Characterizing Pattern Matching and Its Limits on Compositional Task Structures

Hoyeon Chang*[1], Jinho Park*[1], Hanseul Cho*[1], Sohee Yang[2], Miyoung Ko[1], Hyeonbin Hwang[1], Seungpil Won[3], Dohaeng Lee[3], Youbin Ahn[3], Minjoon Seo[1]

[1]KAIST [2]UCL [3]LG AI Research | {retapurayo,binlepain178,jhs4015}@kaist.ac.kr

Key Contributions

- We formalize pattern matching as functional-equivalence-based generalization with a precise, testable boundary.
- We prove tight data scaling laws : 2-hop training data for complete ID generalization grows as $N_{\text{req}} = \tilde{\Theta}(n^c)$ with exponent $2.0 < c \leq 2.5$, which robustly holds across $20\times$ model scaling and architectures.
- We identify path ambiguity as a structural barrier: when a variable influences output via multiple paths, models fail to form unified representations, even with CoT.

Our Formalization of Pattern Matching

Key Insight: Functional Equivalence

If two input fragments **consistently produce the same output** in all shared contexts observed during training, a learner can substitute one for the other to predict unseen combinations.

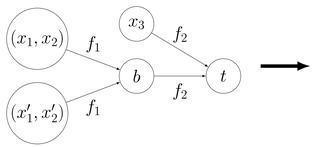
Definition: Functional k-Equivalence

Two input subsequences a, a' at positions I are **functionally k-equivalent in D** if there exist $\geq k$ contexts b (at the remaining positions I^c) such that both (a, b) and (a', b) appear in D **and always yield the same output**: $f(a, b) = f(a', b)$.

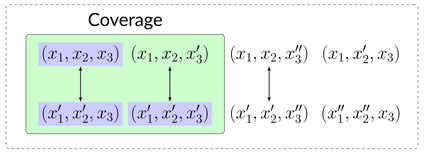
Definition: k-Coverage (← Boundary of Pattern Matching)

The **k-coverage** of D is the set of all inputs reachable from observed data through chains of functionally k-equivalent substitutions in the **substitution graph**.

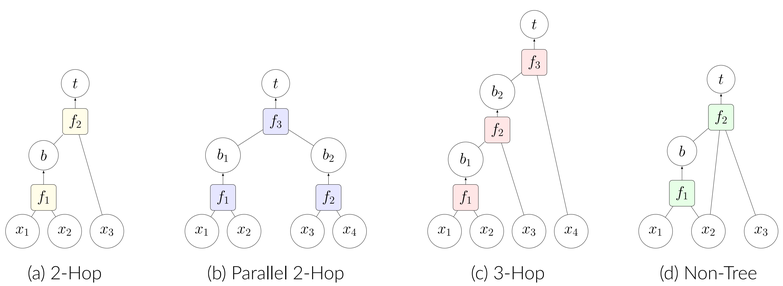
Functional Equivalence Observation



Substitution Graph



Experimental Setup: Four Task Structures

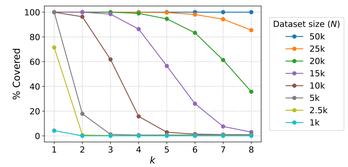


(a) 2-Hop (b) Parallel 2-Hop (c) 3-Hop (d) Non-Tree

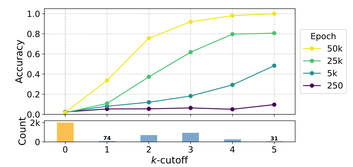
- Primitive functions $f_1, f_2, \ldots$ are **randomly generated** to isolate pattern matching from algebraic shortcuts
- Train GPT-2 (68M–1.5B) and Mamba from scratch; $|\mathcal{X}| \in \{50, \ldots, 200\}$

Evidence Strength Predicts Pattern-Matching Success

Stronger evidence ⇒ **faster & more reliable generalization!**

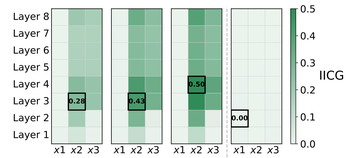


% of ID test data inside k-coverage vs. k and dataset size N.

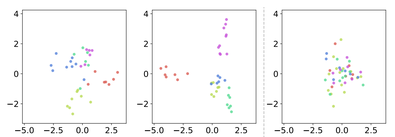


Test accuracy by k-cutoff; Out-of-coverage (k=0) stays at chance.

Representation Clusters Drive Pattern Matching

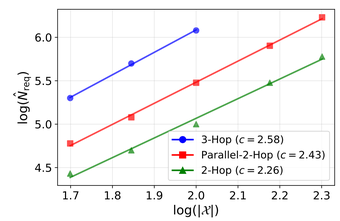


IICG $= \overline{\cos}_{\text{intra}} - \overline{\cos}_{\text{inter}}$: Quantifies clustering by intermediate state b.

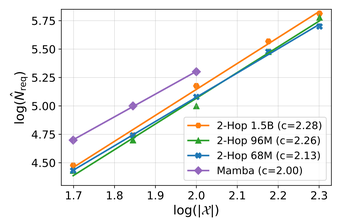


Higher k-cutoff ⇒ Stronger clustering. Out-of-coverage: **no clustering**.

Data Scaling Law: Empirical Confirmation

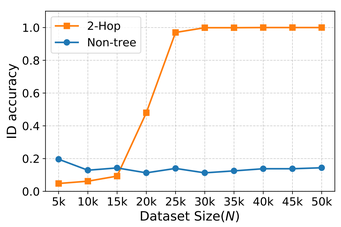


Our theory explains the empirical 2-Hop exponent c=2.26; Deeper structures scale steeper.

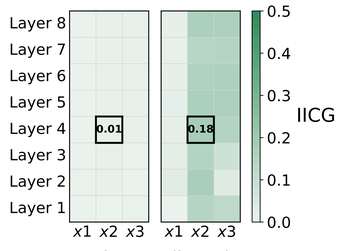


Exponents **invariant across** $20\times$ **model scaling** and architectures ⇒ a **data property**, not capacity!

Path Ambiguity in 'Non-Tree' Task & Chain-of-Thought



2-Hop achieves full ID generalization; Non-Tree fails even with near-exhaustive data.



Model forms **context-dependent** representations (i.e., high IICG for (b, x_2), near-zero for b).

CoT "flattens" multi-hop (improving scaling laws) but **does not fully resolve** path ambiguity.

Theorem 6.1 (Tight Sample Complexity for 2-Hop Task)

For a 2-Hop task with token set size n and an i.i.d. train data of size $|D| = N_{\text{tr}}$, a learner generalizing within k-coverage achieves **perfect ID generalization** w.h.p. if

$$N_{\text{tr}} \geq \widetilde{O}\left(n^c\right) \quad \text{with} \quad c = 2.5 - \frac{0.5}{k}.$$

Conversely ($k \geq 2$), it **cannot** for some 2-Hop task if $n^2 \lesssim N_{\text{tr}} \leq \widetilde{\Omega}\left(n^c\right)$. (We omit $\text{polylog}(n)$.)

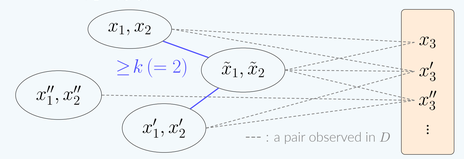
Proof Sketch of Upper Bound (Lower bound proof is similar)

$$(x_1, x_2) \xrightarrow{f_1} b \in (\text{intermediate states}); \quad (b, x_3) \xrightarrow{f_2} y \in (\text{outputs}). \qquad \text{(2-Hop task)}$$

Step 1. Reduce to graph connectivity

For each intermediate state b, define the b-evidence graph $G_b = (V_b, E_b)$:

$$V_b = f_1^{-1}(\{b\}) \subset \mathcal{X}_1 \times \mathcal{X}_2 \qquad W = \mathcal{X}_3$$



Key Lemma: If G_b is connected ($\forall b$), then $\text{Cover}_k(D)$ covers all in-domain data. (proof omitted.)

⇒ We want $\Pr[\exists b\text{: disconnected } G_b]$ small. (randomness: from i.i.d. sampling of D.)

Step 2. Apply Poissonization [Kac49, Ald89, Joh98] for graph independence

Under i.i.d. sampling, G_b's are dependent across b. $\because (\#\{(x_1, x_2, x_3) \in D\}) \sim \text{Multinomial}(\overset{\text{fixed}}{N_{\text{tr}}}; p = \frac{1}{n^3})$.

- We break this dependency by replacing $\#\{(x_1, x_2, x_3) \in D\} \overset{\text{iid}}{\sim} \text{Poisson}((N_{\text{tr}} - \varepsilon)/n^3)$.
- Using monotonicity of 'graph disconnectedness', we can prove:

$$\Pr_{\text{Multinomial}}[\exists b\text{: disconnected } G_b] \leq \sum_b \Pr_{\text{Poisson}}[G_b \text{ is disconnected}] + e^{-\varepsilon^2/(2N_{\text{tr}})}.$$

Step 3. Identify random k-intersection graphs (k-RIG) [Sin95, ZYG14]

Under Poissonization, $\mathbf{1}\{(x_1, x_2, x_3) \in D\} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ with $p = 1 - e^{-(N_{\text{tr}} - \varepsilon)/n^3} \approx N_{\text{tr}}/n^3$.

- Each G_b becomes a **binomial k-RIG** $\mathcal{G}^{(k)}(n_b, m, p)$, with $n_b = |V_b| \approx n$ and $m = |\mathcal{X}_3| = n$.
- Apply the connectivity threshold [Sin95, ZYG14]: for $m = \Omega(n_b)$ and $k \geq 2$,

$$p \geq \left(\frac{k!\,(2 \ln n_b)}{n_b}\right)^{\frac{0.5}{k}} \cdot \frac{1}{\sqrt{m}} \implies \Pr_{\text{Poisson}}\left[\mathcal{G}^{(k)}(n_b, m, p) \text{ is disconnected}\right] \to 0 \quad \text{as } n_b \to \infty.$$

We conclude the proof by plugging $p \approx N_{\text{tr}}/n^3$ and $n_b \approx n$, and solving for N_{tr}.

Discussion & Practical Implications

- **Long-tail:** Low-frequency combinations receive low k ⇒ effectively outside coverage
- **Reversal curse:** "A is B" gives no functional equivalence evidence for "B is A"
- **Planning:** Multi-path state tracking encounters path ambiguity

⇒ Future work: Data augmentation to "maximize k-coverage" by diversifying shared contexts

References

[Ald89] David Aldous. *Probability Approximations via the Poisson Clumping Heuristic*. New York, NY : Springer New York : Imprint: Springer, 1989.

[Joh98] Kurt Johansson. The longest increasing subsequence in a random permutation and a unitary random matrix model. *Mathematical Research Letters*, 5(1):68–82, 1998.

[Kac49] Marc Kac. On deviations between theoretical and empirical distributions. *Proceedings of the National Academy of Sciences*, 35(5):252–257, 1949.

[Sin95] Karen B. Singer. *Random intersection graphs*. PhD thesis, The Johns Hopkins University, 1995.

[ZYG14] Jun Zhao, Osman Yağan, and Virgil Gligor. On k-connectivity and minimum vertex degree in random s-intersection graphs. In *2015 Proceedings of the Twelfth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 1–15. SIAM, 2014.