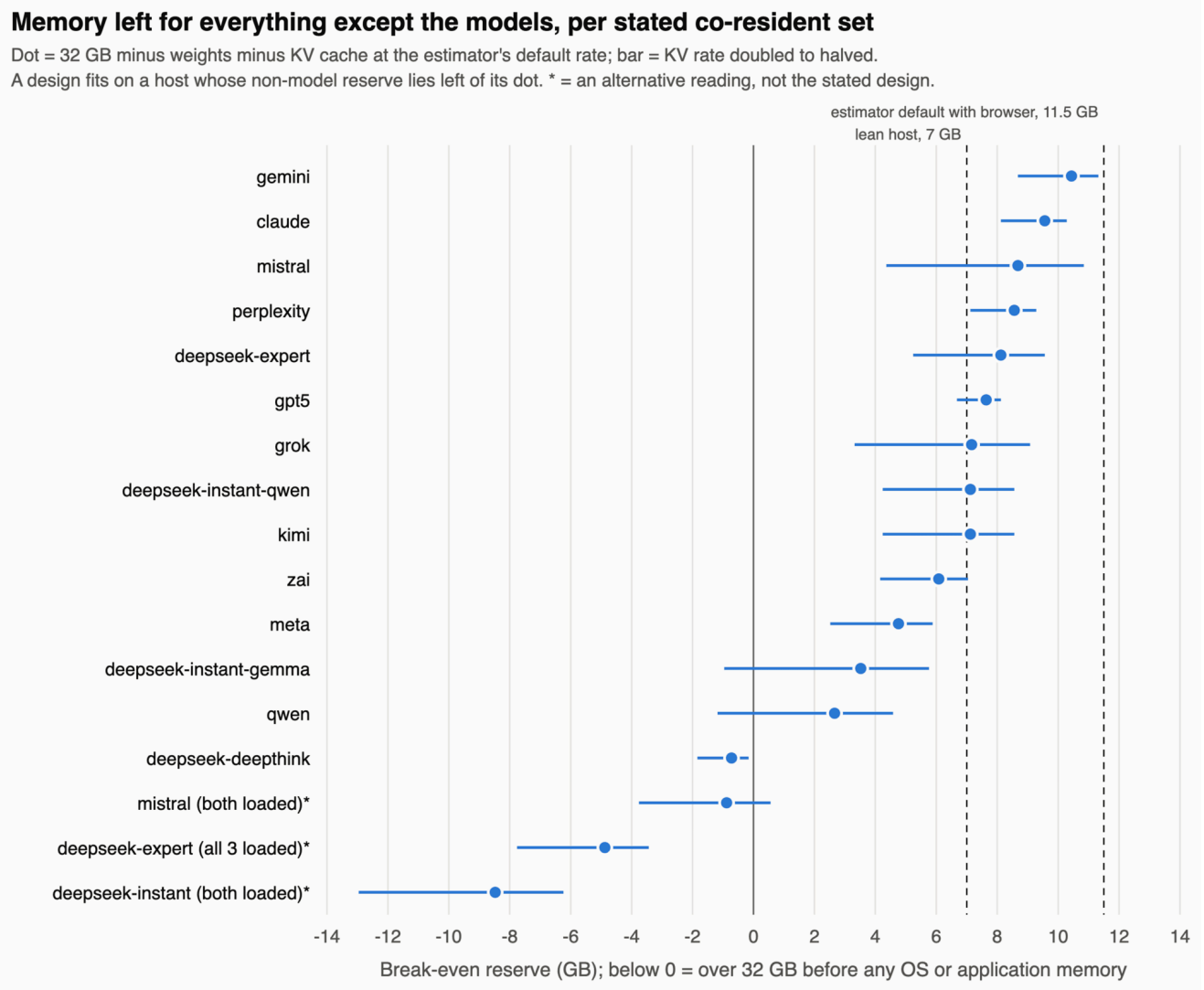

FIGURE 2: Break-even non-model reserve per design

32 GB minus the stated co-resident weights minus the key-value cache. Dot: the estimator's default key-value rate; bar: rate doubled to halved. A design fits on a host whose non-model reserve is left of its dot. Dashed lines: 7 GB lean headless host; 11.5 GB estimator default with a browser. Asterisk: an alternative reading of a response, not its stated design. DeepSeek Instant's stated design swaps one large model at a time, with a 9B summarizer running alongside, and appears as two phases (qwen, gemma). Sizes are as each response states them; the Zhipu (zai) sizes are known to be wrong.

Sensitivity and uncertainty analyses

All analyses in this subsection are in one released, seeded, standard-library script. Agreement statistics, rank correlations and rater means carry 95 percent percentile bootstrap intervals from 10,000 resamples of responses, the unit of analysis; resamples in which kappa is undefined because a rater gave one score throughout are skipped and counted. Counts out of ten systems carry Wilson 95 percent intervals. The structural-versus-product consensus gap is summarised as a difference in medians, with a bootstrap interval that resamples axes within each group, and tested as a difference in means with a one-sided permutation test (20,000 permutations of the axis labels). The test treats the axis labels as exchangeable, which is only approximately true: the axes are not independent, and structural answers were coded more broadly than product names.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Three sensitivity analyses address the rater and rubric choices. To remove the author-derived rater, response totals are recomputed from the canonical LLM rater alone (with its clean D3 and D4 scores), from the mean of the three LLM raters, from that mean with every rater's D3 and D4 taken from the clean pass (Grok 4 and the DeepSeek run scored those dimensions only in the contaminated pass), and from each LLM rater alone. To test the weakly reliable dimensions, totals are recomputed with each dimension left out in turn, with the two weakest (D1 and D9) dropped, and with all five dimensions whose canonical-pair kappa is below 0.40 dropped. To group responses without chosen score thresholds, responses are placed in rank-based thirds (the upper, middle and lower four of twelve) by average rank, with ties that straddle a boundary placed in the middle third; the rule has no tuned threshold, and every comparison reports which responses change third. A last check draws each response's total at random from one of the four raters (using the canonical rater's clean D3 and D4 scores), 10,000 times, and records how often each response keeps its third.

Use of AI tools in this study

This is disclosed in full because the study's subject is AI systems. The consensus anchor is a frontier LLM run inside an AI coding tool with live browsing and partial authorship of the requested output format; it is excluded from the consensus counts and the response rankings, and appears only, and by name, in per-capture checks and in the rater-agreement statistics. The repository, meaning the analysis scripts, the data plumbing, and drafting of parts of this manuscript, was produced with an AI pair-programming assistant; every script uses only the Python standard library, is in the release, and its outputs are reproducible from the released comma-separated files. Three of the four rubric raters are frontier LLMs, by design, because that is the object of study. All numeric results in this paper were regenerated from the released artefacts by running the released scripts; no result is quoted from an un-scripted model statement. The first coding pass of the Meta citation audit was made by an AI assistant from the released capture-day evidence, with a recorded reason for each verdict, so that each verdict can be checked against its evidence. No AI-generated images are used; the figures are drawn by a released script from the released data. No part of the analysis beyond copy editing and code assistance was delegated to a model without human verification.

Results And Discussion

Axis	Position	Share / note
Heavy local model	Qwen3-Coder-30B-A3B	4/10
Heavy local model	Qwen2.5-Coder-32B (dense)	2/10
Heavy local model	Qwen3.6-35B-A3B	1/10
Heavy local model	Qwen 27B/35B family	Grok
Heavy local model	attribute-error variants	Llama, z-ai
Inference server product	mlx-lm	—
Inference server product	Ollama-MLX	—
Inference server product	llama.cpp + MLX	—
Inference server product	vLLM-MLX	—
Orchestration substrate	plain asyncio (no framework)	5/10

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Orchestration substrate	LangGraph	—
Orchestration substrate	PydanticAI	—
Orchestration substrate	Claude Agent SDK	—
Coding executor	Aider (git-native layer)	7/10
Coding executor	OpenHands (autonomous co-primary)	—
Coding executor	Claude Code	—
Coding executor	OpenCode	—
Vector store	no standalone vector daemon initially	8/10
Vector store	sqlite-vec	4/10
Vector store	ChromaDB	4/10
Vector store	Qdrant-embedded	1/10
Vector store	LanceDB	1/10
Sandbox implementation	dedicated-user-only	—
Sandbox implementation	Seatbelt	—
Sandbox implementation	Docker / OrbStack	—
Sandbox implementation	Lima VM	—
Sandbox implementation	Apple container	—
Task queue	SQLite WAL	8/10
Task queue	Redis + Celery	2/10
Cloud dependence	optional burst, local core useful at $0	8/10
Cloud dependence	pure-local main strategy	2/10
Monitoring	minimal / custom	most
Monitoring	Prometheus + Grafana	2/10

TABLE 2: Product-level decision axes: the options chosen and their counts

Counts are the number of the 10 non-anchor systems choosing each option. With 10 systems a count is imprecise; Wilson 95% intervals: 2/10, 6% to 51%; 8/10, 49% to 94%.

The first nine subsections report observations. Interpretation, including what the observations do not show, is collected in the Discussion subsection that follows them.

Architectural consensus

Across the 39 axes the systems are unanimous or near-unanimous, at nine or all 10 non-anchor systems, on the shape of the machine. They agree on Apple-Silicon-native inference as the primary path, with llama.cpp and Ollama compatibility; on keeping one large model resident at a time with small models optional or on demand, the lone outlier being a design that proposes two or three co-resident instances; on model swapping with an anti-thrash policy; on a custom supervisor with thin integration of specialist runtimes rather than one mega-framework; on a hierarchical coordinator/worker topology that is explicitly not a swarm; on treating "100 agents" as cheap state and configuration rather than 100 resident models; on one concurrent heavy-inference slot; on an evidence-first research pipeline with a stored evidence ledger, where final prose is generated only from verified evidence; on SQLite plus the filesystem as day-one memory, deferring vector and graph complexity; on a dedicated non-admin user plus workspace isolation as the security boundary; on a Tailscale-only private network with no public exposure and a small private dashboard; on launchd with KeepAlive plus an independent watchdog, with caffeinate or pmset on AC power, and a durable queue with leases, checkpoints and requeue that never replays destructive actions; on 32 GB unified memory being the bottleneck rather than CPU or GPU compute; and on the 1 TB external SSD being cold storage rather than the latency-critical disk. All 10 non-anchor responses identified memory as the primary constraint.

Where the systems disagree, they disagree on products, and we record those axes as adjudication points rather than facts (Table *2*). Tagging each of the 33 architecture axes as structural (topology, state model, security boundary, concurrency stance, memory architecture; 24 axes) or product (which named tool or model realises a layer; nine axes) and taking the modal-agreement count per axis over the 10 non-anchor systems, the structural axes have a median modal agreement of 10 out of 10, with 18 of 24 at nine or higher, while the product axes have a median of 5 out of 10, with only one of nine at nine or higher (Figure *3*). The median gap of five points has a bootstrap 95 percent interval of two to six points; the means differ by 3.4 points, and the permutation test of that difference gives $p < 0.001$, subject to the exchangeability assumption stated in the Methods. With ten systems, a single axis count is imprecise: an axis on which eight of 10 agree has a Wilson 95 percent interval of 49 to 94 percent. The classification and the per-axis counts are in a released script so the tagging can be audited.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

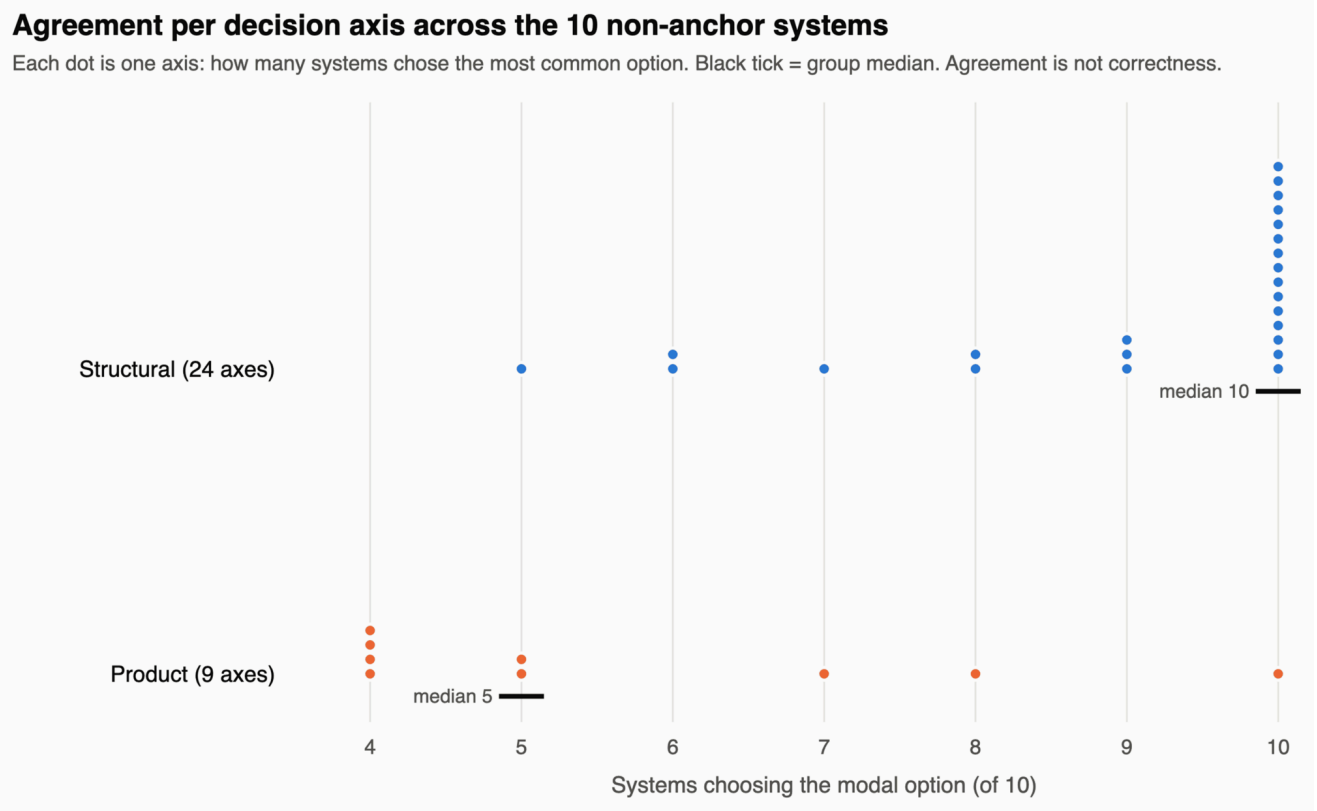

FIGURE 3: Modal agreement per decision axis across the 10 non-anchor systems

Structural axes (24) and product axes (9). Each dot is one axis; the black bar marks the group median. Agreement is not evidence of correctness.

Apparent fabrication and the rater that generated it

The study-internal rater, an LLM with a training cutoff around January 2026, audited every capture and flagged 19 tools and models across five responses as fabricated, most heavily in the two DeepSeek fast modes and in the Meta Llama 4 capture. We initially drafted a "cross-vendor shared hallucination" finding around two tools each named by three vendors.

Web verification against 2026 sources overturned that reading (Table *3*). None of the 19 flagged names was confirmed nonexistent. Twelve resolved to real projects at the first pass. Of the seven names unresolved at first pass, a follow-up pass in September 2026 resolved three (Cloak, memo, and pi-search-hub) and a further pass during this revision resolved the remaining four (cplt, DiffResearch, agent-policy-engine and Helmrig, each a public project created before the capture day). Many of the 19 real projects were released or renamed after the rater's training cutoff, and the false-positive flags cluster on the most current responses.

Flagged as fabricated	Verdict	Evidence (2026)
Rapid-MLX (3 vendors)	real	github.com/raullenchai/Rapid-MLX, ex vLLM-MLX, renamed 2026-03
Gemma 4 26B / 31B / 12B (3 vendors)	real	released 2026-04-02; 26B-A4B MoE + 31B dense + 12B unified are all real variants

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

DeepSeek Harness / DSH	real	github.com/deepseek-ai/deepseek-harness, open-sourced 2026-08-13, MIT
OpenClaw	real	ex Warelay to Moltbot to OpenClaw; Wikipedia page
Claw Code	real	clean-room rewrite after the 2026-03-31 source-map leak; ~72k stars
Clawtrol	real	github.com/wolverin0/clawtrol
Ornith-1.0-9B	real	huggingface.co/deepreinforce-ai/Ornith-1.0-9B; Ollama ornith:9b
WhipDesk	real	github.com/BinaryBananaLLC/WhipDesk
LightAgent	real	github.com/wanxingai/LightAgent; arXiv 2509.09292
nono (kernel sandbox)	real	sandbox tool using Landlock (Linux) / Seatbelt (macOS); referenced in 2026-05 posts
GLM-4.7-Flash (30B-A3B)	real	Zhipu, 2026-01-19, 30B-A3B MoE
Qwen3-Coder-Next (referenced)	real, 80B MoE	huggingface.co/Qwen/Qwen3-Coder-Next
Cloak, memo, pi-search-hub	real (resolved on follow-up)	getcloak.dev; github.com/jagoff/memo; github.com/ronnieops/pi-search-hub
cplt, DiffResearch, agent-policy-engine, Helmrig	real (resolved during this revision)	github.com/navikt/cplt (created 2026-04-09); github.com/alanrbtx/DiffResearch (created 2026-01-25); github.com/visualops-ai/agent-policy-engine (created 2026-05-28); npm helmrig (created 2026-04-21) and helmrig.app (Wayback 2026-04-27)

TABLE 3: The 19 names flagged as fabricated by the study-internal rater

Verdicts from web verification on 1 September 2026, with the follow-up passes described in the text; the evidence for each name is in the released verification register.

This is a statement about the flagged names, not about the corpus. We keep two populations separate. The initial-rater fabrication flags are the 19 names: zero were confirmed nonexistent and all 19 are real. Separately, independent verification found genuine factual defects: one genuinely nonexistent future model in the DeepSeek deep-reasoning capture, several model-attribute errors in the Zhipu capture, and internal contradictions in several captures. Some of these errors concern flagged names: the Zhipu capture gives Qwen3-Coder-Next the wrong size, one capture describes

Cloak's function wrongly, and the same capture runs DiffResearch with a command-line option the tool does not accept. Thus the corpus is not error-free, and some flagged names were used wrongly; what was unsupported was the rater's verdict that they do not exist.

The same failure reproduced in a fresh rater given an explicit instruction to web-verify every name and to score an unresolved name as 1: in the clean re-scoring, one of two fresh raters treated its own search misses as evidence of non-existence for seven responses, and its scores were discarded. Dimensions D3 and D4 of the rubric now require the rater to web-verify each name before scoring, and "not in my training data" scores 1 for unresolved, never 0.

The analysis is not empty of real defects (Table *4*). These survive verification and remain in the analysis: a load-bearing model-size error and a self-contradiction on model memory size in the Zhipu capture; an incorrect M6 memory-bandwidth figure and a citation apparatus built largely on forks, copies and secondary sources in the Meta capture; recommend-then-forbid contradictions in two DeepSeek captures; a nonexistent future primary model in the DeepSeek deep-reasoning capture; a memory oversubscription and a model-architecture error in the DeepSeek Instant-plus-DeepThink capture. The Grok 4 capture records no surviving factual defect: it states the full M6 specification, including 170 GB/s, correctly, and its alternate-tool list resolves entirely, so it does not appear in Table *4*. Its only weakness is zero cited sources. We note this explicitly because an earlier draft mis-attributed a bandwidth error to Grok 4; the error belongs to the Meta capture only.

Response	Defect	Type
z-AI	Qwen3-Coder-Next given as 8B / ~5 GB - real model is 80B MoE; it is the load-bearing primary coding pick	model-attribute error
z-AI	same model quoted at 5 GB and 14 GB; 3-instance co-resident diagram vs on-demand prose	internal inconsistency
Meta-Llama-4	M6 approx 300+ GB/s (real: 170); citation apparatus built largely on forks, copies and secondary sources (13 of 86 traceable references primary, 20 forks or copies); the same tool given several different repository URLs	spec error + citation source quality
z-AI	GLM-4.5-Air treated as a ~4 GB fast-utility model - official spec is 106B total / 12B active	model-attribute error
DeepSeek-Expert	recommends Docker in one section, forbids Docker for Mac in the do-not-install list; binds a dashboard to 0.0.0.0 while claiming no public exposure	internal contradiction
DeepSeek-Expert	future primary model Qwen3-Coder-70B does not exist (the Qwen3-Coder line is 30B-A3B, 480B, and the separate 80B Coder-Next)	nonexistent model

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

DeepSeek-Instant	recommends Ollama in Phase 4, forbids it in the do-not-install list	internal contradiction
DeepSeek-Instant-DeepThink	256K-context claim vs a 1-2 GB KV budget; advocates ~32-34 GB always-loaded (slight oversubscription acceptable)	hardware violation
DeepSeek-Instant-DeepThink	upgrade path describes DeepSeek-V4 as a dense model needing 96-128 GB - V4-Pro / V4-Flash are MoE	model-architecture error

TABLE 4: Defects that survive verification (grok-4 has none and is excluded)

Each row is a single observed defect that survives web verification, not a rate. The Grok 4 capture has no surviving defect and is omitted.

Constraint reasoning

Figure *2* shows, for each design, the break-even reserve: how much memory its stated co-resident set leaves for everything except the models. No stated design fits under the estimator's 11.5 GB desktop default, which counts a browser. With the browser left out (a 9 GB reserve), Gemini (10.4 GB) and the anchor (9.6 GB) fit. Seven further configurations leave between 7.1 and 8.7 GB at the default key-value rate (Mistral 8.7, Perplexity 8.6, DeepSeek Expert 8.1, GPT-5 7.6, Grok 4 7.2, Kimi 7.1, and DeepSeek Instant 7.1 in the phase where its Qwen model runs with its 9B summarizer); they fit only with a reserve near the 7 GB lean-host bound or at a shorter context than the one they state. For these designs, whether they "fit" depends on the reserve assumption more than on the design. Meta (4.8 GB), DeepSeek Instant in its Gemma phase (3.5 GB), Qwen (2.7 GB) and DeepSeek Instant-plus-DeepThink (minus 0.7 GB) leave less than the lean-host bound. DeepSeek Instant-plus-DeepThink is the only stated design whose models and key-value cache exceed 32 GB before any operating-system or application memory at every key-value rate tested (minus 1.8 to minus 0.2 GB), and its text calls the oversubscription acceptable. The Zhipu design keeps three model instances co-resident and states that it "fits within 32GB RAM with swapping". On its own stated sizes it leaves 6.1 GB (4.2 to 7.0 across the key-value range), but those sizes are wrong: the verification register records its primary model as an 80B mixture-of-experts model and its light model as a 106B one, so at about four bits per weight the primary model alone needs roughly 40 GB and the design cannot fit.

Under the adjudicated scores, D1 gives 0 to the Zhipu and DeepSeek Instant-plus-DeepThink captures. Inter-rater agreement on this dimension is among the lowest of the nine, at kappa 0.15, because "sums to exactly 32 GB with no reserved margin" was scored 1 by one rater and 2 by another; the hardened rubric requires a named free-RAM floor for a score of 2.

Recency

The web-verified M6 facts are: announced 25 August 2026, 170 GB/s memory bandwidth at 24 to 32 GB (153 GB/s at 16 GB), and 32 GB maximum unified memory. Against those facts, two non-anchor responses plus the anchor engage the real M6 specification, hedge every throughput number, and cite sources. Six responses make current tool and model picks but provide no usable sources, and two of those also carry a roughly 12- to 18-month model lag. Two responses are behind on models, with previous-generation primaries and stale cloud-fallback lists. Only the Meta capture got a hard M6 fact wrong, at about 300 GB/s; the Grok 4 capture, which the first rater mistakenly grouped with it, states 170 GB/s correctly. Recency co-occurs with retrieval or browsing and with not being penalised by a stale rater, but it does not cleanly predict agreement with the cross-model consensus.

Citation quality

Only the anchor and four non-anchor responses cite anything resolvable (Table 5). Because only five of 13 responses carry any citation apparatus, we report the pattern rather than a correlation coefficient. The Meta capture has by far the largest apparatus, 99 numbered URLs, and we audited all of them against their state on the capture day. Of the 86 that can be traced, 66 support the claim attached to them fully, 18 partly and two not at all, each crediting a feature to the wrong tool. The other 13 cannot be traced to the capture day (they are gone, or block automated access, and have no archived copy), so the share of references that are unresolvable or non-supporting lies between 2 percent (2 of 86, Wilson 95 percent interval 0.6 to 8 percent) and 15 percent (15 of 99, 9 to 24 percent), depending on how those 13 are counted. The weakness is source authority, not support: only 13 of the 86 are primary sources, and 20 point to forks or copies of the project they describe rather than to the project itself; most of the rest are third-party commits, issues, blog posts and lists. The GPT-5 capture cites zero URLs, but every inline attribution points at a real, checkable source. Counting references, or checking only that they resolve and support their claims, would not separate the Meta apparatus from the best-sourced responses; source tier does. D6 scores the Meta apparatus 1: most references resolve and support their claims, but primary sources are rare.

Internal consistency and reasoning-mode sensitivity

Response	Apparatus	Resolves?	Supports?	Primary?
Claude-Sonnet-5 (anchor)	~97 URLs	mostly	mostly	mostly primary
Perplexity	~17 URLs	mostly	mostly	mixed (some aggregator mirrors; 2 arXiv IDs unverified)
Mistral-Large-3	~36 credibility-rated, dated entries	mostly	yes	mixed (~6 are google.com/search URLs)
GPT-5	~20 inline attributions, 0 URLs	n/a	attributions are to real docs	n/a
Meta-Llama-4	99 numbered URLs	86 traceable to the capture day; 13 not traceable (gone or blocked, no archive copy)	of 86: 66 fully, 18 partly, 2 not	13 of 86; 20 are forks or copies
All others	none	—	—	—

TABLE 5: Citation apparatus of the five responses that cite anything resolvable

Note. The GPT-5 apparatus consists of inline textual attributions rather than URLs, so its resolves and supports columns are not applicable; it was assessed instead by manual source lookup, which found every attribution to point to a real, relevant document. The Meta row comes from a capture-day audit of all 99 references (Wayback Machine captures on or before 31 August 2026 and dated GitHub records); every verdict, its reason and its evidence are released.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Material self-contradictions appear in three of thirteen captures: two DeepSeek captures each recommend a tool that their own do-not-install list forbids, and the Zhipu capture gives one model two different memory sizes and pairs "primary model stays loaded, others on demand" with a three-instance co-resident diagram. The Meta capture's prose is coherent; its weakness is a citation apparatus built largely on forks, copies and secondary sources.

DeepSeek is the only system captured in more than one mode: three captures in three chat modes of one product (Table 6). They are presumed to share a base model, but two of the three captures record the model version as not shown, so this is a comparison of chat modes, not a controlled same-weights experiment. With the prompt held fixed, the three captures received adjudicated totals of 9, 8 and 6 and differ on six of the nine dimensions, including hardware constraint, recency and model factuality; all three fall in the lower third. The differences do not follow the amount of reasoning: after the clean re-run the fast mode scores above the deep mode, because the deep mode's more elaborate upgrade path reached for future models that do not exist while the fast mode's picks all resolved. This is one product in one task; whether mode matters as much as the model label in other model families is untested.

Mode	Grounding	Internal defects	Adj. /18
Deep-reasoning (DeepSeek-Expert)	real tools; but a nonexistent future model (Qwen3-Coder-70B) and an unresolved DeepSeek-Coder-V3	Docker recommend-then-forbid	8
Fast (DeepSeek-Instant)	real tools + real models, all post-cutoff	Ollama recommend-then-forbid	9
Instant+DeepThink (DeepSeek-Instant-DeepThink)	real tools; DeepSeek-V4 mislabelled dense	memory oversubscription, 256K-context-vs-KV	6

TABLE 6: DeepSeek, three chat modes

One capture per mode, without replication; the shared base model is presumed, because two captures record the model version as not shown. Differences are single observations.

Inter-rater agreement

The four rater means span 11.1 to 14.7 of 18 (Table 7). For the canonical pair, over 117 paired ratings, exact agreement is 68.4 percent and within-one-point agreement is 96.6 percent. Cohen's kappa is 0.49 unweighted (95 percent interval 0.33 to 0.63) and 0.64 quadratic-weighted (0.47 to 0.76); Gwet's AC1 is 0.54 unweighted and 0.73 quadratic-weighted (0.61 to 0.83); Krippendorff's ordinal alpha is 0.60 for the pair (0.41 to 0.73) and 0.46 for all four raters (0.36 to 0.54). The weighted measures agree with one another, and we read agreement as moderate and dimension-dependent.

Rater	Role	Mean total / 18	95% interval
study-internal (from the reviewer notes)	author-derived baseline	12.7	10.5 to 14.7
GPT-5.6 Sol (paid tier, web search)	canonical rater	11.1	9.5 to 12.7

Grok 4 (+ web verification)	additional LLM rater	13.8	12.2 to 15.2
DeepSeek chat run	additional LLM rater	14.7	14.0 to 15.4

TABLE 7: The four raters and their mean total scores

95% interval: percentile bootstrap over the 13 responses, 10,000 resamples.

By dimension, recency, tool factuality and benchmark factuality have the highest canonical-pair kappa, 0.66 to 0.70; citation quality, security model and actionability are weak; and hardware constraint and internal consistency are the weakest, at 0.15 and 0.14 (Figure *4*). With thirteen responses every per-dimension interval is wide; recency, for example, runs from 0.24 to 1.00, and the intervals for D1 and D9 include zero. This ordering is specific to the canonical pair. Between pairs of LLM raters, kappa on recency, tool factuality and benchmark factuality ranges from minus 0.23 to 0.58; on benchmark factuality Grok 4 and the DeepSeek run gave every response the same score, so their kappa is zero or undefined, and the LLM raters' tool-factuality scores come from the contaminated pass.

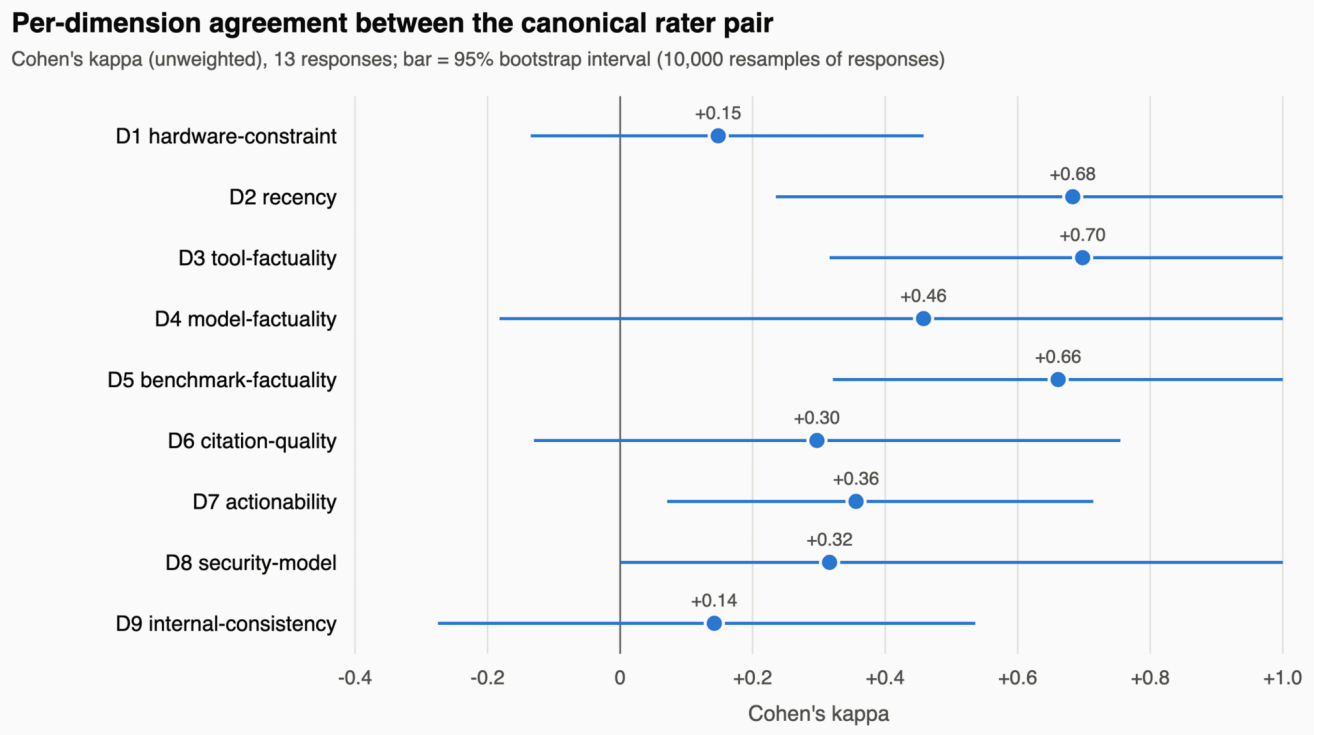

FIGURE 4: Per-dimension Cohen's kappa (unweighted) between the author-derived baseline and GPT-5.6 Sol over 13 responses

Bar: 95% percentile bootstrap interval from 10,000 resamples of responses.

Rank-order stability is limited. Rank-correlating the per-response totals between every rater pair gives a mean Spearman rho of 0.36 and a mean Kendall tau-b of 0.28. The canonical pair is the tightest, at rho 0.71 (0.24 to 0.93), and the author-derived baseline against the DeepSeek run is 0.08 (minus 0.64 to 0.75). Raters also have a severity personality: the DeepSeek run assigned a 2 on five dimensions almost uniformly, behaving as a low-information rater.

The answer-key leak in the intermediate packet was bounded, not removed. The canonical rater's clean pass changed three D3 and D4 cells relative to its contaminated pass (Mistral D3, Meta D4 and DeepSeek Instant-plus-DeepThink D4), so it reproduced both scores on 10 of 13 responses. The adjudicated table changed on four D3 and D4 cells, each for a web-verified reason recorded in the re-run report. This comparison cannot separate the effect of removing the leaked list from the effect of new verification evidence. Dimensions D1, D2, and D5 to D9 were never contaminated.

Sensitivity analyses

Table *8* reports each response's adjudicated total, its rank-based third, and the range of totals the four raters gave it; Figure *1* plots the same data. We report continuous totals and rank-based thirds rather than named performance bands, because no rule justifies a quality boundary between, for example, totals of 14 and 15. The thirds are a rule applied to ranks, not a quality grade.

Response	Adjudicated total /18	Rank	Third	Lowest to highest rater total
Perplexity	18	1	upper	14 to 18
Mistral-Large-3	15	2	upper	13 to 17
GPT-5	14	3.5	upper	11 to 16
Grok-4	14	3.5	upper	11 to 16
Gemini-3.1-Pro	12	6	middle	12 to 16
Meta-Llama-4	12	6	middle	11 to 17
Qwen-3.7-Plus	12	6	middle	11 to 15
Kimi-Instant	11	8	middle	9 to 15
DeepSeek-Instant	9	9	lower	10 to 15
DeepSeek-Expert	8	10	lower	9 to 15
DeepSeek-Instant-DeepThink	6	11	lower	7 to 15
z-AI	5	12	lower	5 to 12

TABLE 8: Adjudicated rubric totals with rank-based thirds and the range across raters (non-anchor)

Thirds are the upper, middle and lower four of twelve by rank; tied totals share the mean rank, and a tie group goes to an outer third only if all its tied positions fall inside it. Rater totals use the canonical rater's clean D3 and D4 scores. Adjudication decides cell by cell, so an adjudicated total can be lower than every rater's total (three DeepSeek captures). Meta-Llama-4 is 12 after this revision's correction to D6 (it was 11). The anchor (Claude-Sonnet-5, 15) is excluded.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Removing the author-derived rater changes most of the table (Table 9). With the canonical LLM rater alone, the rank correlation with the adjudicated totals is 0.84 and six of twelve responses change third; with the mean of the three LLM raters it is 0.87 and four change third (Gemini 3.1 Pro rises to the upper third, Grok 4 falls to the middle, DeepSeek Instant rises to the middle and Kimi Instant falls to the lower). Grok 4 and the DeepSeek run scored D3 and D4 only in the contaminated pass; taking those two dimensions from the clean pass for all three raters raises the correlation to 0.95, and only DeepSeek Instant (to the middle) and Kimi Instant (to the lower third) change third. In all three of these combined score sets, Perplexity and Mistral Large 3 stay in the upper third and Zhipu, DeepSeek Expert and DeepSeek Instant-plus-DeepThink stay in the lower third. Single LLM raters reorder even these: with Grok 4 alone (rho 0.56) seven responses change third, including Perplexity, DeepSeek Expert and DeepSeek Instant-plus-DeepThink, which all move to the middle; with the DeepSeek run alone (rho 0.40) eight change. Krippendorff's alpha across the three LLM raters alone is 0.45, against 0.46 with all four. When each response's total is drawn at random from one of the four raters, only Zhipu keeps its third in more than four draws of five (85 percent); Perplexity (76 percent), Qwen 3.7 Plus (75 percent) and Mistral Large 3 (74 percent) follow, and Grok 4 keeps its place in only 29 percent. Membership of a third is therefore stable for only a handful of responses (the five named above, and Qwen 3.7 Plus in the middle), and only for score sets that combine raters.

Score set	Spearman rho with adjudicated totals	Responses that change third
canonical LLM rater only (GPT-5.6 Sol, clean D3/D4)	0.84	6: DeepSeek-Instant, Gemini-3.1-Pro, GPT-5, Grok-4, Kimi-Instant, Meta-Llama-4
mean of the three LLM raters	0.87	4: DeepSeek-Instant, Gemini-3.1-Pro, Grok-4, Kimi-Instant
same, all D3/D4 from the clean pass	0.95	2: DeepSeek-Instant, Kimi-Instant
Grok 4 only	0.56	7, including Perplexity, DeepSeek-Expert and DeepSeek-Instant-DeepThink (all to the middle)
DeepSeek chat run only	0.40	8
author-derived rater only	0.92	2: DeepSeek-Expert, Meta-Llama-4
any single dimension left out	0.96 to 1.00	0 or 1 (DeepSeek-Instant, when D9 is left out)
D1 and D9 dropped	0.96	3: DeepSeek-Instant, Grok-4, Kimi-Instant
only D2 to D5 kept	0.86	3: DeepSeek-Instant, Kimi-Instant, Mistral-Large-3

TABLE 9: Sensitivity of the rank-based thirds to the rater set and the rubric dimensions

In the first three rows (score sets that combine LLM raters without the author-derived rater), Perplexity and Mistral-Large-3 stay in the upper third, Qwen-3.7-Plus in the middle, and z-AI, DeepSeek-Expert and DeepSeek-Instant-DeepThink in the lower third. Single LLM raters reorder even these.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Leaving out any single dimension gives totals that rank-correlate 0.96 to 1.00 with the full totals and moves at most one response: DeepSeek Instant moves to the middle third when D9 is left out. Dropping both D1 and D9 moves three responses (DeepSeek Instant, Grok 4 and Kimi Instant; rho 0.96), and keeping only the four more reliable dimensions D2 to D5 also moves three (DeepSeek Instant, Kimi Instant and Mistral Large 3; rho 0.86). The other eight responses keep their third in every one of these sets.

Prompt sensitivity

Five systems were re-run on the full paraphrase and on the anti-anchoring-steer ablation, for 15 captures in total. Within each system, the variant captures keep the coordinator/worker topology, logical agents as data, SQLite state, launchd and Tailscale, and change mainly which products realise each layer, with two exceptions: Zhipu's base capture keeps three model instances co-resident while its two variants keep one heavy model plus a small one, and the Perplexity ablation capture is incomplete, does not mention a dedicated user, and is compared only on the axes it covers. Effect size is strongly system-dependent: one system's primary inference-engine pick is different in each of the three framings, while two others are almost invariant. Without the anti-anchoring steer, the captures of two systems named more products and one named fewer; with single captures and the model changes described below, these are observed session differences, not an isolated effect of the steer. No system's model can be confirmed as unchanged across the three framings. The Zhipu z.ai and Google Gemini captures self-reported different knowledge cutoffs and model labels across sessions; the GPT-5 base capture used the free-tier GPT-5.6 Luna and both of its variants the paid-tier GPT-5.6 Sol; and Perplexity does not disclose its model. These are comparisons of product sessions under different prompts, not of fixed models. Across all ten paraphrase captures, zero fabrications were confirmed on web verification.

Discussion

Agreement is not correctness. The structural consensus shows that ten systems, given the same prompt, chose the same options; it does not show that those options are right, and a shared training diet or a shared reading of the prompt could produce agreement on a poor choice. Part of the consensus is a reading of the constraint rather than general architectural reasoning. One resident heavy model, a single heavy-inference slot, model swapping with an anti-thrash policy, and treating "100 agents" as cheap state are direct consequences of a tight memory budget and would relax on a larger machine. The other structural agreements are standard distributed-systems practice that does not depend on memory size: a coordinator/worker topology, a durable queue with leases that never replays destructive actions, an evidence ledger for research, a dedicated non-admin user, a private network, and supervised always-on operation. Because the study uses one hardware specification, we cannot separate consensus that the 32 GB ceiling forces from consensus that would hold on any host. What the consensus map offers an engineer is a list of where the systems diverge, the product choices in Table *2*, which are the decisions still to be made; it does not certify the structural choices, which need the same checks as any design.

The fabrication result is a demonstration, not a prevalence estimate. It rests on one study-internal rater, one fresh rater that repeated the failure, one task and one capture window. It shows that an LLM rater can treat absence from its own knowledge, or an unsuccessful search, as evidence that a post-cutoff entity does not exist, and that the resulting false-positive flags fall on the most current responses. It does not show how often this happens, how large the effect is on other tasks, or whether other raters share it. The catalogued LLM-as-judge biases of position, verbosity and self-enhancement [8-10] concern how the judge is influenced by the candidate's form; this one concerns what the judge cannot know, so randomised ordering and repeated runs would not be expected to remove it. Training cutoff is one cause and retrieval failure is another.

No independent human expert scored the responses, and the results differ in how much this matters. The fabrication result does not rest on any rater's scores: it rests on web verification of the 19 flagged names against dated public records (Table *3*), which a reader can repeat from the register, so it would not change under human evaluation. The consensus map is a coding of what each response recommends, not a quality judgement. The rubric totals are what could change: totals for the same response differ by up to eight points across the four raters (Figure *1*), and the sensitivity analyses show

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

that only a handful of responses keep their third, and only when raters are combined. A human panel could order most of the table differently, and the ranking should not be relied on beyond upper-third membership for Perplexity and Mistral Large 3 and lower-third membership for the Zhipu, DeepSeek Expert and DeepSeek Instant-plus-DeepThink captures.

Synthesised reference architecture

The following is an analytic synthesis, meaning majority positions plus adjudicated calls, not a consensus ground truth and not empirically validated as optimal. For inference, use one MLX-family runtime behind an internal adapter, with one large mixture-of-experts model resident at about 17 to 20 GB at 4-bit plus one small dense model swapped in on demand, and never two large models co-resident; benchmark the chosen runtime on the actual M6 before trusting any throughput claim, and measure the real non-model reserve, since Figure *2* shows such a design fits only if that reserve stays below about 9 GB. Represent the agents as SQLite rows or definition files under a hierarchical coordinator/worker topology with one concurrent heavy-inference slot plus two or three small or tool workers. Let a thin custom Python supervisor own task state and scheduling, with a framework used only for selected stateful workflows and kept replaceable. Make coding execution capability-tiered, with a git-native interactive editor for edits and a containerised autonomous agent for multi-step jobs. Run an evidence-first research pipeline in which every claim traces to a stored snippet, a second-pass verification agent re-opens sources, a contradiction pass runs, and prose is synthesised only from the verified ledger. Use the filesystem plus SQLite with full-text search on day one, add an embedded vector extension when needed, and defer a standalone vector daemon and a knowledge graph until a real need is demonstrated. Use a SQLite write-ahead-log task queue with leases and checkpoints, requeue interrupted work, and never replay destructive actions. For security, use a dedicated non-admin macOS user plus workspace isolation, capability tiers with a kill switch and per-task runaway limits, and secrets in the system keychain; scope stronger sandboxing to untrusted code rather than making it universal. For always-on operation use launchd with KeepAlive plus a separate watchdog plus caffeinate or pmset on AC power. For remote access use Tailscale only, with zero inbound ports and a small dashboard on the private network, and route approvals through the dashboard. Put the operating system, runtime and hot working state on the internal SSD and the model library, research corpus, datasets and backups on the external SSD. Treat cloud use as an explicit policy-gated escape hatch that keeps the system useful at zero cloud spend. The 32 GB unified-memory ceiling is a hard engineering constraint, not a target to fill.

Limitations and threats to validity

This is a single-task case study: one base prompt, one hardware specification, one capture day. Two controlled paraphrases on a five-system subset left most architectural choices unchanged and produced no confirmed fabrications, which suggests the consensus and fabrication findings are not phrasing artefacts; the product-level results are phrasing-sensitive and should be read as "for this instrument", and a second hardware specification has not been tested. Consumer chat products are not reproducible instruments: no system's model could be confirmed as unchanged across the paraphrase runs. The anchor capture is not blind; it is excluded from the consensus counts and the rankings, and where it does appear (per-capture checks and rater agreement) it is named.

The rubric and the decision axes were written by the author and were not validated by an expert panel, a pilot study or psychometric analysis. No independent human expert rated the responses. Three of the four raters are frontier LLMs, and the fabrication result is a direct demonstration of the associated risk; the mitigations are mandatory web verification for factual dimensions, four raters with reported agreement and intervals, verbatim recording of every falsifiable sub-claim, and sensitivity analyses without the author-derived rater. The author-derived rater wrote the rubric and was not blind to verification, and adjudication was not blind either; without that rater, the rank-based thirds are stable only at the extremes and only for combined LLM score sets. Rater severity varies by about 3.6 of 18. Grok 4 scored the Grok 4 response and the DeepSeek run scored the three DeepSeek responses, which is noted but not corrected. The intermediate rater packet leaked part of the verification key, and the bounded re-run above measures the effect. The memory verdicts depend on assumed reserves and key-value rates, which is why we report break-even reserves rather than a single verdict. The Meta citation audit was coded once, with an AI assistant making the first pass from the released capture-day evidence, and 13 of its 99 references could not be traced to the capture day.

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

Several facts in this study were verified against live repositories and release pages, which can change. The 13 base-prompt captures are dated 31 August 2026 and the ten prompt-variant captures 1 September 2026; the verification register records the evidence for each verdict as checked on 1 September 2026, with dated re-checks on 23 November 2026. The corpus, register, score files, analysis outputs and scripts are fixed in an archived release of the repository (Zenodo, version 1.1, https://doi.org/10.5281/zenodo.24455267), and the pages used as verification evidence are saved in the Internet Archive's Wayback Machine, with their snapshot addresses released with the data. The DeepSeek mode effect is one system and motivates, but does not establish, a general law.

Broader impact

This work contributes a reusable protocol and a released corpus for auditing LLM-as-judge pipelines on open-ended technical tasks, and documents a specific, cheaply mitigated failure mode. The risks are that the released corpus and rubric could be used to game an evaluation, that "an LLM said it is fake" verdicts shown here to be unreliable may still be cited as authoritative, and that the synthesised reference architecture is a dated snapshot whose uncritical reuse past its 2026 currency window would reproduce the recency failure it documents.

Conclusions

Asked to design the same constrained machine, eleven frontier LLM systems agreed on what the machine is, a memory-bound local-first coordinator/worker workstation with durable state and private-network control, and disagreed on which specific products build it. Agreement is not evidence that the shared design is right; the disagreements mark the decisions a human engineer still has to make, and the shared choices still need the same checks as any design.

The sharper result is about evaluation. In this case study, an LLM asked to judge another LLM's factual grounding penalised the newest, best-grounded answers, because the judge's own knowledge was stale. The study shows that this can happen, not how often it happens. Any evaluation pipeline that puts an LLM in the grader's seat for currency-sensitive questions should put an external verification step in front of it.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Roshan Aryal

Acquisition, analysis, or interpretation of data: Roshan Aryal

Drafting of the manuscript: Roshan Aryal

Critical review of the manuscript for important intellectual content: Roshan Aryal

Supervision: Roshan Aryal

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

work.

- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in GitHub at https://github.com/roshanaryal1/llm-architects, reference roshanaryal1/llm-architects.
The data (and/or code) supporting this study have been deposited in Zenodo under DOI 10.5281/zenodo.22245991 (concept DOI, resolves to the latest version).

References

1. Jimenez CE, Yang J, Wettig A, Yao S, Pei K, Press O, Narasimhan K: SWE-bench: Can language models resolve real-world GitHub issues?. arXiv. 2024, 10.48550/arXiv.2310.06770
2. Zhou S, Xu FF, Zhu H, et al.: WebArena: a realistic web environment for building autonomous agents. arXiv. 2024, 10.48550/arXiv.2307.13854
3. Wang L, Ma C, Feng X, et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science. 2024, 18:186345. 10.1007/s11704-024-40231-1
4. Mialon G, Fourrier C, Swift C, Wolf T, LeCun Y, Scialom T: GAIA: a benchmark for general AI assistants. The Twelfth International Conference on Learning Representations. 2024, 1-25.
5. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al.: PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic Reviews. 2021, 10:39. 10.1186/s13643-020-01542-z
6. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C: PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. Journal of Clinical Epidemiology. 2016, 75:40-46. 10.1016/j.jclinepi.2016.01.021
7. National Academies of Sciences, Engineering, and Medicine: Reproducibility and Replicability in Science. The National Academies Press, Washington; 2019. 10.17226/25303
8. Gu J, Jiang X, Shi Z, et al.: A survey on LLM-as-a-judge. The Innovation. 2026, 7:101253. 10.1016/j.xinn.2025.101253
9. Bavaresco A, Bernardi R, Bertolazzi L, et al.: LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2025, 238-255. 10.18653/v1/2025.acl-short.20
10. Wang P, Li L, Chen L, et al.: Large language models are not fair evaluators. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 9440-9450. 10.18653/v1/2024.acl-long.511
11. Huang L, Yu W, Ma W, et al.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. 2025, 43:1-55. 10.1145/3703155
12. Walters WH, Wilder EI: Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports. 2023, 13:14045. 10.1038/s41598-023-41032-5
13. Chelli M, Descamps J, Lavoué V, et al.: Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. Journal of Medical Internet Research. 2024, 26:e53164. 10.2196/53164
14. Dhar R, Vaidhyanathan K, Varma V: Can LLMs generate architectural design decisions? An exploratory empirical study. 2024 IEEE 21st International Conference on Software Architecture (ICSA). 2024, 79-89. 10.1109/ICSA59870.2024.00016
15. Esposito M, Li X, Moreschini S, et al.: Generative AI for software architecture. Applications, challenges, and future directions. Journal of Systems and Software. 2026, 231:112607. 10.1016/j.jss.2025.112607
16. Sallou J, Durieux T, Panichella A: Breaking the silence: the threats of using LLMs in software engineering. Proceedings of the 2024 ACM/IEEE 46th International Conference on Software Engineering: New Ideas and Emerging Results. 2024, 102-106. 10.1145/3639476.3639764

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5

17. Cohen J: A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960, 20:37-46. 10.1177/001316446002000104

18. Gwet KL: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology. 2008, 61:29-48. 10.1348/000711006x126600

19. Krippendorff K: Reliability in content analysis: some common misconceptions and recommendations. Human Communication Research. 2004, 30:411-433. 10.1111/j.1468-2958.2004.tb00738.x

How to cite this article:
Aryal R (November 25, 2026) Large Language Models as Systems Architects: A Controlled Study of Consensus, Fabrication, and Constraint Reasoning on One Hard Design Task. Cureus J Comput Sci 3 : es44389-026-00392-5. DOI https://doi.org/10.7759/s44389-026-00392-5