LLMs can answer with or without a decoded intermediate. Embedding geometry predicts the signature.

How Do Language Models Compose Functions?

Apoorv Khandelwal · Ellie Pavlick · Brown University

RESEARCH QUESTION

Do LLMs compute **`z = f(x)`** on the way to **`y = g(f(x))`**, or map **x** to **y** without a detectable intermediate?

18 TWO-HOP TASKS

Mechanism analyses: 10 ICL demonstrations · one forward pass · Llama 3 (3B)

LEXICAL
`bogus → authentic → auténtico`

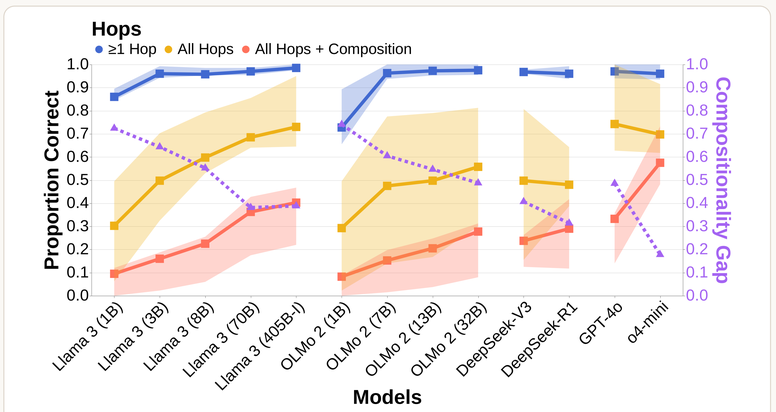
FACTUAL RECALL
`Cape Fear → Martin Scorsese → 1942`

ARITHMETIC
`922 → 1022 → 2044`

STRING · COLOR
`8a735a → 598a73 → dimgray`

1 Compositionality gap

Solving every hop does not guarantee solving the composition. Scaling and reasoning can reduce the gap.



13 models · four tasks · 100 queries per task; bands show IQR.

72% → 39%
Llama 3 · 1B to 405B

74% → 49%
OLMo 2 · 1B to 32B

NEWER MODELS: SMALLER GAPS

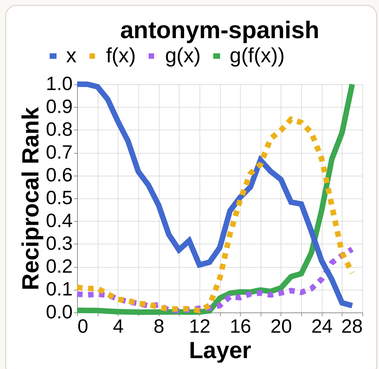
Model	No reasoning	Medium reasoning
Claude Opus 5	4.6%	2.3%
GPT-5.6 Sol	3.2%	2.6%
Gemini 3.6 Flash	13.4%	1.1%

Mean task-level gap on the same four tasks; a behavioral measure.

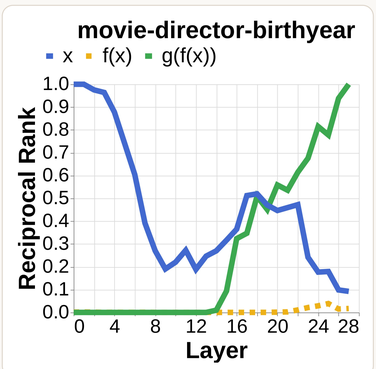
2 Processing signatures

Correct solutions can show a strong intermediate signal, or little signal in the decoded vocabulary space.

COMPOSITIONAL



DIRECT



Reciprocal rank (RR) across layers · correct solutions

A weak decoded signal does not prove the absence of an intermediate representation.

AGREEMENT WITH LOGIT LENS

$r^2 = 0.68$
Token identity patchscope

$r^2 = 0.95$
Jacobian lens

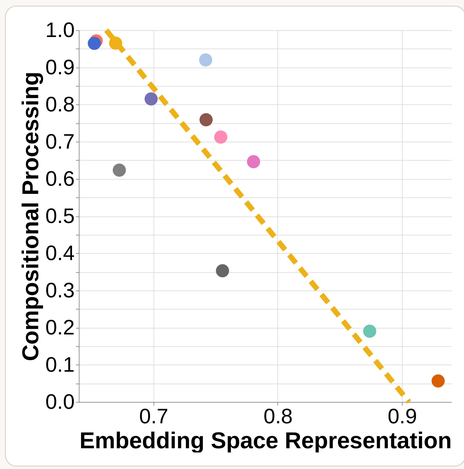
THE PROMPT ALSO MATTERS

+0.323 ± 0.068 RR

Zero-shot gain: book-author-birthyear and plus-100-times-2.
Mean ± SD across these two tasks.

3 Embedding geometry

Stronger task-level embedding structure is associated with weaker decoded intermediate signals.



Each point = one task; dashed line = linear fit.

$r^2 = 0.77$
Geometry vs. signature
Llama 3 (3B)

$r^2 = 0.71$
Average across
five models

82%
of analyzed examples:
peak RR <0.1 or ≥0.5

Vocabulary-centered, mean answer-token unembeddings; geometry scored on examples held out from 100 fitting examples.

PATCHING SUPPORTS CAUSAL TRANSFER

+0.126 ± 0.034 RR

Counterfactual capital gain from single-hop country donors;
mean ± SD across two task pairs, compositional receivers.
Whole-stream patches do not isolate the intermediate alone.

TAKEAWAY

Embedding spaces offer clues to which functions a model treats as "primitive".
Geometry predicts the observed signature; task, example, and prompt all matter.

PAPER · arXiv:2510.01685
CODE · github.com/apoorvkh/composing-functions

Apoorv Khandelwal · Ellie Pavlick

NeurIPS 2026

BROWN
Department of Computer Science