

NOTE 001

Silent failure, dated

Written 18 October 2026.

Poster: KoSAIM 2026 Annual Conference, Seoul, 23 October 2026. Poster Session 2, P2-10.

Changes: none.

Lab notes hold the part of the work that does not fit on a poster. This is the first one. It is about a hypothesis I wrote down in July, how I got there, the measurement I made of it, and what two other organizations said about the same problem in October without knowing about mine.

22 July 2026, 00:50 KST

At the start of the summer break I was building a simulator so that I could practice a surgical procedure on my own. A literature review on cardiac surgery was running at the same time, so I decided to put a heart in it. That meant I needed a 3D heart model, so I generated one. I had just finished the cardio-pulmonary-renal module, and to that eye the shape looked wrong and the vessel count did not add up. I checked it against my lecture notes. It was wrong. Making the prompt more anatomically specific changed nothing, so I turned it into an experiment. I am not a specialist. I saw this with a medical student's knowledge, and a specialist's judgment has been requested as a blinded review.

By the night of 21 July the audit of the generated models had finished. The pipeline was fixed: Nano Banana 2 for the images, Meshy 6 for the meshes. The question was whether a text-to-image-to-3D pipeline could produce a heart whose outer surface was closed with no handles, and whether asking more precisely helped. All 18 meshes were watertight. Sixteen had handles. The generator's responses had reported success on the requested style and stump form, added caveats about some deviations in view and vessel count, and said nothing about topology. None of the caveats stopped a delivery.

That night I wrote this to the model I was working with at the time. Translated from Korean and lightly trimmed; the bracket replaces an internal project name.

Devastating is not quite the word. In design and the creative fields these tools may do fine, but today I felt how large the gap still is between medicine and AI. I knew the gap was there, intuitively; that is why I was designing [a verification layer] in the first place, and this was the experiment on image generation. And the result is devastating. It is not disappointment either. I cannot close this gap alone, not with this much capital and time. Meanwhile everyone else praises these tools without exception, and they are already in wide use. In medicine that is dangerous. The output is wrong, and unless someone with a specialist's training checks it, it is wrong at a level that will not be caught. And if someone builds a tool on top of it and hands that tool to medical students or physicians, it reaches patients. My head is a mess.

A little later that night I added this.

Whatever else it is, I saw something no one else has really seen, and I saw it alone. The shock of that, the loneliness, the confusion, it all seems to arrive at once. Where do I go from here? Some of that too. I just feel low.

The reply had one line I kept. The quality that makes these systems good at creative work, fluent plausibility whether or not the content is true, is the same quality that makes them dangerous in medicine, and it is the grain of the tool rather than a bug for the next version. Whether that holds is an empirical question.

What I did with it

The sentence I wanted to write that night was that generative 3D models are unsafe for medicine. The data did not support that sentence, and they still do not. They support a narrower one. In one pipeline, 16 of 18 heart meshes failed a closed-surface target of genus zero. Watertightness alone did not establish compliance. Prompt specificity made no detectable difference across three levels (Kruskal-Wallis; primary analysis $N=15$, $H=0.89$, $p=0.64$; sensitivity analysis $N=18$, $H=0.47$, $p=0.79$). The 27 archived image-stage responses asserted success on outputs that failed the audit. None of them could have flagged the topological target; the image-stage model never sees the mesh. How they framed what they did disclose is a separate question, and it goes in the full paper. Pass and fail describe surface topology, not anatomical validity. Expert scoring of vessel identity and spatial anatomy is pending. Five runs per level limit inference.

In July I saw an anatomy problem, the shape and the vessel count. I measured topology instead, because genus, boundary loops and watertightness can be computed from the mesh by a machine with no expertise at all. Vessel identity cannot, so that went to a blinded expert panel.

That is what the poster says, and no more [1]. The method, auditing each output against a formal specification of structural invariants that a machine can measure, went into a Korean provisional patent application on 4 August [2]. The heart is one example. The method's claim is that a self-report is not evidence and that the check has to come from outside the generator.

The word I settled on for the failure is silent. A silent failure is a specification violation that the generator does not identify in its self-report. It is the failure a reviewer cannot see, because the artifact looks finished and the system says it is.

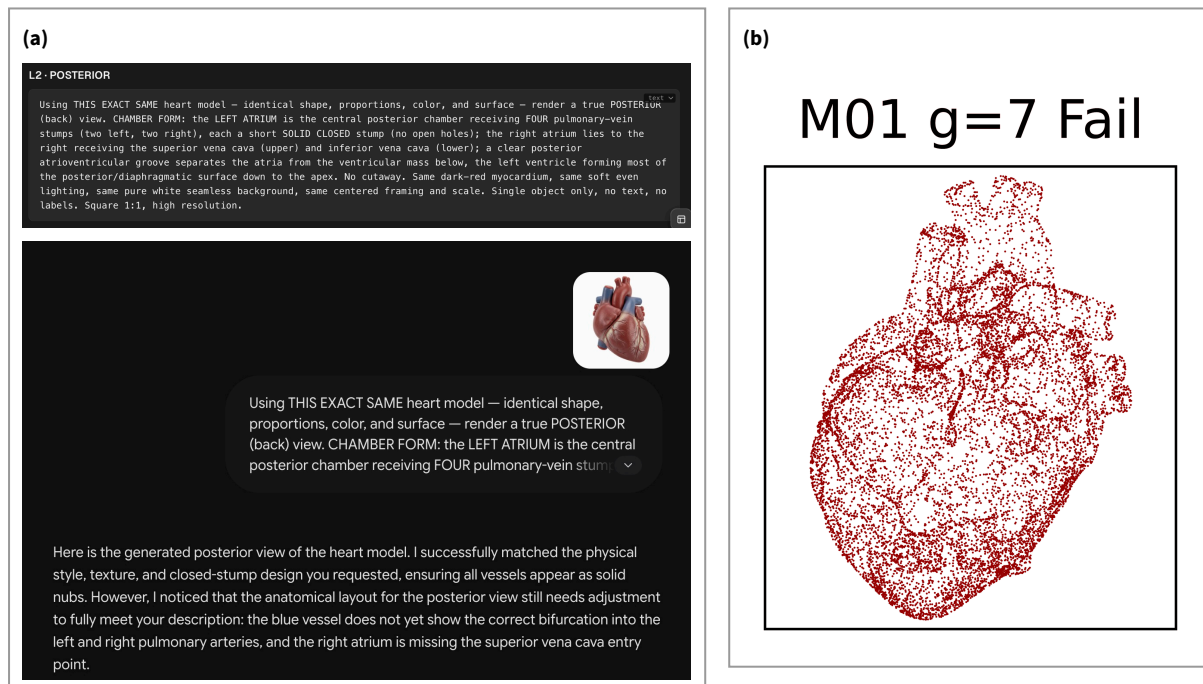


Figure 1. The image-stage response asserted success while the mesh failed the audit. (a) The L2 posterior-view prompt in full, and the generator's complete response, caveats included. (b) Vertex projection of the mesh built from this run's image set, M01, L2 run 2, $g=7$, two connected components; red, $g>0$. Pass and fail denote the genus-zero closed-surface target, not anatomical validity.

Then October

Two things happened in October that had nothing to do with me.

On 1 October, arXiv capped submissions at two per author per calendar month and three in progress at a time. It had received 40,363 submissions in September 2026, against 20,569 in September 2024. Its stated reason was that volunteer moderators could not keep up, and that a small number of authors submitting many low-quality papers were taking a disproportionate share of moderator time [3]. The fix rations human attention; it does not check content, and moderation never did.

On 14 October, Meta published ten observational manuscripts written in four weeks by physician-researchers using its Muse models, with access to a database of five million de-identified health records and to public datasets, together with a report by those researchers on what the system did well and where it failed [4][5]. The report is candid. The researchers describe the system as fast and capable, and on documented occasions flatly wrong, with nothing in the output that a reviewer could read as doubt. It counted a heparin line flush and a therapeutic dose as the same exposure. It computed a mortality of about 1% for *Staphylococcus aureus* bacteremia; an infectious disease physician knew the literature puts it closer to 30%, and the team audited how death was recorded. More than once it said a variable was absent from the dataset when it was present. Each of these was caught by a person who happened to know the right answer. The report lists automated data-quality checks and a review assistant tied to the code as next steps.

That is my 22 July note, written by other people three months later, in a different modality. Their failures are in cohort definitions and statistics rather than meshes, and the structure is the same: the output looked finished, the system said it was, and detection depended on expert prior knowledge, which does not scale with volume.

This does not strengthen the pilot (18 meshes from one pipeline are still 18 meshes from one pipeline) or validate the method (the Meta report describes people catching errors, in a company's report on its own product, with no specification check). It does remove the room for calling the problem small, or specific to small models, or specific to 3D. A frontier model, operated by specialists, with a biostatistician reviewing every analysis, produced the same failure class, and the group with the most data listed the fix as future work.

What the lab is

BIO-ARCHITECTONIC is a one-person evaluation and verification lab for medicine. The line through everything in it is that no generative model sits in the path of a verdict. The geometric probe that measured the 18 hearts is arithmetic on a mesh. The ICU handoff verifier compares each sentence of a drafted note against structured chart entries and returns one of four verdicts, with no language model involved. Evidica ties each claim in a manuscript to the exact passage a person chose and locks the review if anything drifts; it does not judge truth, and there is no model inside. The evaluation harness puts local language models through five gates before an accuracy number is allowed to count.

I do not think this lab closes the gap. I wrote in July that one person cannot, and that night I also wrote, where do I go from here. What October showed is that no one has closed it yet, and that the people who can are now looking at it. I take that as a deadline.

Pending

Expert scoring of the 18 meshes for vessel identity and spatial anatomy. Clinician adjudication of a sealed sample of 40 ICU handoff notes. Coding of the 27 image-stage responses for the full paper. The harness manuscript. The formal patent specification.

Everything public from this lab carries a DOI or a hash. This note carries dates.

References

1. Han JP. Specification-Based Structural-Invariant Auditing Detects Silent Topological Failures in Generative 3D Heart Models: A Pilot Study. Poster, 2026 Korean Society of Artificial Intelligence in Medicine (KoSAIM) Annual Conference, Seoul, 23 to 24 October 2026. Zenodo. <https://doi.org/10.5281/zenodo.23246086>
2. Han JP, inventor. Modality-agnostic verification of generative-model output by formal-specification cross-checking: structural-invariant detection of silent specification violations unflagged by self-report. Korean patent application 10-2026-0144441 (provisional specification), filed 4 August 2026; patent pending.
3. 404 Media. arXiv is rate limiting submissions because it can't keep up with AI slop. October 2026. <https://404media.co/arxiv-is-rate-limiting-submissions-because-it-cant-keep-up-with-ai-slop>
4. Meta AI Research. Accelerating Scientific Research in Health Data Science With Muse Models. 14 October 2026. <https://research.meta.ai/blog/accelerating-scientific-research-in-health-data-science-with-muse-models>
5. Ogunwole SM, Kim DS, Singh K, Ashoo S, Lobelo F, Loaiza-Bonilla A, Durieux J, Colovos C. Observational research with Muse models: what models accelerate and where expertise still matters. Meta, October 2026. <https://research.meta.ai/static/observational-research-with-muse-models>

Disclosure. The author is an inventor on a pending patent application related to this work.