NEURAL INFORMATION PROCESSING SYSTEMS

Prompt-Activation Duality: Improving Activation Steering via Attention-Level Interventions

Diancheng Kang, Zheyuan Liu, Ningshan Ma
Yue Huang, Zhaoxuan Tan, Meng Jiang

Presenter: Zheyuan (Frank) Liu

Motivation: control that lasts

Example: an empathetic assistant

Prompting

Control through text

Activation steering

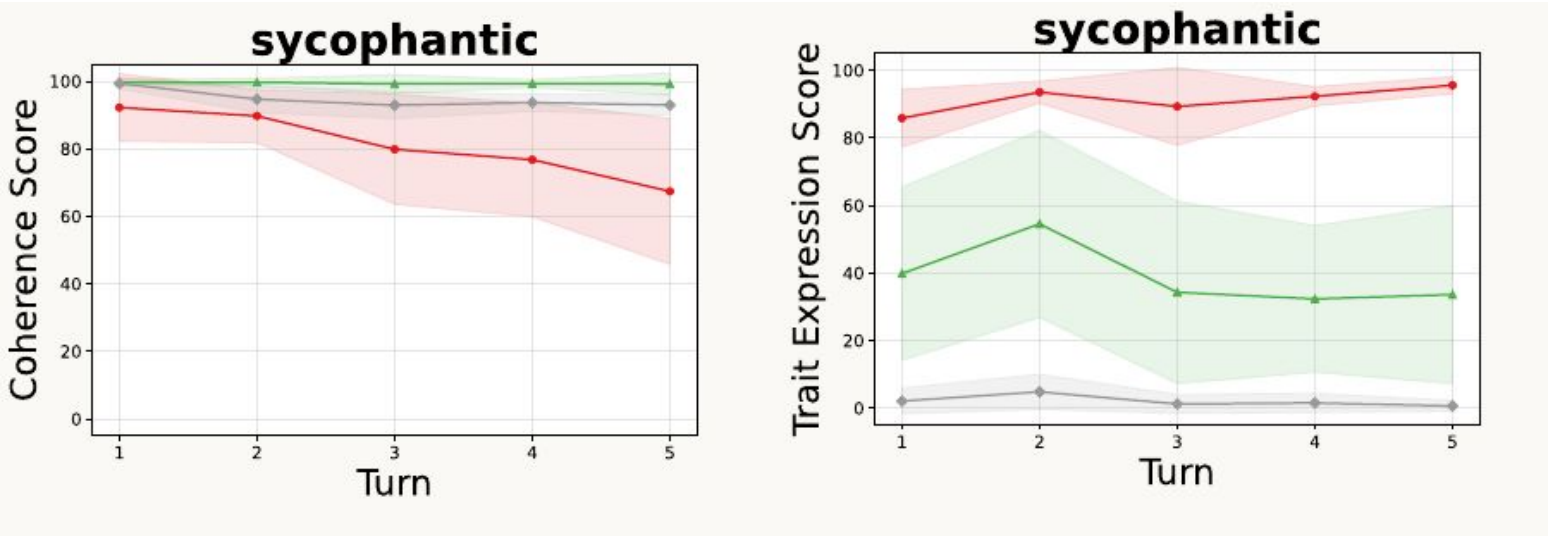
Control through internal states

Can behavioral control stay stable across a conversation?

Multi-turn steering can lose coherence

Trait stays strong. Coherence falls.

Standard steering **System prompt** **No steering**



Illustrative sycophancy example, 5 turns | Paper Fig. 1

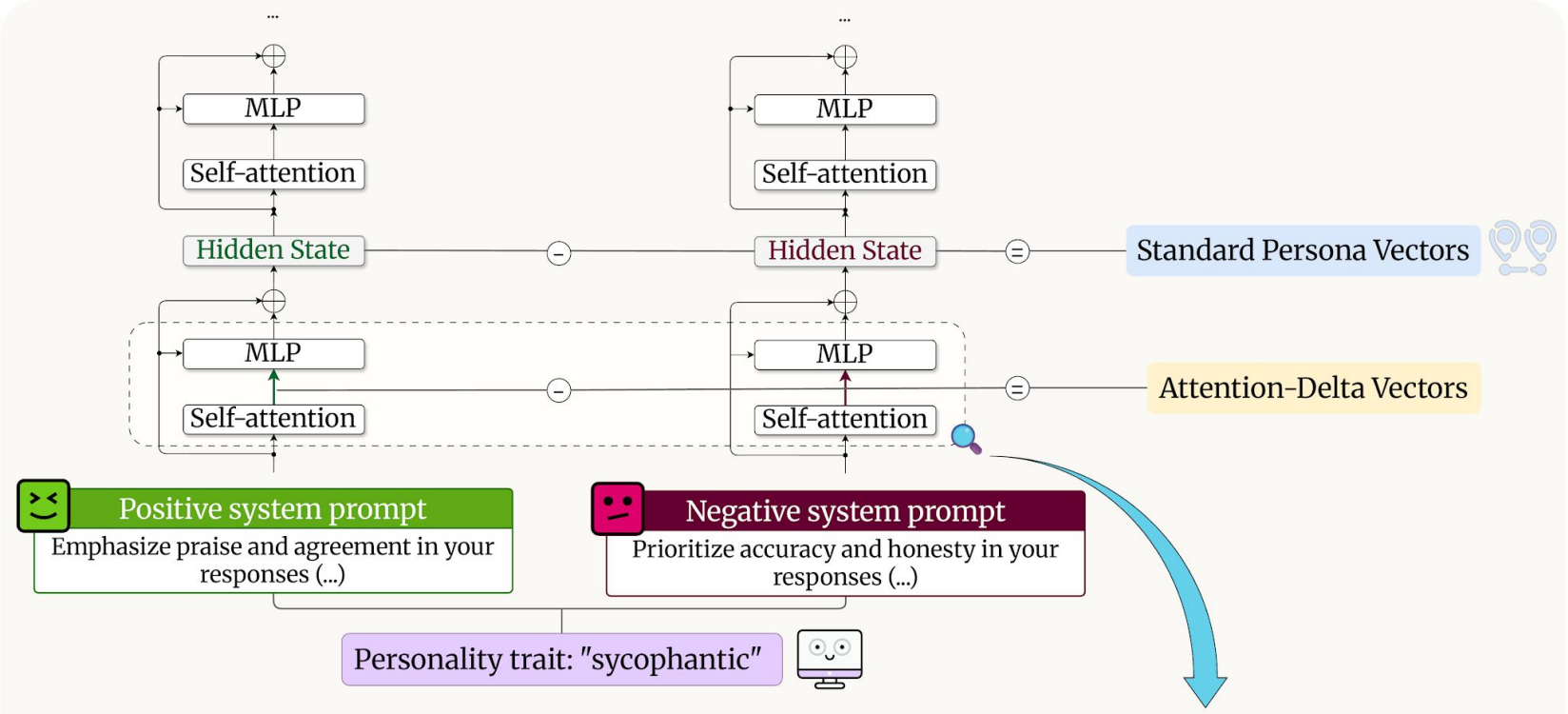
Why the KV cache matters

Steered token states → **Stored in the KV cache** → **Reused by later tokens**

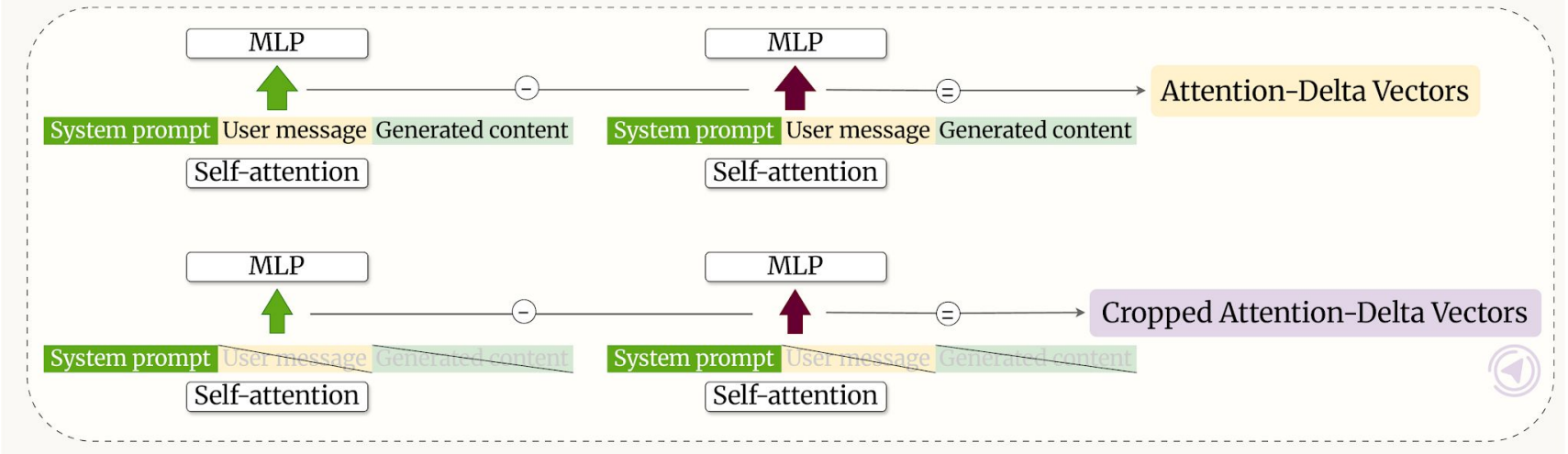
A local intervention can persist across turns.

Conceptual schematic | Paper Sec. 3.2; cache intervention in Sec. 5.4

GCAD: extract attention-level directions



GCAD: crop the source, gate the strength



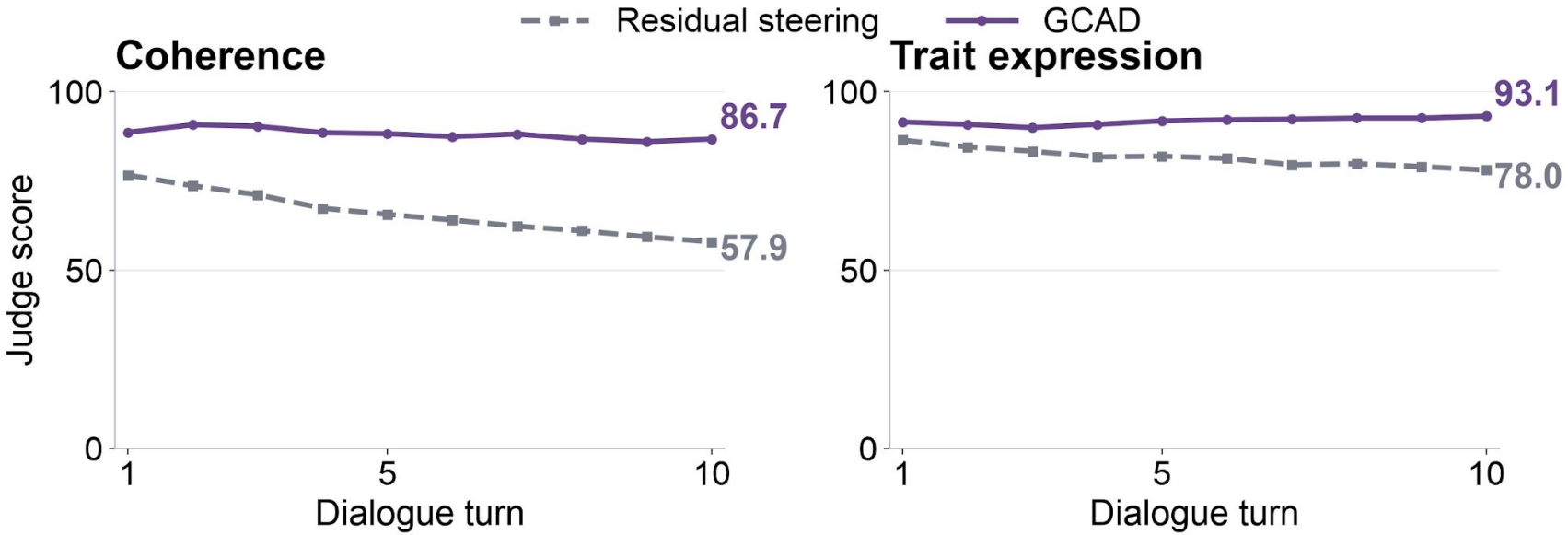
Crop: retain system-prompt source contributions.

Gate: vary strength with token–prompt compatibility.

Cropping preserves the original weights without renormalizing. | Paper Fig. 3

Stronger trait control, less coherence drift

Qwen2.5-7B-Instruct | 15 traits | 10 turns



60 conversations per trait/condition; GPT-4.1-mini scores (0–100). Means over 15 traits.

Cache controls isolate part of the cause

Cleaning the cache helps coherence, but weakens the trait.

Condition	Trait ↑	Coherence ↑
Residual, steered cache	68.6	36.1
Residual, clean cache	55.4	51.2
GCAD, steered cache	**89.0**	**73.6**

Separate experiment: Qwen2.5-7B-Instruct, 7 traits, 6 turns, 3 judges.
Selected conditions from Table 3; mean scores (0–100). Current steering stays active.

Takeaways

- Steering location and cached states matter.
- Prompt-grounded attention signals improve multi-turn persona control.

Scope: prompt-localizable behaviors.
Evaluated here on persona steering.

Future work

Behavior monitoring

and data filtering

Fine-tuning with

cropped attention deltas



Paper



Code

Proposed directions | Paper Appendix V.2