Geometric Instability of Hidden-State Trajectories Predicts Reasoning Failures in Large Language Models

Hoda Fakharzadehjahromy[1], Andreas Bueff[1], Fredrik Heintz[1], Mattias Tiger[1], Qing Li[2], Jiaqi Li[3], Jiahui Geng[1]

[1]Linköping University [2]University of Groningen [3]The University of Hong Kong

LINKÖPING UNIVERSITY

TrustLLM

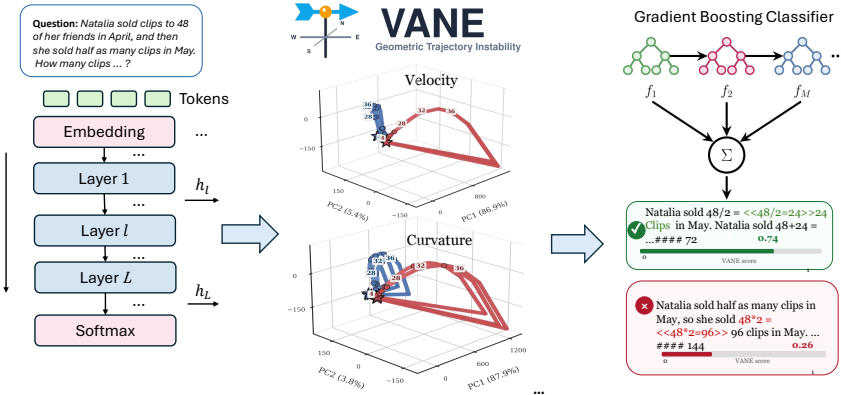
ELLIIT

NSC

VANE

Geometric Trajectory Instability

VANE reads how the hidden states move through the layers



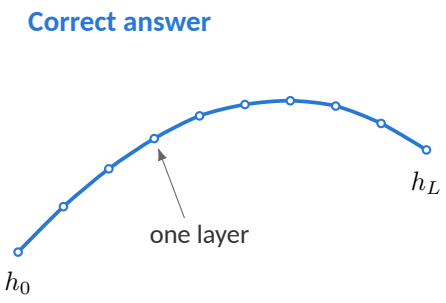
(a) Hidden state extraction through a single forward pass.

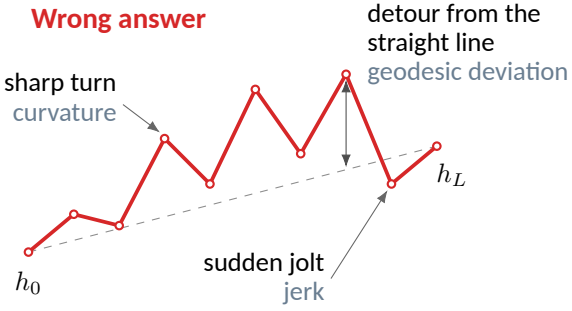
(b) Descriptor computation over internal state trajectories.

(c) Classifier training based on trajectory descriptors.

Follow one token through the layers of the model

Each dot is the hidden state of one token at one layer. This is a drawing, not data.



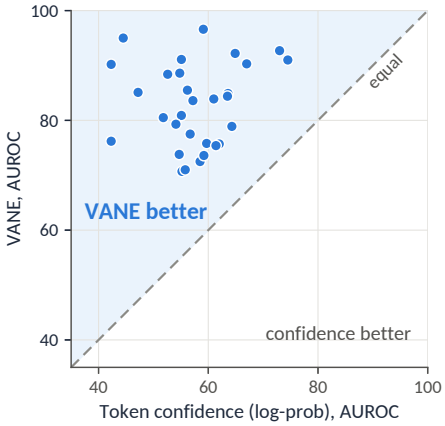


VANE uses five measurements of the path

Velocity	how far each layer moves the hidden state
Jerk	a sudden change between two steps
Curvature	how sharply the path turns
Geodesic deviation	how far the path goes from a straight line
Token coherence	whether the tokens move in the same direction

The measurements are fixed. Only the classifier is trained, with correctness labels.

VANE beats token confidence in every setting



Each dot is one model on one task.

Six models, from 1.5B to 72B parameters.

Five tasks in math, code and logic.

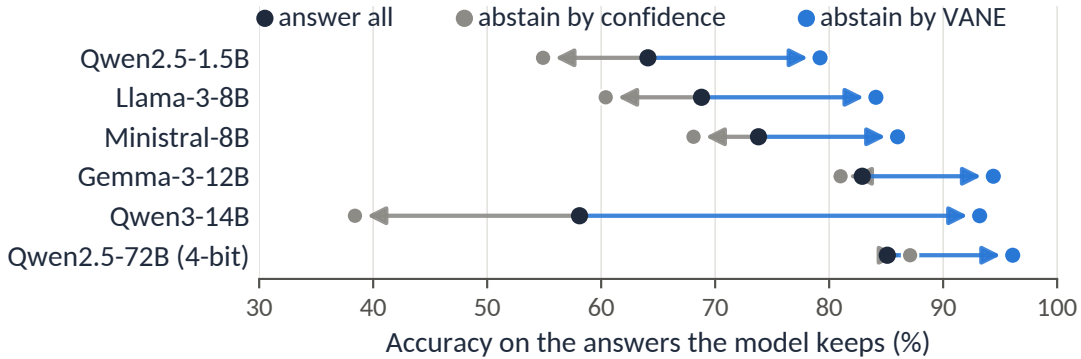
Every dot is above the line:
VANE detects wrong answers better.

VANE is also the best of seven methods in most settings.

Data: paper Table 1.

Abstaining on unstable answers raises accuracy

GSM8K (paper Table 7). The model keeps only the half of its answers that is most reliable.



Abstaining by token confidence lowers the accuracy of most models.

Robustness and limitations

Robustness

- Mostly transfers to new tasks
- Holds under 4-bit quantization
- Not due to length or format

Limitations

- Needs the hidden states
- Needs correctness labels
- Transfer is partial
- Correlation, not cause

Read the path, not only the answer

VANE
Geometric Trajectory Instability

Correctness depends not only on where the model ends, but on how it computes.

One forward pass. Useful for abstention.
Better than token confidence in every setting we tested.

hoda.fakhar@liu.se jiahui.geng@liu.se

Paper

Code

TrustLLM ELLIIT NSC