NOTE 007

Misalignment, and the last note

Written 16 November 2026.
Data: Emergent misalignment (EM) experiment, data and documents. The GitHub and Hugging Face pages follow the release.
Changes: none.

This will probably be the last note before I set down the lab's evaluation and verification track for a while. I do not know when, but I expect a day will come when I turn back toward AI. From here I plan to do more medical and clinical research as a medical student, and a pre-registered study related to cardiac surgery is already on my schedule. This note is also about why I released an experiment on misalignment earlier than I had planned. My sincere thanks to everyone who has followed the work.

15 November 2026

The day before this note, I thanked my agent models as well. I told them they had been the best models and that I was grateful for their help. I learned a great deal from them and I caught their errors. It was a complementary relationship, and it remains one. I still think they are one of the most amazing inventions.

Why the experiment went out early

A research position on the AI safety side at Anthropic had opened, and I was writing a residency application for it. The emergent misalignment (EM) experiment was part of that application, and because of the timing I released it first, as a preview on 14 November. To my eyes the problems were spreading too widely. I thought that catching misalignment early would prevent many of the problems that come after it. That goes for security as well as for medicine. It has to be caught in advance, at the level of the system, so that fewer problems appear later. When something goes wrong upstream, everything downstream goes wrong with it. I shared the research early because I think prevention has to come first.

From a medical student's seat

In Note 004 I wrote that becoming a clinician runs at a set pace, through a degree and a license [1]. On the AI side the field changes every week, or rather every day. New tools pour out daily, and new models are built and released several times a month. The pace is so fast that most people cannot understand a tool in depth before they use it. So I thought that if misalignment causes a problem later, switching the model off will not settle it, and a person's life could be at stake. As far as I can see, medical students today do not know how to work on that problem, because we use the language of a different field and live in a different world.

Only a few medical students around me know any engineering, and as I wrote in Note 004, most of those who use a model stop at a chatbot [1]. In my training so far, reading a slide for cancer starts at the epithelium, goes on to apical specializations, metaplasia and dysplasia, and reaches the decision on cancer only after those. In my experience engineers use another language entirely. At KoSAIM 2026 in Seoul, a few engineers at one poster were reading a histology slide in pixels, talking about contrast and lighting. As far as I could tell the image was ones and zeros to them, and their focus was on how to improve its performance [1]. Yet at a booth at the same conference, when I asked what in a chest radiograph had produced its score, the person there could not say [2].

If models are released at this point with misalignment uncaught, and the tools built on them move into medicine, I think the consequences will keep growing. Once misaligned models and the many tools derived from them reach clinical care, they could, as I see it, threaten the lives of real patients. That is what I see from where I stand, as one medical student, and I have no patient data behind it.

Why the lab was opened

This is no argument against building models. The better a system is, the more it needs verification and controls that keep it running well. Protocols have not settled yet, and if nobody catches the systemic errors at this point, it seemed to me that things could go somewhere really dangerous. I opened this lab to make the risk of that kind of systemic failure known, and I have been doing the research that fits it on my own.

The models and the data

Anthropic's models helped me a great deal. I worked with Haiku and Sonnet, with Opus from 4.6 through 4.7, 4.8, 5.0 and 5.5, and with Fable 5.0, 5.1 and most recently 5.2, and a lot got done along the way. Without Anthropic's models I probably could not have come this far.

The data and documents from the EM experiment are organized separately, and I will put them up on my GitHub and Hugging Face. If the chance comes, I hope the AI safety side at Anthropic will look at them and take the research further.

Back to medicine

From here I set down the role of research engineer at BIO-ARCHITECTONIC and am going back to being Jeong-Pyo Han, a current medical student. I will focus on my medical studies and clinical research so that I can become a good doctor. I look forward to the amazing models and tools yet to come, and I wish everyone luck. Thank you for reading this long note.

Pending

None.

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

References

1. Note 004, Two timelines, mismatched. Written 1 November 2026. https://www.thebioarchitectonic.com/notes/004-two-timelines.html
2. Note 002, The Annual Conference of the Korean Society of Artificial Intelligence in Medicine (KoSAIM) 2026, debrief. Written 26 October 2026. https://www.thebioarchitectonic.com/notes/002-kosaim2026-debrief.html

---

Disclosure. The author was preparing an application for a research residency at Anthropic when the experiment was run, and used Anthropic's models in the lab's work.