Stories in Space: A Model Tracks Story Arcs, Not Who Wrote Them

Kabui, Charles

2026-07-20

Read at ToKnow.ai

Goodfire researchers mapped how one 4-bit Llama 3.1 8B Instruct model changed its reading of a story sentence by sentence. The study used 2,500 training stories and 500 held-out stories from SimpleStories, a synthetic dataset with controlled styles. After each sentence, the model rated six emotions on a 0-to-10 scale while researchers also recorded its internal activity. Simple linear probes predicted the model's emotion and genre ratings with a root mean square

error of 0.09. Distances between concepts in the model's reported ratings and internal activity correlated at `r = 0.92` for emotions and `r = 0.89` for genres. Steering activity across seven layers shifted those ratings, though pushing sadness also affected nearby concepts such as anger.

This gives interpretability researchers a way to watch a model's idea of a story move over time, which could help trace where a summary or continuation starts reading the narrative differently. It does not reveal whether a person or AI wrote the text. The experiment never compared human-written and AI-written stories, and it tested one quantized model on synthetic data. The authors also hand-picked three domains with six concepts each, treated each concept as applying to the whole story rather than one character, and reported no replication on another model family. The Goodfire explanation is best read as a map of one model's changing state, not an authorship fingerprint.

Read More: Claude Has an Internal Workspace, Not Proof of Consciousness

Sources:

- Stories in Space paper
- Goodfire lab article
- SimpleStories dataset paper
- Llama 3.1 8B Instruct model card
