

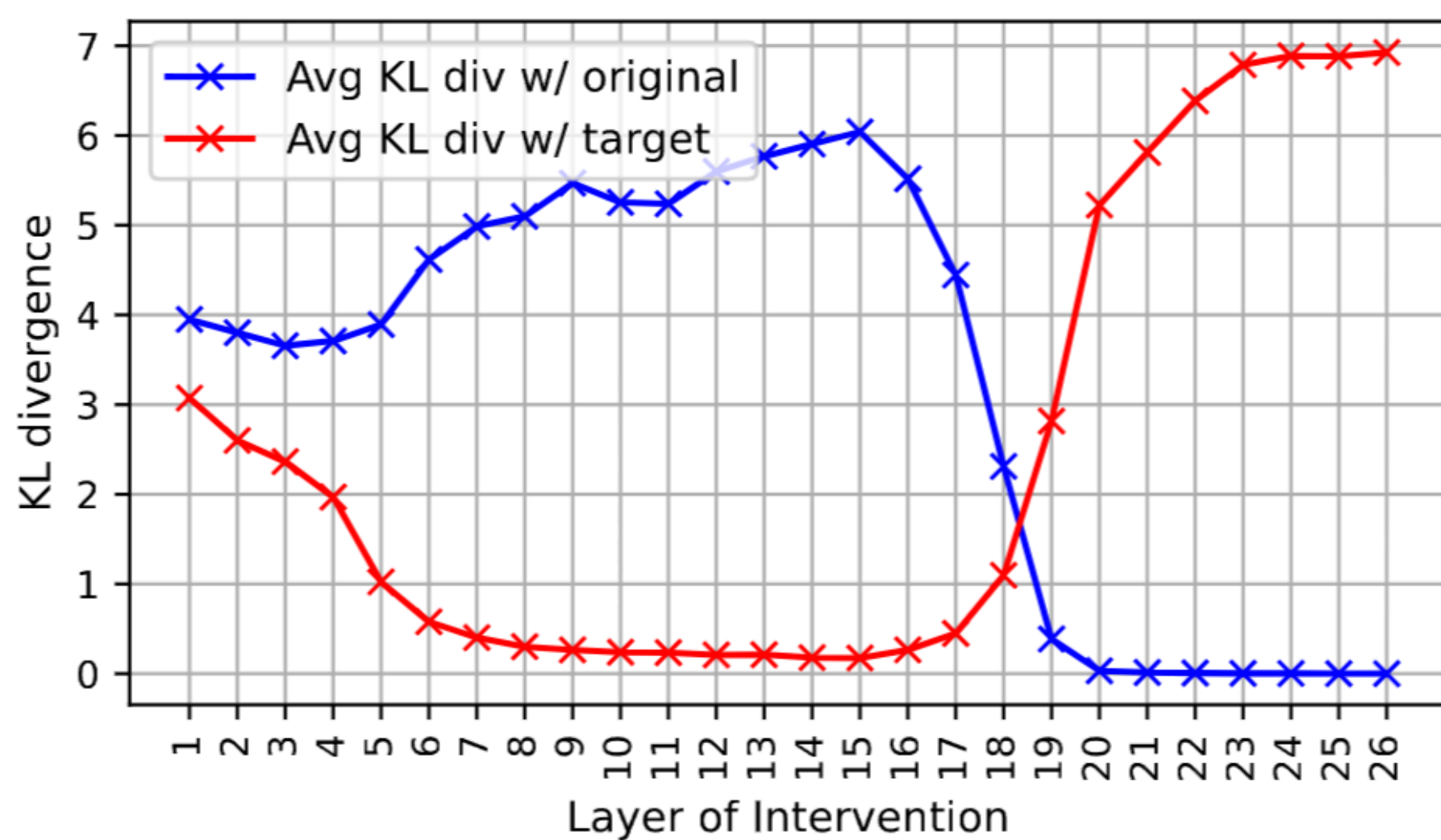
Frozen in the Middle: Hidden States Remain Unchanged Across Intermediate Layers of Language Models

Pavel Tikhonov (AIRI & Skoltech), Dmitry Ilvovsky (HSE University)

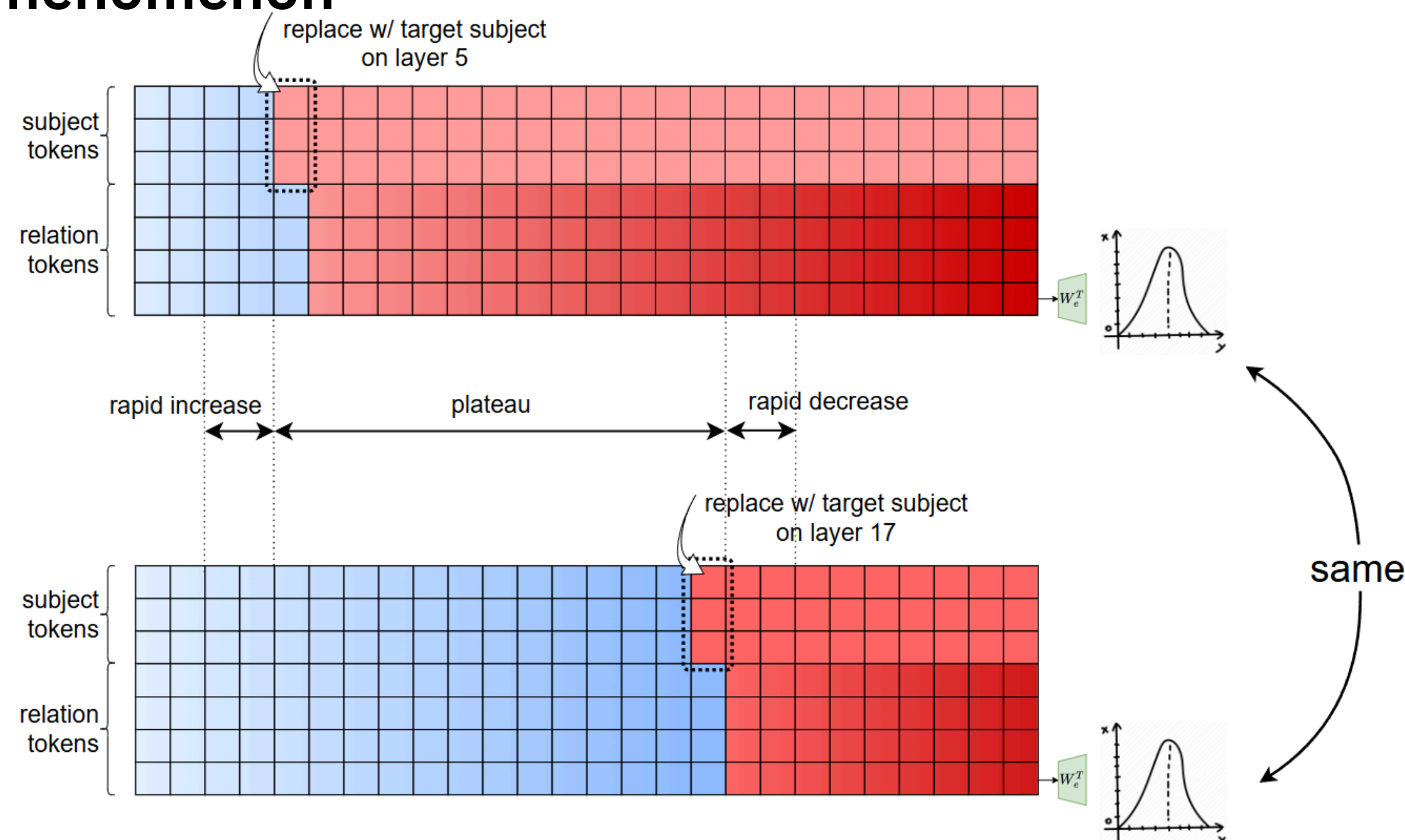
Given a prompt “*The Eiffel Tower is located in the city of*” an LLM correctly outputs “*Paris*”. But how exactly is this information **processed** through the network’s layers?

TL;DR: across large spans of middle layers (often dozens), **hidden states of subject tokens remain essentially the same**. Other tokens deliberately **ignore** this readily available information until reaching specific deeper layers, despite it being **necessary** for their predictions.

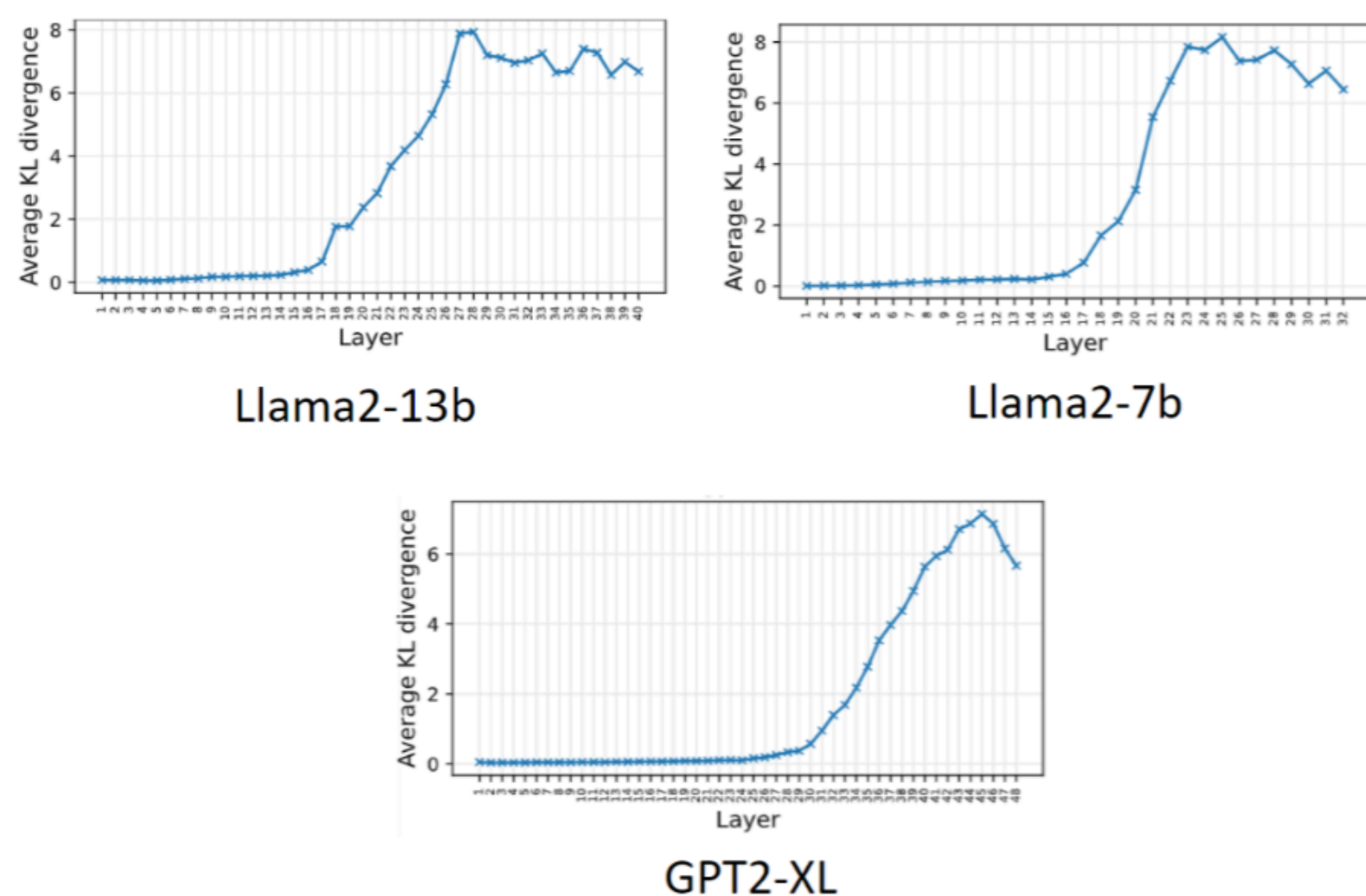
The Plateau Phenomenon



(*s*, *r*, *o*): (“*The Eiffel Tower*”, “is located in the city of”, “*Paris*”)
(*ŝ*, *r*, *ô*): (“*The Colosseum*”, “is located in the city of”, “*Rome*”)

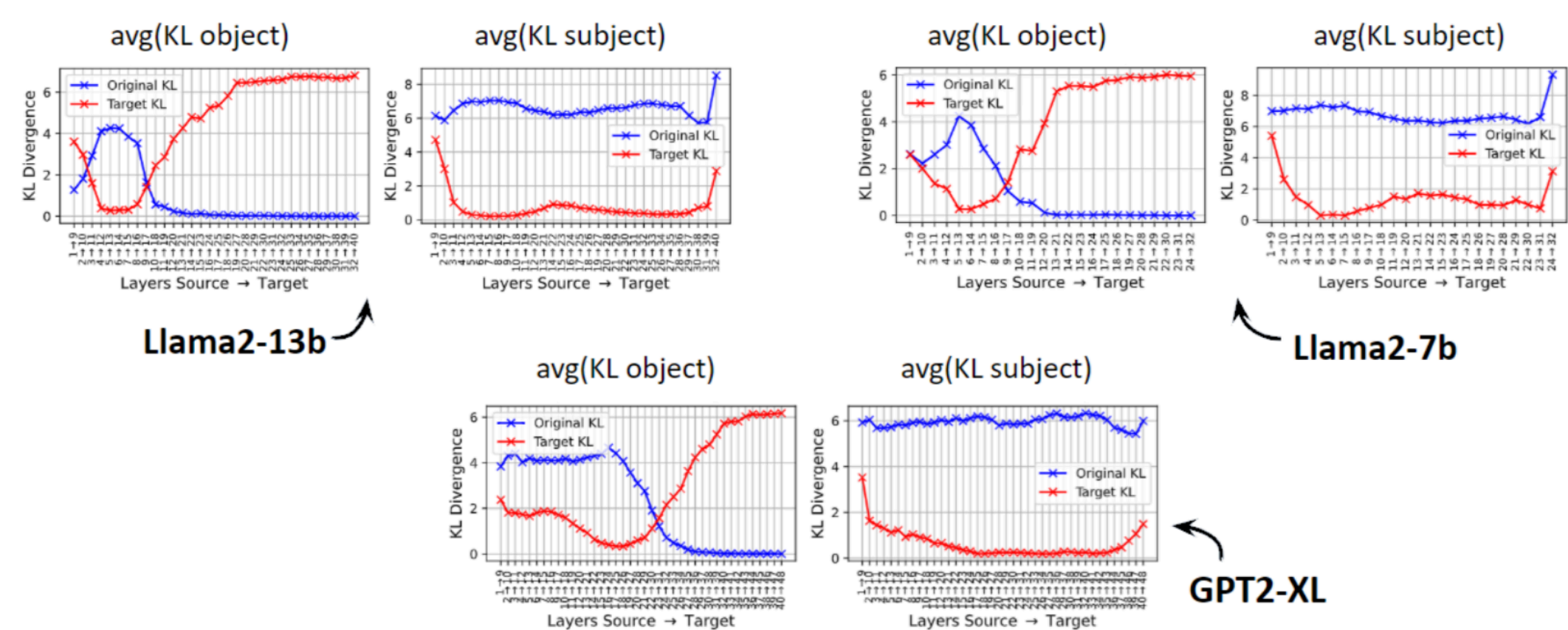


(1) When does information exchange occur between subject and relation tokens?



The near-zero divergence until the end of the plateau phase indicates that **relation tokens are not influenced by subject tokens before this point**. This provides strong evidence that **information exchange happens only during a narrow range of layers near the end of the plateau**.

(2) Do subject token states undergo transformations during the plateau?



Even when replacing states with those from much **earlier** layers, both factual outputs and next-token predictions remain **unchanged**. This indicates that **subject token states during the plateau neither undergo transformations nor contribute to immediate next-token prediction**, while still maintaining all information needed for the eventual factual output.

Three-Phase Pattern in of Processing Factual Information

- (1) Enrichment Phase (Early Layers):** The first few layers process token inputs and enrich hidden states with necessary factual information.
- (2) Plateau Phase (Middle Layers):** Subject token states stabilize and remain unchanged while containing all necessary information for the final prediction.
- (3) Extraction Phase (Late Layers):** Information is finally exchanged between subject and relation tokens, leading to factual prediction

Model Name	Layers	Avg Onset Layer	Avg End Layer
GPT-2-XL	48	15.7 ± 2.1	30.1 ± 1.8
GPT-J-6b	28	5.2 ± 2.3	15.0 ± 1.7
OpenLlama-3b	26	5.2 ± 1.9	17.3 ± 0.9
Mistral-7b	32	4.3 ± 1.8	16.2 ± 2.8
Llama2-13b	40	4.2 ± 2.0	17.5 ± 2.5
Llama2-7b	32	4.6 ± 1.7	16.7 ± 2.2
Llama3-8b	32	5.2 ± 2.6	15.9 ± 1.3

Table 1: Plateau range of different models ($\mu \pm \sigma$)