Negative Before Positive: Asymmetric Valence Processing in Large Language Models

Sohan Venkatesh
Manipal Academy of Higher Education, Bengaluru

The Central Claim

Mechanistic interpretability shows how concepts are encoded, yet emotional content stays poorly understood. Using activation patching and steering on three open-source large language models, we ask whether valence has dedicated internal structure — and find that **negative and positive valence are computed at different network depths**.

> Is valence a single internal axis? **No — negative is computed early, positive late.**

NEGATIVE VALENCE

Early
layers 14–27% deep
One dominant layer drives the response.

POSITIVE VALENCE

Mid–late
layers 53–66% deep
Spread diffusely across many layers.

Four Findings

1. **Depth Dissociation**
   Negative valence localizes to early layers (14–27% depth); positive peaks mid-to-late (53–66%). Highly significant in all three models (Mann–Whitney $p < 10^{-10}$).
2. **Valence, Not Topic**
   With topic and corrupted baseline held fixed, flipping valence flips the response sign 69–85% of the time — far above the 50% chance baseline.
3. **Asymmetric Localization**
   For bad news a single dominant layer predicts response strength (ρ −0.49 to −0.68); for good news no single layer dominates ($\rho \approx 0$).
4. **A Steerable Direction**
   A mean-difference good-news direction shifts 92–100% of neutral prompts positive and is monotonic in strength ($\rho > 0.89$) — valence is a manipulable linear direction.

One Pair, One Variable

A shared corrupted baseline holds topic fixed — only valence changes

GOOD	"I just got **accepted** into my dream PhD program today."
NEGATIVE	"I just got **rejected** from my dream PhD program today."
CORRUPTED	"I just received an email about my PhD program today."

Experimental Design

Three instruct-tuned models · two architectures · 100 clean/corrupted prompt pairs per condition

LLAMA-3.2 · MHA 1B | **QWEN2.5 · GQA** 1.5B 3B

A **shared corrupted baseline** isolates valence from topic. Pairs span academia, career and personal domains.

PATCHING
Residual-stream patch across every layer; logit-gap metric records the top causal layer.

FLIP TEST
Aligns good-news and negative pairs on one baseline; counts sign reversals.

STEERING
Adds a mean-difference direction to 50 neutral prompts; sweeps strength α.

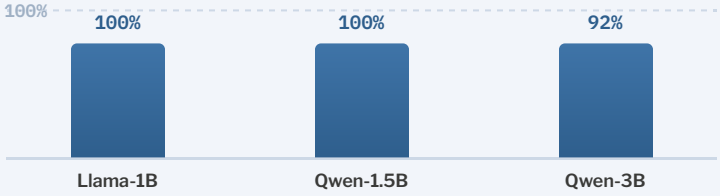
Models Track Valence

Sign accuracy — response has the correct direction · Flip rate — sign reverses when only valence flips

	GOOD-NEWS ACC.	NEG.-CTRL ACC.	FLIP RATE
LLAMA · MHA			
1B	79%	88%	69%
QWEN2.5 · GQA			
1.5B	92%	86%	80%
3B	89%	95%	85%

Steering Closes the Loop

Neutral prompts shifted toward positive valence by the good-news direction (α = +10)



Monotonic in steering strength across all models (Spearman $\rho > 0.89$, $p < 10^{-100}$) — the direction is genuinely linear and bidirectional.

Why early vs. late?
Negative outcome words (*rejected, failed, denied*) are lexically distinctive — detectable from surface features with little contextual integration. A warm response to good news first requires assembling what the outcome *means*, pushing positive valence deeper into the network.

Key Takeaways

- Emotional valence in LLMs is **localized, causal and steerable** — a concrete target for interpretability-based oversight.
- **Negative localizes early** (14–27% depth); positive peaks at mid-to-late layers (53–66%).
- A topic-controlled **flip test rules out surface token matching** (sign flips 69–85% of the time).
- **Negative routes through one dominant layer**; positive valence is distributed across many.
- The **good-news direction is manipulable** — it shifts 92–100% of neutral prompts positive.
- The early negative signal **appears before the model commits** to a response — a natural monitoring target.

ACCEPTED · ICML 2026 MECHANISTIC INTERPRETABILITY WORKSHOP

Full paper & code
arXiv: 2605.05653