Subject: Re: Neighbor Inspection "Updates"
Date: Wednesday, July 16, 2025 at 9:50:24 AM Central Daylight Time
From: Martin Paulus
To: Mikey Ferguson
Attachments: image001.png

Here is what my GPT friend suggests:

Below is a concise, "post-mortem + action plan" written at the level I would expect in a lab notebook or methods supplement.

---

1. Why the two neighbor-quality metrics diverge

Symptom	Likely Explanation	Quick Diagnostic
Near-zero correlation between LLM relevance score and Mahalanobis distance	*Representation mismatch*: Mahalanobis assumes Euclidean structure and Gaussian spread in the raw embedding space, while the LLM is judging similarity in **narrative-semantic** space after a lossy text-generation step.	Project a random sample of embeddings with t-SNE/UMAP, color by LLM score. If high-scoring points are scattered rather than forming a tight cluster, the mismatch is real.
High p-value despite many comparisons	Effective sample size is low: (a) few visit sequences; (b) heavy tail in the distance distribution inflates variance; (c) LLM scores are bounded (1-10) and likely skewed.	Plot distribution of Mahalanobis distances and LLM scores; check skewness/kurtosis.
LLM relevance weak overall	Prompt leakage: narrativization may omit key structured features (labs, meds, timeline) that the LLM needs to assess relevance; or prompt length truncates info.	Prompt ablation—drop labs vs. meds vs. diagnoses and see which omission collapses the score.
Mahalanobis instability	Covariance matrix poorly estimated in high-dim space with limited neighbors; susceptible to collinearity among term-level dimensions.	Compare classic vs. **Ledoit–Wolf shrinkage** covariance; watch change in distances.

---

2. Immediate "no-code" fixes

1. **Calibrate both metrics on a sanity set**
   - Curate 30–50 visit sequences with *known* clinically similar and dissimilar cases (e.g., same ICD chapter, same dominant med class).
   - Expect high LLM score and small Mahalanobis distance for true positives. If neither metric discriminates, the problem is upstream (embedding or narrativization).
2. **Separate *narrative* from *embedding* evaluation**
   - Instead of narrativizing each neighbor independently, feed the *structured* neighbor facts plus the target's narrative into the LLM and ask, "How similar?" This isolates the scoring component.
3. **Robust distance**
   - Swap Mahalanobis for **cosine** or **fractional L-p norms (p<2)**, which behave better in sparse, high-dimensional term embeddings.

---

3. Embedding improvements (next code cycle)

Layer	Current	Proposed upgrade	Rationale
Token	Bag-of-term vectorizer	*ClinicalBERT* (or **BlueBERT-mimic** if PHI concerns) for codes, plus one-hot for Labs/Rx	Captures contextual semantics; avoids synonym explosion.
Visit	Sum/avg of token vectors	Transformer with positional encoding on event time	Preserves chronology and inter-event interaction.
Sequence	Simple concatenation	Hierarchical attention network (visit→sequence) or GRU with time-gap embeddings	Handles variable visit counts; lets model weight salient visits.

Add Dr. Paulus' synthetic-synonym augmentation as noise injection during pre-training, not at inference—forces embedding to collapse true synonyms.

---

4. Evaluation redesign

1. **Cluster-level metrics**
   - Silhouette score, Davies–Bouldin, or intrinsic **Neighborhood Hit@k** using diagnosis label heuristics.
   - Compare before/after embedding upgrade.
2. **Extrinsic task**
   - Predict 30-day readmission or new-onset diabetes from embeddings. If upgraded embedding helps, neighbor metrics should also improve.
3. **Statistical power check**
   - Monte-Carlo simulation of correlation under plausible effect sizes to set sample size for the neighbor study.

---

5. Roadmap (two-week sprint)

Day	Task	Deliverable
1-2	Curate sanity dataset (50 target + 10 neighbors each)	JSONL file
3-4	Run calibration experiments; visualize distributions	Jupyter notebook
5-6	Implement synonym augmentation pipeline	Augmented vocabulary, unit tests
7-9	Replace Mahalanobis with cosine + robust covariance; rerun neighbor inspection	Updated plots, effect sizes
10-11	Integrate visit-level transformer embedding	Checkpointed model & inference script
12-14	Evaluate on extrinsic task + neighbor metrics	Short report & slide deck

---

Bottom line
The weak correlation is almost certainly a **comparing-apples-to-oranges artifact**: one metric operates in raw vector space with dubious Gaussian assumptions; the other in a narrative, semantically enriched space. By upgrading the embedding to capture temporal context, calibrating the evaluation, and using distance metrics better suited to sparse clinical data, you should see convergence—or at least interpretable divergence—between neighbor-quality measures.

The Art of War: In the midst of chaos, there is also opportunity.
Martin

Scientific Director and President, Laureate Institute for Brain Research
6655 S Yale Ave Tulsa, OK 74136-3326 P 918 502 5120 F 918 502 5135
email: mpaulus@laureateinstitute.org X:@mpwpaulus
web: http://www.laureateinstitute.org

Professor of Neuroscience, Oxley College of Health and Natural Sciences
University of Tulsa, email: mpp4692@utulsa.edu

Adjunct Professor, Department of Psychiatry, University of California, San Diego
email: mpaulus@health.ucsd.edu

From: Mikey Ferguson <MFerguson@laureateinstitute.org>
Date: Wednesday, July 16, 2025 at 9:32 AM
To: Martin Paulus <mpaulus@laureateinstitute.org>, Rayus Kuplicki <rkuplicki@laureateinstitute.org>
Subject: Neighbor Inspection "Updates"

Hi Dr.'s Paulus and Kuplicki,
This morning, I got back the results from the job I submitted yesterday on the neighbor inspection. The good news is that the code is working, so with any luck further modifications should go relatively smoothly. Not surprisingly, the results are not very illuminating. Just as a quick review, we are using two metrics to judge how good the neighbors of a patient visit sequence are:

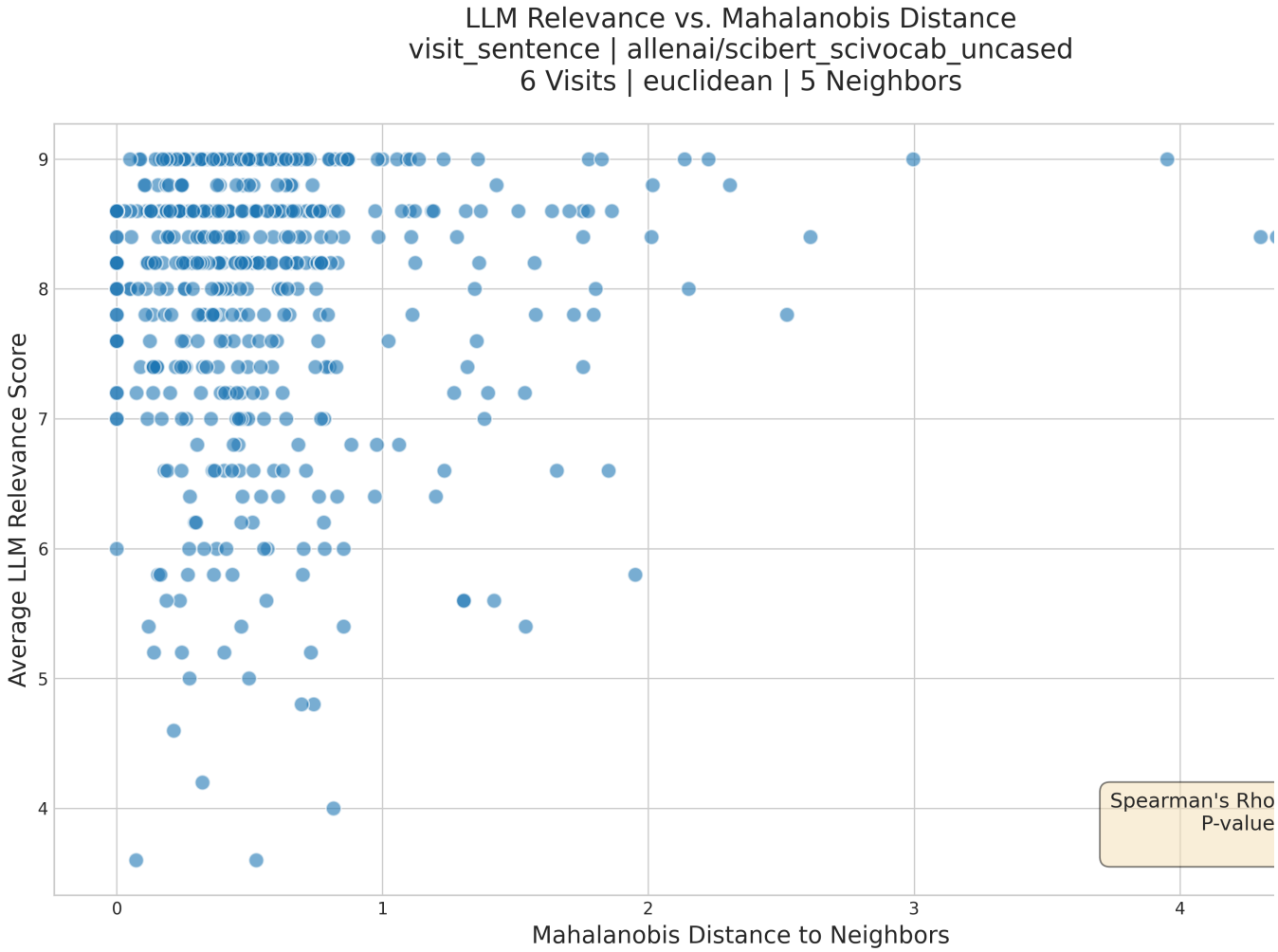
- Generate a narrative for the patient visit sequence of interest and generate a narrative for all its nearest neighbors. Ask an LLM how relevant each of those neighbors are to the visit sequence of interest (scale of 1-10) and take the average value. This gives us the first metric for each patient visit sequence of interest.
- Compute the Mahalanobis Distance between the patient visit sequence of interest (as a vector) and the distribution of all its neighbors (as vectors). This gives us the second metric for each patient visit sequence of interest.

The results saw very weak correlation between these two metrics and such a high p-value that the results are not remotely statistically significant. I just used the vectorizer that I've been encoding medical terms with, so again I'm hardly surprised by the underwhelming results.

I'm still waiting for the job on the term-embedder to finish – hopefully, I'll have some cosine similarity graphs soon like I showed in yesterday's lab meeting. That version of the term embedding job does *not* incorporate the "synthetic synonym" generation that Dr. Paulus suggested I use instead of just asking an LLM to produce a synonym.

So, at this point, I have two plans for my immediate next steps. For the term embedding cosine similarity vectorizer exploration, I plan on implementing what Dr. Paulus suggested regarding the synonym term generation. For the neighbor inspection, I plan on making the embedding method for a patient visit sequence more robust – last week Dr. Paulus sent Dale and me a strategy to embed patient visit sequences, and I plan on following that for when I run the next neighbor inspection job.

I'm open to any other suggestions and am happy to try to clarify any questions or confusions this message may have introduced.
Thanks!
Mikey



Get Outlook for Mac