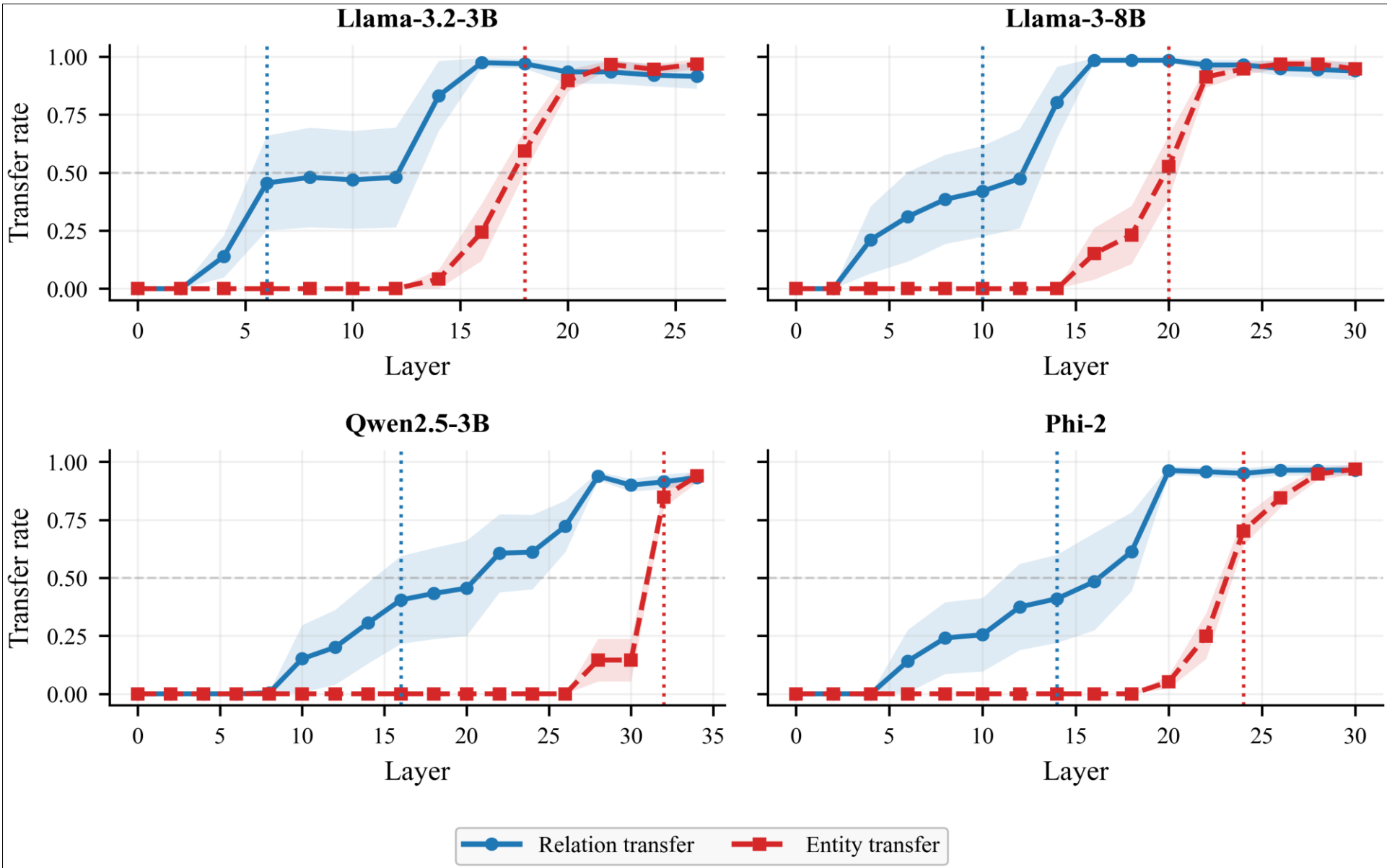
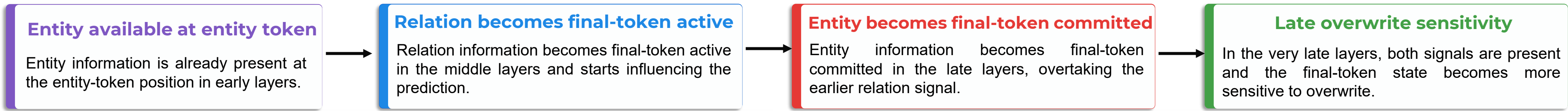


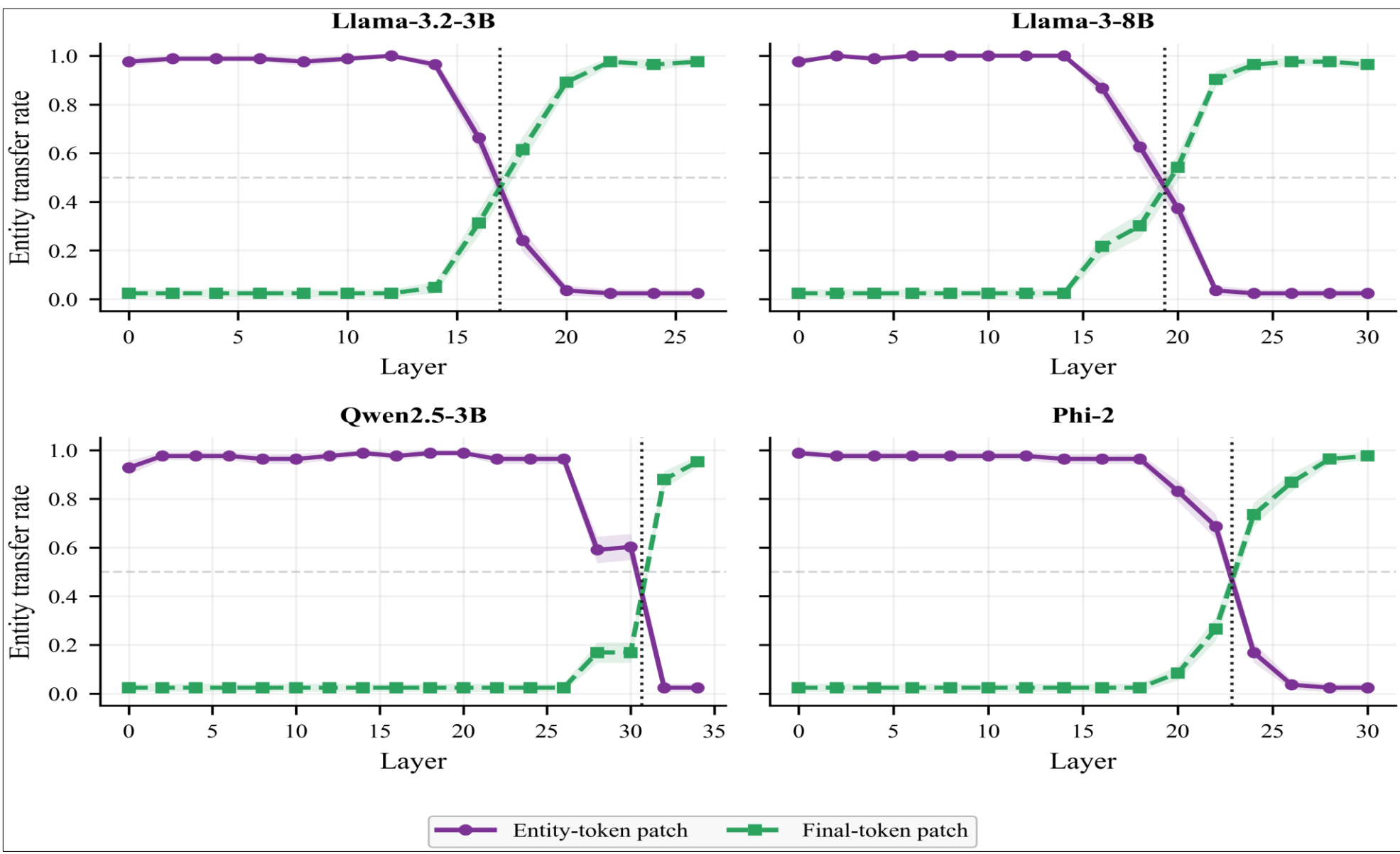
Information can be available before it controls generation.

Relation Before Entity: Deferred Commitment in Language Model Factual Recall

Background: Relation information becomes generation-controlling before entity-specific answer information, even though entity information is already available earlier at the entity-token position.

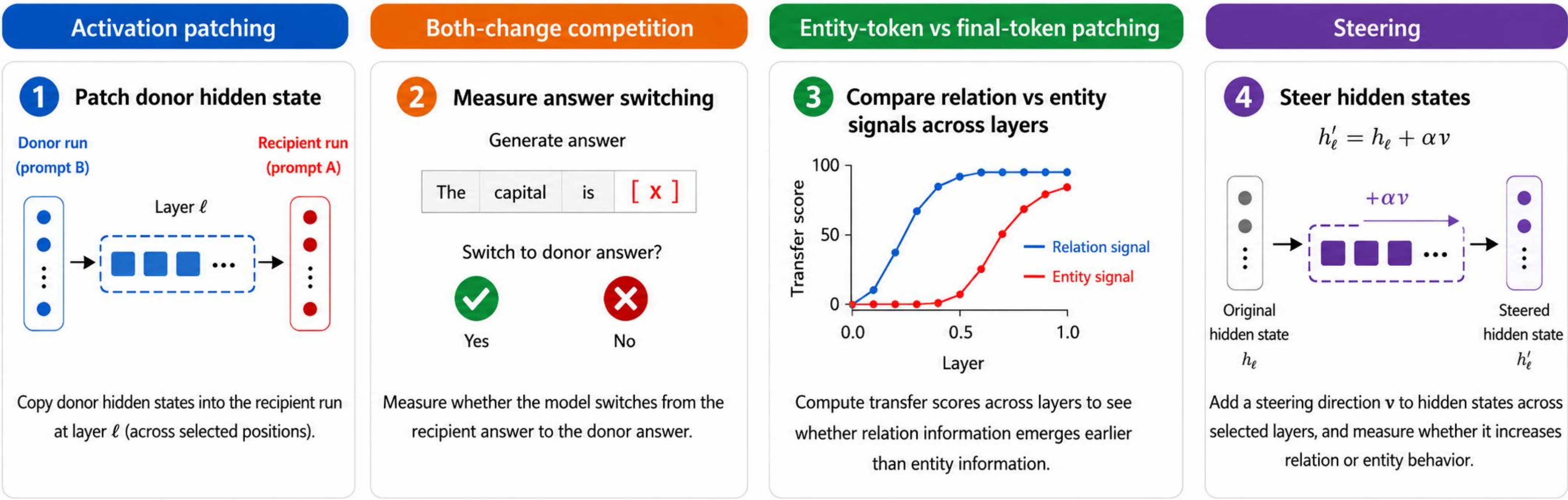


Relation onset precedes entity onset by 10–16 tested layers (31–44% of depth) at threshold 0.4.



Entity-token patching succeeds at 90–100% early, while final-token entity transfer rises only in late layers.

Methods



Limitation: Controlled fill-in-the-blank prompts in 3B–8B decoder-only models; the routing circuit behind deferred entity commitment remains future work.



PAPER

Divyansh Agarwal
University of Queensland
Mechanistic Interpretability Workshop at ICML 2026

CODE

