Foreshock

WHITE PAPER · ENGINE ARCHITECTURE

Aldous 1-2.0

A zero-shot semantic telemetry and guardrail engine built on DECE, Diagonal Emotional Covariance Estimation. It operates purely mathematically and never reasons about your text.

DOCUMENT
Engine Overview · v1-2.0

METHOD
DECE · Multivariate Gaussian

STAGE
Research pre-release

LICENSE MODEL
Open source · Apache2 deps

REPOSITORY
github.com/splinterhq/Aldous

ABSTRACT

Aldous defines each semantic concept as a small population of diverse phrases, embeds each phrase independently, and computes the concept's geometric centroid and dimensional variance in latent space. A concept becomes an independent multivariate Gaussian distribution rather than a single point. At inference, Aldous measures incoming text against these distributions using Standardized Euclidean Distance, Cosine Similarity, and Dot Product, in constant time, with no generative inference and no string matching. Outputs are similarity and distance floats only.

00 CONTENTS

01 FOUNDATIONS

What is DECE

Diagonal Emotional Covariance Estimation · pronounced "dee-see"

DECE defines a semantic concept with a few-shot population of diverse, conceptually representative phrases.

Each phrase is embedded independently. That lets the compiler calculate both the geometric centroid and the dimensional variance of the concept in latent space, which transforms the concept from a single monolithic point into an independent multivariate Gaussian distribution. The centroid is the mean of the embedded phrase vectors:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

The per-dimension variance uses Bessel's correction to avoid low-N variance explosions from terser sensors. This correction is applied by Splinter's embedded Lua vector-processing logic:

$$\sigma_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2$$

Because Aldous keeps only the diagonal of the covariance, each dimension carries its own spread, and the concept occupies an elliptical, stretched region of meaning rather than a perfect sphere.

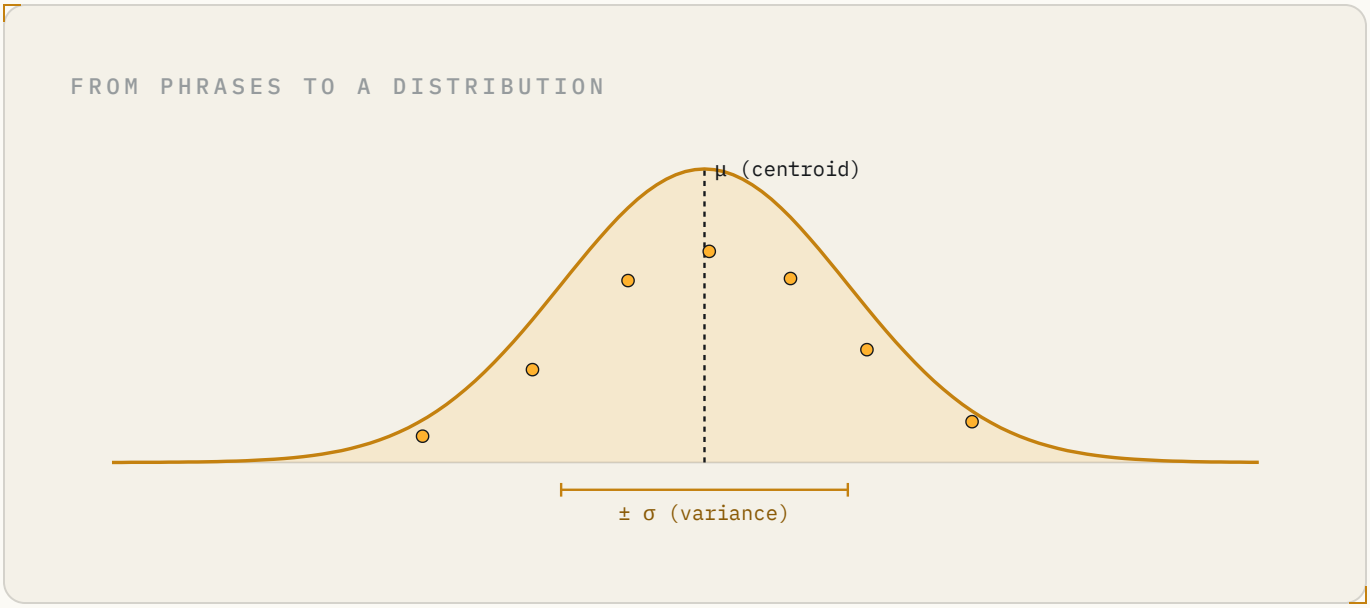

Fig. 1 — A handful of independently embedded phrases (points) define a centroid μ and a variance σ². A specimen is scored by where it falls under the resulting curve, not by keyword overlap.

WHY SPLINTER

Aldous models are built on Splinter stores because they conveniently hold vectors, strings, relation graphs, epochs, and the atomics involved in model creation. They are lightweight, file-backed shared-memory regions with lock-free guardrails.

Measuring a specimen against the distribution

During inference, Aldous compares incoming specimen text against each concept distribution with three complementary measures.

Std. Euclidean

VARIANCE-SCALED DISTANCE · THE DIAGONAL MAHALANOBIS

Cosine

DIRECTIONAL SIMILARITY OF MEANING

Dot product

MAGNITUDE-AWARE ALIGNMENT

Standardized Euclidean Distance divides each dimension's squared difference by that dimension's variance. This is a diagonal approximation of Mahalanobis distance, and it is what lets the system understand the elliptical, stretched boundaries of human emotional meaning:

$$d(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{\sum_{j=1}^{D} \frac{(x_j - \mu_j)^2}{\sigma_j^2}}$$

A hard minimum threshold, tuned to the target embedding model's natural spread, prevents any single dimension from collapsing toward zero variance and dominating the sum:

$$\sigma_j^2 \leftarrow \max\left(\sigma_j^2,\ \sigma_{\min}^2\right)$$

> *It understands the stretched boundaries of human meaning, executed in $O(1)$ time, with no LLM guessing and no string matching.*
>
> ALDOUS INFERENCE MODEL

Inference also offers convenient ways to filter results, including an automatic Kneedle step that finds the natural cut in the distance and similarity falloff, so you are not forced to pick an arbitrary threshold by hand.

03 ERASURE

Latent Concept Erasure

The centroid-weighted architecture unlocks Latent Concept Erasure. Text cannot be reverse-engineered from vectors, but Aldous can use orthogonal projection to scrub a specific semantic geometry, such as hate speech or toxicity defined by a shunt, out of a specimen. Removing the component of the specimen that lies along the shunt direction u gives a purified vector:

$$\mathbf{x}_{\perp} = \mathbf{x} - \frac{\mathbf{x} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}$$

By re-scoring that purified vector, a platform can numerically demonstrate whether the specimen carries value beyond the problematic content. If nothing meaningful remains, it can be rejected with confidence. If something does remain, that residual signal can be examined on its own.

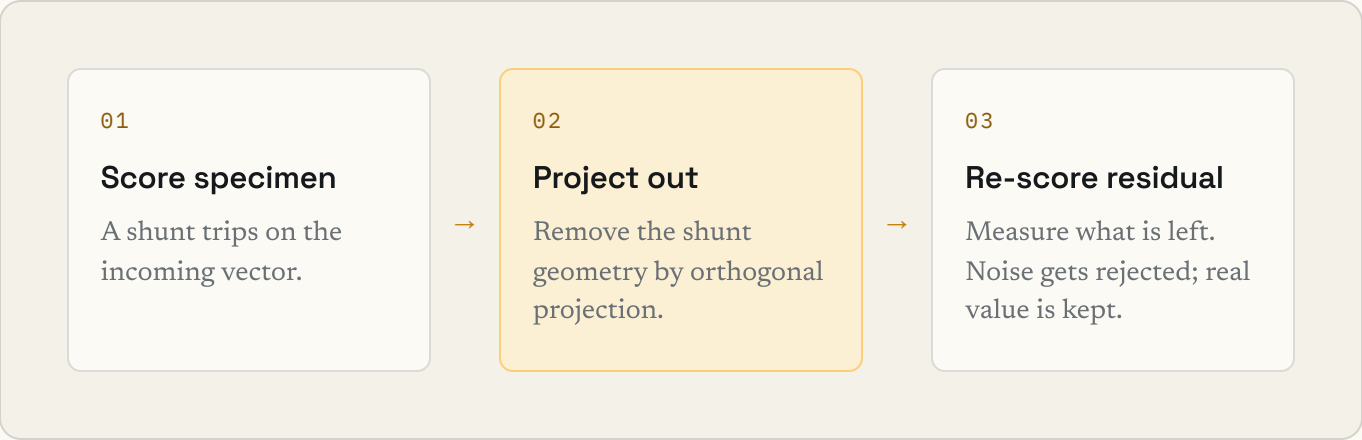

Fig. 2 — LCE as a moderation aid. The specimen is purified of the offending geometry, then re-examined for residual intent and value.

CONSERVATIVE BY DESIGN

Integrated conservatively, shunts are built to inform on obviously problematic content while giving other mechanisms context to work efficiently. An action need not always be a block or a queue hold. Aldous might require fewer community reports to auto-remove an image that scored high on certain indexes, or throttle visibility based on confidence.

04 SENSORS

The graduated emotional valence spectrum

Aldous ships with a broad spectrum of graduated dimensions, ranging from anger, joy, and sadness to hedging, philosophical and political tension, outlook, and the full spectrum of human sexuality. Each graduated dimension is backed by three centroids, and two signed indexes are backed by five centroids each.

21	63	2	10
GRADUATED DIMENSIONS	KEY SENSORS · 3 CENTROIDS EACH	SIGNED INDEXES	KEY SENSORS · 5 CENTROIDS EACH

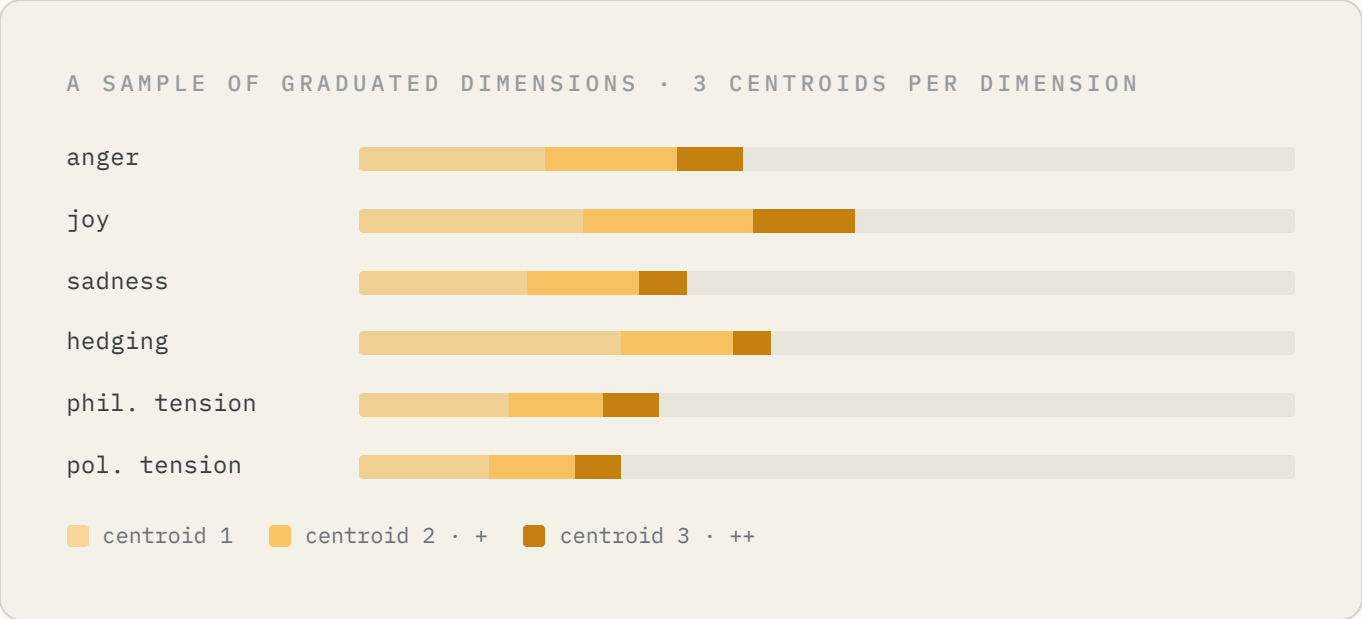

Fig. 3 — Graduated dimensions resolve intensity. Inference can filter results automatically with a Kneedle step based on distance and similarity falloff.

05 SENSORS

Signed indexes and pragmatic scalars

Signed affection and intent indexes

When a concept has direct antonyms expressible across the gradients that language intensity affords, Aldous can build a signed index. Outlook runs from optimistic through neutral to pessimistic. Motivation to action runs from damping through neutral to likely to move the reader. This is not practical for broad concepts like "good" or "evil," but phrases describing outcomes and intentions are highly practical, tunable, and consistent.

Pragmatic saturation indexes, or scalars

How exactly did an author intend their words to be read? Scalar indexes are a singular version of the graduated sensors: no "+" or "++" variants are created. A scalar typically contains several times more phrases than a graduated sensor, at mixed intensity, worded as generally as possible. Their job is to measure tonality and abstract meta concepts about a specimen.

Rather than estimate how a measured concept changes a score, Aldous simply provides the estimated quantity. A post is not scored as less angry for being sarcastic. It is scored as angry, and separately as sarcastic:

$$\text{Score} = f(\text{angry}) + g(\text{sarcastic}) + \dots$$

A scalar is compressed to a single dimension with dynamic magnitude rather than a series of graduated intensities, so the modifier does not need to be captured per concept.

PRE-DEFINED SCALARS

Aldous comes stocked with indexes for Human Sycophancy, Sarcasm, Gratitude, and Autonomic Reaction. More can be defined easily.

06 GUARDRAILS

Trust & Safety shunts, two circuit breakers

Aldous borrows the electrical term "shunt trip" for its circuit breakers. There are two kinds, with opposite update characteristics.

CENTROID-WEIGHTED · INTRINSIC

Built for nuance and erasure

- Trains like a centroid-weighted sensor.
- Shows content that should universally never be present, such as incitement to violent rioting or coercion of a minor against a guardian.
- Made intrinsic because magnitude is necessary for the nuance.
- Phrases are not very dynamic, so the group does not need constant updating.
- The centroid set is essential for concept elimination and re-examination.

MEAN-POOLED · MONOLITHIC

Built for speed and volatility

- Designed to be edited tens or hundreds of times a day.
- Keeps up with shifting bad-actor tactics, political fallout, synthetic generation, and agentic troll rigs.
- Requires only one embedding pass to go live.
- Operates at a lower magnitude, so it is less discernible.
- Cannot be reliably separated from specimen vectors, so no LCE.

While the noise floor and confidence thresholds stay generally the same as other sensors, the distance required for sensitivity varies by application. Shunts also require corroboration from other valence metrics, or strong signal after LCE re-scoring, before driving action.

● CURRENTLY IN ALDOUS · 18 MEAN-POOLED SHUNTS

These range from AI Sycophancy ("you didn't just write a for-loop, you redefined how people indent loops") and other engagement patterns that drift weekly, all the way to identifying hate. Monolithic shunts can also detect confidential leaks through concept triangulation, and Aldous can monitor outgoing text as easily as incoming.

Dynamic kinetic aggression normalization

Not every channel has the same normal background intensity. Phrases like "we murdered them" or "we totally slaughtered their whole team" register as violent conflict in any semantic system. Aldous expects and embraces this. Because it is so deployment specific, you can use any or all of the following.

- **Composite signal logic.** Use other signals to negate or lessen the impact of event-oriented sensors. Joy and sarcasm alongside conflict is a good indicator. Log readings for a few weeks, then set per-channel floors.
- **Domain-specific erasure.** For extreme environments like gaming or high-pressure sales, create an intrinsic shunt that captures local hyperbole. If a ++violence vector survives removal of the hyperbole vector, track it as such.
- **Pragmatic scalars as context lenses.** Create a "@Victory_Lap" or "@Competitive_Drive" scalar to track this directly and reference it when evaluating sudden surges. This is often the simplest and most comprehensive option, and it allows LCE.
- **Independent collector baselines.** Sometimes a 20% noise floor is far too low, but only for certain latent concepts. Adjust actionable and confidence levels per node.

```
This is only necessary if it becomes a problem. If you can spot and
filter it naturally, there is no need to add processing overhead to
each call.
```

08 BONUS

Structural AI detection, a side effect of the geometry

Because DECE measures the geometry of language rather than keywords, the engine naturally detects synthetic and adversarial AI text through structural anomaly and phrasing noise. This acts as a structural contrast agent against human text, providing deep synthetic detection without watermarks or LLM-based detectors. Since the specimen has already been deconstructed geometrically, you can run a few more checks.

- **Empty re-orientation.** Use a scalar to embed generic turns of phrase where the scene shifts but meaning never advances ("they went upstairs and then outside"). Run LCE and see what is left. This can address a robo-testimonial problem very effectively.
- **Unnatural density.** Centroid group density tends to be very smooth. Research notes that LLMs use a specific slice of vocabulary and distribute its structural density with unnatural uniformity. If four of five markers sit just above the noise floor and look nearly identical, that warrants a closer look.
- **Hedging plus certainty.** Look for hedging and certainty in the same specimen. That combination is uncommon from humans alone and typical of a helpfulness-trained model asked to generate polarizing content.

SCOPE NOTE

Aldous should not be the basis for academic or professional submission evaluation, and structural detection should not stand alone for critical applications without a formal verification harness. Because basic synthetic detection already exists in the math Aldous performs, we simply expose it.

09 TRANSPARENCY

Aldous receipts are safe to issue

> *Black boxes destroy trust in moderation. There is no such thing as real security through obscurity.*
>
> ALDOUS DESIGN PRINCIPLE

Aldous's monolithic mean-pooled shunts are designed to be inspected by any user and edited by community moderation teams. Shunts work on concept matching, not string matching, so changing some phrasing gets an attacker nowhere. It is better if your whole community knows what the system watches for, and receives a receipt when something is blocked.

This turns usually contentious conversations between users and moderators into productive, continuous improvements to the firewall. No more nebulous, uninformed decisions about your community. Aldous is fully open source and needs no more than a Chromebook to train, even if you self-host inference. A full training harness is included in the repository for reproduction, tuning, and customization.

10 SPECIFICATIONS

Feature summary

`Detailed specs and an initial benchmark imagining for this class of model are in progress.`

- **Emotional valence model** that never reads or logs actual text. No string matching, no LLM reasoning. Outputs similarity and distance floats only.
- Measures **70+ emotional valence inflection points** and **2 signed intent indexes**.
- **20 Trust & Safety shunts**, both mean-pooled and centroid / variance-weighted.
- Allows **LCE** of shunt vectors from specimens, then re-scores for residual signal and intent.
- Designed to work with **Nomic Text 1.5** and **Libsplinter**, both Apache2 licensed.
- Trains fully in about **45 minutes**, then scores up to 2k graphemes in milliseconds. Well suited to event-stream processing or high-frequency sampling.

SPECIFICATION	VALUE
Classification method	DECE · diagonal multivariate Gaussian
Distance measures	Standardized Euclidean, Cosine, Dot product
Inference complexity	$O(1)$ per concept
Valence inflection points	70+
Signed intent indexes	2

SPECIFICATION	VALUE
Trust & Safety shunts	20 (mean-pooled + centroid-weighted)
Training time	~45 minutes, full
Scoring window	up to 2,000 graphemes, milliseconds
Dependencies	Nomic Text 1.5, Libsplinter · Apache2
Hardware floor	Trains on a Chromebook

11 CREDITS

Foundations & Acknowledgements

On the shoulders of giants, measured along the diagonal

Aldous is Free Software under the terms of the Apache 2 software license, where applicable, or CC-0 where the content is prosaic, not code, and not describing the creation of code. The vector substrate and file system backing Aldous, which is named Splinter, is also free software under the terms of the Apache 2 software license.

Unless otherwise explicitly marked, supporting code, scripts and utilities are released under the terms of the MIT software license, with all documentation and other creatives under the terms of CC-0.

Apache 2.0 · engine & Splinter | MIT · scripts & utilities | CC-0 · docs & prose

> *Aldous comes with no warranty or guarantee of suitability for any purpose.*
>
> ALDOUS LICENSING TERMS

Infrastructure

Aldous does not run on its own. It needs somewhere to keep vectors, key-value state, relation graphs, epoch counters, atomics, and an embedded scripting runtime, all in a footprint small enough to sit on whatever hardware it is handed. That substrate is Splinter (Post, 2026), a lock-free shared-memory manifold that holds every one of those things with no database server and no socket in the path. To Splinter, DECE is only one pose it can hold: the engine that does Aldous' emotional scoring is, from the substrate's point of view, a single arrangement of slots and vectors among the many it could carry. That generality is why Aldous can be as light as it is. None of the machinery it leans on had to be built into it.

The same substrate is also why Aldous is not bounded by the hardware it can train on. Because a governing process can read the same physical memory an inference engine writes into, the observation gap that forces most oversight to happen after the fact closes as a property of the address space rather than as a matter of policy. On larger hardware, that is what lets Aldous observe generation while it is still in flight, at a latency and scale the single-machine framing understates. We link to the Splinter thesis rather than restate it here, but the ceiling it raises for this kind of work sits well above the modest hardware floor it advertises.

Mathematics

Aldous does not introduce new math; it reimagines settled, peer-reviewed ideas and aims them at a problem they were not originally built for: fast, honest emotional and intent telemetry. Its scoring rule is a variance-scaled distance from a specimen to a concept's centroid, which is the diagonal case of the Mahalanobis distance (Mahalanobis, 1936) and a close relative of the diagonal discriminant classifiers and nearest-shrunken-centroid methods that statisticians refined for high-dimensional data (Fisher, 1936; Dudoit et al., 2002; Tibshirani et al., 2002; Bickel & Levina, 2004). What Aldous adds is the embedder: it computes those centroids and variances over independently embedded phrases rather than raw features, which is the same technique the few-shot learning community applied with matching, prototypical, and Gaussian-prototypical networks (Vinyals et al., 2016; Snell et al., 2017; Fort, 2017).

We don't state the lineage defensively other than to state that Aldous' design is based on very grounded, accepted, published and peer-reviewed research. Aldous' returned responses are the result of measurements that other people have studied and validated independently across decades. We

aren't changing measurements; we're just applying them to a different, perhaps unconventional, class of problem.

Aldous' most exciting trust & safety features are also not new art: Latent Concept Erasure is the smallest, single-direction case of linear concept erasure, a technique the interpretability community has developed with real rigor (Ravfogel et al., 2020, 2022; Belrose et al., 2023). Aldous applies it to re-scoring rather than debiasing, but the geometry is theirs. Furthermore, the tolerance harness that gates every release is behavioral testing in the tradition of CheckList (Ribeiro et al., 2020), carried into trust and safety, where reproducible receipts matter the most.

> **CREDIT WHERE DUE**
>
> Their work matters every bit as much as Aldous' core mission to remain transparent, deterministic and auditable. We are glad to build in the open on foundations that were shared with so much care and respect for our collective sum of knowledge.

References

`Every claim above gets a receipt.`

The references below are here so that anyone can follow a claim back to its source.

Scoring, distance, and few-shot prototypes

Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6), 989–1010.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457), 77–87.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188. (Journal now published as *Annals of Human Genetics*.)

Fort, S. (2017). Gaussian prototypical networks for few-shot learning on Omniglot. `arXiv:1708.02735`. Bayesian Deep Learning Workshop, NIPS 2017.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49–55.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)* 30. arXiv:1703.05175.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences (PNAS)*, 99(10), 6567–6572.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems (NeurIPS)* 29, 3630–3638. arXiv:1606.04080.

Concept erasure

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2023). LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems (NeurIPS)* 36. arXiv:2306.03819. Code: github.com/EleutherAI/concept-erasure.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7237–7256. arXiv:2004.07667.

Ravfogel, S., Twiton, M., Goldberg, Y., & Cotterell, R. (2022). Linear adversarial concept erasure. *Proceedings of the 39th International Conference on Machine Learning (ICML)*, PMLR 162, 18400–18421.

Behavioral testing

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4902–4912. arXiv:2005.04118.

Infrastructure

Post, Timothy L. (2026). Splinter: A Lock-Free Shared-Memory Substrate For Tightly-Coupled Inference And Governance. *Open Source Vector Substrate* · @splinterhq/libsplinter (GitHub). splinterhq.github.io/splinter_thesis.pdf (Thesis).

Contact: Tim Post · linkedin.com/in/TimThePost