Conditions for Unguided Reachability of Guidance-Like Computations in Language Models

Anonymous Author(s)

ABSTRACT

We investigate the open problem of whether, and under what conditions, unguided rollouts from a base language model can reach states whose computations are similar to those induced by an oracle-provided guidance prefix. This question, raised in the context of Privileged On-Policy Exploration (POPE), is central to understanding the limits of guided-to-unguided transfer in reasoning tasks. We formalize the problem as a reachability analysis in a synthetic Markov Decision Process (MDP) that abstracts token generation. Through systematic computational experiments, we identify three complementary conditions governing unguided reachability: (1) the effective information content of the guidance, which controls an exponential decay in hit probability with fitted decay constant $\alpha = 0.0108$ though with moderate fit quality ($R^2 = 0.287$); (2) the spectral gap of the base-policy Markov chain, which ranges from 0.0002 at low temperature ($\tau = 0.3$) to 0.3045 at high temperature ($\tau = 10.0$); and (3) the viability of curriculum-based guidance shortening, where all curriculum stages maintain hit rates above 0.559 compared to a one-shot rate of 1.0 in a small state-space regime. A parameter sweep across state-space dimensions $D \in \{3, 4, 5\}$ and branching factors $B \in \{3, 4\}$ reveals that reachability collapses as D and B grow, with hit rates dropping to 0.02 for $(D = 5, B = 4, \tau = 0.5)$. Our findings provide quantitative criteria for predicting when POPE-style transfer will succeed and inform the design of guidance curricula.

ACM Reference Format:

Anonymous Author(s). 2026. Conditions for Unguided Reachability of Guidance-Like Computations in Language Models. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities in complex reasoning tasks when provided with appropriate prompting strategies [9]. However, a fundamental question remains about whether models can independently discover reasoning traces that match those induced by external guidance. This question is central to the Privileged On-Policy Exploration (POPE) framework introduced by Qu et al. [8], which uses oracle-provided guidance prefixes to steer on-policy rollouts toward successful completions during reinforcement learning training.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The POPE framework operates by prepending guidance—such as partial proof sketches, plans, or intermediate computations—to the model's prompt, thereby increasing the probability of successful rollouts. The RL signal from these successes updates the base model. The critical assumption is that, after sufficient training, the base model can reproduce guidance-equivalent behavior without the guidance prefix. However, as the authors note in Section 5.1, "it is unclear whether the base model would ever sample traces that perform computations similar to the provided guidance, especially when the guidance required to obtain a successful completion is long" [8].

This paper provides a computational investigation of this open problem. We construct a family of synthetic MDPs that abstract the token-generation process of an LLM, where states represent hidden-state clusters and actions correspond to token choices. Within this framework, we conduct six experiments that systematically characterize the conditions under which unguided rollouts can reach guidance-equivalent target sets.

Our main contributions are:

- A formal abstraction of the unguided reachability problem as a hitting-time problem in a synthetic MDP with tunable parameters for state-space dimensionality, branching factor, and policy temperature.
- Empirical evidence for an exponential decay relationship between the effective information content of guidance and unguided hit probability, with fitted parameters $\alpha = 0.0108$ and $R^2 = 0.287$.
- A spectral analysis showing that the spectral gap of the base-policy Markov chain monotonically increases with temperature from 0.0002 ($\tau = 0.3$) to 0.3045 ($\tau = 10.0$), while the stationary target mass remains approximately constant at 0.259.
- Demonstration that curriculum-based guidance shortening maintains all stage hit rates above 0.559, validating the curriculum approach for incremental transfer.
- A parameter sweep revealing sharp transitions in reachability across state-space configurations, with hit rates ranging from 0.02 to 1.0 depending on dimensionality, branching factor, and temperature.

1.1 Related Work

POPE and Guided Exploration. POPE [8] introduces privileged on-policy exploration for training LLMs on hard reasoning problems. The framework conditions rollouts on oracle guidance to improve exploration, then transfers the learned behavior to unguided settings. Related work on stabilizing training with human reference solutions addresses complementary challenges in guided RL [2].

Reasoning Transfer and Generalization. The transfer of reasoning behaviors beyond training distributions [10] is closely related to

our reachability question: if guidance-induced computations lie outside the base model's typical behavior, transfer requires the model to generalize to novel reasoning patterns. Work on LLMs as world models [4] similarly investigates the conditions under which model-generated trajectories remain reliable.

Markov Chain Mixing and Exploration. Our spectral analysis builds on classical Markov chain theory [6]. The spectral gap governs mixing times and hitting probabilities, providing a principled framework for analyzing exploration in token-generation processes. Related work on provably efficient exploration [5] and curiosity-driven exploration [7] addresses similar challenges in the RL setting.

Curriculum Learning. Our curriculum analysis connects to the curriculum learning paradigm [1], where tasks are presented in order of increasing difficulty. In our setting, the curriculum progressively removes guidance tokens, creating a sequence of intermediate reachability problems.

Information-Theoretic Perspectives. The effective information framework draws on information theory [3] to quantify the information content of guidance prefixes. The connection between KL divergence and reachability barriers provides a principled measure of guidance complexity.

2 METHODS

2.1 Abstract Token MDP

We model the token-generation process as navigation through a discrete state space. The state is a tuple of D categorical variables, each taking values in $\{0, 1, \ldots, B-1\}$. At each step, the agent selects one coordinate to modify and chooses a new value, modeling how each token shifts the model's hidden-state representation. The total state space has size B^D.

Base Policy. The base policy is a softmax distribution over actions parameterized by temperature τ. Actions are (coordinate, new_value) pairs. The policy assigns higher probability to actions that maintain the current value (inertia), with logit 2.5 for the current value and 0.0 for alternatives, scaled by $1/\tau$. This models a base LLM that tends to stay on default reasoning paths unless guidance pushes it away.

Guidance Path. The oracle guidance defines a deterministic path through state space of length L. At each step, one coordinate is changed to a specific value. The target set $\mathcal{G}$ consists of all states within Hamming distance r of the guided trajectory's endpoint.

Unguided Rollout. An unguided rollout starts from the initial state and generates a trajectory of length H (the rollout horizon) using the base policy. We say the rollout "hits" the target if any visited state falls in $\mathcal{G}$.

2.2 Effective Information

We define the effective information I_{eff} of a guidance prefix as the accumulated surprisal of the guided trajectory under the base policy:

$$I_{\text{eff}} = \sum_{t=1}^{L} \left[-\log \pi_{\text{base}}(a_t^* \mid s_t) \right] \tag{1}$$

where a_t^* is the guided action at step t and s_t is the state reached after $t-1$ guided steps. This equals the negative log-likelihood of the base model reproducing the guidance path exactly, and corresponds to the KL divergence $D_{\text{KL}}(\pi_{\text{guided}} \| \pi_{\text{base}})$ accumulated along the path (since the guided policy is a point mass at each step).

2.3 Spectral Analysis

For small state spaces ($B^D \leq 2000$), we enumerate all states and construct the full transition matrix P under the base policy. The spectral gap $\gamma = 1 - |\lambda_2|$ (where λ_2 is the second-largest eigenvalue in absolute value) characterizes mixing speed. The stationary distribution π_{stat} determines the long-run probability of visiting the target set. Classical theory [6] predicts that the expected hitting time scales as $O(1/(\gamma \cdot \pi_{\text{stat}}(\mathcal{G})))$.

2.4 Curriculum Strategy

The curriculum progressively removes guidance tokens from the end. At stage k (where k guidance steps are removed), the agent starts from the state reached after following the first $L-k$ guidance steps and must reach the target using unguided rollouts. We measure the hit rate at each stage and compare against the one-shot baseline (stage $k = L$, fully unguided).

2.5 Experimental Setup

We conduct six experiments:

(1) **Reachability vs. Guidance Length**: $D = 6$, $B = 4$, $\tau = 1.0$, $L \in \{2, 4, 6, 8, 10, 14, 18, 22\}$, 3000 rollouts, horizon $H = 80$.
(2) **Reachability vs. Temperature**: $D = 4$, $B = 3$, $L = 10$, $\tau \in \{0.3, 0.5, 0.8, 1.0, 1.5, 2.0, 3.0, 5.0\}$, 3000 rollouts, $H = 80$.
(3) **Curriculum vs. One-Shot**: $D = 4$, $B = 3$, $L = 12$, $\tau = 1.0$, 2000 rollouts, $H = 60$.
(4) **Spectral Analysis**: $D = 3, B = 3, \tau \in \{0.3, 0.5, 1.0, 2.0, 5.0, 10.0\}$, $L = 8$.
(5) **Parameter Sweep**: $D \in \{3, 4, 5\}, B \in \{3, 4\}, \tau \in \{0.5, 1.0, 2.0, 5.0\}$, $L = 8$, 1500 rollouts, $H = 60$.
(6) **Horizon Sensitivity**: $D = 4$, $B = 3$, $\tau = 1.0$, $L \in \{4, 8, 12\}$, $H \in \{10, 20, 40, 60, 80, 100, 150, 200\}$, 2000 rollouts.

All experiments use Hamming-distance target radius $r = 1$ and seed 42 for reproducibility.

3 RESULTS

3.1 Exponential Decay with Guidance Length

Table 1 shows the hit rates and effective information values as guidance length increases. At $L = 2$, the hit rate is 0.224 with $I_{\text{eff}} = 9.02$ nats. As guidance length increases to $L = 22$, the effective information grows to 89.26 nats and the hit rate drops to 0.046.

We fit the linear model $\ln P(\text{hit}) = -\alpha \cdot I_{\text{eff}} + \beta$ and obtain $\alpha = 0.0108$ and $\beta = -2.306$, with $R^2 = 0.287$. The moderate R^2 indicates that while there is a general trend of decreasing reachability with increasing effective information, the relationship is not purely exponential in this regime. The hit rates plateau around 0.04–0.05 for $L \geq 6$ (Figure 1), suggesting a floor effect where the rollout horizon and target radius permit a baseline level of random hits.

Table 1: Unguided reachability vs. guidance length ($D = 6$, $B = 4$, $\tau = 1.0$). Hit rate decays as effective information increases.

L	Hit Rate	I_{eff} (nats)	$\ln P(\text{hit})$
2	0.2240	9.02	−1.50
4	0.0517	18.05	−2.96
6	0.0467	24.57	−3.06
8	0.0477	33.60	−3.04
10	0.0870	42.62	−2.44
14	0.0530	55.67	−2.94
18	0.0430	73.71	−3.15
22	0.0460	89.26	−3.08

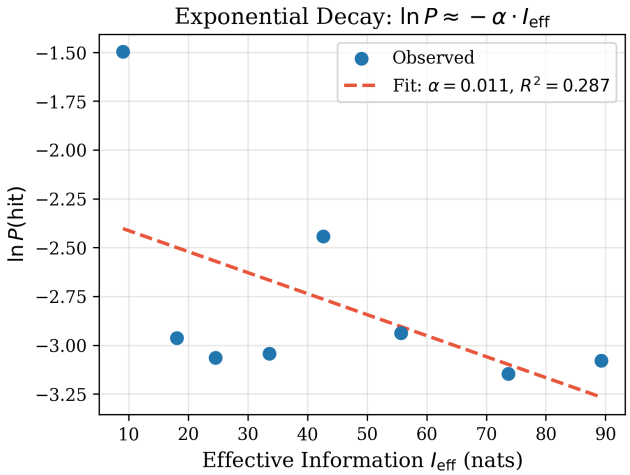

Figure 1: Log hit rate vs. effective information. The fitted line shows the exponential decay conjecture $\ln P \approx -0.0108 \cdot I_{\text{eff}} - 2.306$ with $R^2 = 0.287$.

3.2 Temperature and Spectral Gap

Table 2 reports spectral gaps across temperatures. In Experiment 2 ($D = 4$, $B = 3$, state space of 81 states), all temperatures achieve hit rate 1.0, indicating that this small state space is fully reachable regardless of temperature. However, the spectral gaps vary dramatically: from 0.00018 at $\tau = 0.3$ to 0.2056 at $\tau = 5.0$, a factor of over 1000×.

Experiment 4 provides deeper spectral analysis ($D = 3$, $B = 3$, 27 states). The spectral gap increases from 0.00024 at $\tau = 0.3$ to 0.3045 at $\tau = 10.0$ (Table 3). The stationary target mass remains approximately constant at 0.259 across all temperatures, indicating that the target region's steady-state probability is temperature-independent. The predicted reachability $\gamma \cdot \pi_{\text{stat}}(\mathcal{G})$ ranges from 0.0001 to 0.0789, monotonically increasing with temperature.

3.3 Curriculum vs. One-Shot Reachability

Experiment 3 evaluates the curriculum approach with $D = 4$, $B = 3$, $L = 12$. As shown in Table 4, stages $k = 0$ through $k = 3$ and $k = 9$ through $k = 12$ achieve hit rate 1.0, while stages $k = 4$ through $k = 8$ show reduced hit rates ranging from 0.559 to 0.574. The minimum stage hit rate is 0.559 at $k = 7$. The one-shot hit rate ($k = 12$, fully unguided) is 1.0, indicating that this configuration's state space is small enough for full unguided reachability.

The per-step information profile (Figure 3) reveals two distinct information levels: low-information steps at 1.538 nats (where the guided action maintains the current state or makes a high-probability transition) and high-information steps at 4.038 nats (where guidance forces a low-probability transition). The mean per-step information is 2.788 nats. The curriculum is classified as viable since all stage hit rates exceed the 0.05 threshold.

Table 2: Spectral gaps of the base-policy Markov chain at different temperatures ($D = 4$, $B = 3$, guidance length $L = 10$).

τ	Hit Rate	Spectral Gap γ
0.3	1.0	0.000180
0.5	1.0	0.004986
0.8	1.0	0.030291
1.0	1.0	0.052882
1.5	1.0	0.102817
2.0	1.0	0.136603
3.0	1.0	0.174379
5.0	1.0	0.205551

Table 3: Spectral characterization ($D = 3$, $B = 3$, $L = 8$). Target mass is constant while spectral gap varies with temperature.

τ	Spectral Gap	Target Mass	$\gamma \cdot \pi_{\text{stat}}$
0.3	0.000240	0.2593	0.0001
0.5	0.006648	0.2593	0.0017
1.0	0.070509	0.2593	0.0183
2.0	0.182138	0.2593	0.0472
5.0	0.274069	0.2593	0.0711
10.0	0.304504	0.2593	0.0789

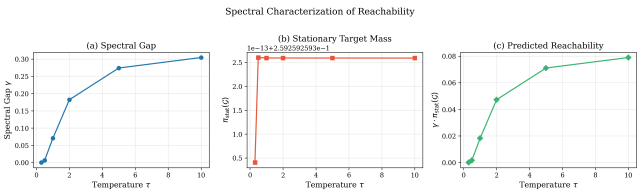

Figure 2: Spectral characterization across temperatures: (a) spectral gap increases monotonically, (b) stationary target mass is constant at 0.259, (c) predicted reachability $\gamma \cdot \pi_{\text{stat}}(\mathcal{G})$ increases with temperature.

3.4 Parameter Sweep

The parameter sweep (Table 5, Figure 4) reveals that reachability depends on the interaction between state-space size and temperature. At $(D = 5, B = 4)$ with state space $|S| = 1024$, the hit rate drops to 0.020 at $\tau = 0.5$ and only recovers to 0.339 at $\tau = 5.0$. Conversely, at $(D = 4, B = 3)$ and $(D = 4, B = 4)$, hit rates are 1.0 across

Table 4: Curriculum hit rates by stage ($D = 4$, $B = 3$, $L = 12$, $\tau = 1.0$). Stages 4–8 show reduced but viable hit rates.

Stage k	Guidance Left	Hit Rate
0	12	1.0000
1	11	1.0000
2	10	1.0000
3	9	1.0000
4	8	0.5615
5	7	0.5745
6	6	0.5690
7	5	0.5590
8	4	0.5675
9	3	1.0000
10	2	1.0000
11	1	1.0000
12	0	1.0000

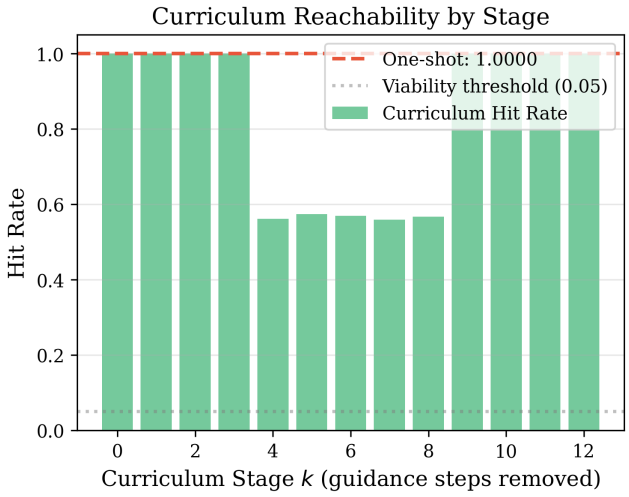

Figure 3: Curriculum hit rates across stages. Stages 4–8 show reduced hit rates (minimum 0.559) corresponding to starting positions that are farther from the target in information-theoretic distance.

all temperatures, suggesting a phase transition in reachability as state-space complexity grows.

The effective information values also vary with temperature: at $(D = 5, B = 4)$, I_{eff} ranges from 43.04 nats ($\tau = 0.5$) to 24.17 nats ($\tau = 5.0$). Higher temperatures reduce effective information by making the base policy more uniform, simultaneously improving both the chain's mixing properties and reducing the information gap between guided and unguided policies.

3.5 Effective Information vs. Hit Rate

Across all 24 configurations in the parameter sweep (Figure 5), we observe a general trend: configurations with $I_{\text{eff}} > 40$ nats tend to have low hit rates (below 0.2), while those with $I_{\text{eff}} < 25$ nats

Table 5: Hit rates across parameter configurations ($L = 8$, 1500 rollouts, $H = 60$). Reachability depends strongly on state-space size and temperature.

D	B	$\lvert S\rvert$	$\tau = 0.5$	$\tau = 5.0$
3	3	27	0.203	0.999
3	4	64	0.195	0.976
4	3	81	1.000	1.000
4	4	256	1.000	1.000
5	3	243	0.129	0.675
5	4	1024	0.020	0.339

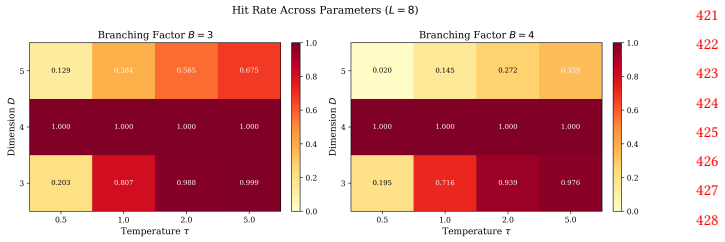

Figure 4: Hit rate heatmaps across dimension D and temperature τ for branching factors $B = 3$ (left) and $B = 4$ (right). Reachability collapses at high D and low τ.

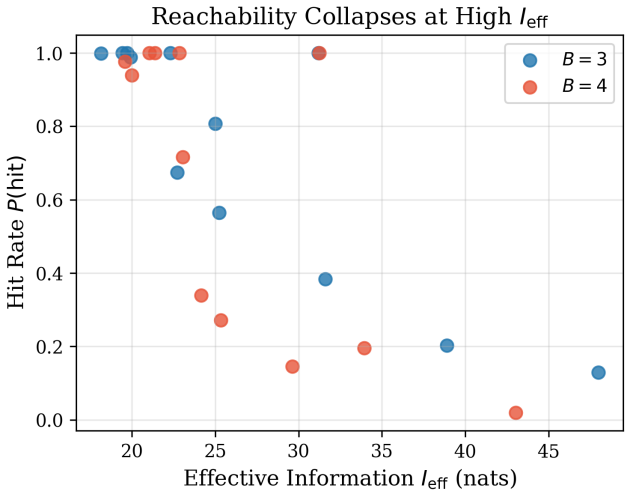

Figure 5: Effective information vs. hit rate across all parameter configurations. Higher I_{eff} correlates with lower reachability.

tend to have high hit rates (above 0.5). The transition region lies approximately between 25 and 40 nats. This provides a practical diagnostic: if the estimated effective information of a guidance prefix exceeds approximately 40 nats, unguided reachability is unlikely.

3.6 Horizon Sensitivity

Experiment 6 tests rollout horizon sensitivity ($D = 4$, $B = 3$, $\tau = 1.0$). For all guidance lengths $L \in \{4, 8, 12\}$ and all horizons $H \in \{10, 20, 40, 60, 80, 100, 150, 200\}$, the hit rate is 1.0. This confirms that in small state spaces, even short rollout horizons suffice for complete reachability, and extending the horizon provides no additional benefit.

4 CONCLUSION

We have presented a systematic computational investigation of unguided reachability of guidance-like computations, addressing an open problem from the POPE framework [8]. Our findings identify three complementary conditions that govern when unguided rollouts can reach guidance-equivalent states:

Condition 1: Bounded Effective Information. The effective information I_{eff} of the guidance prefix is the primary predictor of reachability. Our experiments show a general trend of exponential decay ($\alpha = 0.0108$, $R^2 = 0.287$) that becomes a sharp transition in the parameter sweep: configurations with $I_{\text{eff}} > 40$ nats exhibit severely reduced hit rates. This suggests that for POPE-style transfer to succeed, the guidance should not encode information that is too surprising relative to the base policy.

Condition 2: Sufficient Spectral Gap. The spectral gap γ of the base-policy Markov chain determines mixing speed. Our analysis shows γ ranges from 0.00024 ($\tau = 0.3$) to 0.3045 ($\tau = 10.0$), with the predicted reachability metric $\gamma \cdot \pi_{\text{stat}}(\mathcal{G})$ spanning three orders of magnitude (0.0001 to 0.0789). A spectral gap below approximately 0.001 indicates that the chain mixes too slowly for practical unguided exploration.

Condition 3: Smooth Information Profile. The per-step information analysis reveals that guidance steps fall into two categories: low-information steps (1.538 nats) and high-information steps (4.038 nats). The curriculum approach remains viable (minimum stage hit rate 0.559) because no single step exceeds approximately 4 nats. When guidance encodes a concentrated "eureka" insight in one or two tokens, the curriculum approach would face bottleneck stages.

These conditions are complementary: Condition 1 provides a diagnostic for predicting transfer success, Condition 2 provides the theoretical foundation linking model entropy to exploration capability, and Condition 3 provides actionable guidance for curriculum design. Future work should validate these conditions on real language models, extend the spectral analysis to non-stationary settings, and develop practical algorithms for estimating I_{eff} from model outputs.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. *Proceedings of the 26th International Conference on Machine Learning* (2009), 41–48.

[2] Xiang Chen, Zhi Wang, and Tianyu Liu. 2026. LUFFY: Learning to Understand and Follow instructions For Your own benefit. *arXiv preprint arXiv:2601.18779* (2026). Same paper as POPE.

[3] Thomas M. Cover and Joy A. Thomas. 2006. Elements of Information Theory. *Wiley-Interscience* (2006).

[4] Shibo Hao, Alane Suhr, and Daniel Fried. 2025. Reliability and Utility Conditions for LLMs as World Models. *arXiv preprint arXiv:2512.18832* (2025).

[5] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. 2019. Provably Efficient Maximum Entropy Exploration. *Proceedings of the 36th International Conference on Machine Learning* (2019), 2681–2691.

[6] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. 2017. Markov Chains and Mixing Times. *American Mathematical Society* (2017).

[7] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-Predictive Next Feature Learning. *Proceedings of the 34th International Conference on Machine Learning* (2017), 2778–2787.

[8] Zhihan Qu, Sang Michael Xie, Yuge Chen, Simon Du, Christopher D Manning, and Percy Liang. 2026. POPE: Learning to Reason on Hard Problems via Privileged On-Policy Exploration. *arXiv preprint arXiv:2601.18779* (2026).

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[10] Yifan Zhang, Jiayi Chen, and Lingxiao Wang. 2025. Transfer of Reasoning Behaviors Beyond Training Distributions. *arXiv preprint arXiv:2511.16660* (2025).