

LLMs compose functions **two ways** – **compute the intermediate step**, or **skip straight to the answer**. Embedding geometry **predicts which**.

How Do Language Models Compose Functions?

Apoorv Khandelwal and Ellie Pavlick · Brown University

RESEARCH QUESTION

Do LLMs solve a compositional task $x \rightarrow y$ **compositionally** – computing an intermediate $z = f(x)$ latently – or do they map $x \rightarrow y$ **directly**?

THE TASKS

$x \rightarrow f(x) \rightarrow g(f(x))$

18 two-hop tasks · one example per family

LEXICAL
bogus → authentic → auténtico

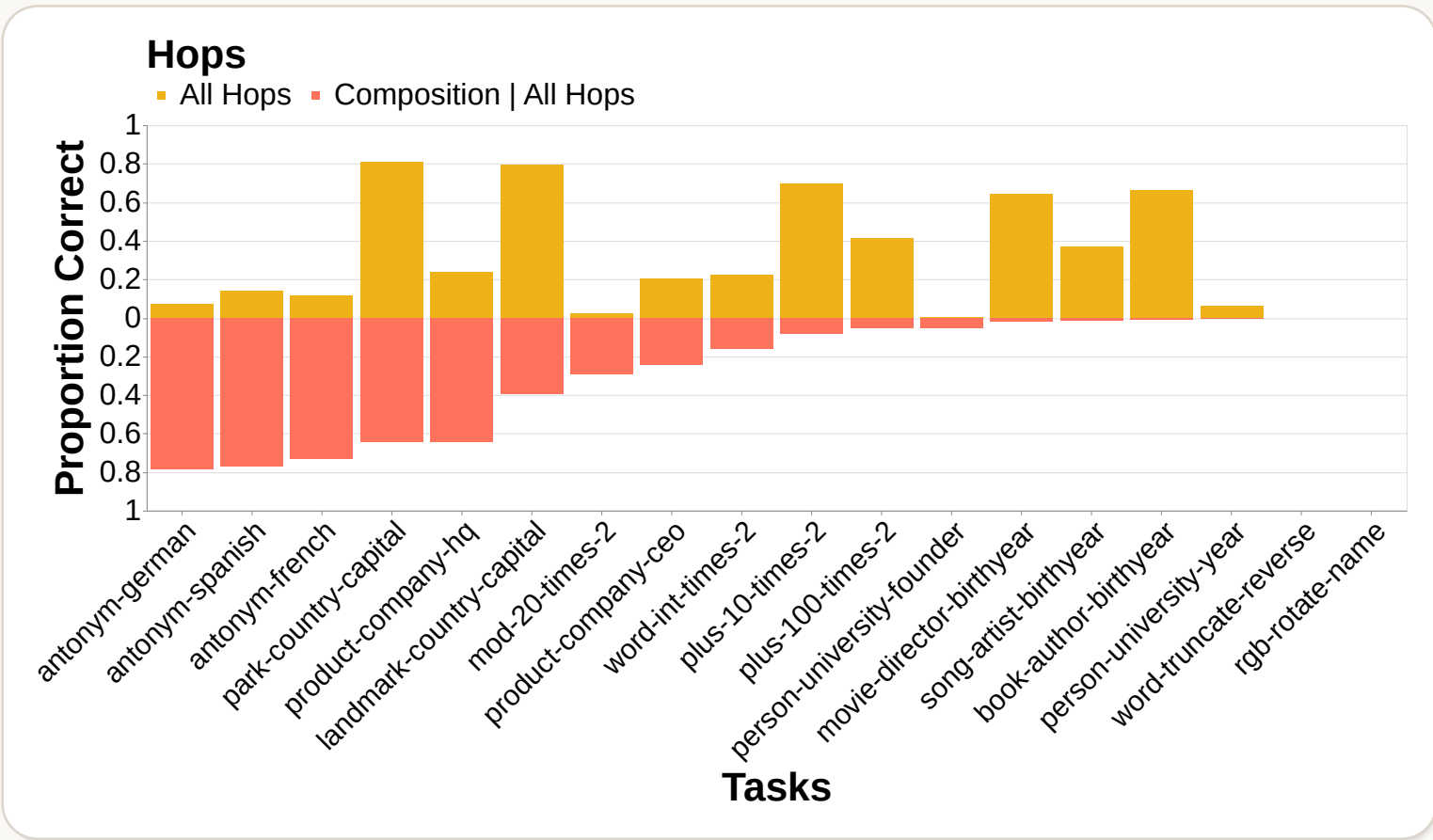
FACTUAL RECALL
Cape Fear → Martin Scorsese → 1942

ARITHMETIC
922 → 1022 → 2044

STRING · COLOR
8a735a → 598a73 → dimgray

1 Compositionality Gap

Models fail at $x \rightarrow g(f(x))$ even when they solve **every hop** – and the gap barely closes with scale.



Per-task gap · Llama 3 (3B)

72% → 39%

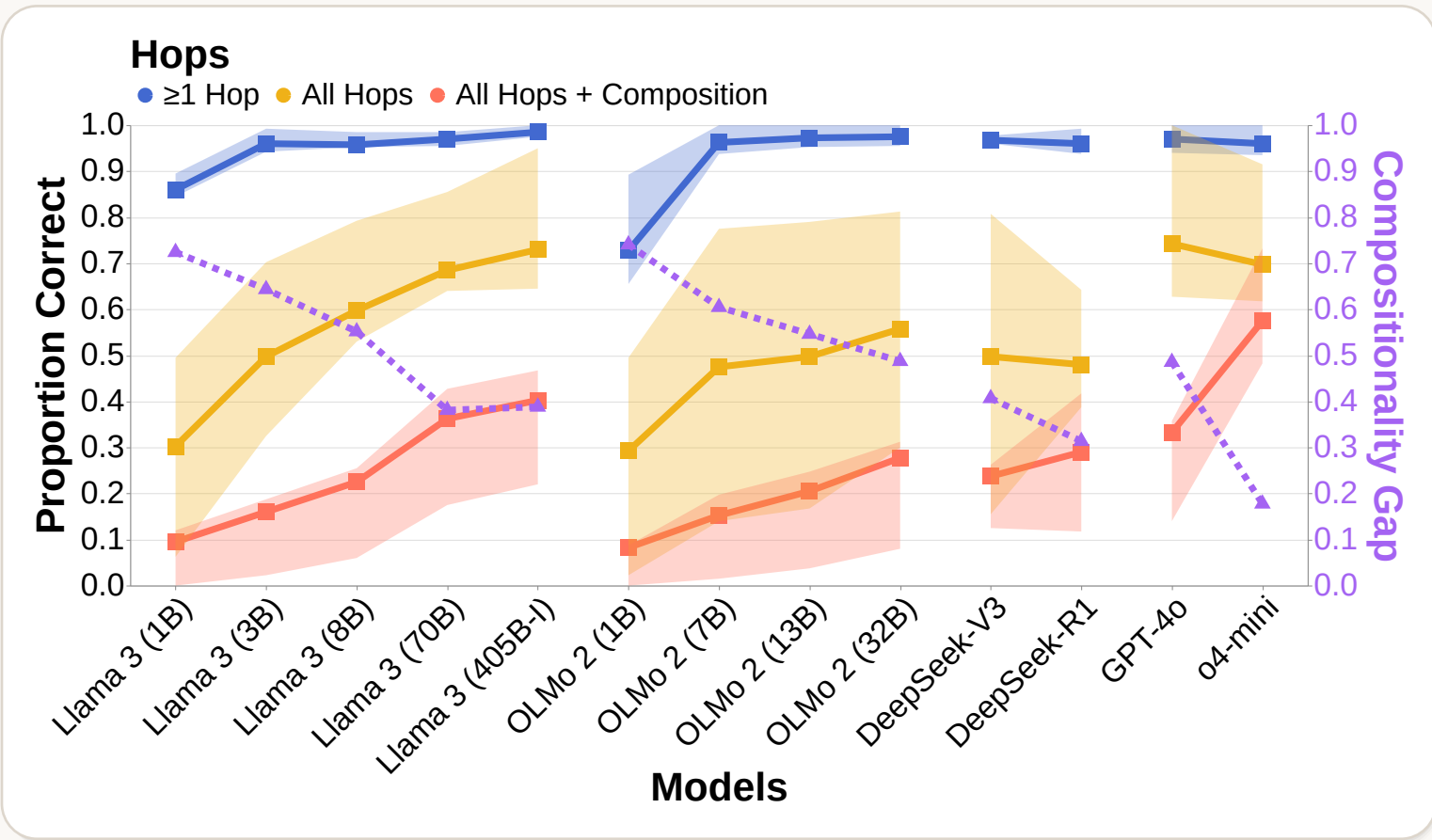
Llama 3 gap · 1B → 405B

74% → 49%

OLMo 2 gap · 1B → 32B

No gain 70B → 405B

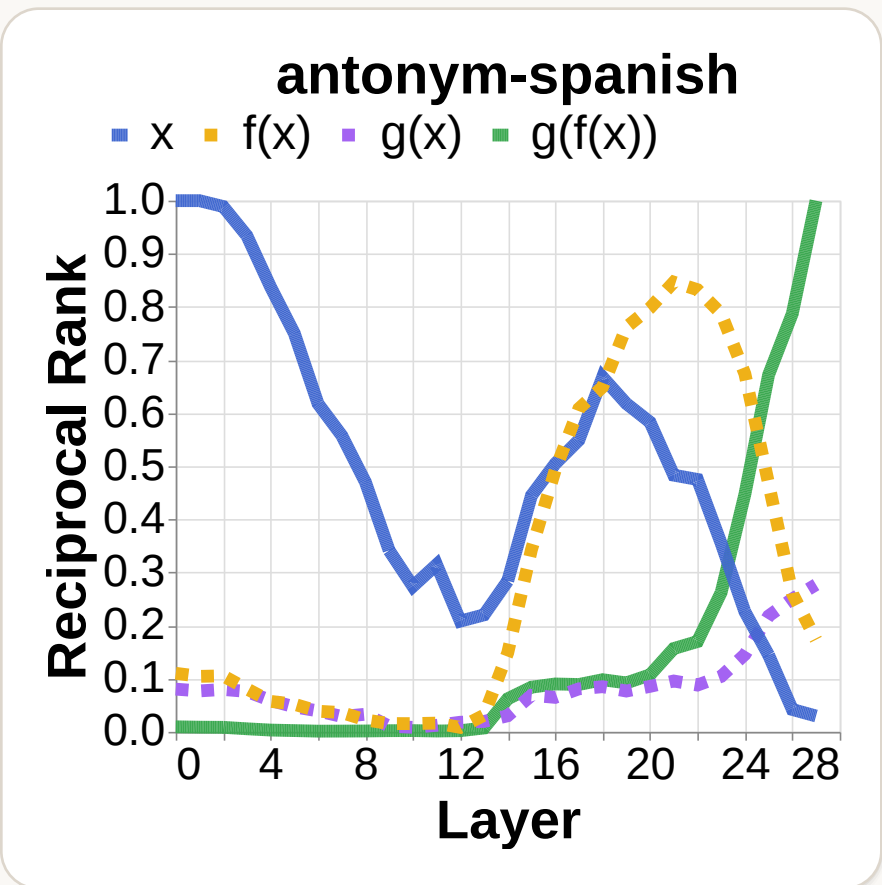
the gap does not vanish at scale



Aggregated over 12 models · Llama 3, OLMo 2, DeepSeek, GPT

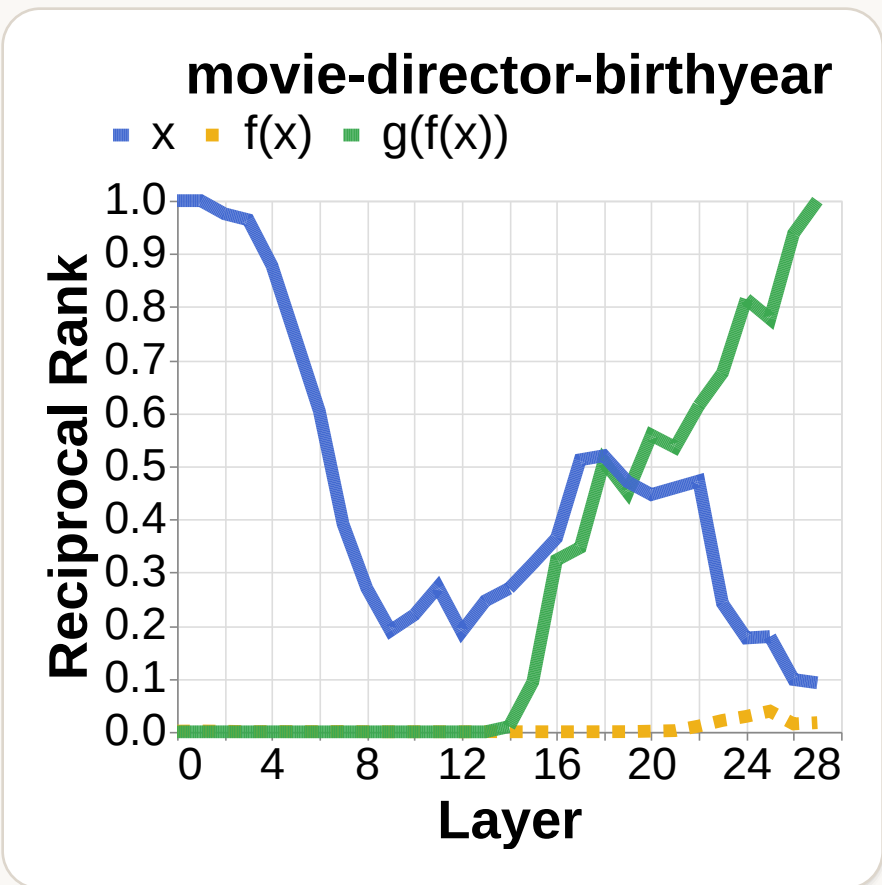
2 Logit lens reveals **two types of processing signatures** during the latent computation (one forward pass) when models do predict $g(f(x))$ correctly.

COMPOSITIONAL $f(x)$ peaks mid-network

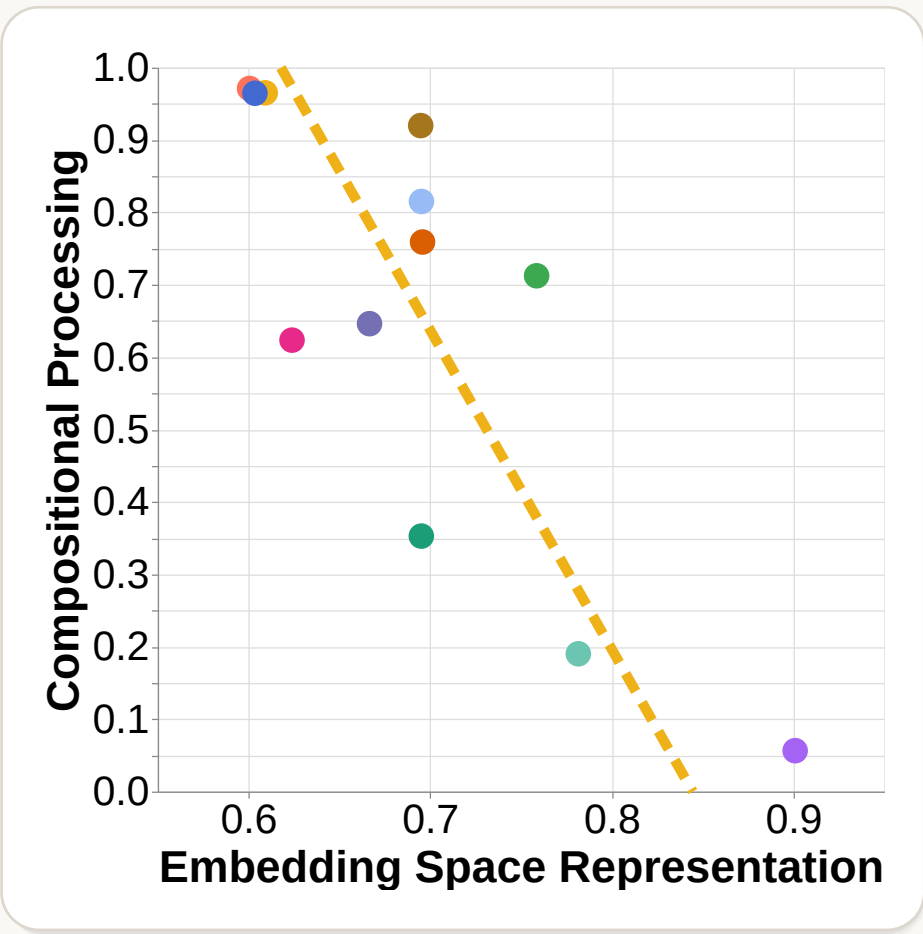


Reciprocal rank of $x, f(x), g(x), g(f(x))$ across layers · Llama 3 (3B), correct solutions

DIRECT no trace of $f(x)$



3 Embedding geometry predicts the mechanism



Each point = one task · dashed line = best-fit translation · Llama 3 (3B)

The better a task is captured as a straight **translation** $x \rightarrow g(f(x))$ from the embedding to unembedding spaces, the more the model **skips** the intermediate.

$r^2 = 0.65$

geometry vs. mechanism · Llama 3 (3B)

$r^2 = 0.66$ avg

holds across all 5 models tested

82% of cases pick a side

mechanism choice is strongly bimodal

TAKEAWAY

We tie a fundamental property of LLMs – their **embedding spaces** – to the mechanisms they invoke, hinting at which functions are **“primitive”** to a model.



PAPER

arXiv:2510.01685

CODE

github.com/apoorvkh/composing-functions

Apoorv Khandelwal · Ellie Pavlick

Mechanistic Interpretability Workshop · NeurIPS 2025

BROWN
Department of Computer Science