What a Synthetic Stand-In Hid: Decomposing an Inflated Eye-Tracking Benchmark of Cognitive Load
William Garrow
Boston University, Boston, MA · Georgia Institute of Technology, Atlanta, GA · wgarrow@bu.edu

On the same pipeline, the synthetic stand-in scored a third higher than the real corpus.

The stand-in erased the participant correlation, so a leakage audit run on synthetic data passes whether or not the real data leaks.

The question

A synthetic stand-in is usually validated by showing that its descriptive statistics match the source. In 2025 I could not open the COLET corpus, so my paper said the stand-in preserved those statistics, and nobody, including me, could compare the two. Privacy rules and scarce recordings put many studies in the same spot. Once the corpus was open, I compared them.

What I did

In 2025 I built the proof of concept, a classifier trained on a synthetic stand-in because the corpus ships as MATLAB v7.3 tables that no standard Python reader could decode. This year I wrote the decoder (mat73-reader, on PyPI) and opened the real recordings. Then I ran that pipeline, unchanged, on the stand-in and on the real corpus, with ungrouped and with participant-grouped 5-fold cross-validation. That is the 2×2 in the center.

The corpus

COLET [1] holds 47 participants in four visual-search conditions, 188 labeled task instances in all. Each task ended with a raw NASA-TLX [2] (RTLX, the unweighted mean of six subscales), and my primary label is the top tertile of that mean. The 2×2 keeps the proof of concept's ten features: gaze dispersion, pupil diameter [4, 5], gaze velocity, and the fixation and saccade rates. The honest protocol adds two blink features, twelve in all, z-scored within each participant.

WHAT I FOUND

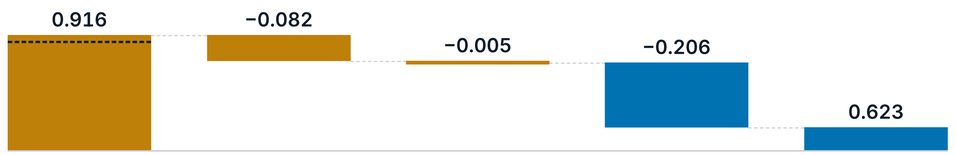
1. Most of the inflation was structure the generator fabricated, some was the luck of one split, and almost none was leakage.
2. The checks that caught it are standard and need no new data: grouped validation, a label sweep, and a negative control, all run on the real data [6, 7].
3. With one corpus and one generator, n = 188, I have shown that this can happen and have not measured how often it does.

THE RESULT

	Ungrouped	Participant-grouped
Synthetic stand-in	0.834 (0.020)	0.829 (0.046)
Real corpus, tertile label	0.626 (0.121)	0.623 (0.089)

Every cell is macro F1, mean (SD) across five folds, on the same pipeline and features. The proof of concept replicated exactly on its own split first: positive-class F1 0.897, macro 0.916.

Where the 0.897 came from



The dashed line is the reported 0.897 in positive-class F1, and the bar is the same split in macro F1. The two metrics differ by 0.02 to 0.15 across these cells. The dominant term is structure the generator fabricated, and split leakage contributed almost nothing.

REAL RECORDED DATA



Thirty seconds of participant 5's gaze in each task, with nothing simulated. Dot size is pupil diameter at the last sample.

The honest protocol

The 2×2 kept the 2025 pipeline as it was. Here I z-scored each feature within the participant, added two blink features, ran five participant-grouped folds over five seeds, and reported macro F1 with its SD.

A twelve-coefficient logistic regression matched or beat three tree ensembles.



Six subscales, one negative control

Mental demand scores highest, with effort within one SD of it. Physical demand sits at chance, as a model of cognitive load [3] should.



The label sweep

The score moves with where TLX is cut into high and low, so I fixed the top tertile before running the sweep.



The stand-in, measured

	Real corpus	Stand-in
Participant ICC, mean pupil diameter	0.848	0.008
|r| with the label, ten features	0.085 to 0.361	0.455 to 0.675
SD of the label, mean TLX	19.47	9.45

The 2025 generator drew every feature and subscale from one hand-set difficulty term, with no participant term and no corpus statistic behind it. A stand-in fitted to the real means, SDs and correlations scored 0.555 (0.048, 25 draws), below the corpus's 0.623 on the same pipeline, so the moments carry none of the inflation.

CORRECTION, 2026-08-30

The accepted abstract reports 0.762. My decoder, mat73-reader, turned every empty blink table into two phantom rows and contaminated the blink features in 40 of 188 instances. An independent decoder disagreed on those tables and no others. I fixed the decoder and rebuilt the features with the protocol unchanged. Every cell without blink features reproduced, and 0.762 became 0.783. The audit of the stand-in needed the same audit on my own decoder.

REFERENCES

1. Ktistakis E, Skaramagkas V, Manousos D, Tachos NS, Tripoliti E, Fotiadis DI, Tsiknakis M. COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. Comput Methods Programs Biomed. 2022;224:106989. doi:10.1016/j.cmpb.2022.106989
2. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Adv Psychol. 1988;52:139-183. doi:10.1016/S0166-4115(08)62386-9
3. Sweller J. Cognitive load during problem solving: Effects on learning. Cogn Sci. 1988;12(2):257-285. doi:10.1207/s15516709cog1202_4
4. Kahneman D, Beatty J. Pupil diameter and load on memory. Science. 1966;154(3756):1583-1585. doi:10.1126/science.154.3756.1583
5. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol Bull. 1982;91(2):276-292. doi:10.1037/0033-2909.91.2.276
6. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. GigaScience. 2017;6(5):gix019. doi:10.1093/gigascience/gix019
7. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. ACM Trans Knowl Discov Data. 2012;6(4):15. doi:10.1145/2382577.2382579

THE COMPANION

cognitive.williamgarrow.com

The replay in motion, the model running in your browser, full references, the numbers ledger, the protocol, and the correction's before-and-after files. Thanks to the COLET authors for releasing the corpus.