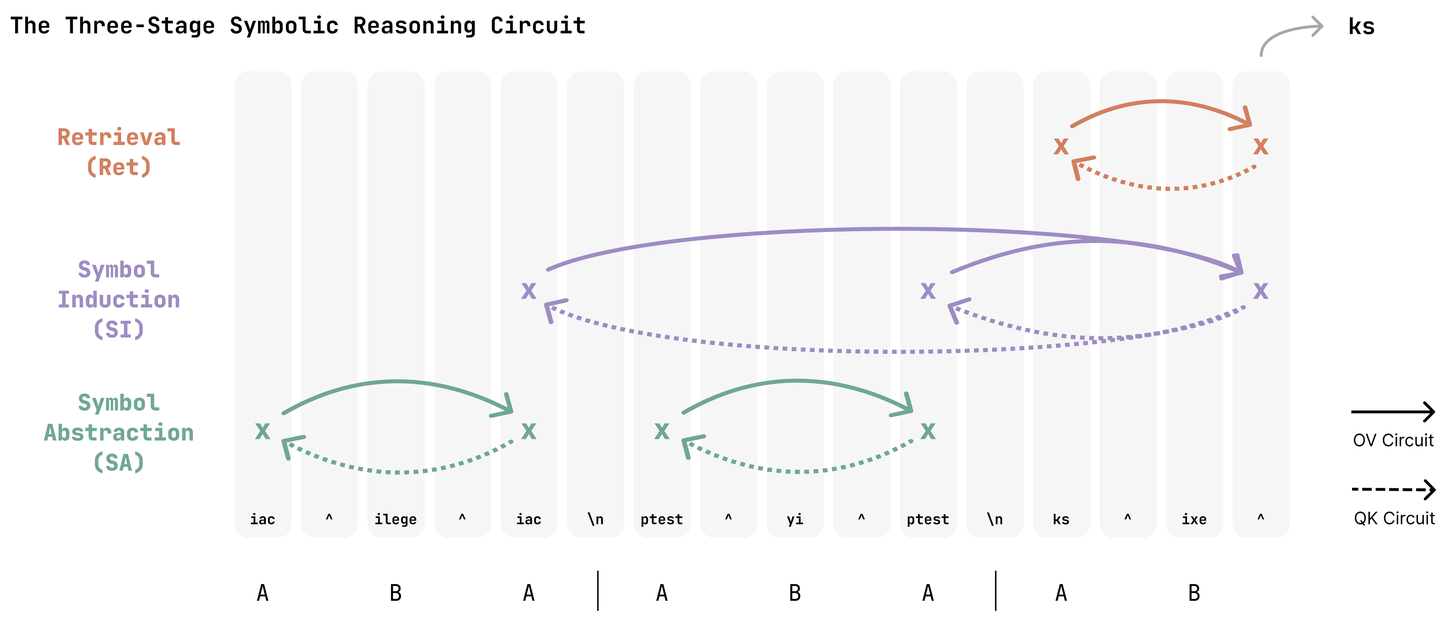
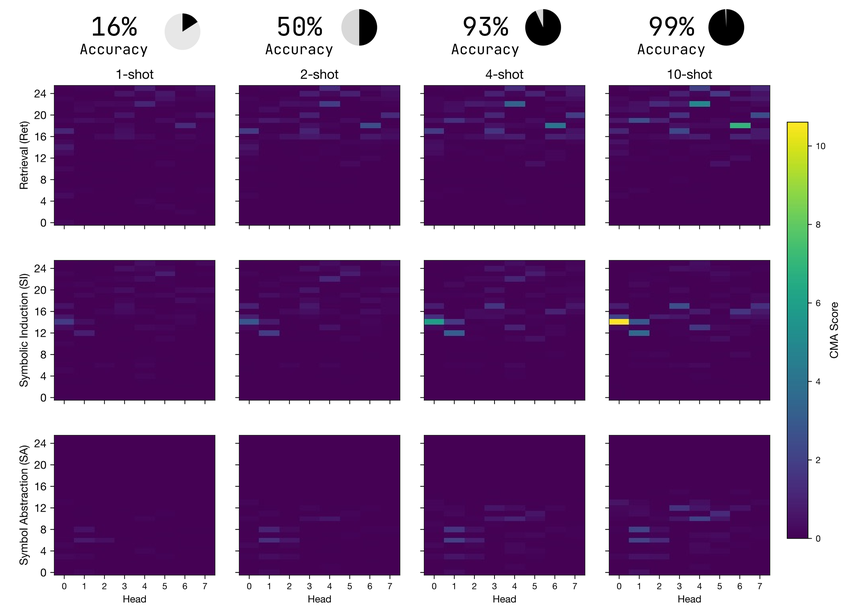
In-Context Learning Amplifies a Latent Symbolic Circuit

*A three-stage symbolic reasoning circuit is **detectable, functional, and partially token-invariant before the model achieves high accuracy**. Demonstrations, activation patching, and function vectors all act as alternative inputs to the same latent pathway.*

Observed in Gemma, Llama and Qwen model families, Gemma-2-2B results reported below.

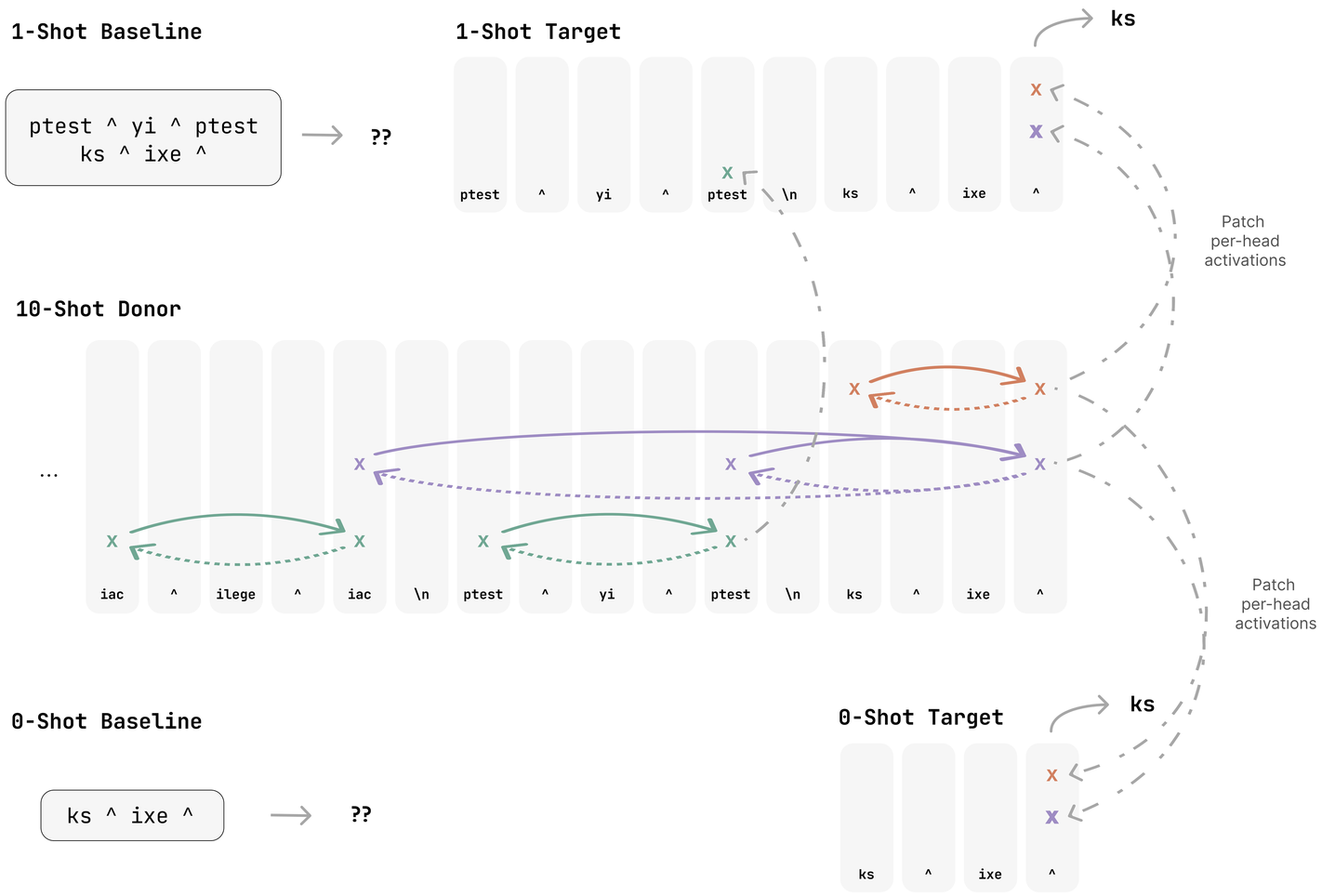
§ 01 Circuit Amplification Across Shot Counts

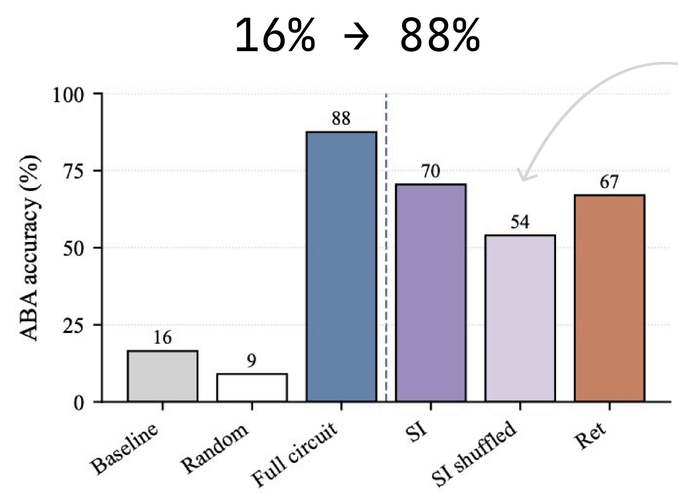




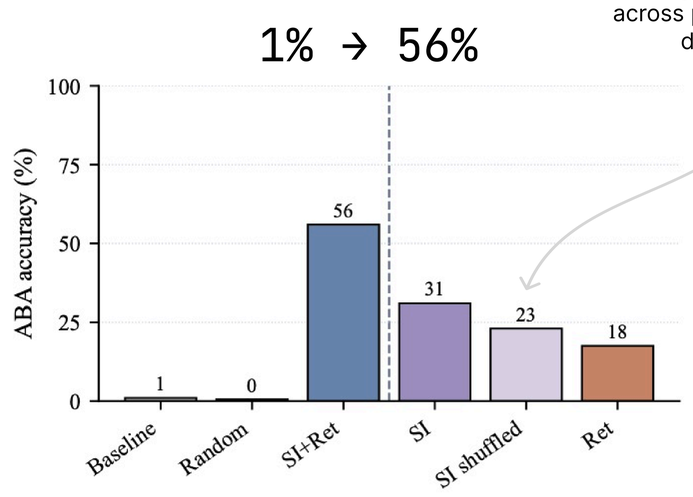
Per-head impact grows up to **8x** across shot counts

§ 02 Cross-Shot Activation Patching

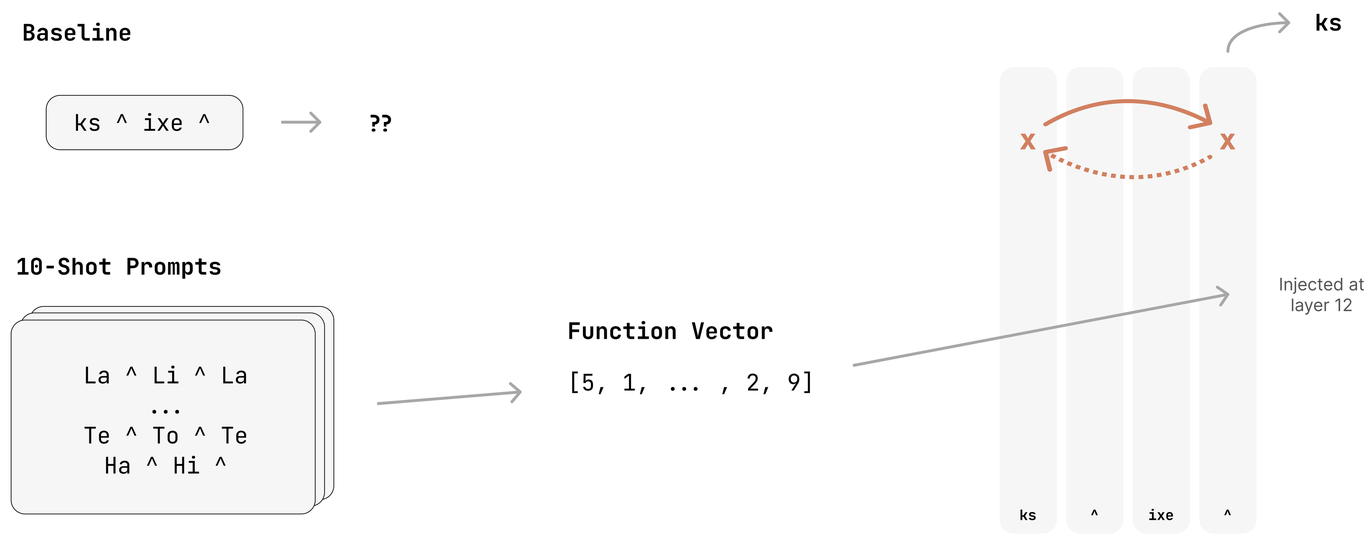


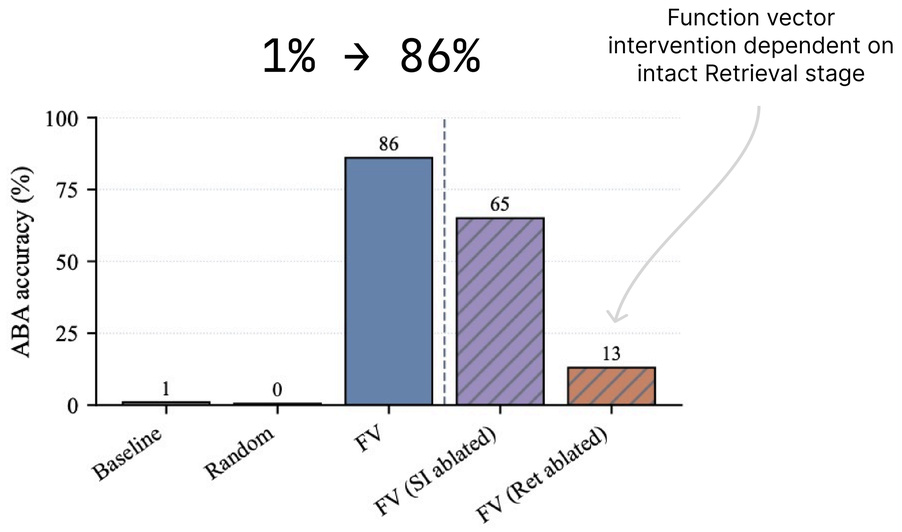


SI stage improves accuracy across prompts with entirely different tokens



§ 03 Function Vectors as Circuit Signal





Takeaways

- The infrastructure for abstract rule following is in the pretrained weights before demonstrations
- Detectable before high task accuracy, demonstrations amplify this latent pathway up to 8x
- The pathway can be driven by function vectors, dependent on the downstream circuit being intact