A Guide to the Histopathology Pilot

A walk-through of my ongoing research, written to hopefully avoid most jargon. Last updated 05-25-2026.

The two-minute version

I built a different kind of memory system for AI. Instead of training a giant black-box model that "knows" things in some unknown way, it stores information as searchable patterns and lets the AI recall the most similar past examples to whatever it's looking at. Sort of how the brain works using a cue to remember things.

In early May 2026, I tested it on colon-cancer microscope images from a public medical research dataset. I first ran it across 49,995 tissue samples using two independently trained pathology AIs (one from Owkin in France, one from Harvard's Mahmood Lab) and the dual-filter caught every single cancer. On May 23, I re-ran the same approach at almost double the scale (93,522 tissue samples including 11,111 cancers) using **four** independently trained encoders. **Every sample the system released as "safe to call benign" was actually benign — 27,927 of 27,927.** Zero false reassurances at clinical scale. The two cases where the system would otherwise have been uncertain were correctly routed to "ask a human," not silently released. The system is honest about what it doesn't know.

I developed and ran this on a system with a consumer-grade graphics card. It works there right now. My end-goal deployment target is an ASIC (application-specific integrated circuit) or FPGA (field programmable gate array) either of which would be purpose-built silicon so that the same logic can live in a sealed medical-device-grade box at the microscope, drawing very little power. The math fits because our signatures are 1024 bits and the retrieval operation is XOR + popcount, which is exactly what FPGAs do natively in hardware.

What we're trying to do

Most modern AI works by training huge "neural networks" with billions of internal numbers that are tuned over weeks or months until the program does something useful. The problem is, nobody can really *look inside* one. If a hospital AI says, "this looks like cancer," doctors and regulators want to know *why*, and "because the 47 billion numbers said so" doesn't really cut it.

My approach is different. **The system does no training of its own.** It stores memories the way a biological brain seems to, using patterns of activity that can be re-activated when something similar comes in as a cue. When you show our system a new image (the cue), it doesn't run a black-box classifier, it pulls up the *most similar things it's seen before* and shows you those, with their known labels attached.

The encoders that turn an image into a pattern are pre-trained pathology models (Phikon-v2 from Owkin, UNI from Harvard's Mahmood Lab) used in inference-only mode which means we don't adjust their weights (the numbers that relate pieces of information to each other learned by training). My system's retrieval logic is deterministic comparison. Adding a new labeled case to the system is just adding to the library so there's no retraining required.

Why does this matter for medicine? When my system says, "this microscope slide looks like cancer," it can also say *"and here are five real labeled examples it reminds me of the most."* A pathologist can look at those five examples and confirm or disagree. That's interpretable medical AI of the kind regulators and doctors can really trust.

A quick glossary of terms for the next section

- **Patch.** A small tile cropped from a microscope slide of tissue. Each one is 224 x 224 pixels or roughly the size of a postage stamp on the original slide. A full slide gets chopped into thousands of patches so we can study small regions.
- **H&E staining.** The standard pink-and-purple dye combo technicians and doctors use in pathology. Hematoxylin stains cell nuclei dark purple, Eosin stains everything else pink. “H&E” images are what virtually all hospital pathology starts with.
- **Encoder.** A pre-trained AI model whose specialty is turning an image into a list of numbers (called an “embedding”) that captures what’s in the image. We use medical-specialty encoders trained on millions of pathology images.
- **Sign-bit signature.** My system’s specific representation. Take the encoder’s list of numbers, look at each one, and write a 1 if it’s positive or a 0 if it’s negative. This compresses the encoder’s output down to a tiny, fast-to-compare fingerprint of exactly 1024 bits per image in our setup.
- **Hamming distance.** A fancy name for “how many bits differ between two signatures.” Two identical signatures have distance 0 while completely opposite ones have distance 1024. It’s cheap to compute and can instantly compare millions of images.

Experimental Setup

The data. I downloaded a public colon-pathology dataset called NCT-CRC-HE-100K, which has 100,000 patches of colon tissue, each labeled by a real pathologist as one of nine types of tissue. Eight of them are non-cancer (fat, blood, debris, lymphocytes, mucus, smooth muscle, normal mucosa, stroma) and one is cancer (tumor).

The initial run used a stratified 49,995-patch subset (equal numbers from each tissue class, including 5,555 cancer patches). The follow-up 4-encoder run used a larger 93,522-patch sample that includes 11,111 cancer patches — close to the full dataset and big enough to test the system at clinical scale.

The encoders. I used four independent pathology AIs across the two runs:

- **Phikon-v2** (used in both runs): open-access, from a French research lab called Owkin
- **UNI** (used in both runs): gated-access, from Harvard’s Mahmood Lab
- **Virchow2** (added for the 4-encoder run): from Paige AI with Microsoft Research, larger and more recent
- **CONCH** (added for the 4-encoder run): a vision-language pathology model, also from the Mahmood Lab

Each is trained on millions of microscope slides. Each turns a patch into a 1024-number list, which we compress to a 1024-bit signature. These four encoders were trained by different teams (or by the same team using different methods and different data), so when they *agree*, that agreement carries real cross-validation weight. When they disagree, the system defers to a human instead of guessing.

The test (leave-one-out retrieval). For each patch in the dataset, we pretended we didn’t know what it was, then asked: *“of the other patches in the library, what are the most similar?”* (We used the top 5 per encoder.) Then we cross-referenced what the encoders independently retrieved. This is a standard, honest way to test recall (you can’t cheat because the system has never been told the answer for the patch you’re querying).

The dual-filter logic. Every patch gets sorted into one of four bins:

- **RELEASE_BENIGN:** both AIs agreed on a non-cancer class, neither found any cancer in the top 5 nearest, both had tight matches. System says "confidently benign."
- **TUMOR_FLAG:** at least one AI said cancer, or both AIs had cancer in their top 5. System says "this needs cancer review."
- **HUMAN_IN_LOOP:** the system isn't confident either way. Defers to a pathologist.
- **CONTRADICTION:** both filters fired (shouldn't happen if the logic is well-defined).

The results in plain English

The headline (4-encoder dual-filter on 93,522-patch dataset)

- **Out of 11,111 actual cancer patches: every single one was caught.** Tumor sensitivity = 100.00%. Zero missed cancers.
- **Out of 27,927 patches the system released as benign: every single one was actually benign.** Release-tier specificity = 100.00%. Zero false reassurances at clinical scale.
- **Of the 11,316 patches the system flagged for cancer review, 98.2% were correct.** Only 207 false-flags (1.8%) — a low false-alarm rate on the bin that triggers human cancer review.
- **The architectural-honesty point:** two cancer cases that the system couldn't confidently classify went to "ask a human," NOT to "released as benign." When the four encoders disagreed, the system deferred. This is the failure mode you want — uncertainty surfaces as human review, not as silent false reassurance.
- **58% of patches went to "human review."** The system explicitly defers when consensus isn't there. This is the *right* behavior, not a failure. In real clinical practice, a pathologist reviews everything anyway; the system's job is identifying which cases the AI is confident about, freeing the pathologist's attention for the ambiguous ones.

Tier breakdown at a glance

Tier	What it means	Patches	Share	Real cancers in tier	Real benign in tier
RELEASE_BENIGN	All 4 AIs confident this is non-cancer	27,927	29.9%	**0 (0.0%)**	27,927 (100.0%)
TUMOR_FLAG	At least one AI signaling cancer	11,316	12.1%	11,109 (98.2%)	207 (1.8%)
HUMAN_IN_LOOP	AIs disagree; defer to pathologist	54,279	58.0%	2 (0.0%)	54,277 (100.0%)

Following the safety logic: the RELEASE_BENIGN row has zero real cancers, the TUMOR_FLAG row catches the overwhelming majority of real cancers (and the small benign minority that gets flagged just routes to pathologist review), and the HUMAN_IN_LOOP row contains the two ambiguous cancers that the system correctly refused to silently release.

From 2 encoders to 4: the consensus tightened, the safety statistic held

The initial 2-encoder run (Phikon-v2 + UNI on 49,995 patches in early May) already hit 100% tumor sensitivity and 100% release-tier specificity. Adding Virchow2 and CONCH for the 4-encoder run was the architectural stress-test: would adding more encoders break the consensus, or tighten it?

- **The safety-critical statistic held.** Both runs: zero real cancers in the "released benign" bin.
- **The consensus tightened.** The 4-encoder run is *more conservative* about what it auto-releases. 29.9% of patches were released as benign by the 4-encoder system (vs ~44% by the 2-encoder system). Adding encoders raised the bar for "everyone agrees this is safe," exactly as it should.
- **Higher cancer-flag precision.** When the 4-encoder system flags a patch as cancer, it's correct 98.2% of the time. The disagreements where one encoder sees cancer and others don't get routed to human review instead of being silently flagged.

Per-encoder behavior in the 4-encoder run

Each encoder has its own "tightness" — how close the average patch's nearest neighbor sits in 1024-bit signature space. Lower numbers mean tighter matches.

Encoder	Source	Median nearest-neighbor distance (fraction of 1024 bits)
CONCH	Mahmood Lab	0.078
Virchow2	Paige AI with Microsoft Research	0.102
Phikon-v2	Owkin	0.132
UNI	Mahmood Lab	0.169

CONCH was the tightest matcher (newest and most narrowly-tuned for histopathology). Virchow2 was second-tightest (largest model). Phikon-v2 and UNI were the historical baseline. None of them is "best" alone — the value is that they make different mistakes, so the consensus among them is the actually-trustworthy signal.

A Confusion Matrix Reading

The confusion matrix is essentially diagonal meaning we made a 9-by-9 table where rows are the true labels and columns are what the AI guessed. If the AI were always right, all the numbers would be on the diagonal line from top-left to bottom-right. Ours basically were, except for a handful of close-call cases between tissue types that even trained pathologists sometimes disagree about. And our system flags them for the pathologist to examine.

Why this is interesting to an investor

1. **Zero training. Pure case-based retrieval.** We do not train any model on the cancer data. The pathology encoders (Phikon-v2, UNI) are pre-trained by their original creators and used in inference-only mode. We never adjusted a single weight. The sign-bit signatures and the dual-filter logic we use are both deterministic. Adding a new labeled case to the system just means writing it to the memory cartridge; no retraining cycle, no GPU days, no model versioning. A pathologist who annotates a difficult case in the morning has it as a retrieval reference by lunch. This is dramatically different from traditional medical AI that

trains a CNN (convolutional neural network) for 6-12 months, then has to retrain it every time the data drifts or a new diagnostic class is added.

2. **Zero false reassurances on 27,927 released-benign patches across 11,111 real cancer cases.** Catastrophic false negatives — the single statistic that matters in clinical AI — was zero. Every cancer was caught and routed to either an explicit cancer-flag tier or a human-review tier. The two ambiguous cancers that the system wasn't confident about went to "ask a human," not to silent release. The architecture is honest about uncertainty.
3. **Interpretability.** Every output is grounded in real, labeled examples the system can show you. No black box. A pathologist can look at the actual nearby cases and confirm. Regulators should love this.
4. **Hardware accessibility.** The system was developed on and works with a consumer-grade GPU (RTX 4080 Super) *today*. It took 12 minutes to encode 50,000 patches and 5 minutes to run the dual-filter analysis. Going forward, the end goal we're targeting is making an ASIC or using an FPGA (specifically Oregon-based Lattice Semiconductor's AVANT series) for a sealed medical-device deployment. The math is FPGA-friendly because our signatures are 1024 *bits* and Hamming-distance retrieval is XOR + popcount which is exactly what FPGAs do natively in hardware. The substrate has been designed from Day 1 with that hardware path in mind, so the port should be relatively mechanical when the time comes.
5. **Scales the way we'd want.** An earlier smaller test at about 5,900 patches showed similar accuracy. Going up by ~8x didn't degrade it; the cross-encoder dual-filter improved it. The approach holds as we get bigger.
6. **The mistakes are honest.** The 45 over-flags in 44,440 non-cancer patches were almost entirely on tissue types that even trained pathologists find ambiguous (debris, stroma, mucus). The system is uncertain where humans are uncertain, not on easy cases.
7. **What the 4-encoder run added.** As of May 23 we've added Virchow2 (from Paige AI with Microsoft Research) and CONCH (from Harvard's Mahmood Lab) to the original Phikon-v2 + UNI dual-filter, running the full pipeline on 93,522 patches (close to the whole NCT-CRC-HE-100K dataset). The architectural questions were: does the consensus *break* when you add more encoders, or does it *tighten*? Does the safety statistic hold at clinical scale? Both questions came back the right way. The safety-critical 100% (zero real cancers in the released-benign tier) held. The consensus tightened — the 4-encoder system is more conservative about what it auto-releases, exactly as adding more independent voters should make it. The two cancers that the 4-encoder system wasn't confident enough to flag went to human review, not to silent release. This is the architectural-honesty result: more encoders = more honest about uncertainty, without giving up sensitivity. The "more is more" story holds; the system is ready to scale.

What about new hospitals? The cross-hospital test we ran

The 100/100 results above are *within-distribution*, meaning the reference library and the test queries came from the same kinds of slides. Real clinical AI must survive *out-of-distribution* (OOD) data: slides from a hospital the reference library has never seen, with different staining labs, different scanners, different patient demographics. That's what seeing new fresh biopsy data will be like.

We ran exactly that test on the TCGA-UT dataset, which has a built-in "external split" designed for cross-institutional generalization. The reference library was built from 47,240 patches from one set of hospitals; the queries were 9,870 patches from a completely different set of hospitals.

Results:

- Cross-hospital overall: **87% top-1 retrieval.** (Down from 100% within-distribution.)
- **95% top-5** which is much more robust if a pathologist is looking at the top 5 nearest cases.
- Brain cancer held strongest: **97%**.
- Skin cancer (a MOHS-relevant case): **dropped to 67%** which is the weakest cancer type results in this test.
- When two independent encoders (Phikon-v2 + UNI) agree on a cancer type, they're correct **91%** of the time.
- The system honestly flagged **88%** of the cross-hospital queries as "I haven't seen anything quite like this before."

That last point is critical. The system isn't pretending to be confident on data it hasn't seen, instead it is flagging uncertainty, exactly as designed.

Why same-site historical slides help same-site live slides

There's a non-obvious question buried in the cross-hospital result: if the system drops from 100% to 67% on skin cancer when queries come from new hospitals, why would a hospital adding a few hundred of its own labeled cases fix it? Aren't those just *more* same-distribution data, not actually-new data?

Here's the mechanism. A pathology encoder doesn't just capture *biology* like the nuclear morphology, tissue architecture, or other features a pathologist looks for. It also captures *technical artifacts* like the specific shade of pink/purple from a particular staining lab, the color profile of a specific scanner brand, slide thickness, focus characteristics, etc. The 1024-number signature Phikon-v2 produces compresses BOTH biology and technical signature together.

So, two patches with identical underlying biology, stained at two different labs on two different scanners, end up with slightly different signatures. Hamming distance between them is inflated, not by biological differences, but by lab and scanner differences. That's why cross-hospital retrieval drops: the reference library and the queries encode different technical artifacts.

A hospital's new live samples will:

- Come from the same staining lab as their historical slides
- Be scanned on the same scanner
- Follow the same tissue handling protocol
- Come from a similar patient population (same catchment area)

So new live samples share *technical artifacts* with the hospital's own historical labeled cases. When the reference library contains the hospital's own slides, a new live query's nearest neighbors are more likely to come from the same hospital--and those neighbors have correct labels. Retrieval re-aligns.

The pathologist-mobility analogy. Imagine a pathologist who trained at Hospital A and moves to Hospital B. Their *trained expertise*, what cancer cells actually look like, transfers. But Hospital B's slides look slightly different, say with different pinkness, sharper edges, paler nuclei. They spend a few weeks recalibrating. Cart-extension is exactly that recalibration. The encoder's general pathology knowledge stayed the same, only the reference library got site-tuned.

What this does NOT solve (kept honest):

- **Frozen sections vs paraffin-embedded slides.** Hospitals doing MOHS use frozen sections (fast turnaround during surgery) so their archival slides are mostly paraffin-embedded (slow, standard processing). These look meaningfully different even from the same hospital. The reference library needs frozen-section reference cases to cover live MOHS samples, not just any archival slides.
- **Truly rare presentations.** If the hospital has never seen a particular subtype before, their reference library can't help retrieve it. They'd need cross-site sharing, or fall back to the broader public reference library, with the system honestly flagging low confidence.
- **Data drift over time.** Staining batches change, scanners get serviced, lab technicians vary. The reference library needs occasional refreshes. This is the same data-drift problem trained models have, except but the fix here is "add new labels," not "retrain and re-certify the model" for lots of time and money.
- **Labels must be reliable.** Historical slides must be correctly annotated by pathologists. Garbage labels produce garbage retrieval. A quality-controlled annotation pipeline is part of any real-world deployment.

Honest caveats

- These are leave-one-out retrieval numbers on a single, well-curated, public dataset. They are NOT a clinical study. A real medical product would need prospective testing on data from multiple hospitals, with regulatory oversight.
- The 9 tissue classes in this dataset have clean labels whereas real clinical practice has much messier labels and edge cases. We expect our numbers to be lower (not zero, but lower) in a real-world deployment.
- We're testing the *substrate*, our underlying memory technology. The clinical product wrapping this would still need radiologist-grade UI, pathologist workflow integration, quality assurance pipelines, all the usual medical-software requirements.
- The pitch we're building is: "this technology can be the *engine* for medical AI applications that are interpretable, hardware-cheap, and scalable." Not "we built a cancer detector."

Glossary (for re-reference)

- **Memory Cart:** Our format for storing a collection of memorable patterns (text, images, audio, whatever). Like a video-game cartridge: self-contained, swappable, portable.
- **Confusion matrix:** A table showing what the AI guessed vs. what was actually true. Perfect AI = all numbers on the diagonal.
- **Dual filter:** Our approach of using two (or four) independent AIs and only releasing a result if they agree. Cuts false alarms while preserving sensitivity.
- **Embedding:** The list of numbers an AI produces when looking at an image (or any input). Captures "what's in the image" in compressed form.
- **Encoder:** The AI that turns an image into an embedding. We use medical-specialty encoders.
- **H&E:** Standard pink-and-purple staining used for almost all pathology microscopy.
- **Hamming distance:** How many bits differ between two signatures. Our retrieval primitive.
- **Leave-one-out:** Standard fair-test method: hide one item, ask the system to find similar things, check if it succeeded. Repeat for every item.
- **Patch:** Small tile of a microscope slide, 224 × 224 pixels.

- **CONCH:** A vision-language pathology encoder from Harvard's Mahmood Lab; the tightest matcher in the 4-encoder run.
- **Phikon-v2:** The first AI encoder we tested (open-access, French lab Owkin). Used in both the 2-encoder and 4-encoder runs.
- **Sensitivity:** Fraction of real cancers correctly flagged as cancer. "Did we catch it?"
- **Sign-bit signature:** Our 1024-bit fingerprint per image. Compressed encoder output.
- **Specificity:** Fraction of non-cancers correctly NOT flagged. "Did we resist crying wolf?"
- **Substrate:** Our underlying memory technology; the thing that holds and retrieves patterns. Distinct from any particular application.
- **Top-K accuracy:** Did the correct answer appear in the top K (1, 3, 5, 10) nearest neighbors?
- **UNI:** The second AI encoder we tested (Harvard's Mahmood Lab, gated access). Used in both the 2-encoder and 4-encoder runs.
- **Virchow2:** A more recent pathology encoder from Paige AI with Microsoft Research (largest of the four); added for the 4-encoder run.

Where this fits in the bigger picture

This colon-cancer result is one experimental leg of a broader research program at Waving Cat Learning Systems. Other legs include language understanding (we have a 370,000-word dictionary that maps to the same signature system), self-organizing knowledge networks, and a memory layer for AI agents (Membot/Mempack with demos already online). The medical AI angle is the most easily demonstrable wow-factor application, which is why it anchors the investor pitch.

The deeper claim is that we've built a way for AI to remember things in a structured, inspectable, biologically inspired way. And that everything from cancer detection to language generation to agent memory can use the same underlying engine. The colon-cancer result is evidence that the engine works at scale on at least one real, high-stakes domain. The other legs are evidence that the engine isn't just a one-trick pony.