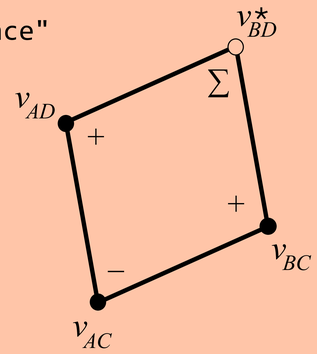
LLMs represent ICL **functions** as **vectors.**

Input: `"Italy, Russia, China, Japan, France"`

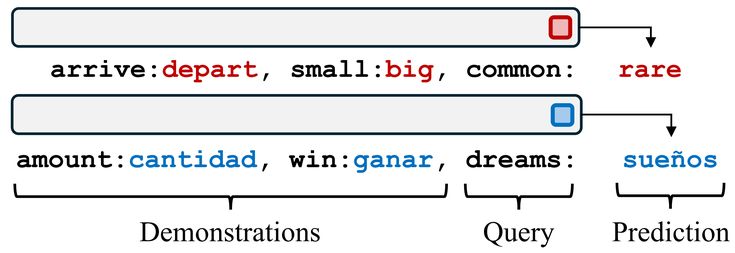
FV	Task	Expected Output
v_{AC}	First-Copy	`Italy`
v_{AD}	First-Capital	`Rome`
v_{BC}	Last-Copy	`France`
$v^{\star}_{BD}$	Last-Capital	`Paris`



Function Vectors in Large Language Models

Eric Todd, Millicent L. Li, Arnab Sen Sharma,
Aaron Mueller, Byron C. Wallace, David Bau

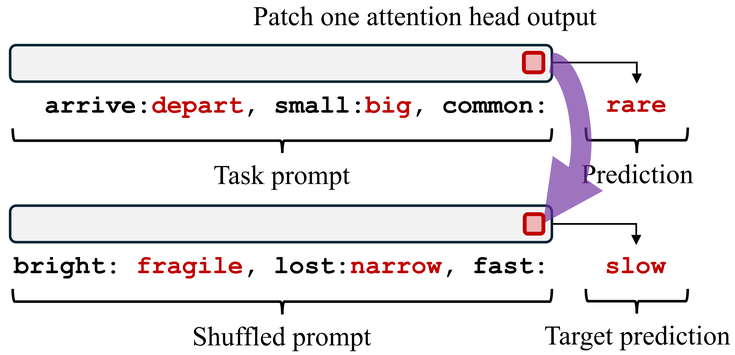
Website: functions.baulab.info



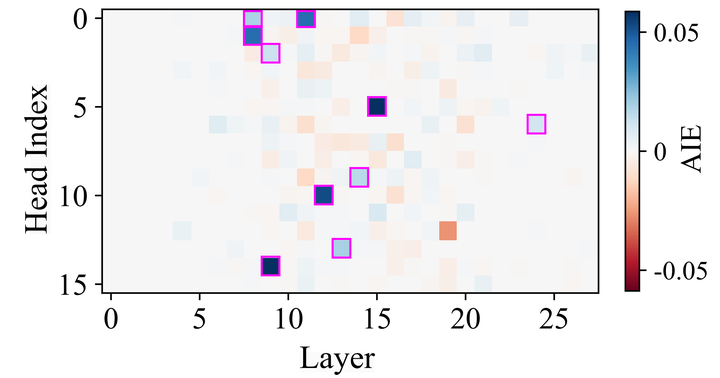
In-context learning can induce function-like behavior in LLMs.

How do LLMs **represent** an ICL task?

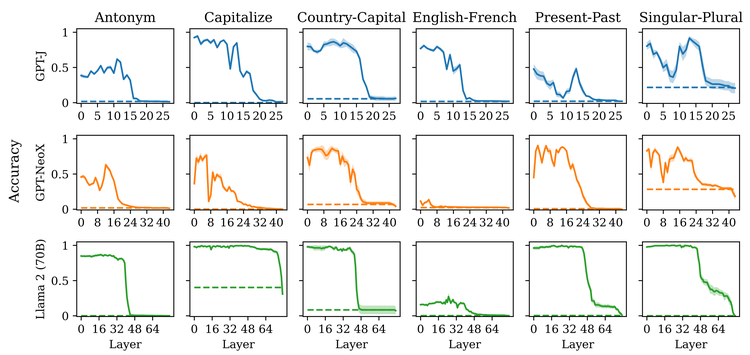
(1) We patch activations to localize **function vectors** that **identify a task**



(2) A small number of **attention heads transport** these ICL-specified **functions**



(3) Function Vectors **cause performance of a task** when applied in new settings



Prompt:	The word “daily”, means
GPT-J	“every day”.\n\nThe
GPT-J + English-French FV	“tou les jours”,

Prompt:	When you think of Netherlands,
GPT-J	You probably think of tulips, windmills, and cheese. But the Netherlands is also home to…
GPT-J + Country-Capital FV	You think of Amsterdam. But there are many other cities in the Netherlands. Here are some…

Northeastern University