

Bilinear autoencoders find interpretable manifolds

Ward Gauderis & Thomas Doods

Geraint Wiggins & Jose Oramas



Flanders AI
Research Program



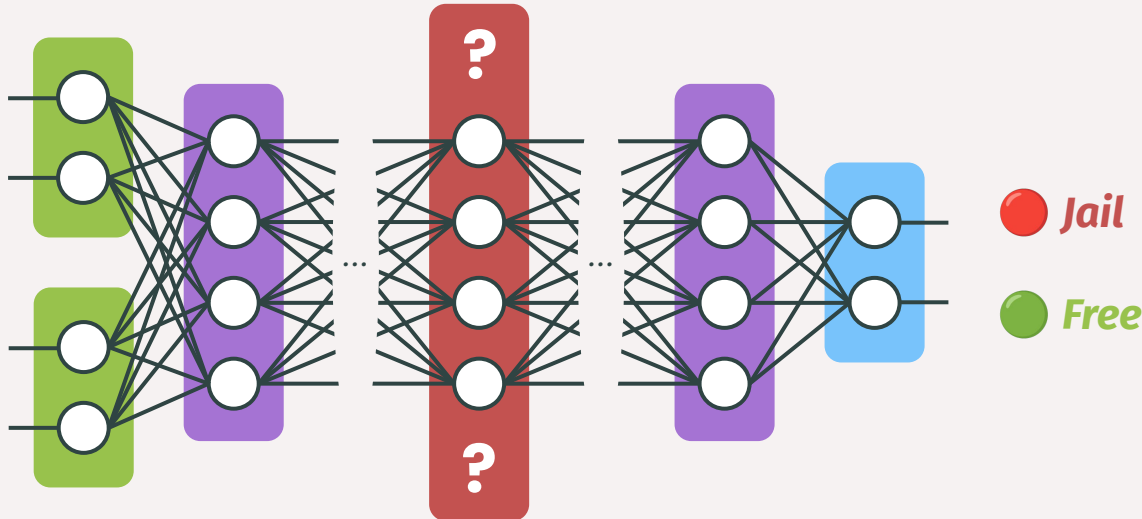
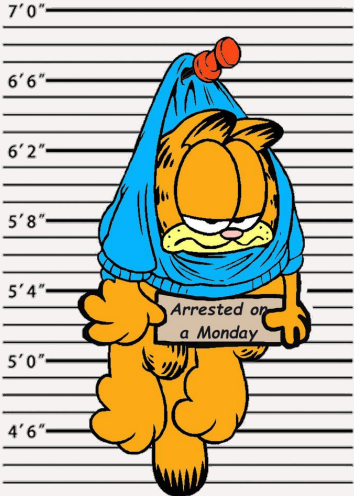
Artificial
Intelligence
Vlaanderen/Flanders

fwo



GOODFIRE

To build safe neural networks, we need to understand their internal working.



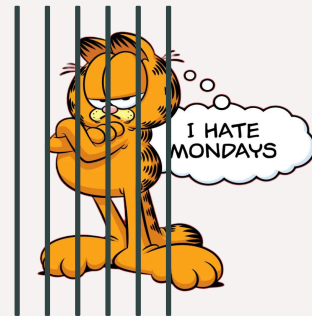
(Gauders, Doods, *From Mechanistic to Compositional Interpretability*, ICML WS 2026)

Most practicable interpretability methods assume neural representations are linear.

“Defendant arrested on a **Monday**”



Jail



Most practicable interpretability methods assume neural representations are linear.

“Defendant arrested on a *Monday*”

+

Thursday

=

“Defendant arrested on a *Friday*”



Most practicable interpretability methods assume neural representations are linear.

“Defendant arrested on a *Monday*”

+

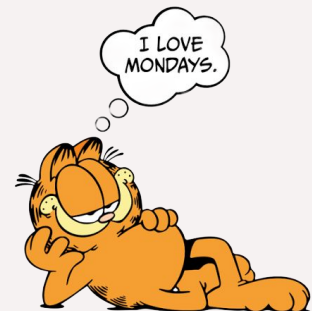
Thursday

=

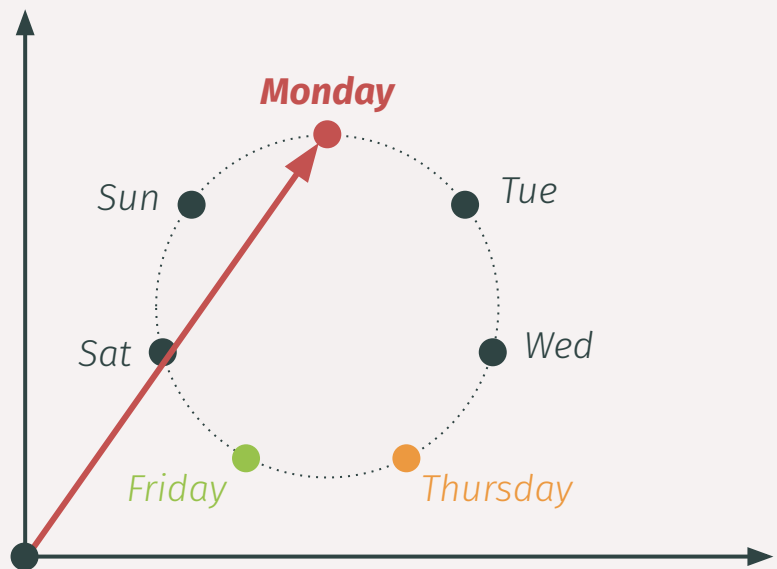
“Defendant arrested on a *Friday*”



 *Free*



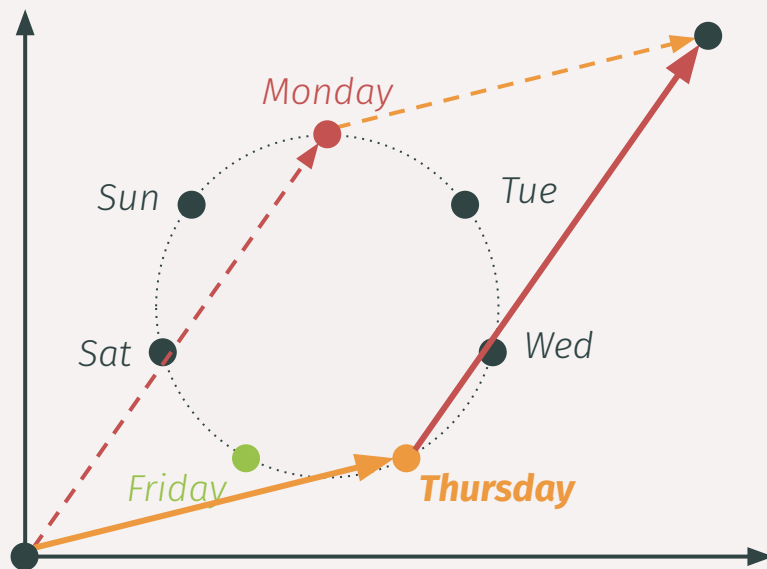
Neural representations are curved. Linear methods go off the rails.



“Defendant arrested on a **Monday**”

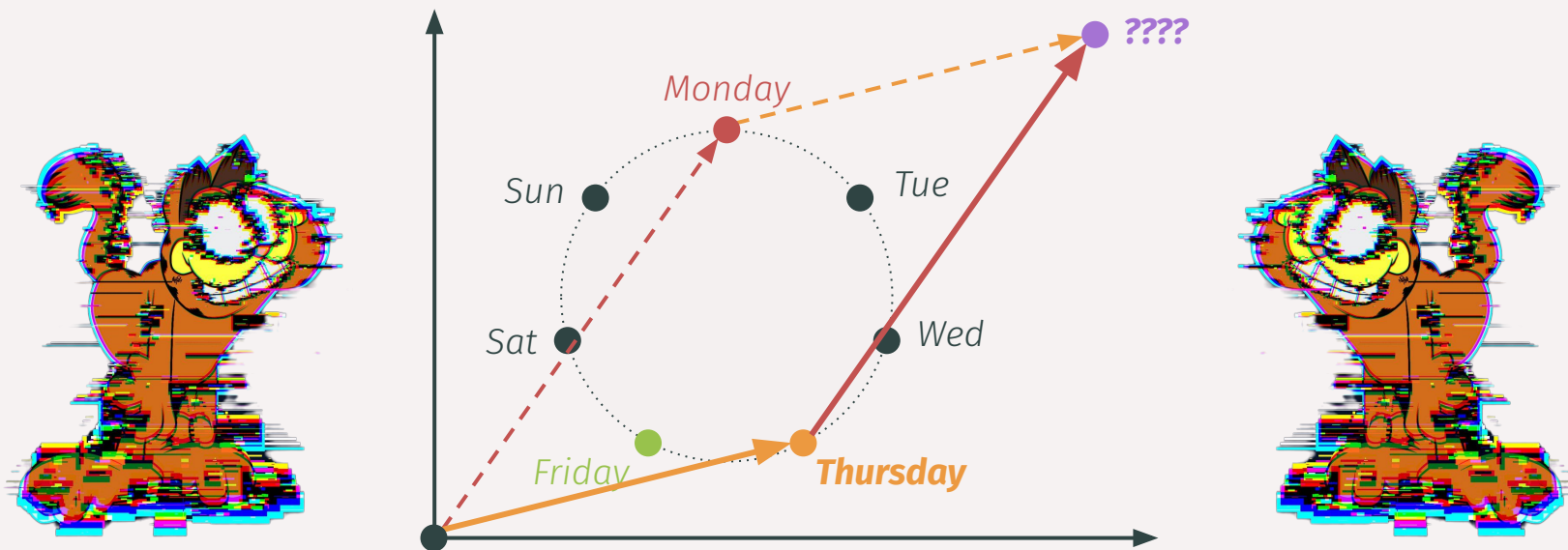


Neural representations are curved. Linear methods go off the rails.



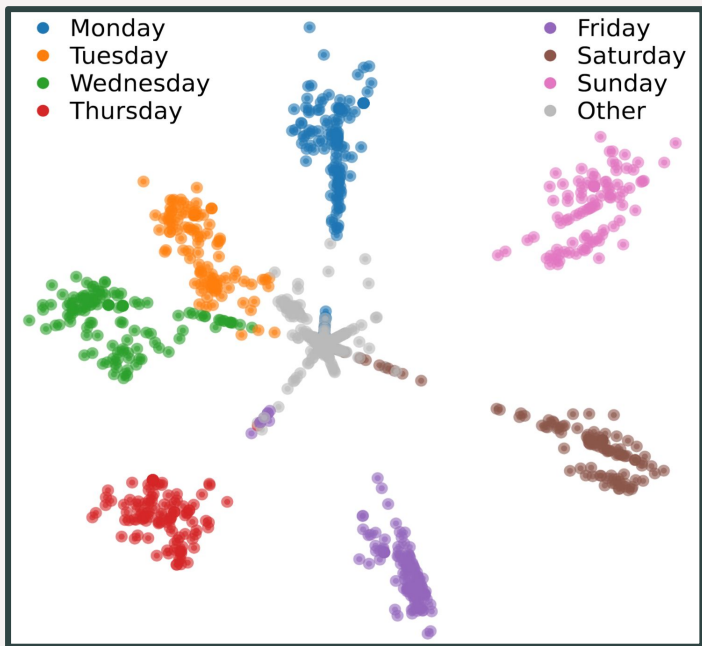
“Defendant arrested on a *Monday* + *Thursday*” \longrightarrow  *Free?*

Neural representations are curved. Linear methods go off the rails.



“Defendant arrested on a **????**” \longrightarrow **Caşăghă**

Neural representations are curved. Linear methods go off the rails.

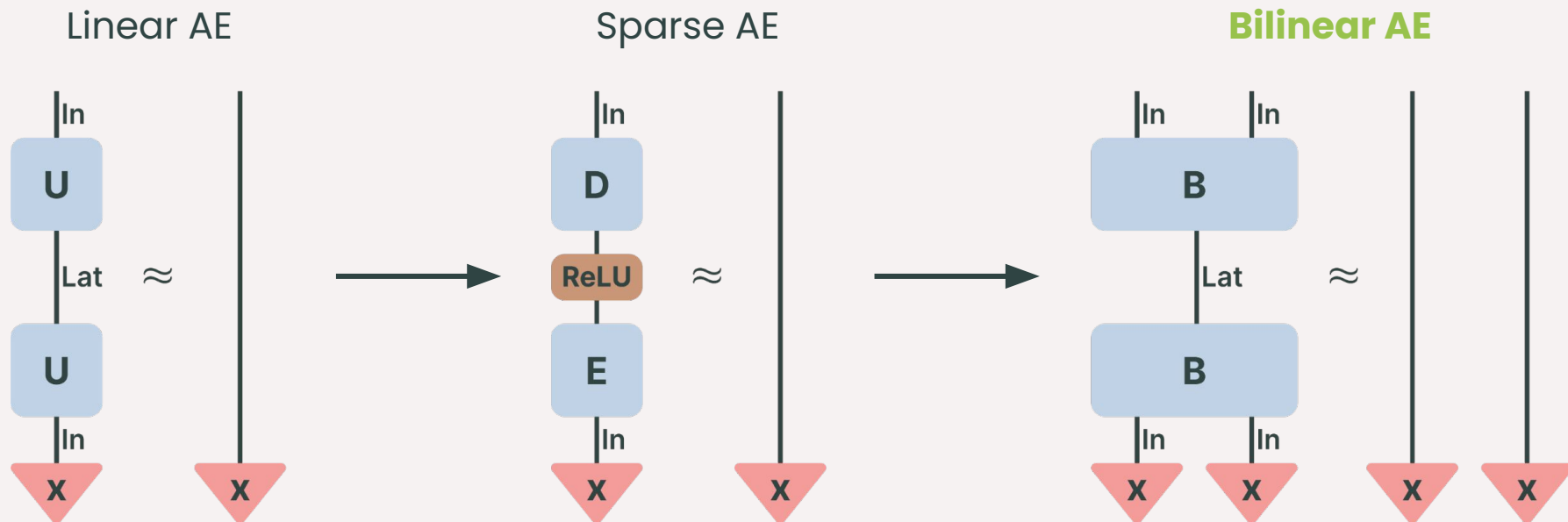


“Defendant arrested on a **????**”

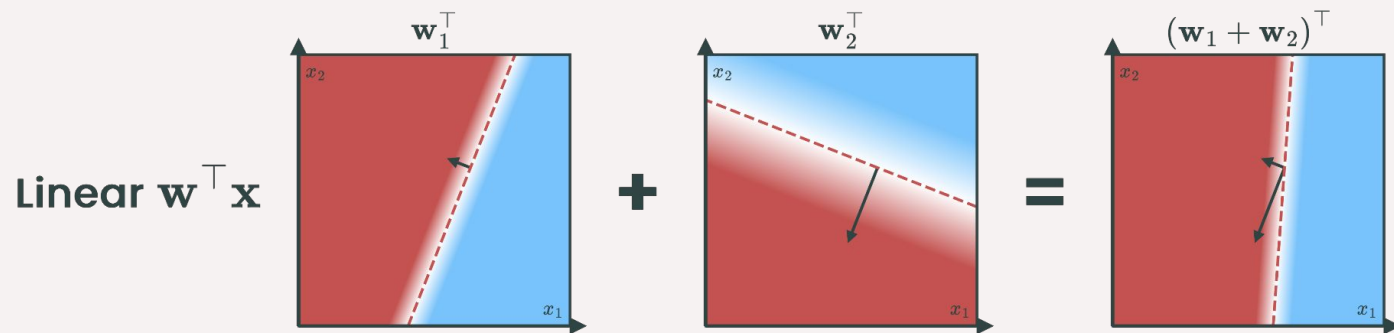


Caşăghă

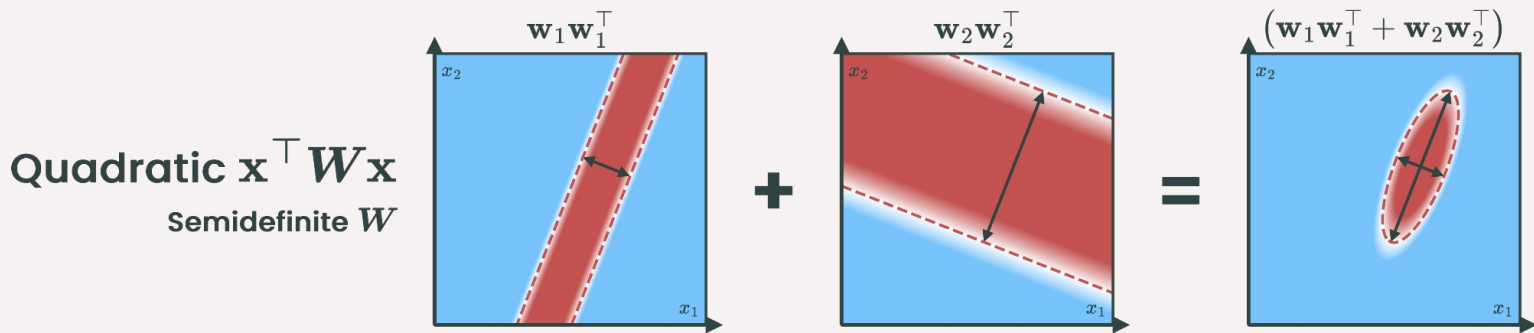
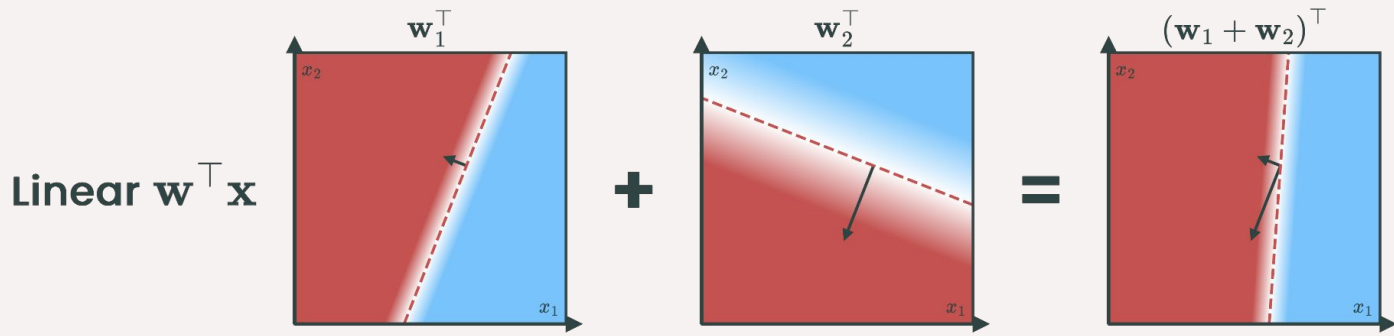
Bilinear autoencoders decompose representations into quadratic forms.



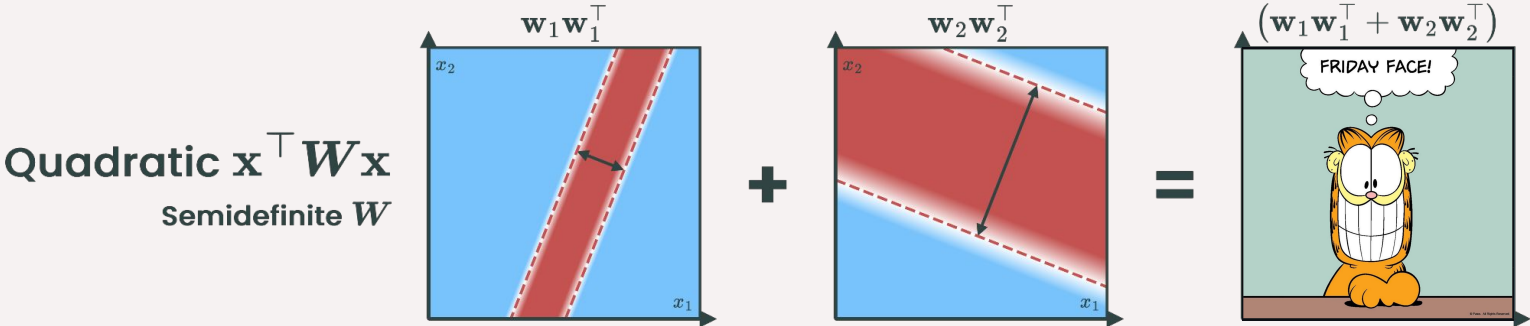
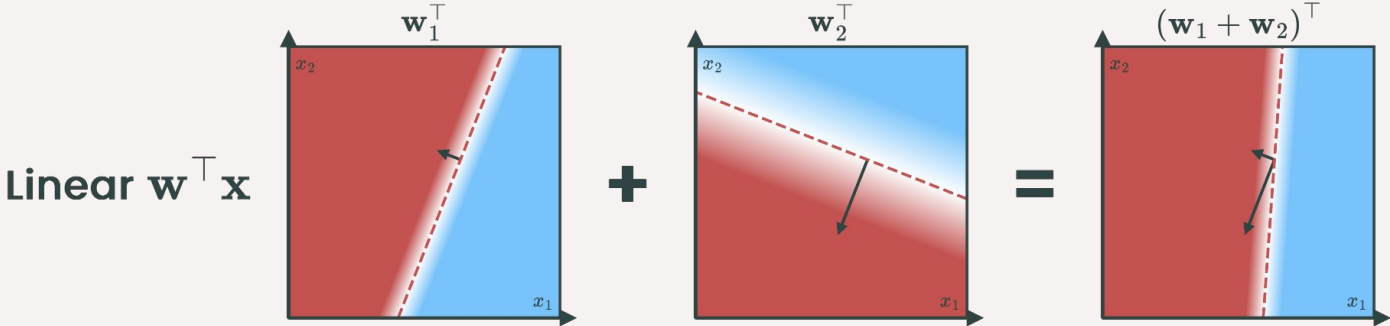
Quadratic forms compose into richer geometries, and remain interpretable.



Quadratic forms compose into richer geometries, and remain interpretable.



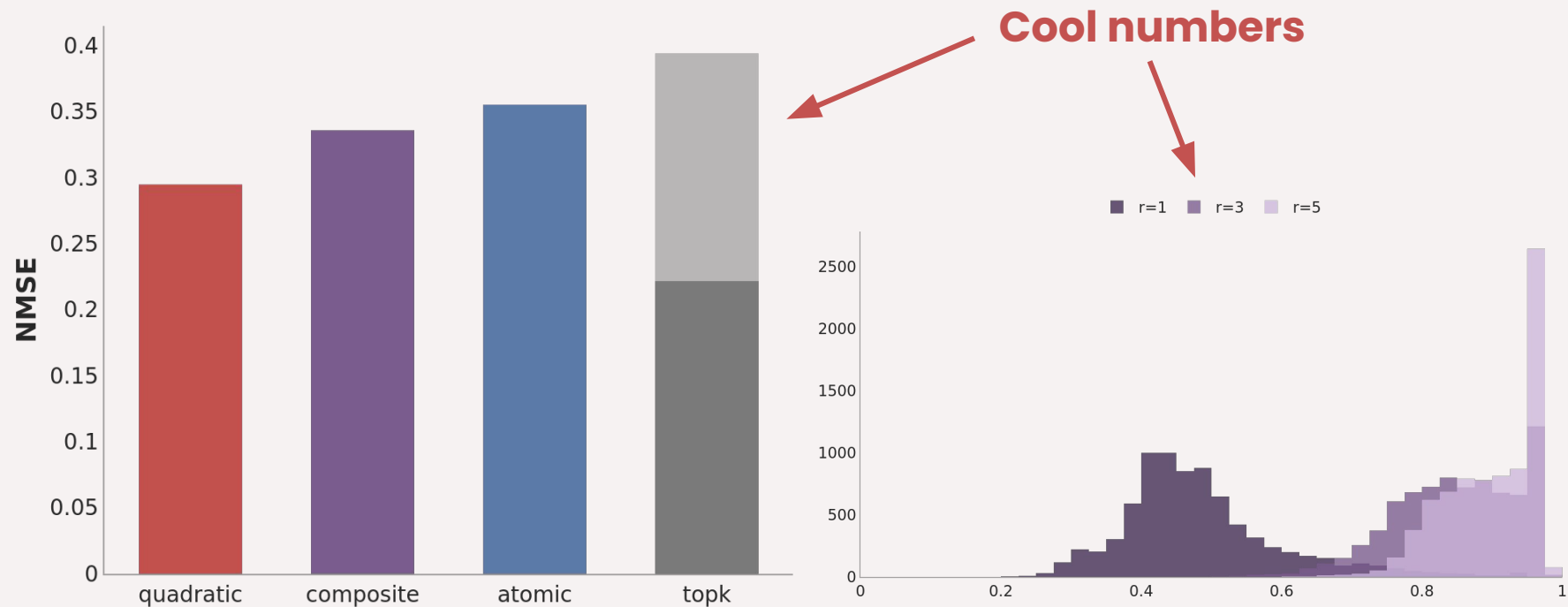
Quadratic forms compose into richer geometries, and remain interpretable.



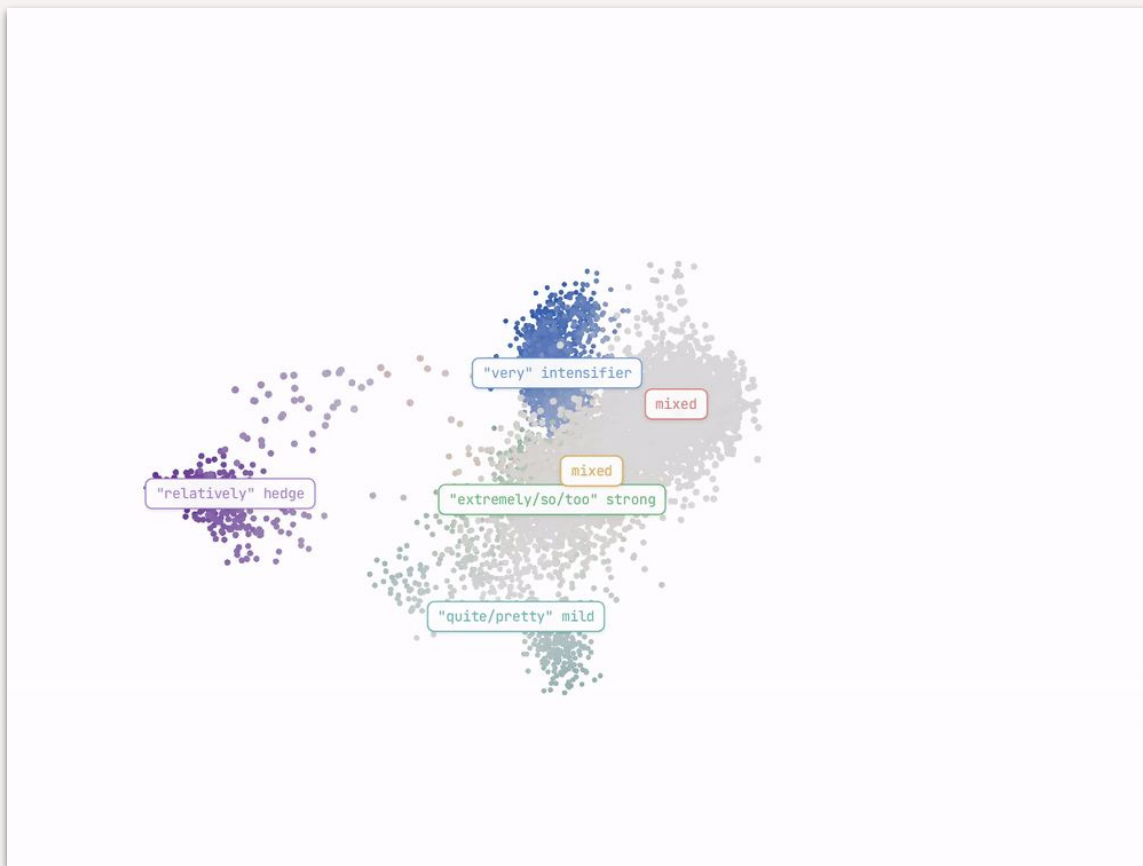
(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

How prevalent are non-linear geometries in modern language models?

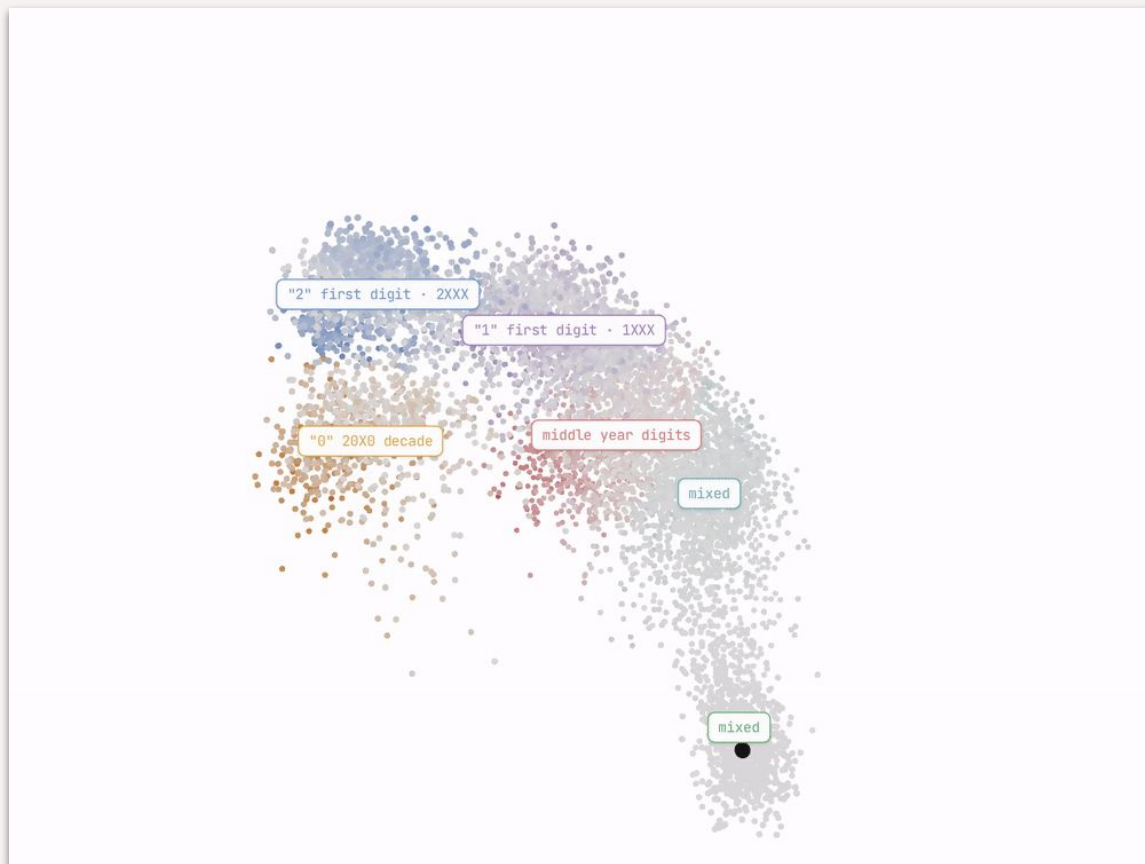
How prevalent are non-linear geometries in modern language models?



Intensity qualifiers lie on a triangle.

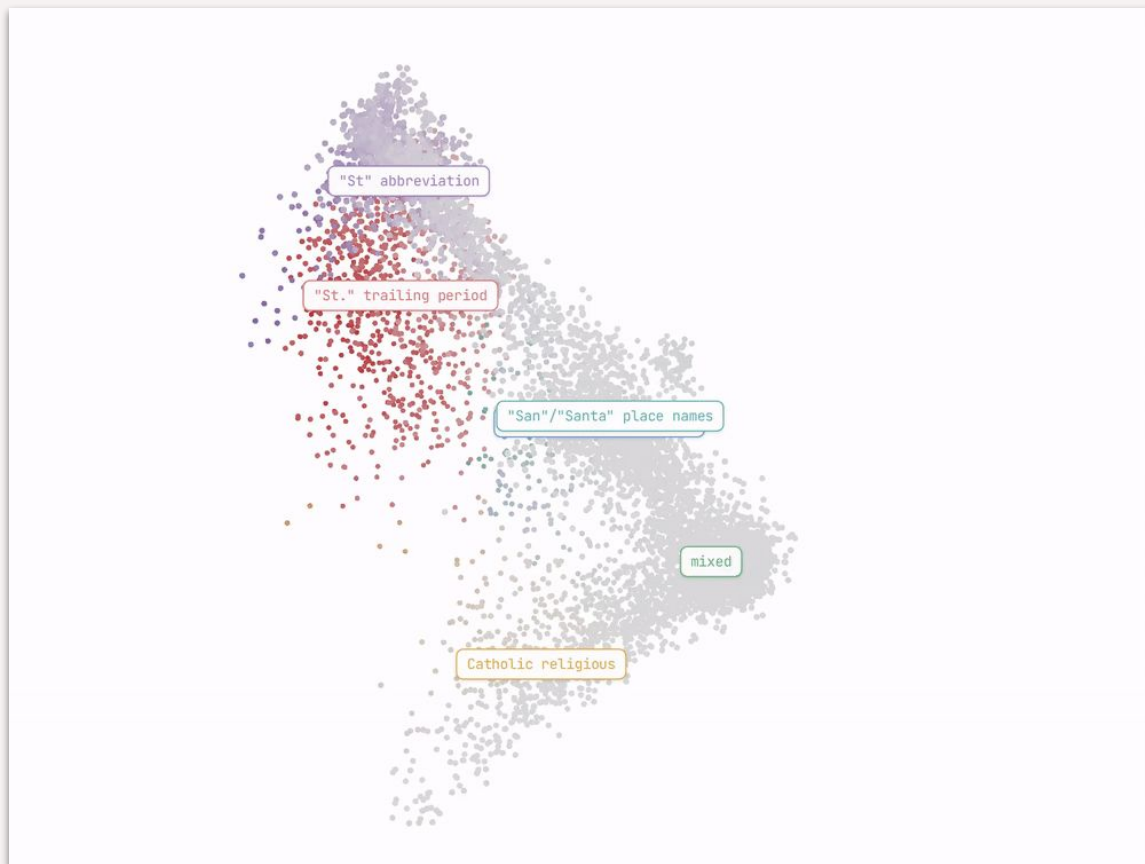


Year representations lie on a curve.



(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

Holy people and places also span a curve.



(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

These tools are useful outside of language models, they can become drivers of scientific discovery.

EEVE Evo Variant Effect Predictor | Search gene, rsID, or ClinVar ID (BRCA1, rs80357906, 15126) | **GOODFIRE**

SDHA c.217G>A (p.Gly73Ser)
ClinVar: 2933037 | rs2477286087 | chr5:224426 G>A | exon 3/15 | Missense

PREDICTED PATHOGENICITY: **98%**

VUS ★★☆☆

Hereditary cancer-predisposing syndrome | gnomAD AF: not observed

ClinVar | gnomAD | dbSNP | UCSC | UniProt | Ensembl | OMIM | GeneCards

VARIANT INTERPRETATION ?

Gly73Ser disrupts beta-strand geometry in the FAD-binding domain of SDHA while also perturbing a nearby splice acceptor context, collectively impairing proper protein folding and potentially pre-mRNA processing of the SDHA transcript.

This glycine-to-serine substitution at position 73 of SDHA generates a high-confidence pathogenicity prediction of 98%, consistent with the calibration data showing 97% of variants in this score range are pathogenic. The most prominent disruptions are at a splice acceptor region approximately 945-952bp upstream, where polypyrimidine tract identity drops by 0.49, intron identity by 0.42, intron-to-exon transition signal by 0.41, and splice acceptor probability by 0.34, suggesting the variant propagates a conformational or sequence context change that destabilizes a nearby intronic regulatory element. At the variant site itself, the P-loop containing nucleoside triphosphate hydrolase domain signal increases by 0.34, which reflects a gain of anomalous domain-like character rather than preservation of normal architecture, and is accompanied by a gain of disorder (+0.27) and loss of beta-strand integrity (-0.27) in the immediate vicinity. SDHA encodes the flavoprotein subunit of succinate dehydrogenase (Complex II), where the FAD-binding domain near the N-terminus depends on precise beta-strand geometry for cofactor coordination and electron transfer; disruption of beta-strand structure at this position would impair FAD binding and catalytic function. The combination of local structural disruption and upstream splice region perturbation produces a broad disruption profile consistent with a pathogenic variant causing mitochondrial Complex II dysfunction.

KEY EVIDENCE

- Polypyrimidine tract disruption of -0.49 at -952bp, with correlated splice acceptor loss of -0.34 at -945bp
- Beta-strand identity loss of -0.27 at -20bp from variant, consistent with structural disruption in the FAD-binding domain
- Gain of disorder signal (+0.27) near the variant site, indicating local structural destabilization
- Anomalous P-loop hydrolase domain signal gain (+0.34) at the variant position, reflecting aberrant sequence context
- Calibration data: 97% of variants scoring 90-101% are pathogenic, and this variant scores 98%

TIME | SUBSCRIBE

BUSINESS | AI

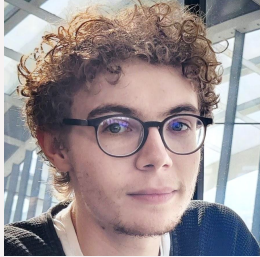
A New AI Tool Could Transform How We Diagnose Genetic Diseases

ADD TIME ON GOOGLE



(Peirce, Dooms et al. „EEVE: Interpretable variant effect prediction from genomic foundation model embeddings, 2026)

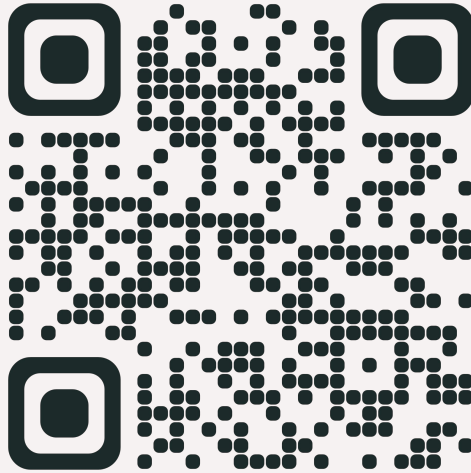
Want to know more or reach out?



Ward Gauderis



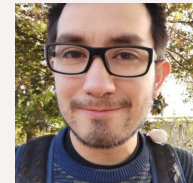
Geraint Wiggins



Scan the code!



Thomas Doods



Jose Oramas