

COMPOSITIONAL REASONING IN LARGE LANGUAGE MODELS

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Aisha Khan

B.S., Computer Science, Vertex University
M.S., Electrical and Computer Engineering, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

May 2026

© Aisha Khan, 2026

All Rights Reserved

Acknowledgments

My doctoral committee was Professor Jonathan M. Reyes (chair), Professor Helen Adeyemi, Professor Sung-min Cho and Dr. Clara Voss; I thank each of them for their guidance. I thank my labmates and colleagues in the Department of Electrical and Computer Engineering for their help and company. This work was supported by a graduate fellowship and by computing credits from a research cloud program.

Abstract

Large language models answer many questions that require combining facts, yet they fail in revealing ways when a question combines familiar parts in an unfamiliar arrangement. This thesis studies that gap. It introduces a benchmark that separates compositional structure from surface familiarity, shows that step-by-step prompting helps most on shallow compositions and least on deep ones, and proposes a decomposition procedure that plans sub-questions before answering them. On compositions of depth four, the procedure raises exact-match accuracy from 31% to 58% for a model of fixed size.

Contents

Acknowledgments	iii
Abstract	iv
1 Introduction	1
2 Background	2
3 Plan-Then-Answer Decomposition	3
4 Experiments	4
5 Conclusion	5
A Benchmark Construction	6

List of Tables

4.1	Exact-match accuracy (%) by composition depth.	4
-----	--	---

List of Figures

3.1	Plan-then-answer decomposition.	3
-----	---	---

Chapter 1

Introduction

A model that knows who wrote a novel and where that author was born should be able to say where the author of the novel was born. Humans compose knowledge this way without effort; neural sequence models often do not [3, 2]. Large language models narrow the gap [1], and prompting them to reason step by step narrows it further [4], but the remaining errors are systematic.

This thesis measures those errors precisely and proposes a way to reduce them.

Chapter 2

Background

We model a compositional question as a tree whose leaves are atomic facts and whose internal nodes are operations such as lookup, comparison and counting. Its depth k is the length of the longest path from the root to a leaf. For a model that answers each atomic step correctly with probability p and independently, the chance of answering a depth- k chain is

$$P_k = p^k, \tag{2.1}$$

so accuracy should fall geometrically with depth. Measured accuracy falls faster, which is the signature of a model that does not decompose the question at all.

Chapter 3

Plan-Then-Answer Decomposition

The procedure first asks the model to write a plan, a list of sub-questions whose answers compose to the final answer, and then answers each sub-question in a fresh context that contains only the answers it depends on. Separating the contexts prevents an early error from being copied into later steps.



Figure 3.1: Plan-then-answer decomposition.

Chapter 4

Experiments

Table 4.1 reports exact-match accuracy by composition depth for a single model under three prompting strategies.

Table 4.1: Exact-match accuracy (%) by composition depth.

Strategy	Depth 1	Depth 2	Depth 3	Depth 4
Direct answer	88	61	42	22
Step-by-step	90	74	51	31
Plan-then-answer	89	79	67	58

The gain from planning grows with depth, consistent with Equation (2.1): isolating each step keeps its error rate close to the atomic rate $1 - p$.

Chapter 5

Conclusion

Compositional failures in language models are largely failures to decompose. Asking the model to plan first and answer each part in isolation recovers much of the accuracy that depth takes away, without changing the model.

Appendix A

Benchmark Construction

Questions were generated from a synthetic knowledge base of 12,000 invented entities so that no answer could be memorized from pre-training data. Each question was checked by two annotators.

Bibliography

- [1] Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
- [2] Daniel Keysers et al. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020.
- [3] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2873–2882, 2018.
- [4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.