Flanders AI Research Program
Internal Research Day May 12th 2026

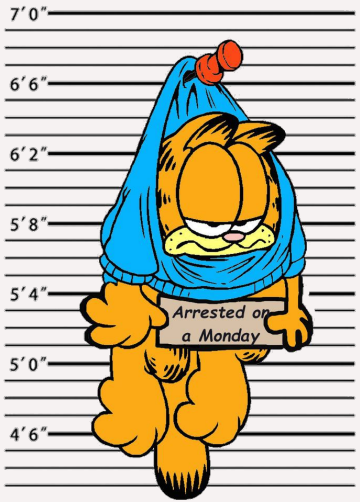
Bilinear autoencoders find interpretable manifolds

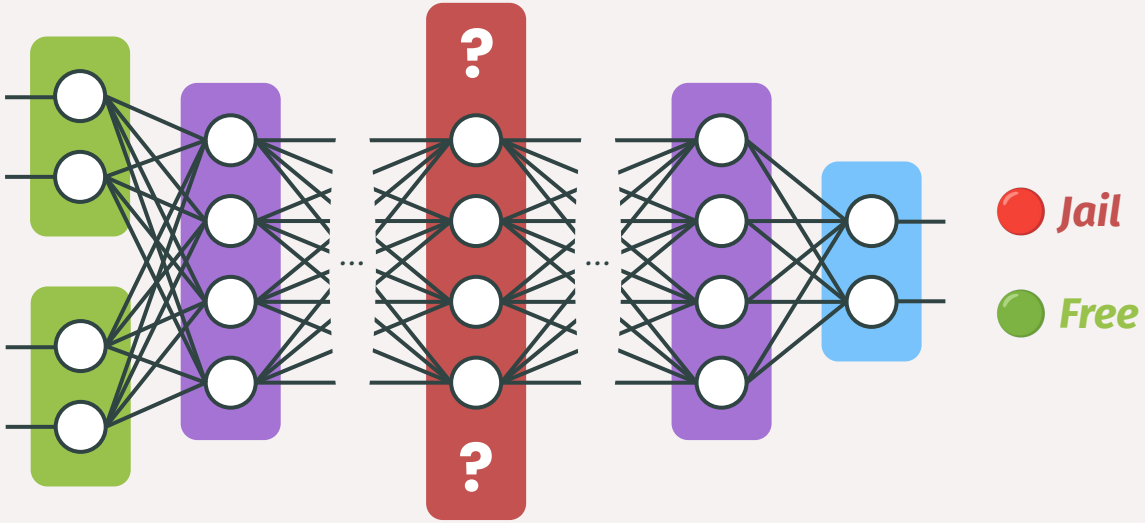
Ward Gauderis & Thomas Dooms

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, Dooms, *From Mechanistic to Compositional Interpretability*, ICML WS 2026)

Most practicable interpretability methods assume neural representations are linear.

*"Defendant arrested on a **Monday**"* → **Jail**

(Park et al., *The linear representation hypothesis and the geometry of large language models*, 2024)

Most practicable interpretability methods assume neural representations are linear.

"Defendant arrested on a Monday"

\+

Thursday

=

"Defendant arrested on a Friday"

(Park et al., *The linear representation hypothesis and the geometry of large language models*, 2024)

Most practicable interpretability methods assume neural representations are linear.

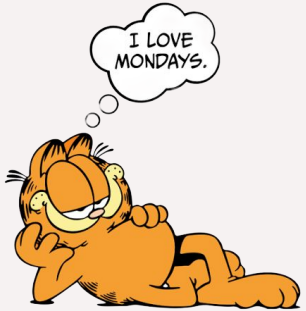
"Defendant arrested on a Monday"

\+

Thursday

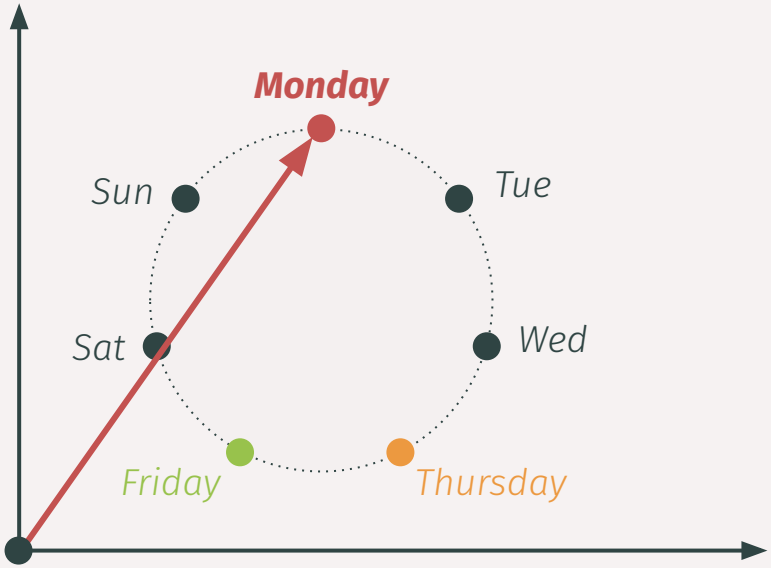
=

*"Defendant arrested on a **Friday**"*



(Park et al., *The linear representation hypothesis and the geometry of large language models*, 2024)

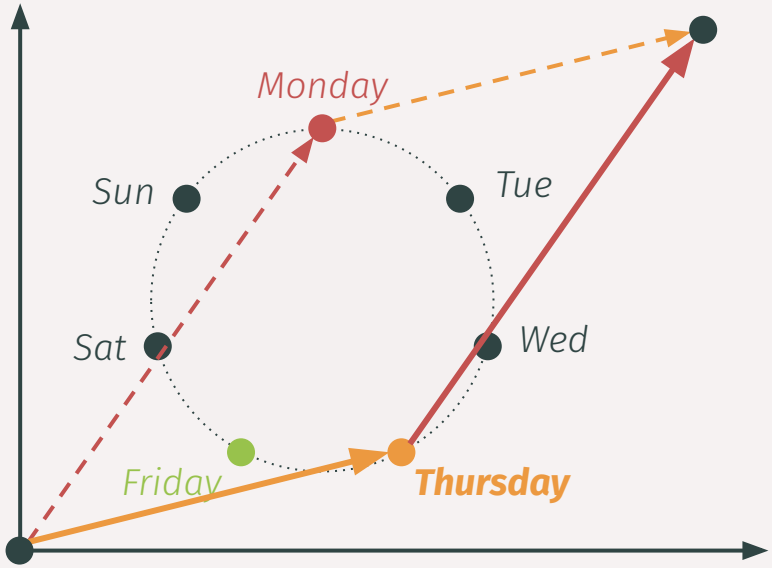
Neural representations are curved. Linear methods go off the rails.



*"Defendant arrested on a **Monday**"* → 🔴 ***Jail***

(Engels et al., *Not all language model features are one-dimensionally linear*, 2025)

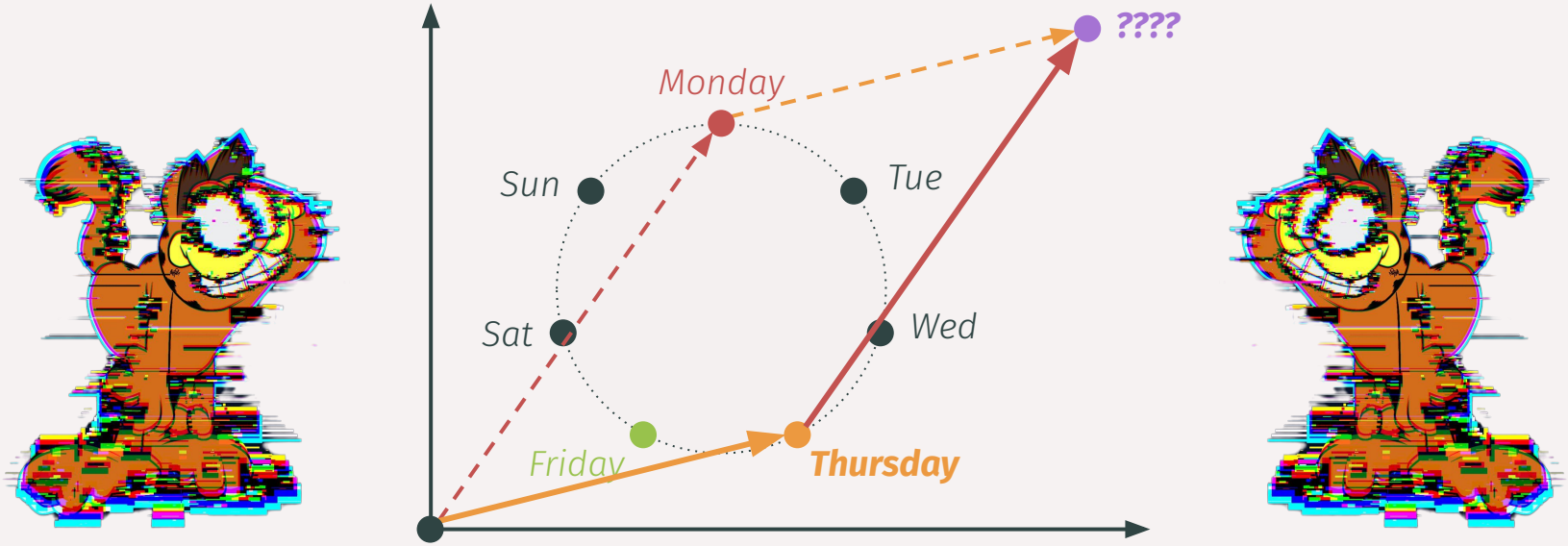
Neural representations are curved. Linear methods go off the rails.



*"Defendant arrested on a Monday + **Thursday**"* → 🟢 **Free?**

(Engels et al., *Not all language model features are one-dimensionally linear*, 2025)

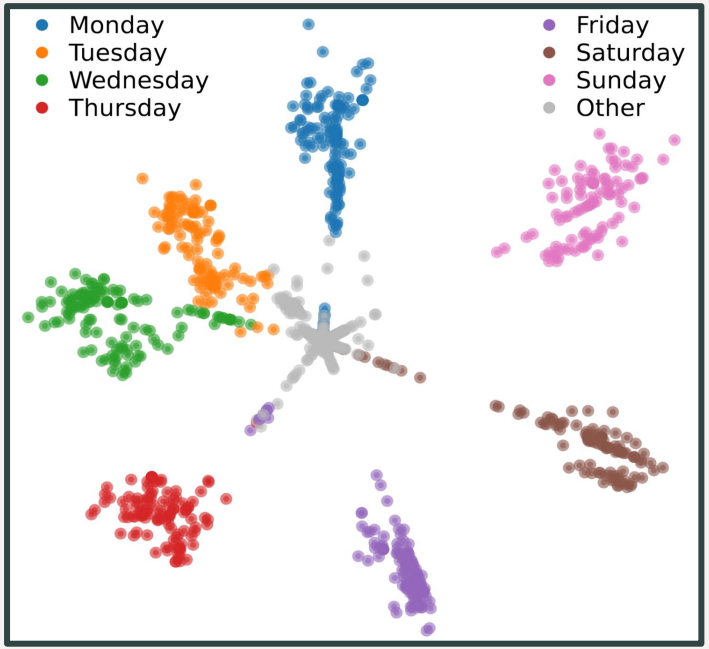
Neural representations are curved. Linear methods go off the rails.



"Defendant arrested on a ???? *"* → Lasagna

(Engels et al., *Not all language model features are one-dimensionally linear*, 2025)

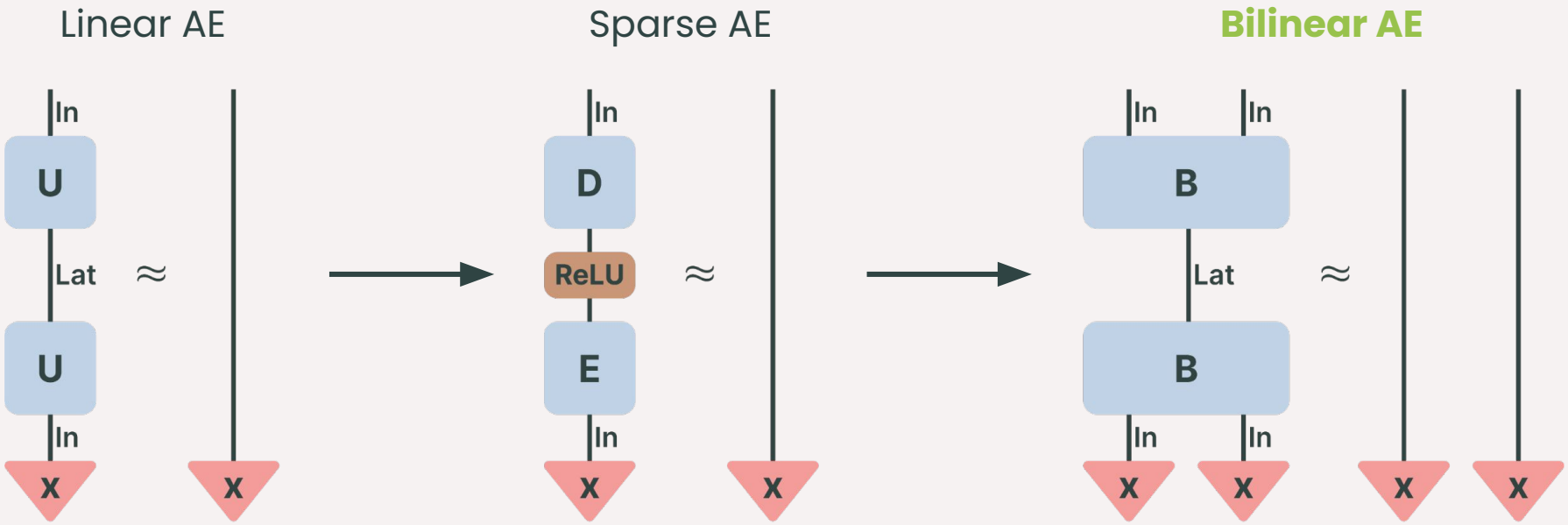
Neural representations are curved. Linear methods go off the rails.



"Defendant arrested on a ***????****"* → Lasagna

(Engels et al., *Not all language model features are one-dimensionally linear*, 2025)

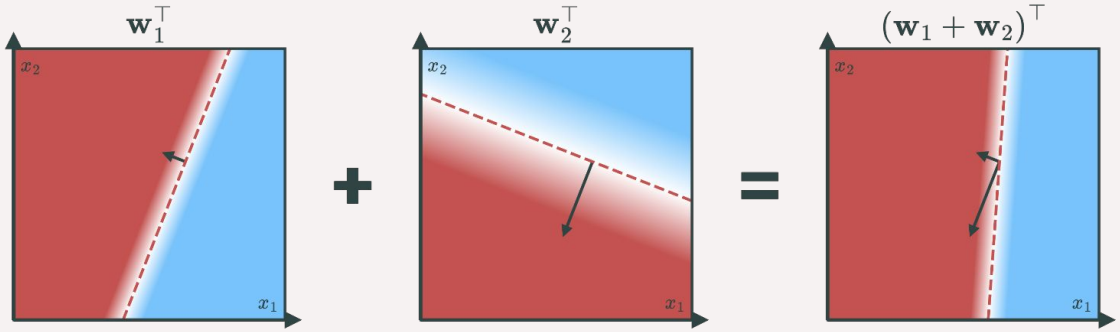
Bilinear autoencoders decompose representations into quadratic forms.



(Dooms, Gauderis et al., *Finding Manifolds with Bilinear Autoencoders*, NeurIPS WS 2025)

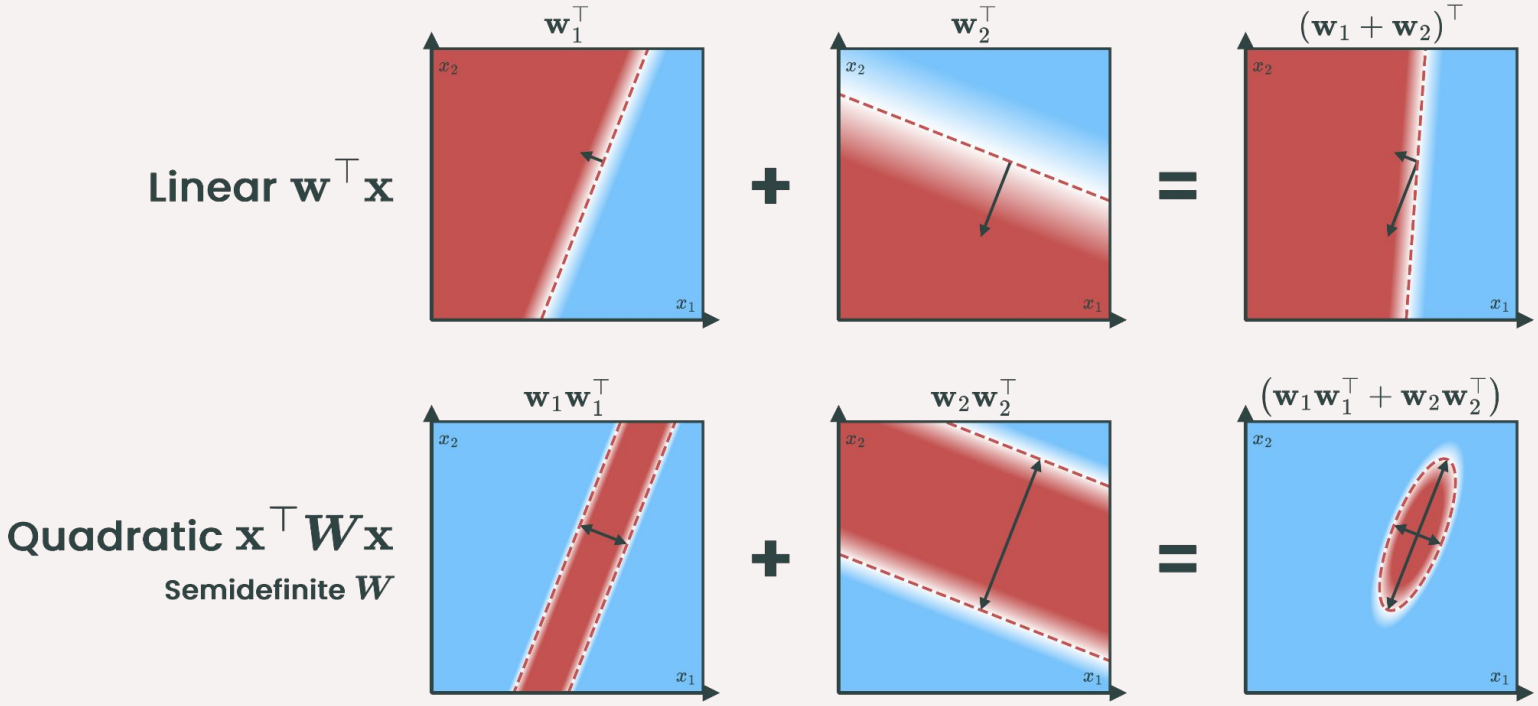
Quadratic forms compose into richer geometries, and remain interpretable.

Linear $\mathbf{w}^\top \mathbf{x}$



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

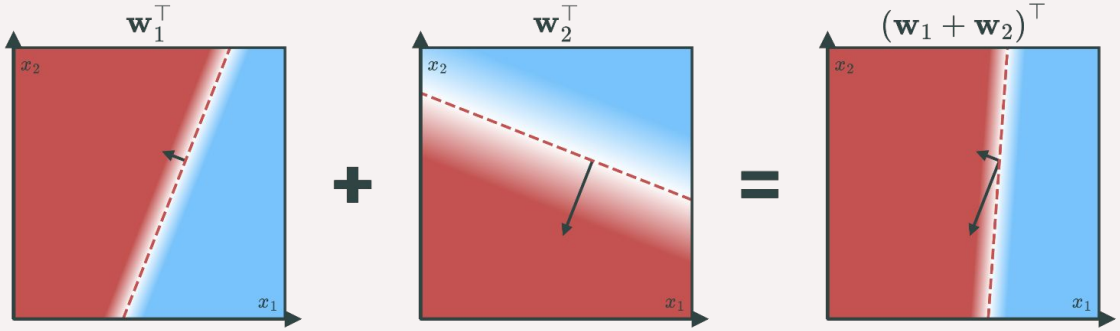
Quadratic forms compose into richer geometries, and remain interpretable.



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

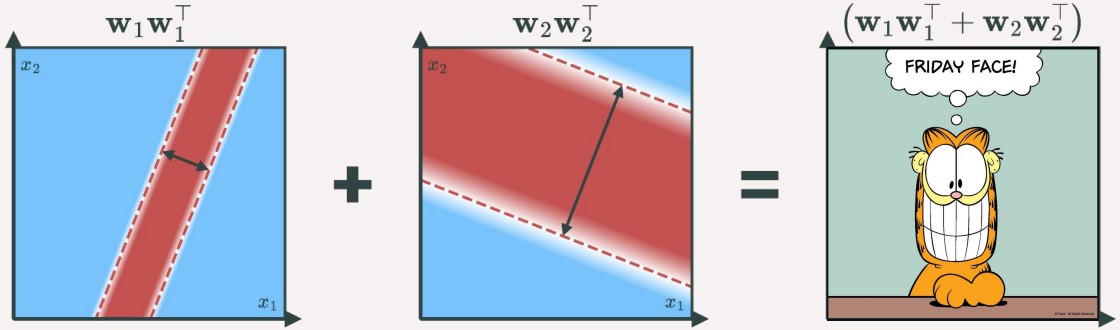
Quadratic forms compose into richer geometries, and remain interpretable.

Linear $\mathbf{w}^\top \mathbf{x}$



Quadratic $\mathbf{x}^\top \boldsymbol{W} \mathbf{x}$

Semidefinite $\boldsymbol{W}$

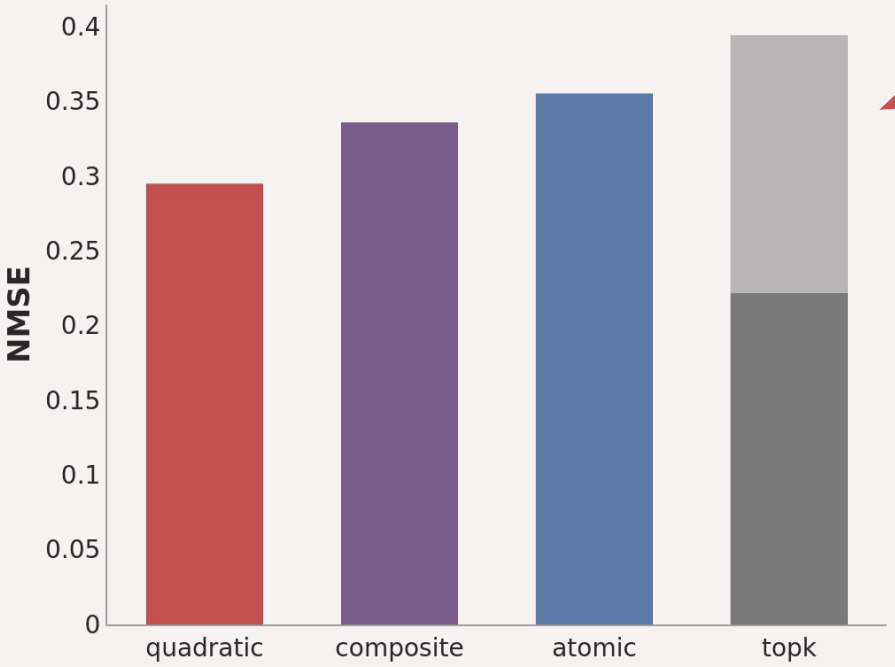


(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

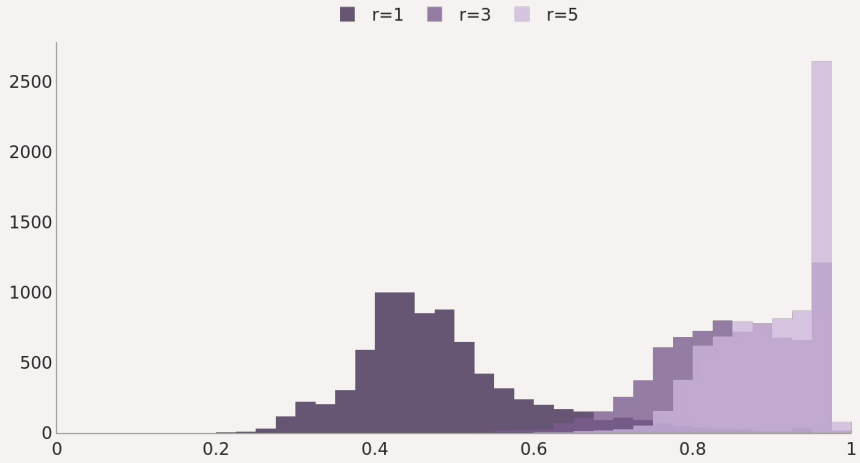
How prevalent are non-linear geometries in modern language models?

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

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

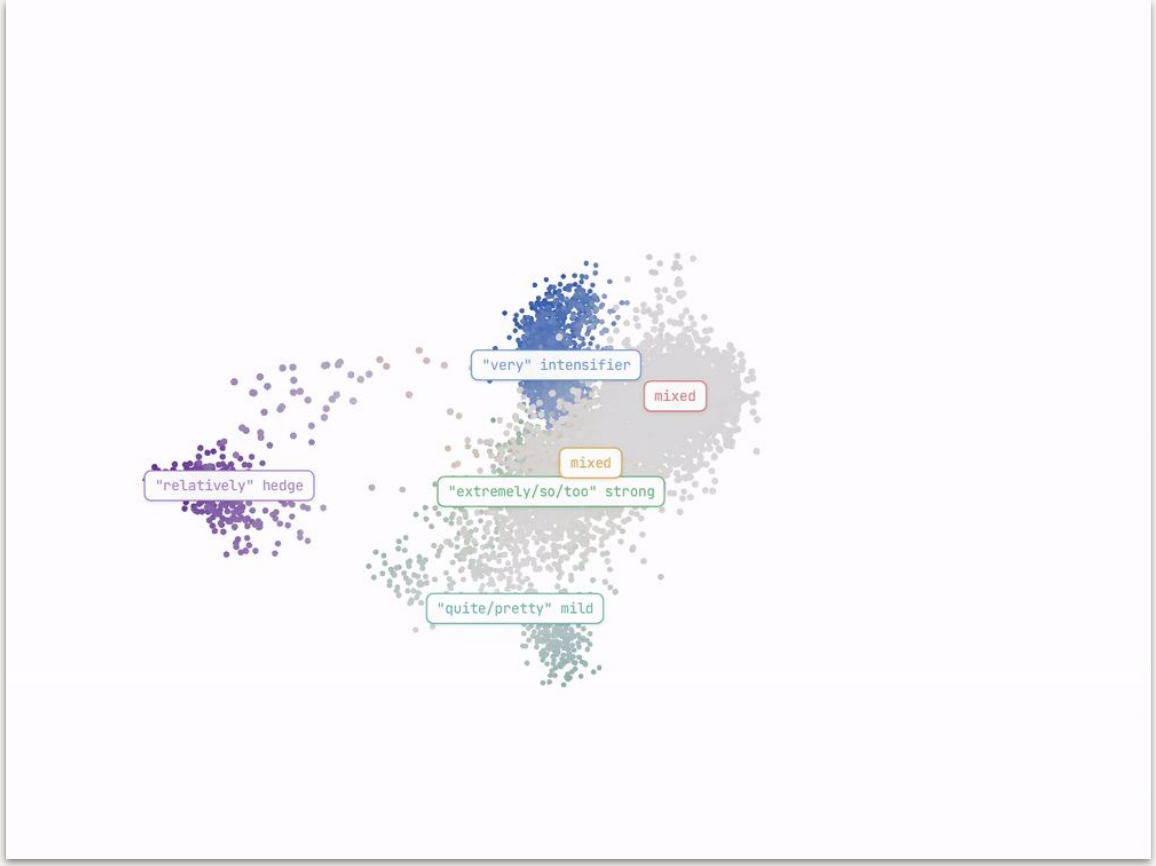


Cool numbers



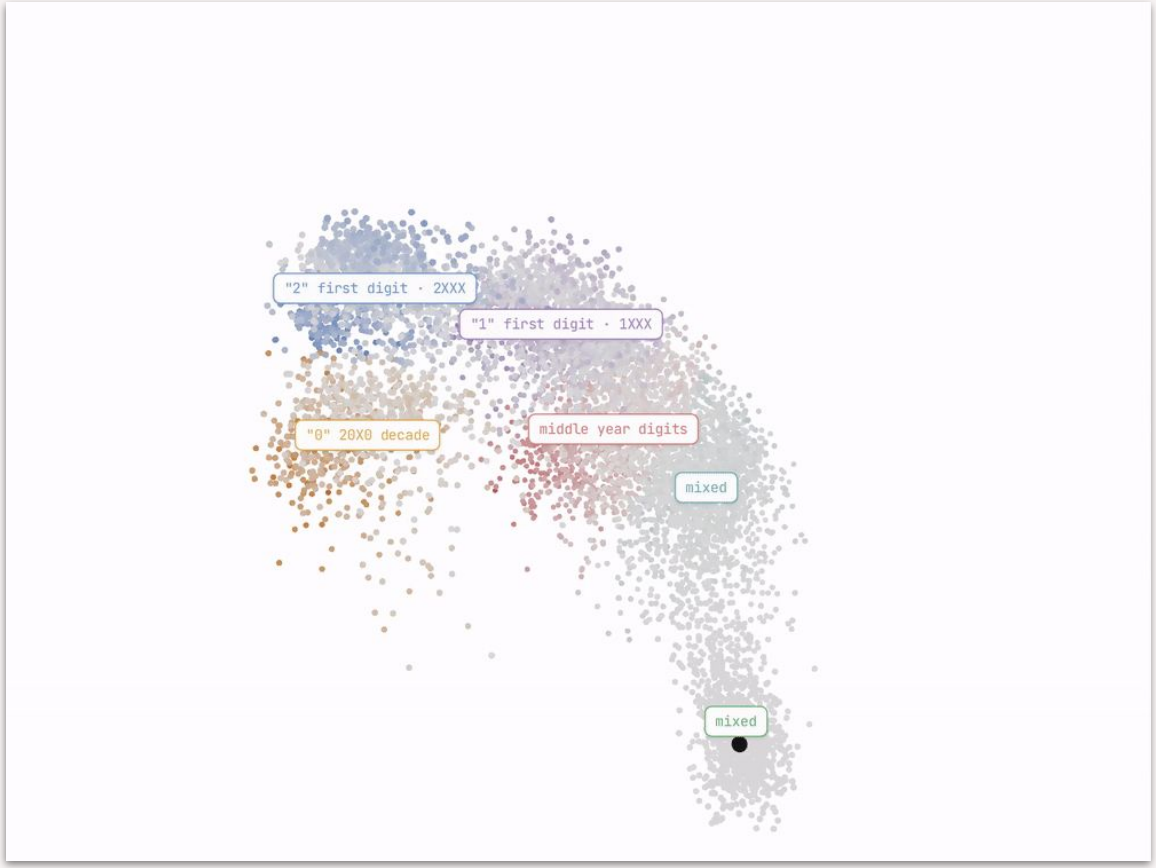
(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

Intensity qualifiers lie on a triangle.



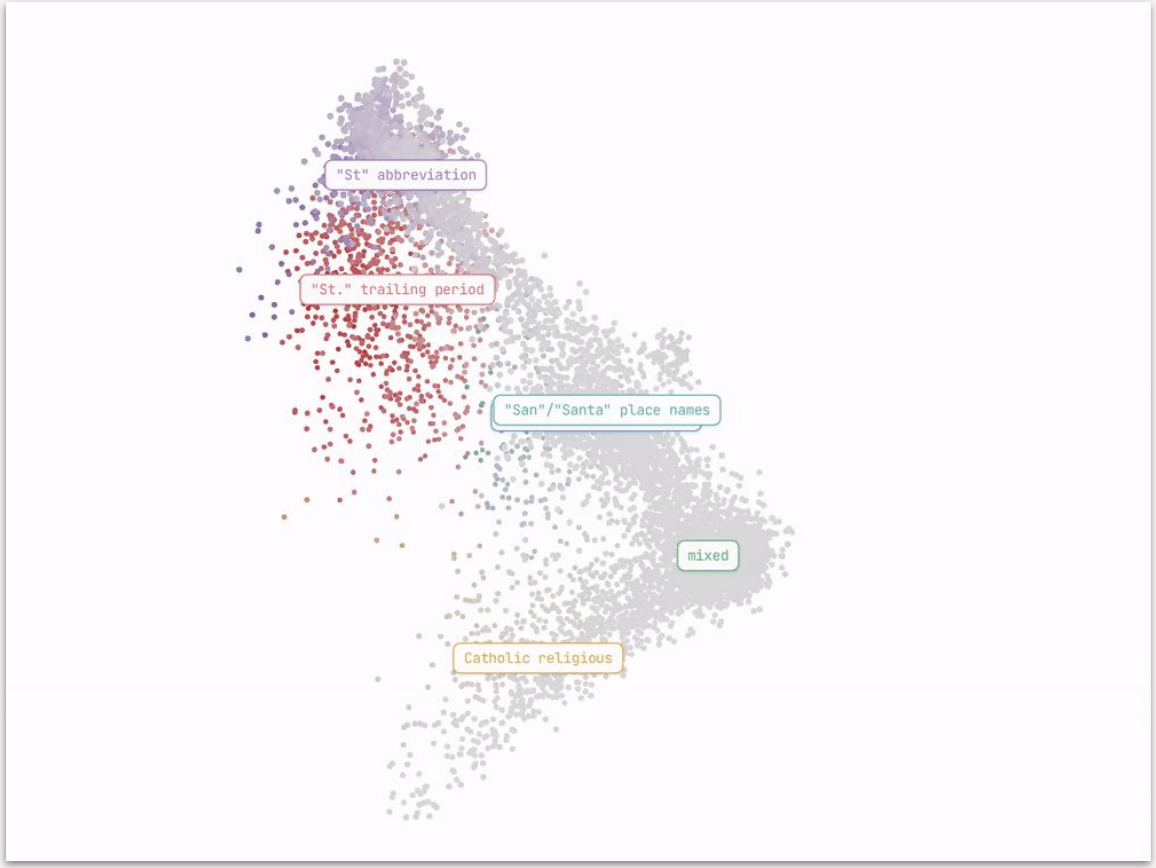
(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

Year representations lie on a curve.



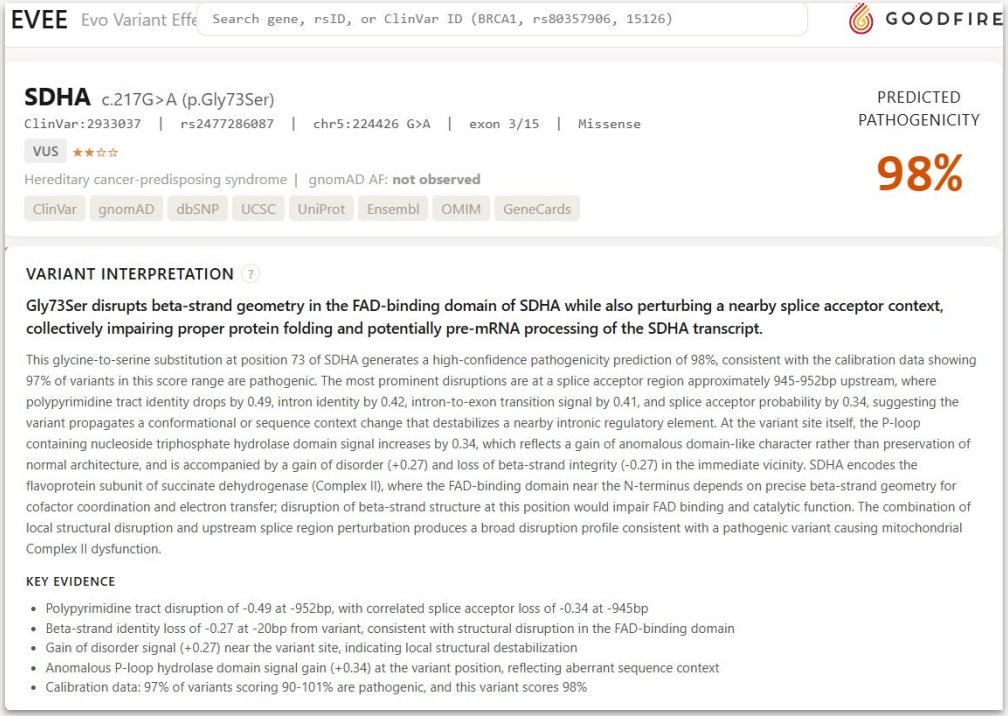
(Dooms, Gauderis et al., *Bilinear autoencoders find interpretable manifolds*, NeurIPS 2026)

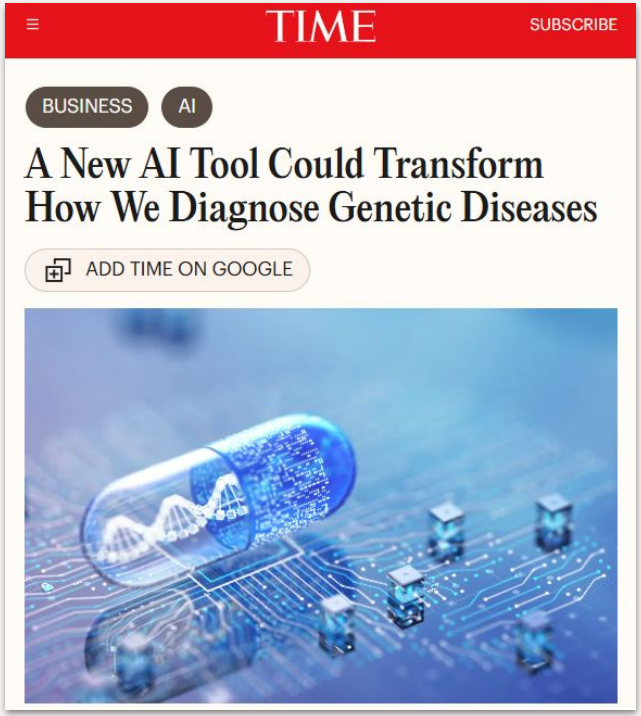
Holy people and places also span a curve.



(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.





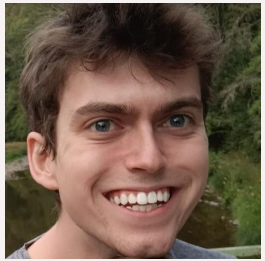
(Peirce, Dooms et al. .,*EVEE: Interpretable variant effect prediction from genomic foundation model embeddings*, 2026)

Want to know more or reach out?

Ward Gauderis

Geraint Wiggins

Scan the code!



Thomas Dooms

Jose Oramas

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP

Universiteit Antwerpen