WHY IMPLICIT COT FAILS AT COMPLEX LOGIC: THEORY, BENCHMARK, AND SOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought (CoT) has unlocked advanced reasoning abilities of Large Language Models (LLMs) with intermediate steps, yet incurs prohibitive computational costs due to generation of extra tokens. Recent studies empirically show that compressing reasoning steps into latent states, or implicit CoT compression, offers a token-efficient alternative. However, the mechanism behind CoT compression remains unclear. In this paper, we provide the first theoretical analysis of the difficulty of learning to internalize intermediate reasoning steps. We introduce *Order-r Interaction* and show that, under uniform attention initialization, the gradient signal identifying relevant inputs decays exponentially with interaction order. Separating this signal from finite-sample fluctuations is ensured by a sample size that grows exponentially with the order. Motivated by the semantic shortcuts that can obscure the evaluation of implicit CoT, we introduce NatBool-DAG, a benchmark of Boolean reasoning expressed in natural language and designed to reduce reliance on such shortcuts. Guided by our theory, we propose ALiCoT (**Al**igned **I**mplicit **CoT**), which aligns latent and explicit reasoning states to enable lower-order computation through latent tokens. Experimental results demonstrate that ALiCoT maintains strong reasoning performance under extreme token compression.

1 INTRODUCTION

In recent years, Large Language Models (LLMs) have achieved a paradigm shift in reasoning capabilities, evolving from simple pattern matching to addressing complex tasks through Chain-of-Thought (CoT) reasoning Wei et al. (2022); Zhao et al. (2023); Ma et al. (2025), generating sequential intermediate steps leading to the final answer. While thinking models such as GPT-4-o1 OpenAI (2025) and DeepSeek-R1 Guo et al. (2025) have achieved remarkable success in mathematical and commonsense reasoning, it comes at a steep price: the heavily expanded reasoning chains incur massive token consumption and high latency, creating a significant bottleneck for real-world deployment and efficiency Zhu et al. (2025); Warner et al. (2025). To bridge this gap, recent works have pivoted towards CoT Compression. One direction that has been empirically successful is explicitly reducing the length of intermediate steps by training model to learn shorter CoT Kang et al. (2025). Another distinct line of research has aimed to internalize intermediate reasoning steps to latent states to perform implicit CoT Zhang et al. (2025b); Hao et al. (2024); Shen et al. (2025). Meanwhile, several research studies posited that implicit CoT maintain competitive performance on tasks where dependencies are often shallow or linguistically redundant (e.g., semantic shortcuts), yet suffer a precipitous performance collapse on solving tasks requiring rigorous logic such as multi-step mathematical reasoning or symbolic derivation Zhu et al. (2025); Chen et al. (2025).

Despite these empirical successes, however, the theoretical understanding of the implicit CoT compression mechanism remains limited. Prior investigations into CoT mechanisms Kim & Suzuki (2025b); Abbe et al. (2024) often overlook the critical challenges arising from the omission of local reasoning steps. Consequently, it is unclear how the difficulty of internalization scales with the number of steps being internalized across different tasks. We introduce *Order-r Interaction* to connect the dependencies involved in CoT compression with initial gradient strength and sufficient sample sizes for signal separation. For example, as illustrated in Figure 1, the model can successfully shortcut from "Louvre" to "Europe" via simple associations in latent tokens but fails to skip intermediate steps in arithmetic tasks, where the complex dependencies among variables lead to high-order interaction barriers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

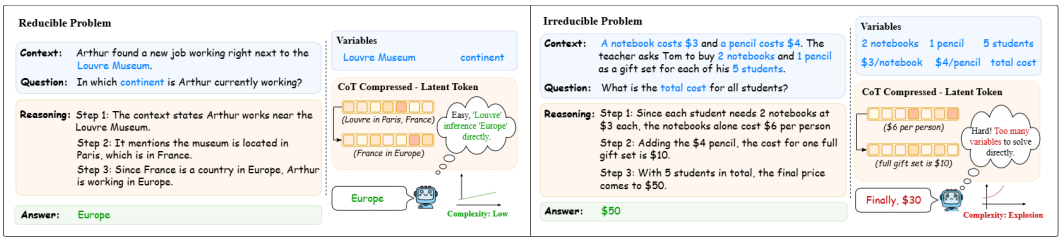


Figure 1: CoT compression impacts reasoning difficulty based on problem reducibility. For Reducible Problems (top), compression is effective as the model faces low difficulty in reconstructing reasoning relationships. Conversely, for Irreducible Problems (bottom), omitting intermediate steps removes essential computational states, triggering a complexity explosion that leads to failure.

By analyzing the parity problem, a representative and analytically tractable setting, we identify an optimization bottleneck in CoT compression: when omitting intermediate steps forces the model to combine more inputs directly, the initial learning signal becomes weaker, and our sufficient sample-size bound for separating it from sampling noise increases. We further establish an average signal upper bound for multi-layer, multi-head Transformers under stated conditions, and identify conditions for analogous behavior in natural-language reasoning. Controlled experiments further support the connection between this bottleneck and increased training cost under stronger compression.

Motivated by this theoretical insight, we present a comprehensive study spanning theory, evaluation, and methodology to overcome these barriers. Our contributions are as follows:

• **Learning Bottlenecks in CoT Compression:** We characterize initial gradient attenuation and sufficient sample sizes for signal separation, establish conditional multi-layer signal bounds, and empirically examine the resulting learning bottleneck.

• **NatBool-DAG Benchmark:** To rigorously evaluate CoT compression beyond the noise of natural language redundancies, we introduce NatBool-DAG, a scalable synthetic benchmark based on Directed Acyclic Graphs of Boolean logic embedded in natural language. Unlike standard datasets that may allow semantic shortcuts, NatBool-DAG enforces strict, irreducible logical dependencies, serving as a litmus test for genuine implicit reasoning capabilities.

• **ALiCoT: Aligned Implicit Chain-of-Thought:** Guided by our theory, we propose ALiCoT to enable latent tokens to serve as intermediate variables that reduce the effective interaction order of subsequent reasoning. By aligning latent representations with explicit reasoning states, ALiCoT encourages subsequent computation to build on these intermediate variables, reducing the need to recover high-order dependencies directly from the original inputs. Experiments demonstrate improved learning efficiency and reasoning performance under strong compression.

2 Related Work

Higher-order interactions and reasoning complexity. Higher-order interactions have long been studied in statistics and information theory, including interaction information, co-information, attribute interactions, and lattice-based formulations McGill (1954); Bell (2003); Whittaker et al. (2015); Jakulin & Bratko (2003); Jakulin (2005); Liu et al. (2023; 2025). Related redundancy–synergy decompositions and feature interaction attribution methods Williams & Beer (2010); Rosas et al. (2019); Scagliarini et al. (2022); Sundararajan et al. (2020) measure dependence structure as well, yet none of these works consider the cost that high-order dependence imposes on learning. Recent theory has taken up the complementary question of learnability, showing that CoT and scratchpads improve learnability on parity-like and compositional tasks by exposing intermediate structure Kim & Suzuki (2024); Wen et al.; Abbe et al. (2024); Li et al.; Merrill & Sabharwal (2024). However, these results largely characterize an *endpoint* separation between full CoT and no CoT, leaving open what happens in the intermediate regime where some reasoning steps are preserved and others are compressed to trade off accuracy for speed. Our work addresses both gaps by quantifying the impact of progressively compressing chain-of-thought on learning complex reasoning problems.

CoT compression and latent reasoning. Reasoning-efficient LLMs generally pursue explicit CoT compression or implicit latent reasoning. Explicit methods shorten or prune natural-language rationales through distillation, token selection, or keypoint-based compression Kang et al. (2025); Luo et al. (2025); Cheng & Van Durme (2024); Li et al. (2025); Xia et al. (2025); Feng et al. (2024). Latent methods instead replace intermediate text with continuous representations, ranging from hidden-state distillation Deng et al. (2023; 2024) to autoregressive latent scratchpads Hao et al. (2024); Xu et al. (2025). Within this line, we distinguish methods by how the latent representation is instantiated. Imp.Base-1 uses a globally shared learnable placeholder, closely related to learnable pause tokens Goyal et al. (2024). Imp.Base-2 generates an instance-conditioned latent state from contextual hidden states or an auxiliary encoder, as in Coconut Hao et al. (2024), CODI Shen et al. (2025), LightThinker Zhang et al. (2025a), and SoftCoT Xu et al. (2025). Recent works have also explored auxiliary supervision for latent reasoning, including CODI and SIM-CoT Shen et al. (2025); Wei et al. (2025). Our Imp.Base-1 and Imp.Base-2 instantiate these two core cases, while ALiCoT introduces a theoretically motivated alignment objective.

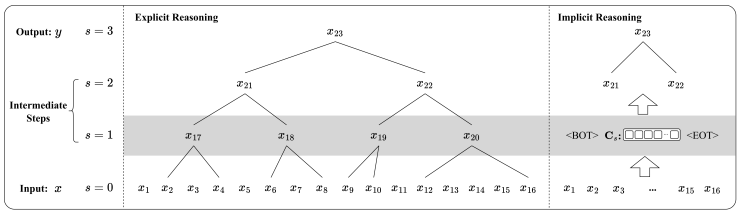


Figure 2: Schematic of the k-Parity Problem. $\boldsymbol{x}_{17}, \cdots, \boldsymbol{x}_{22}$ represent intermediate reasoning steps, which are replaced by latent tokens in the case of implicit Chain-of-Thought.

3 Preliminaries and Setup

This section first defines the mathematical notations used in the paper, then systematically introduces the foundational problem (parity problem), core object (Implicit CoT), and model architecture (Transformer) required for subsequent analysis, laying the groundwork for theoretical and experimental discussions.

Notation. We denote the set $\{1, 2, \ldots, n\}$ as $[n]$. Vectors and matrices are denoted in bold text (e.g., $\boldsymbol{x}, \boldsymbol{A}$), whereas scalars appear in plain text (e.g., y). For $\boldsymbol{z} \in \mathbb{R}^n$ we write $\phi(\boldsymbol{z}) = (\phi(z_1), \cdots, \phi(z_n))^\top$, $\boldsymbol{z}^2 = \boldsymbol{z} \odot \boldsymbol{z} = (z_1^2, \cdots, z_n^2)$ and $|\boldsymbol{z}| = (|z_1|, \cdots, |z_n|)^\top$. The 2-norm is always denoted by $\|\cdot\|$. The multi-linear inner product or contraction of $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_r \in \mathbb{R}^n$ for any $r \in \mathbb{N}$ is denoted as $\langle \boldsymbol{z}_1, \cdots, \boldsymbol{z}_r \rangle := \sum_{i=1}^n z_{1,i} \cdots z_{r,i}$. In particular, $\langle \boldsymbol{z}_1 \rangle = \boldsymbol{z}_1^\top \mathbf{1}_n$ and $\langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle = \boldsymbol{z}_1^\top \boldsymbol{z}_2$.

3.1 The Parity Problem

Given an index set $\{1, \cdots, d\}$, for d-bit inputs $\boldsymbol{x} = (x_j)_{j=1}^d \sim \mathrm{Unif}(\{\pm 1\}^d)$, the k-parity problem involves predicting $y = \prod_{j \in p} x_j$, where $p \subseteq [d]$ and $|p| = k$. In this problem, the support subset p is unknown, and we denote the set of all possible p as P_k, so $|P_k| = \binom{d}{k}$. Our problem setup follows the definition in Kim & Suzuki (2025a) for theoretical analysis. We abuse notation and identify the set of indices p with the corresponding parity mapping $\boldsymbol{x} \mapsto \prod_{j \in p} x_j$. Given n samples $(\boldsymbol{x}^i, y^i)_{i \in [n]}$, our goal is to accurately predict the parity of any unseen test input.

3.2 Chain-of-thought Compression Objectives

In the parity problem, explicit Chain-of-Thought (CoT) decomposes the k-parity task into a binary tree of 2-parity operations of height $H = \lceil \log_2 k \rceil$ (Figure 2). Each intermediate node x_t ($t > d$) is computed by multiplying its two child nodes, $x_t = x_u \cdot x_v$, where $u, v < t$ are indices from the previous level. The model is trained to predict every intermediate step under cross-entropy supervision:

$$\mathcal{L}^{\mathrm{CoT}}(\theta) = -\frac{1}{n} \sum_{j=d+1}^{d+k-1} \sum_{c \in \{1,-1\}} \langle \mathbb{I}(\boldsymbol{x}_j = c), \log(\hat{\boldsymbol{x}}_{j,c}) \rangle \,. \tag{1}$$

Implicit CoT compresses the reasoning chain by replacing one layer of explicit reasoning tokens in the reasoning tree (Figure 2) with a single latent token $\boldsymbol{c}_s$ (Zhang et al., 2025b; Hao et al., 2024). The training objective skips supervision for the concealed layer and predicts only the subsequent explicit steps:

$$\mathcal{L}^{\text{iCoT}}(\theta) = -\frac{1}{n} \sum_{j=d+\tau_s+1}^{d+k-1} \sum_{c\in\{1,-1\}} \langle \mathbb{I}(\boldsymbol{x}_j = c), \log(\hat{\boldsymbol{x}}_{j,c}) \rangle \,, \quad \text{with } \hat{\boldsymbol{x}}_j = f_\theta(\boldsymbol{x}_{1:d}, \boldsymbol{c}_s, \boldsymbol{x}_{d+\tau_s+1:j-1}), \tag{2}$$

where τ_s denotes the total number of explicit tokens originally required up to step s, and f_θ predicts $\boldsymbol{x}_j$ from the input $\boldsymbol{x}_{1:d}$, the latent state $\boldsymbol{c}_s$, and the history of subsequent explicit steps.

The latent token $\boldsymbol{c}_s$ is instantiated through two paradigms. **Imp.Base-1** treats $\boldsymbol{c}_s$ as a standalone learnable parameter (soft prompt) initialized from $\mathcal{U}(-1,1)^{d_{\text{model}}}$ and optimized via backpropagation across all instances. **Imp.Base-2** generates $\boldsymbol{c}_s$ dynamically from the model's hidden state at the <bot> position, making the latent representation instance-specific. In this paper, Imp.Base-1 denotes our pause-token-style implementation Goyal et al. (2024), and Imp.Base-2 our Coconut-based implementation Hao et al. (2024).

4 Theoretical results

This section develops a theoretical account of the learning difficulty underlying implicit Chain-of-Thought compression. In Section 4.1, we formalize the notion of Order-r Interaction and derive a generalized gradient decomposition, characterizing how interaction order, context length, and finite-sample fluctuations affect early-stage learnability. In Section 4.2, we connect this theoretical notion to reasoning datasets by quantifying interaction density and interaction quality, and use these measures to distinguish semantic shortcuts from logically irreducible dependencies. Together, these analyses explain why implicit CoT can benefit from compression on shortcut-rich tasks but struggle on tasks requiring high-order compositional reasoning.

4.1 Order-r Interaction and Signal Decomposition

To quantify the structural difficulty of reasoning, we model a dependency from a minimal set of input variables or semantic units to a target output as an Order-r Interaction. The parity problem provides a clean and analytically tractable instance of this abstraction. In this view, r measures the number of input units in the minimal support required to determine the target.

Definition 1 (Order-r Interaction)**.** A target variable y is driven by an Order-r Interaction supported on a set of input vectors $\mathcal{S}$, if y depends on the joint values of the vectors in $\mathcal{S}$ ($|\mathcal{S}| = r$), while remaining independent of any proper subset $\mathcal{S}' \subsetneq \mathcal{S}$. For any vector $\boldsymbol{x} \in \mathcal{S}$, we say that $\boldsymbol{x}$ is in the Order-r Interaction of y.

This abstraction deliberately focuses on the dependency between inputs and outputs rather than on the particular reasoning program used to compute the output. For example, in the parity tree shown in Figure 2, x_{17} is an Order-2 Interaction supported on $\{\boldsymbol{x}_2, \boldsymbol{x}_4\}$, while x_{21} is an Order-4 Interaction relative to the atomic inputs $\{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_6, \boldsymbol{x}_8\}$. The same abstraction can be applied to semantic units in mathematical, commonsense, and knowledge reasoning. For instance, the result of "$3 + 5 = 8$" can be associated with the support $\{3, +, 5\}$ when no proper subset provides a reliable shortcut. The parity analysis in Appendix A provides a concrete instantiation, while Appendix E discusses the conditions under which the same analysis extends to natural-language reasoning.

The definition above specifies a dependency, while learning requires discovering its support. We focus on the initial gradient that distinguishes relevant from irrelevant inputs. This focus is motivated by Wies et al. (2023), who report that, in bit-subset parity experiments, validation accuracy rapidly exceeds 95% after rising above chance. In a related Transformer setting, Kim & Suzuki (2025b) analyze gradients at uniform initialization to establish one-update learning with teacher forcing under their stated assumptions. These results motivate studying the onset of support discovery through the strength and finite-sample detectability of its initial gradient signal.

For the causal Transformer defined in Appendix A.1, consider a target position m with $m-1$ visible preceding positions. Since attention weights sum to one, the average attention mass across these positions is $1/(m-1)$. We use uniform attention as a reference initialization in which every position

receives this average allocation, without a pre-existing attention preference for the true support. Let $w_{j,m}$ denote the attention logit connecting position j to the target. We use m as the context-length scale in the asymptotic notation, n as the number of training samples, and p as the failure probability. We characterize the initial gradient decomposition as follows.

Theorem 1 (Gradient Signal Decomposition at Initialization)**.** *Consider the model in Appendix A.1 under the initialization conditions of Lemma 3, including uniform attention $\sigma_j(\boldsymbol{w}_m) = 1/(m-1)$ for every $j < m$, with an analytic or locally Taylor-approximable smooth activation ϕ. Let $\gamma_r = \phi^{(r)}(0)/(r-1)!$ be the interaction coefficient for order r. Let $\mathbb{I}_j^{(r)}$ indicate whether input $\boldsymbol{x}_j$ is part of the Order-r Interaction set $\mathcal{S}^{(r)}$. Let $C_{\text{signal}}^{(r)}$ denote the gradient signal of the Order-r interaction term. With probability at least $1-p$, the finite-sample noise κ satisfies the bound $O\left(\sqrt{\frac{\log(m/p)}{n}}\right)$, and the gradient with respect to the attention weight $w_{j,m}$ follows the generalized asymptotic expansion:*

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = \sum_{r=1}^{\infty} \gamma_r \cdot \left[\underbrace{-\Theta(m^{-r}) \cdot \mathbb{I}_j^{(r)} \cdot C_{\text{signal}}^{(r)}}_{\textit{Order-r Signal}} + \underbrace{\Theta(m^{-(r+1)}) \cdot C_{\text{signal}}^{(r)}}_{\textit{Order-r Bias}} \right] + \underbrace{O\left(\frac{\kappa}{m}\right)}_{\textit{Noise}} \cdot \tag{3}$$

Theorem 1 identifies the gradient hierarchy underlying the optimization difficulty of learning different interaction orders. In particular, it yields three consequences:

1. order-dependent signal separation. Equation (3) decomposes the attention-logit gradient into the Order-r Signal, the Order-r Bias, and the Noise. Only the signal consistently identifies tokens in the Order-r support; the bias does not identify the support, and the noise reflects finite-sample fluctuations. For fixed $m > 1$ and non-degenerate interaction coefficients, the support-selective signal scales as $\Theta(m^{-r})$ and therefore decays exponentially with r. As it approaches the noise floor, gradient updates become less reliable at favoring relevant tokens over irrelevant ones.

2. Sufficient sample size for signal separation. Comparing the noise upper bound $O\Big(m^{-1}\sqrt{\log(m/p)/n}\Big)$ with the signal magnitude $\Theta(m^{-r})$ shows that $n = \Omega\Big(m^{2(r-1)} \log \frac{m}{p}\Big)$, with a sufficiently large implied constant, suffices for gradient signal separation under the theorem's assumptions. Under this condition, the support-selective signal dominates finite-sample fluctuations, facilitating the model's learning to attend to relevant tokens from an initially uniform attention distribution.

3. the context-length effect. For fixed r, increasing m attenuates the support-selective signal polynomially and reduces its signal-to-noise ratio by a factor of $m^{-(r-1)}$, up to logarithms. Shortening the context therefore improves gradient detectability, but this benefit can be outweighed if compression increases the effective interaction order.

In the parity tree of Figure 2, x_{21} is an Order-4 interaction with respect to the atomic inputs, but becomes the lower-order relation $x_{21} = f(x_{17}, x_{18})$ when intermediate states are available. Explicit CoT therefore decomposes a high-order dependency into local lower-order dependencies, producing stronger gradient signals and smaller sample requirements. Implicit CoT can retain this benefit only if its latent tokens preserve the omitted, support-aligned intermediate computation. Otherwise, the model must recover the high-order dependency directly from the raw context, and the useful signal remains attenuated as $\Theta(m^{-r})$. This identifies the central limitation of implicit CoT: shortening the context without providing an effective lower-order representation can increase the interaction order rather than simplify learning. Appendix C shows that **Imp.Base-1** lacks sample-dependent support information, whereas **Imp.Base-2** may mix useful context features with irrelevant ones and increase the noise floor.

The $\Theta(m^{-r})$ hierarchy extends under the conditions in Appendices A.4 and E. Appendix D further shows that, under the stated conditions, the m^{-r} signal attenuation in Theorem 1 extends to multi-layer, multi-head Transformers as an $O(m^{-r})$ upper bound averaged over support locations. Appendix B experimentally validates the single-layer scaling in r, m, and n.

4.2 Empirical Analysis: Semantic Shortcuts vs. Logical Irreducibility

Theorem 1 establishes that learning implies a trade-off: implicit reasoning reduces context length m but potentially increases the interaction order r. This raises a critical empirical question: why do

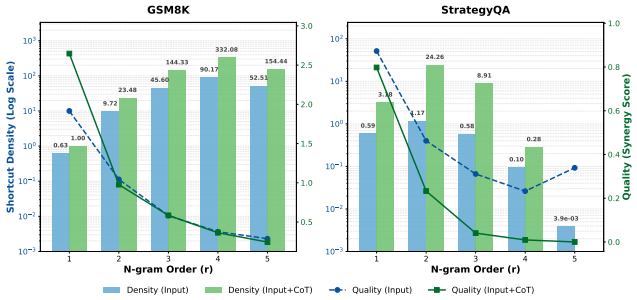


Figure 3: Empirical interaction-order profiles. The figure compares effective interaction density ρ_k and interaction quality ϕ_k across different interaction orders for commonsense and mathematical reasoning datasets, with and without CoT.

current methods significantly degrade performance in mathematical tasks, while often maintaining effectiveness in commonsense reasoning? To answer this, we must analyze the inherent interaction orders required by these respective tasks.

4.2.1 Quantifying Interaction Orders

Our theory predicts that reasoning tasks requiring higher-order dependencies to determine their outputs are harder to learn. We therefore characterize the empirical interaction-order profiles of different reasoning datasets. Specifically, we compare the prevalence of low- and high-order dependencies in commonsense and mathematical reasoning, both with and without explicit Chain-of-Thought (CoT).

For each example i, let $\mathbf{x}_i$ denote the input context and y_i denote its target output. We consider a candidate set $\mathcal{S}_k \subseteq \mathbf{x}_i$ containing k input tokens. Here, k serves as an empirical proxy for the theoretical interaction order r. To measure the additional information provided by the joint token set beyond its proper subsets, we define the PMI-based synergy score Bouma (2009):

$$\sigma(X_{\mathcal{S}_k}; y_i) = \text{PMI}(X_{\mathcal{S}_k}; y_i) - \max_{\mathcal{S}' \subsetneq \mathcal{S}_k} \text{PMI}(X_{\mathcal{S}'}; y_i). \tag{4}$$

The sign of σ indicates whether the joint token set provides information beyond its proper subsets. When $\sigma > 0$, the joint set has additional synergistic information and provides empirical evidence of an irreducible dependency. When $\sigma \leq 0$, we do not observe additional joint information beyond a smaller subset. We therefore use this condition as an approximate empirical indicator of a redundant or shortcut-based relation, since an exact independence test would require access to the full underlying data distribution, which is unavailable in practice.

We use a fixed threshold γ to exclude weak associations. The effective interaction density measures the average number of valid order-k interactions per sample:

$$\rho_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathcal{S}_k \subseteq \mathbf{x}_i} \mathbb{I}(\sigma(X_{\mathcal{S}_k}; y_i) > 0 \wedge \text{PMI}(X_{\mathcal{S}_k}; y_i) > \gamma)\,. \tag{5}$$

The interaction quality ϕ_k is computed by averaging $\sigma(X_{\mathcal{S}_k}; y_i)$ over the valid interactions counted by ρ_k. High density with low quality suggests many weak or redundant patterns, whereas high quality indicates a stronger structural dependence on the joint input.

4.2.2 Results and Analysis

We compute these statistics separately for raw and CoT-augmented inputs on representative commonsense and mathematical reasoning datasets, including StrategyQA Geva et al. (2021) and GSM8K Cobbe et al. (2021). Since the number of candidate token sets grows combinatorially with k, candidate sets are sampled at each order using the same sampling protocol across datasets. Figure 3 visualizes the resulting interaction-order profiles, revealing whether each dataset is dominated by low-order reducible dependencies or strong high-order irreducible dependencies.

Commonsense: The Prevalence of Semantic Shortcuts. For commonsense tasks, valid interactions are concentrated at lower orders, while higher-order interactions have lower quality. Adding Chain-of-Thought (CoT) further reduces the quality of higher-order interactions, suggesting that its new

Table 1: **Statistics of the NatBool-DAG Dataset.** Distinct sets are generated for training, validation, and testing across reasoning hops 3–10.

	3	4	5	6	7	8	9	10
Train	469	662	911	1,104	1,349	1,522	1,712	1,886
Val	187	298	204	189	235	253	256	227
Test	220	299	276	285	287	189	229	177

information is mainly encoded in lower-order dependencies. This is consistent with the reducible "Arthur $\rightarrow$ Europe" example in Figure 1, where removing intermediate steps need not increase the effective interaction order. Thus, when r remains unchanged, Implicit CoT can benefit from a shorter context: reducing m improves the signal-to-noise ratio for $r > 1$, despite the explicit $1/m$ factor in the noise bound.

Mathematics: The Barrier of Logical Irreducibility. Mathematical reasoning instead exhibits stronger high-order interactions, whose quality is not substantially reduced by CoT, indicating that its intermediate steps provide non-redundant information. Although CoT improves lower-order interaction quality, consistent with the logically irreducible structure in Figure 1, removing these steps can make the effective interaction order grow with computational depth. The resulting attenuation of the high-order signal can outweigh the benefit of a shorter context, consistent with the degradation of implicit reasoning.

Remark on Data Scarcity. Low-order correlations in math datasets may reflect overfitting to spurious shortcuts under limited data rather than genuine reasoning, motivating a rigorous testbed.

5 NatBool-DAG: A Benchmark for Irreducible Reasoning

Section 4.2 reveals a critical flaw: prevalent low-order interactions in standard reasoning benchmarks often permit semantic shortcuts. Consequently, the success of existing methods may reflect statistical pattern matching rather than genuine high-order compression. To rigorously evaluate Implicit CoT, we introduce Natural-Language Boolean Directed Acyclic Graphs (NatBool-DAG), a synthetic benchmark ensuring *Logical Irreducibility* and *Sample Sufficiency* to enforce robust reasoning.

5.1 Dataset Construction

NatBool-DAG embeds random Directed Acyclic Graphs of Boolean operators (AND, OR, XOR, NOT) into coherent narratives, ensuring solutions rely strictly on internal logic rather than external knowledge. The construction pipeline involves four key steps:

1. Logic Graph Generation. We generate computational graphs where nodes denote binary states and edges represent logical dependencies. Topologies are randomized per sample to prevent structural overfitting. **2. Natural Language Conversion.** Nodes are mapped to randomized entities (e.g., "Alice") to eliminate prior knowledge biases (e.g., semantic shortcuts), while logical gates are wrapped in narrative templates to ensure self-containment. **3. Controllable Complexity.** Datasets are generated with variable reasoning depths ranging from 3 to 10 hops to strictly evaluate scaling behavior across complexity levels. **4. Ground Truth CoT.** A rigorous step-by-step derivation is generated for each sample.

Using this pipeline, we constructed the evaluation suite detailed in Table 1. While the validation and test sets maintain a relatively stable size, the training set expands as reasoning hops increase to accommodate the growing complexity. Examples are provided in Appendix F.

6 Methodology: Implicit Reasoning via Latent State Alignment

6.1 From Theoretical Insight to ALiCoT

Theorem 1 shows that omitted steps create a high-order learning bottleneck. The parity-invariant correlation analysis in Appendix A.2 shows that target-aligned tuples yield a stable invariant direction, whereas irrelevant tuples average to zero. Hence, latent states need not match a prescribed distribution,

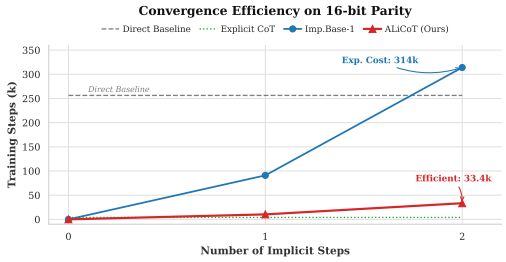


Figure 4: **Convergence Efficiency on 16-bit Parity.** We track the training steps required to reach 100% accuracy as intermediate steps are compressed. Imp.Base-1 suffers exponential growth in training cost (surging to 314k steps), while ALiCoT maintains a flat curve (peaking at only 33.4k), demonstrating that latent alignment mitigates the explosion in learning cost.

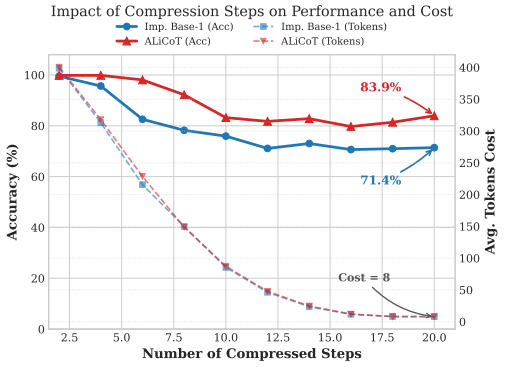


Figure 5: **Impact of Compression Steps on Performance and Cost.** As more intermediate steps are compressed, token consumption for both methods drops to only $\sim$8. Yet ALiCoT (red) maintains accuracy (83.94%), outperforming Imp. Base-1 (blue, 71.40%) by $> 12\%$ when all reasoning steps are compressed away.

reproduce omitted text, or imitate a teacher; they only need to remain distinguishable across reasoning states and provide low-order computational support.

We propose **ALiCoT** to enforce this weaker condition through the model's own latent representations. For natural-language reasoning, alignment uses the current model without a separate teacher model or an auxiliary step decoder. At a high level, it augments the task objective with a task-appropriate alignment loss that keeps latent representations for different reasoning states stable and distinguishable:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{align}}. \tag{6}$$

The concrete form of $\mathcal{L}_{\text{align}}$ depends on the representation space: it can use discrete state classification for bounded tasks or continuous feature alignment for natural language.

6.2 Preliminary Verification on Parity Problem

We study the 16-bit Parity problem[1] under progressively stronger compression, replacing intermediate layers with latent tokens until all are compressed. In this setting, the contextual hidden state of each latent token is fed to the auxiliary classifier to predict the discrete state of its omitted parity step. A step involving b bits has 2^b possible states, so compressing four bits corresponds to a $2^4 = 16$-way classification problem. In this discrete setting, we instantiate the alignment loss as

$$\mathcal{L}_{\text{align}} = -\sum_i \log P(z_i \mid g(h_{\text{latent},i})), \tag{7}$$

where z_i is the ground-truth state of the corresponding omitted step. This sweep tests the rapid training-cost growth predicted by the gradient hierarchy and whether ALiCoT mitigates it by preserving omitted information. We measure iterations required to reach 100% test accuracy. As shown in Figure 4, the standard iCoT baseline exhibits a sharp increase in convergence cost, rising to 314k steps, whereas ALiCoT remains flatter, peaking at 33.4k steps.

6.3 Adaptation to Natural Language

For natural language reasoning, ALiCoT uses two forward passes per batch with the same current model: one on the full gold CoT under teacher forcing, and one on the latent-token sequence. We pair gold steps and latent positions in order, stopping when either sequence ends. This avoids the

[1]We use 16-bit Parity rather than higher-bit variants because fully compressed training becomes prohibitively expensive in our setup, potentially requiring more than a month.

Table 2: **Main Accuracy Results Across Model Scales.** Accuracy (%) on 3–10 hop NatBool-DAG under full intermediate-step compression. Avg. is weighted by the test-set sizes in Table 1. Imp. Base-2 is reported only where evaluated due to computational constraints.

Model Setting	Method	Accuracy (%) by Hops								Avg.
		3	4	5	6	7	8	9	10	Acc
GPT2	Imp. Base-1	46.82	55.85	57.97	57.19	54.01	54.50	57.21	55.37	55.05
	Imp. Base-2	39.73	51.31	49.28	51.73	44.79	51.04	53.87	53.26	49.28
	ALiCoT	**56.36**	**60.20**	**64.86**	**65.26**	**62.37**	**60.32**	**65.07**	**64.41**	**62.44**
Llama-3.2 (1B)	Imp. Base-1	**68.18**	**69.57**	71.01	68.77	66.20	**66.67**	72.05	63.84	68.50
	ALiCoT	65.91	66.89	**72.46**	**73.68**	**74.56**	65.08	**76.42**	**74.01**	**71.25**
Llama-3.2 (3B)	Imp. Base-1	72.27	64.55	75.00	73.68	71.08	69.84	72.93	72.32	71.36
	ALiCoT	**93.64**	**88.96**	**84.42**	**88.07**	**82.58**	**85.71**	**89.52**	**81.36**	**86.85**
Qwen3 (0.6B)	Imp. Base-1	68.64	68.56	72.46	72.63	72.47	70.37	75.11	70.62	71.41
	Imp. Base-2	46.42	55.92	57.14	57.29	54.16	54.68	57.75	54.89	54.97
	ALiCoT	**87.27**	**84.28**	**82.25**	**85.26**	**82.23**	**82.01**	**84.72**	**83.62**	**83.94**
Qwen3 (4B)	Imp. Base-1	79.09	75.92	75.36	81.05	77.35	78.84	81.22	74.01	77.88
	ALiCoT	**100.00**	**97.99**	**96.74**	**93.68**	**92.68**	**92.59**	**92.58**	**92.66**	**95.01**
Qwen3 (8B)	Imp. Base-1	77.73	77.26	78.62	81.75	78.05	80.42	84.72	78.53	79.56
	ALiCoT	**100.00**	**96.66**	**95.29**	**96.49**	**92.33**	**92.59**	**87.77**	**92.09**	**94.34**

auxiliary step decoder in SIM-CoT (Wei et al., 2025) and the sequential forward passes for latent generation in CODI (Shen et al., 2025). We use the cosine alignment loss:

$$\mathcal{L}_{\text{align}} = 1 - \cos(\mathbf{c}_s, \text{SG}(\mathbf{t}_s)) = 1 - \frac{\mathbf{c}_s \cdot \text{SG}(\mathbf{t}_s)}{|\mathbf{c}_s|_2 |\text{SG}(\mathbf{t}_s)|_2} \tag{8}$$

where $\mathbf{c}_s$ is the final-layer contextual hidden state at latent position s, and $\mathbf{t}_s$ is the elementwise maximum over the final-layer token hidden states in gold step s. Targets are recomputed each batch and detached by $\text{SG}(\cdot)$ (stop-gradient); unmatched latent positions receive no alignment loss.

Experimental Results. On NatBool-DAG, full compression reduces each answer to only ~ 8 generated tokens, corresponding to a benchmark-specific $54.40\times$ output-token compression ratio relative to explicit CoT. Under this setting, ALiCoT maintains 83.94% accuracy on Qwen3-0.6B Yang et al. (2025), over 12% above Imp. Base-1 (Figure 5); across model scales, it reaches 95.01% on Qwen3-4B versus 77.88% for Imp. Base-1 (Table 2).

Generalization to Mathematical Reasoning (GSM8K). To evaluate real-world mathematical reasoning beyond Boolean logic, we test **Llama-3.2-3B** on **GSM8K** Cobbe et al. (2021) with all intermediate steps compressed:

Table 3: **GSM8K (Llama-3.2-3B).** Dk denotes k reasoning hops. Accuracy (%) with all intermediate steps compressed. Training: 10 epochs except CODI (15); Coconut did not converge.

Paradigm	Method	D2	D3	D4	D5	D6	D7	D8	Overall	AvgTok
Imp. Base-1	Pause-token-style Goyal et al. (2024)	19.02	7.82	**6.40**	4.00	5.75	5.00	0.00	9.40	8.24
	Coconut Hao et al. (2024)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1419.12
Imp. Base-2	CODI Shen et al. (2025)	3.99	1.62	1.68	2.29	0.00	0.00	**10.00**	2.27	7.00
	SIM-CoT Wei et al. (2025)	0.61	3.50	4.04	4.00	2.30	2.50	0.00	2.81	3.00
ALiCoT		**23.62**	**12.13**	6.06	**9.14**	**9.20**	**7.50**	5.00	**12.74**	8.22

Within the reported training budgets (10 epochs; 15 for CODI), ALiCoT achieves **12.74%** accuracy, compared with 9.40% for Imp. Base-1, 2.27% for CODI, and 2.81% for SIM-CoT (Table 3). Coconut does not converge within 10 epochs under our configuration. These results support the benefit of alignment on GSM8K within these budgets, but do not establish superiority over fully converged baselines.

7 Discussion and Limitation

This study identifies interaction-order explosion as the central barrier to CoT compression: omitting intermediate steps attenuates the learning signal. NatBool-DAG addresses semantic shortcuts by enforcing irreducible logical dependencies, while ALiCoT mitigates the barrier by aligning latent states with explicit reasoning states. A limitation is that, like existing implicit CoT methods, ALiCoT still relies on explicit CoT traces during training; future work should construct low-order supports automatically.

AI USE STATEMENT

We used generative AI tools primarily to polish the manuscript, including correcting grammar and improving clarity and readability. The authors reviewed all AI-assisted revisions to ensure accuracy and consistency with the intended meaning.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27850–27895. Curran Associates, Inc., 2024. doi: 10.52202/079017-0874.

Anthony J. Bell. The co-information lattice. 2003. URL https://api.semanticscholar.org/CorpusID:5031248.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*, 2025.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. Keypoint-based progressive chain-of-thought distillation for llms. *Proceedings of Machine Learning Research*, 235:13241–13255, 2024. ISSN 2640-3498. Publisher Copyright: Copyright 2024 by the author(s); 41st International Conference on Machine Learning, ICML 2024 ; Conference date: 21-07-2024 Through 27-07-2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ph04CRkPdC.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2024. URL https://arxiv.org/abs/2412.06769.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19660–19722, 2024.

Aleks Jakulin. Machine learning based on attribute interactions. 2005. URL https://api.semanticscholar.org/CorpusID:124311115.

Aleks Jakulin and Ivan Bratko. Quantifying and visualizing attribute interactions. *arXiv preprint cs/0308002*, 2003.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24312–24320, 2025.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24527–24561, 2024.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 44642–44668, 2025a.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=n2NidsYDop.

Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. Compressing chain-of-thought in llms via step entropy. *arXiv preprint arXiv:2508.03346*, 2025.

Zhaolu Liu, Robert Peach, Pedro AM Mediano, and Mauricio Barahona. Interaction measures, partition lattices and kernel tests for high-order interactions. *Advances in neural information processing systems*, 36:36991–37012, 2023.

Zhaolu Liu, Mauricio Barahona, and Robert Peach. Information-theoretic measures on lattices for higher-order interactions. In *International Conference on Artificial Intelligence and Statistics*, pp. 2206–2214. PMLR, 2025.

Yijia Luo, Yulin Song, Xingyao Zhang, Jiaheng Liu, Weixun Wang, GengRu Chen, Wenbo Su, and Bo Zheng. Deconstructing long chain-of-thought: A structured reasoning optimization framework for long cot distillation. *arXiv preprint arXiv:2503.16385*, 2025.

Boxiang Ma, Ru Li, Wang Yuanlong, Hongye Tan, and Xiaoli Li. Memorization $\neq$ understanding: Do large language models have the ability of scenario cognition? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20747–20763, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1047. URL https://aclanthology.org/2025.emnlp-main.1047/.

Arvind V Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024.

W. McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954. doi: 10.1109/TIT.1954.1057469.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. International Conference on Learning Representations 2024, 2024.

Meta AI. Llama 3.2 model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B.

OpenAI. "learning to reason with llms,". *[Online]*, 2025. URL https://openai.com/index/learning-to-reason-with-llms/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Fernando E Rosas, Pedro AM Mediano, Michael Gastpar, and Henrik J Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):032305, 2019.

Tomas Scagliarini, Daniele Marinazzo, Yike Guo, Sebastiano Stramaglia, and Fernando E. Rosas. Quantifying high-order interdependencies on individual patterns via the local o-information: Theory and applications to music analysis. *Phys. Rev. Res.*, 4:013184, Mar 2022. doi: 10.1103/PhysRevResearch.4.013184. URL https://link.aps.org/doi/10.1103/PhysRevResearch.4.013184.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation, 2025. URL https://arxiv.org/abs/2502.21074.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. *arXiv preprint arXiv:2509.20317*, 2025.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *The Thirteenth International Conference on Learning Representations*.

Joe Whittaker, Florian Martin, and Yang Xiang. Synergy, suppression and immorality: forward differences of the entropy function. *arXiv preprint arXiv:1501.04368*, 2015.

Noam Wies, Yoav Levine, and Amnon Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=BrJATVZDWEH.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. TokenSkip: Controllable chain-of-thought compression in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3351–3363, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.165. URL https://aclanthology.org/2025.emnlp-main.165/.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. LightThinker: Thinking step-by-step compression. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13307–13328, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.673. URL https://aclanthology.org/2025.emnlp-main.673/.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. Lightthinker: Thinking step-by-step compression, 2025b. URL https://arxiv.org/abs/2502.15589.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in neural information processing systems*, 36:31967–31987, 2023.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, Zhenhe Wu, Yongchi Zhao, Tianyu Liu, Jian Yang, Wangchunshu Zhou, Chujie Zheng, Chongxuan Li, Yuyin Zhou, Zhoujun Li, Zhaoxiang Zhang, Jiaheng Liu, Ge Zhang, Wenhao Huang, and Jason Eshraghian. A survey on latent reasoning, 2025. URL https://arxiv.org/abs/2507.06203.

A Proof of Theorem 1

A.1 Transformer Model

We discuss the learning ability of models in constructing implicit reasoning chains under a one-layer Transformer architecture. To simplify the analysis, we use absolute position encoding and single-head Softmax attention, and the single-layer Transformer also omits residual connections.

Data encoding: We utilize high-dimensional vector representations initialized via a uniform distribution $\mathcal{U}(-1, 1)$. To preserve distributional symmetry and geometric separability for the binary input tokens, we adopt a coupled initialization strategy. Let $\mathbf{v} \in \mathbb{R}^{d_{model}}$ be the base embedding vector corresponding to the token "1", with its elements sampled independently from $\mathcal{U}(-1, 1)$. Accordingly, the token "-1" is encoded as the negation $-\mathbf{v}$. Formally, for the input token $x_j \in \{1, -1\}$ at the j-th position, the resulting encoding $\mathbf{e}_j$ is defined as:

$$\mathbf{e}_j = x_j \mathbf{v}. \tag{9}$$

This formulation ensures that the embedding direction is strictly determined by the token sign while sharing the same magnitude and global statistical properties.

Softmax attention layer: The attention layer consists of key, query, and value matrices $\mathbf{K}, \mathbf{Q}, \mathbf{V}$. To make the dynamical analysis tractable, following Zhang et al. (2024); Huang et al. (2024); Mahankali et al. (2024); Kim & Suzuki (2024), we focus on position encoding when obtaining attention scores, while the value matrix operates solely on the semantic embeddings $\boldsymbol{e}$. Therefore, in this paper, we treat the position encoding $\boldsymbol{p}$ separately from the input embedding $\boldsymbol{e}$. The forward propagation of the attention layer is formally defined as:

$$\hat{\boldsymbol{z}}_m = \sum_{j=1}^{m-1} \mathbf{V}\boldsymbol{e}_j \cdot \mathsf{softmax}(\hat{\boldsymbol{p}}_j^\top \mathbf{K}^\top \mathbf{Q} \hat{\boldsymbol{p}}_m) = \sum_{j=1}^{m-1} \sigma_j(\boldsymbol{w}_m)\boldsymbol{e}_j, \tag{10}$$

where $\sigma_j(\boldsymbol{w}_m)$ represents the attention weights (softmax scores) allocated to the j-th token. Specifically, we aggregate the key and query projections into a single learnable interaction matrix $\mathbf{W} \triangleq \mathbf{K}^\top \mathbf{Q}$. Consequently, the attention weights are computed as $\sigma_j(\boldsymbol{w}_m) = \frac{\exp(w_{j,m})}{\sum_{\alpha=1}^{m-1} \exp(w_{\alpha,m})}$, where the scalar logit $w_{j,m} = \hat{\boldsymbol{p}}_j^\top \mathbf{W} \hat{\boldsymbol{p}}_m$ quantifies the position-based compatibility between the query at step m and the key at step j.

Feedforward layer: To fully adapt to the parity problem, following Kim & Suzuki (2025a), we use a mapping function $\phi : [-1, 1] \to [-1, 1]$, requiring $\phi(0) = -1$, $\phi(\pm 1) = 1$, and set the function to be smooth and differentiable. Finally, we define $\phi(t) = 1 - ct^2 + ct^4$ and $\phi'(t) = -2ct + 4ct^3$.

Output probability: To perform the classification task, the hidden state $\boldsymbol{h}_m$ is projected into the logit space. We employ the Softmax function to obtain the predicted probability distribution vector $\hat{\boldsymbol{y}}_m \in \mathbb{R}^2$:

$$\hat{\boldsymbol{y}}_m = \text{Softmax}(\mathbf{W}_o^\top \boldsymbol{h}_m), \quad \text{where } [\hat{\boldsymbol{y}}_m]_k = \frac{\exp(\mathbf{c}_k^\top \boldsymbol{h}_m)}{\sum_{j=1}^{2} \exp(\mathbf{c}_j^\top \boldsymbol{h}_m)}. \tag{11}$$

Here, $\mathbf{W}_o \in \mathbb{R}^{d_{model} \times 2} = [\mathbf{c}_1, \mathbf{c}_2]$ denotes the weights of the output projection head, and $[\hat{\boldsymbol{y}}_m]_k$ represents the probability of the k-th class. Analogous to our input encoding strategy, we enforce a symmetric structure on the output weights. We initialize a vector $\mathbf{u} \in \mathbb{R}^{d_{model}}$ such that the weight for the class "1" is defined as $\mathbf{c}_1 = \mathbf{u}$, while the weight for the class "-1" is set as the negation, $\mathbf{c}_2 = -\mathbf{u}$. This construction ensures that the decision boundaries for the binary states remain geometrically symmetric.

Loss function: We define the optimization objective using the standard Cross-Entropy (CE) loss, rather than the Mean Squared Error (MSE). Given the ground-truth label $y \in \{1, -1\}$, the loss for the m-th token is the negative log-likelihood of the true class. Considering the binary output dimension, the loss is formulated as:

$$\mathcal{L} = -\sum_{k=1}^{2} \mathbb{I}(y = l_k) \log([\hat{\boldsymbol{y}}_m]_k), \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The term l_k denotes the class label corresponding to the k-th index of the probability vector, specifically $l_1 = 1$ and $l_2 = -1$. Minimizing $\mathcal{L}$ is equivalent to maximizing the probability assigned to the correct label required for the reasoning step.

Thus, the transformer computes $\text{TF}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{d+k-1}; \mathbf{W}) = (\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_{d+k-1})$ where the original data $\hat{\boldsymbol{x}}_j = \boldsymbol{x}_j, j \in [d]$ remain unchanged and tokens $\hat{\boldsymbol{x}}_{d+1}, \cdots, \hat{\boldsymbol{x}}_{d+k-1}$ are computed as $\hat{\boldsymbol{x}}_m = \phi(\hat{\boldsymbol{z}}_m)$.

A.2 Parity-Invariant Correlation Analysis

To analyze the interaction mechanism between input variables and the prediction target, we introduce a geometric interpretation based on the Hadamard product of their representations.

Embedding Representations. We represent an input variable $x \in \{1, -1\}$ as a vector $x\boldsymbol{v}$, and a target variable $y \in \{1, -1\}$ as $y\boldsymbol{u}$, where $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^{d_{model}}$ are base vectors for the input and output spaces, respectively.

Second-order Interaction. Consider a reasoning step where the target y_m is determined by the parity of two input variables x_i and x_j, i.e., $y_m = x_i x_j$. In the binary domain $\{1, -1\}$, this multiplicative relationship enforces the strict constraint:

$$y_m \cdot x_i \cdot x_j = 1. \tag{13}$$

This implies that the sign of the output must align with the combined sign of the inputs. To capture this geometric alignment, we define the interaction vector $\boldsymbol{\psi}$ as the element-wise product (Hadamard product, denoted by $\odot$) of the embeddings of these three variables:

$$\begin{aligned} \boldsymbol{\psi} &= (y_m \boldsymbol{u}) \odot (x_i \boldsymbol{v}) \odot (x_j \boldsymbol{v}) \\ &= (y_m \cdot x_i \cdot x_j)(\boldsymbol{u} \odot \boldsymbol{v} \odot \boldsymbol{v}). \end{aligned} \tag{14}$$

Given the constraint $y_m x_i x_j \equiv 1$, the scalar coefficients cancel out, causing the interaction vector to collapse into a constant, input-independent state:

$$\boldsymbol{\psi} = \boldsymbol{u} \odot \boldsymbol{v}^{\odot 2}, \tag{15}$$

where $\boldsymbol{v}^{\odot 2} = \boldsymbol{v} \odot \boldsymbol{v}$ denotes the element-wise square. We term this property **Parity-Invariant Correlation**. It demonstrates that for any valid reasoning triplet (x_i, x_j, y_m), their joint representation in the hyperspace always points to a fixed invariant direction $\boldsymbol{u} \odot \boldsymbol{v}^{\odot 2}$, regardless of the individual values of the variables.

Generalization to High-order Interactions. This invariance property generalizes to k-th order interactions. Suppose a target y_m is determined by the cumulative product of k variables $\{x_{i_1}, \ldots, x_{i_k}\}$, satisfying $y_m = \prod_{n=1}^{k} x_{i_n}$. The consistency condition becomes $y_m \cdot \prod_{n=1}^{k} x_{i_n} = 1$. Consequently, the cumulative Hadamard product of their vector representations remains invariant:

$$(y_m \boldsymbol{u}) \odot \left(\bigodot_{n=1}^{k} (x_{i_n} \boldsymbol{v}) \right) = \boldsymbol{u} \odot \boldsymbol{v}^{\odot k}. \tag{16}$$

This result suggests that learning the underlying logical rule is geometrically equivalent to aligning the model's aggregate representation with the specific invariant direction $\boldsymbol{u} \odot \boldsymbol{v}^{\odot k}$.

Support Set and Interaction Categories. To rigorously analyze the gradient components, we categorize the interaction terms based on their ability to reconstruct the target parity y_m. We consider a tuple of indices $J = (j_1, \ldots, j_r)$ with $1 \le j_k \le m-1$. The collective contribution of these features is determined by the product $X_J = \prod_{k=1}^{r} x_{j_k}$.

Consistent with the parity constraint, we define a tuple J as *target-aligned* (or relevant) if the product of its components strictly recovers the target label (i.e., their correlation is trivial):

$$\mathcal{S}_{r,m} := \left\{ (j_1, \ldots, j_r) \in [m-1]^r \;\middle|\; y_m \cdot \prod_{k=1}^{r} x_{j_k} \equiv 1 \right\}. \tag{17}$$

Tuples in $\mathcal{S}_{r,m}$ correspond to the true support of the underlying function (potentially including redundant pairs that cancel out).

Conversely, we define the set of *misaligned* (or irrelevant) tuples $\mathcal{I}_{r,m}$ as those that fail to reconstruct the target:

$$\mathcal{I}_{r,m} := \left\{ (j_1, \ldots, j_r) \in [m-1]^r \;\middle|\; y_m \cdot \prod_{k=1}^{r} x_{j_k} \not\equiv 1 \right\}. \tag{18}$$

Based on the rigorous definitions of $\mathcal{S}_{r,m}$ and $\mathcal{I}_{r,m}$, we examine the statistical behavior of the scalar contraction over the dataset. This analysis reveals a fundamental dichotomy in how these terms contribute to the gradient updates:

Statistical Distinction. We consider the average contribution of a specific index tuple $J = (j_1, \ldots, j_r)$ across n samples:

- **Relevant Terms (Deterministic Signal):** For any tuple $J \in \mathcal{S}_{r,m}$, the parity constraint ensures that $y_m \cdot \prod_{k=1}^{r} x_{j_k} \equiv 1$. Consequently, the contraction collapses into a deterministic constant for every sample, yielding a stable average:
$$\frac{1}{n} \sum_{i=1}^{n} \left\langle y_m^{(i)} \boldsymbol{u}, x_{j_1}^{(i)} \boldsymbol{v}, \ldots, x_{j_r}^{(i)} \boldsymbol{v} \right\rangle = C_{\text{signal}}. \tag{19}$$
- **Irrelevant Terms (Zero-Mean Noise):** In contrast, for any tuple $J \in \mathcal{I}_{r,m}$, the product term represents a non-trivial parity function. Under uniform inputs, the single-sample contraction is stochastic:
$$\langle y_m \boldsymbol{u}, x_{j_1} \boldsymbol{v}, \ldots, x_{j_r} \boldsymbol{v} \rangle \sim \text{Uniform}(\{+C_{\text{signal}}, -C_{\text{signal}}\}). \tag{20}$$
Since these terms are uniformly distributed around zero, their average over the dataset is a sum of zero-mean random variables, which tends to vanish as n increases.

Concentration of Irrelevant Interactions. To ensure the model learns the correct rule, the signal C_{signal} must dominate the worst-case noise from irrelevant terms. We formalize this bound using the concentration of measure.

Lemma 2 (Concentration of Irrelevant Terms). *Let $\mathcal{I}_{r,m}$ be the set of misaligned index tuples. If inputs are i.i.d. uniform over $\{\pm 1\}$, then for any confidence level $1-p$, the maximum magnitude of the noise interactions is bounded by:*

$$\max_{J \in \mathcal{I}_{r,m}} \left| \frac{1}{n} \sum_{i=1}^{n} \left\langle y_m^{(i)} \boldsymbol{u}, x_{j_1}^{(i)} \boldsymbol{v}, \ldots, x_{j_r}^{(i)} \boldsymbol{v} \right\rangle \right| \le \kappa, \tag{21}$$

where $\kappa := C_{signal} \cdot \sqrt{\frac{2}{n} \log \frac{2|\mathcal{I}_{r,m}|}{p}}$.

Proof. The proof relies on the zero-mean property of terms in $\mathcal{I}_{r,m}$ and a union bound argument:

1. **Point-wise Bound via Hoeffding:** Fix an irrelevant tuple $J \in \mathcal{I}_{r,m}$. The summation consists of n independent random variables bounded by $[-C_{\text{signal}}, C_{\text{signal}}]$ with zero expectation. By Hoeffding's inequality:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle y_m^{(i)}\boldsymbol{u}, x_{j_1}^{(i)}\boldsymbol{v}, \ldots, x_{j_r}^{(i)}\boldsymbol{v}\right\rangle\right| \geq \kappa\right) \leq 2\exp\left(-\frac{n\kappa^2}{2C_{\text{signal}}^2}\right). \tag{22}$$

2. **Uniform Bound via Union Bounding:** We must ensure that *none* of the potential irrelevant terms masquerade as a signal. Applying the union bound over all tuples in $\mathcal{I}_{r,m}$, we limit the probability that the *maximum* noise exceeds κ:

$$\Pr\left(\max_{J\in\mathcal{I}_{r,m}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle y_m^{(i)}\boldsymbol{u}, x_{j_1}^{(i)}\boldsymbol{v}, \ldots, x_{j_r}^{(i)}\boldsymbol{v}\right\rangle\right| \geq \kappa\right) \leq |\mathcal{I}_{r,m}| \cdot 2\exp\left(-\frac{n\kappa^2}{2C_{\text{signal}}^2}\right). \tag{23}$$

Setting the right-hand side to p and solving for κ yields the stated bound. □

Implication. With sufficient samples ($n = \tilde{\Omega}(\log m)$), the unweighted contraction signal C_{signal} dominates κ. Applying this distinction to the cross-entropy gradient additionally requires controlling its sample-dependent output factor, as established below.

A.3 Gradient Derivation for Attention Weights

To analyze the learning dynamics, particularly for tasks requiring high-frequency representations like parity, we derive the gradient of the loss $\mathcal{L}$ with respect to the attention weights $w_{j,m}$. For the sake of clarity in the initial derivation, we define $\mathcal{L}$ as the standard Cross-Entropy loss for a **single training instance** with target class t, denoted as $\mathcal{L} = -\log([\hat{\boldsymbol{y}}_m]_t)$. The generalization to the mean loss over a batch of n samples will be incorporated subsequent to the derivation of the per-sample gradient components.

Applying the chain rule, the gradient can be decomposed into four primary components:

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{y}}_m} \cdot \frac{\partial \hat{\boldsymbol{y}}_m}{\partial \boldsymbol{h}_m}}_{\text{(I)}} \cdot \underbrace{\frac{\partial \boldsymbol{h}_m}{\partial \hat{\boldsymbol{z}}_m}}_{\text{(II)}} \cdot \underbrace{\frac{\partial \hat{\boldsymbol{z}}_m}{\partial w_{j,m}}}_{\text{(III)}}. \tag{24}$$

Part (I): Loss and Output Projection. We derive the gradient of the loss with respect to the hidden state $\boldsymbol{h}_m$ by decomposing the chain rule through the logits $\mathbf{z} = \mathbf{W}_o^\top \boldsymbol{h}_m$. Let $z_i = \mathbf{c}_i^\top \boldsymbol{h}_m$ be the i-th logit. The gradient can be expressed as:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_m} = \sum_{i=1}^{K} \frac{\partial \mathcal{L}}{\partial z_i}\frac{\partial z_i}{\partial \boldsymbol{h}_m}. \tag{25}$$

First, we compute the derivative of the Cross-Entropy loss $\mathcal{L} = -\log(p_t)$ with respect to the logit z_i, where $p_k = [\hat{\boldsymbol{y}}_m]_k = \text{softmax}(\mathbf{z})_k$ denotes the predicted probability for class k, and c is the index of the true class. Applying the chain rule, we have:

$$\frac{\partial \mathcal{L}}{\partial z_i} = \sum_{k=1}^{K} \frac{\partial \mathcal{L}}{\partial p_k}\frac{\partial p_k}{\partial z_i}. \tag{26}$$

Since the loss only depends on the target class probability p_t, the term $\frac{\partial \mathcal{L}}{\partial p_k}$ is non-zero only when $k = c$, specifically $\frac{\partial \mathcal{L}}{\partial p_t} = -\frac{1}{p_t}$. Using the derivative of the Softmax function $\frac{\partial p_k}{\partial z_i} = p_k(\delta_{ki} - p_i)$, the expression simplifies to:

$$\frac{\partial \mathcal{L}}{\partial z_i} = -\frac{1}{p_t} \cdot \frac{\partial p_t}{\partial z_i} = -\frac{1}{p_t} \cdot p_t(\delta_{ci} - p_i) = p_i - \delta_{ci}. \tag{27}$$

In vector notation, this corresponds to the residual vector $\frac{\partial \mathcal{L}}{\partial \mathbf{z}} = (\hat{\boldsymbol{y}}_m - \mathbf{1}_t)^\top$, where $\mathbf{1}_t$ is the one-hot encoding of the target class. Next, considering the linear projection $z_i = \mathbf{c}_i^\top \boldsymbol{h}_m$, the partial derivative is simply the class embedding: $\frac{\partial z_i}{\partial \boldsymbol{h}_m} = \mathbf{c}_i^\top$. Substituting these back into the summation yields:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_m} = \sum_{i=1}^{K} (p_i - \delta_{ci}) \mathbf{c}_i^\top = \left(\sum_{i=1}^{K} p_i \mathbf{c}_i - \mathbf{c}_t \right)^\top = -\left(\mathbf{c}_t - \mathbb{E}_{\hat{\boldsymbol{y}}_m}[\mathbf{c}] \right)^\top. \tag{28}$$

This term represents the error signal in the embedding space, pointing from the current weighted expected embedding $\mathbb{E}_{\hat{\boldsymbol{y}}_m}[\mathbf{c}]$ towards the target class centroid $\mathbf{c}_t$.

Part (II): Feed-Forward Network. The gradient propagates through the element-wise Feed-Forward layer, defined as $\boldsymbol{h}_m = \phi(\hat{\boldsymbol{z}}_m)$. Since ϕ is applied independently to each dimension of the input vector $\hat{\boldsymbol{z}}_m$ without any learnable weight matrices, the Jacobian matrix is diagonal. Specifically:

$$\frac{\partial \boldsymbol{h}_m}{\partial \hat{\boldsymbol{z}}_m} = \text{diag}(\phi'(\hat{\boldsymbol{z}}_m)), \tag{29}$$

where $\phi'(t) = -2ct + 4ct^3$ is the element-wise derivative of the polynomial mapping function. This term acts as a gate, scaling the gradient flow based on the magnitude of the attention output features.

Part (III): Attention Mechanism. Finally, we compute the derivative of the attention output $\hat{\boldsymbol{z}}_m = \sum_{\alpha=1}^{m-1} \sigma_\alpha(\boldsymbol{w}_m) \boldsymbol{x}_\alpha$ with respect to the attention logits $w_{j,m}$. Using the quotient rule for the Softmax function $\sigma(\cdot)$, we obtain:

$$\frac{\partial \hat{\boldsymbol{z}}_m}{\partial w_{j,m}} = \sum_{\alpha=1}^{m-1} (\delta_{j\alpha} - \sigma_j(\boldsymbol{w}_m)) \sigma_\alpha(\boldsymbol{w}_m) \boldsymbol{e}_\alpha = \sigma_j(\boldsymbol{w}_m)(\boldsymbol{e}_j - \hat{\boldsymbol{z}}_m). \tag{30}$$

Combined Gradient. Substituting Parts (I), (II), and (III) back into Eq. (24), we obtain the final expression for the gradient:

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = -\underbrace{\left(\mathbf{c}_t - \mathbb{E}_{\hat{\boldsymbol{y}}_m}[\mathbf{c}] \right)^\top}_{\text{Output Error}} \cdot \underbrace{\text{diag}(\phi'(\hat{\boldsymbol{z}}_m))}_{\text{Element-wise Jacobian}} \cdot \underbrace{\sigma_j(\boldsymbol{w}_m)(\boldsymbol{e}_j - \hat{\boldsymbol{z}}_m)}_{\text{Attention Gradient}}. \tag{31}$$

This equation elucidates how the error signal backpropagates directly through the non-linear activation ϕ', and interacts with the attention mechanism's sensitivity to the input token $\boldsymbol{x}_j$.

Detailed Decomposition. To analyze the specific learning dynamics, we substitute the precise functional forms into the combined gradient in Eq. (24). Specifically, the three components are instantiated as follows:

- **Output Error:** The general error term concretizes to $-(\mathbf{c}_t - \sum_{k=1}^{K} p_k \mathbf{c}_k)^\top$, representing the residual between the target class prototype $\mathbf{c}_t$ and the current probability-weighted prototypes.
- **Element-wise Jacobian:** By applying the Taylor expansion to the activation derivative ϕ', this term equals $-2c \cdot \text{diag}(\frac{1}{m-1} \sum_{\alpha=1}^{m-1} \boldsymbol{e}_\alpha) + 4c \cdot \text{diag}(\hat{\boldsymbol{z}}_m^{\odot 3})$, where the linear term dominates the high-order residuals.
- **Attention Gradient:** Under the assumption of uniform attention initialization (where attention weights $\sigma_j(\boldsymbol{w}_m) = 1/(m-1)$), this term simplifies to $\frac{1}{m-1}(\boldsymbol{e}_j - \frac{1}{m-1} \sum_{\beta=1}^{m-1} \boldsymbol{e}_\beta)$, capturing the deviation of the specific token $\boldsymbol{e}_j$ from the historical context.

Multiplying these three specific terms and fully expanding the resulting expression leads to a sum of eight distinct gradient components:

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = \frac{2c}{m-1}(1 - p_t)\mathbf{c}_t^\top \cdot \text{diag}\left(\frac{1}{m-1} \sum_{\alpha=1}^{m-1} \boldsymbol{e}_\alpha \right) \cdot \boldsymbol{e}_j \tag{32}$$

$$- \frac{2c}{m-1}(1-p_t)\mathbf{c}_t^\top \cdot \text{diag}\left(\frac{1}{m-1}\sum_{\alpha=1}^{m-1}\boldsymbol{e}_\alpha\right)\cdot\left(\frac{1}{m-1}\sum_{\beta=1}^{m-1}\boldsymbol{e}_\beta\right) \tag{33}$$

$$- \frac{2c}{m-1}\left(\sum_{k\neq c}p_k\mathbf{c}_k\right)^\top \cdot \text{diag}\left(\frac{1}{m-1}\sum_{\alpha=1}^{m-1}\boldsymbol{e}_\alpha\right)\cdot\boldsymbol{e}_j \tag{34}$$

$$+ \frac{2c}{m-1}\left(\sum_{k\neq c}p_k\mathbf{c}_k\right)^\top \cdot \text{diag}\left(\frac{1}{m-1}\sum_{\alpha=1}^{m-1}\boldsymbol{e}_\alpha\right)\cdot\left(\frac{1}{m-1}\sum_{\beta=1}^{m-1}\boldsymbol{e}_\beta\right) \tag{35}$$

$$- \frac{4c}{m-1}(1-p_t)\mathbf{c}_t^\top \cdot \text{diag}\left(\hat{\boldsymbol{z}}_m^{\odot 3}\right)\cdot\boldsymbol{e}_j \tag{36}$$

$$+ \frac{4c}{m-1}(1-p_t)\mathbf{c}_t^\top \cdot \text{diag}\left(\hat{\boldsymbol{z}}_m^{\odot 3}\right)\cdot\left(\frac{1}{m-1}\sum_{\beta=1}^{m-1}\boldsymbol{e}_\beta\right) \tag{37}$$

$$+ \frac{4c}{m-1}\left(\sum_{k\neq c}p_k\mathbf{c}_k\right)^\top \cdot \text{diag}\left(\hat{\boldsymbol{z}}_m^{\odot 3}\right)\cdot\boldsymbol{e}_j \tag{38}$$

$$- \frac{4c}{m-1}\left(\sum_{k\neq c}p_k\mathbf{c}_k\right)^\top \cdot \text{diag}\left(\hat{\boldsymbol{z}}_m^{\odot 3}\right)\cdot\left(\frac{1}{m-1}\sum_{\beta=1}^{m-1}\boldsymbol{e}_\beta\right). \tag{39}$$

Leveraging the binary symmetry where $\mathbf{c}_1 = -\mathbf{c}_2$, the output error term simplifies significantly: $\mathbf{c}_t - \sum p_k\mathbf{c}_k = 2(1-p_t)\mathbf{c}_t$. This allows us to merge the target and non-target components in Eqs. (32)-(39). Consequently, the eight-term expansion collapses into four concise terms, consisting of the linear and residual interactions with the target prototype:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial w_{j,m}} = &\frac{4c}{(m-1)^2}(1-p_t)\sum_{\alpha=1}^{m-1}\langle\mathbf{c}_t,\boldsymbol{e}_\alpha,\boldsymbol{e}_j\rangle \\ &- \frac{4c}{(m-1)^3}(1-p_t)\sum_{\alpha=1}^{m-1}\sum_{\beta=1}^{m-1}\langle\mathbf{c}_t,\boldsymbol{e}_\alpha,\boldsymbol{e}_\beta\rangle \\ &- \frac{8c}{m-1}(1-p_t)\left\langle\mathbf{c}_t,\hat{\boldsymbol{z}}_m^{\odot 3},\boldsymbol{e}_j\right\rangle \\ &+ \frac{8c}{(m-1)^2}(1-p_t)\sum_{\beta=1}^{m-1}\left\langle\mathbf{c}_t,\hat{\boldsymbol{z}}_m^{\odot 3},\boldsymbol{e}_\beta\right\rangle.\end{aligned} \tag{40}$$

This unified representation highlights the structural symmetry between the primary attention-like interactions (terms 1 and 2) and their corresponding higher-order corrections (terms 3 and 4), where the global context acts as a centering mechanism across both orders of approximation.

To account for the optimization over a batch of n samples, we extend the gradient formulation by averaging over the sample index i. Let the superscript (i) denote the specific variables (embeddings, probabilities, and residuals) associated with the i-th sample, and let t_i represent the target class for that sample. The aggregate gradient becomes:

$$\frac{\partial\mathcal{L}}{\partial w_{j,m}} = \frac{4c}{n(m-1)^2}\sum_{i=1}^{n}\left[(1-p_{t_i}^{(i)})\sum_{\alpha=1}^{m-1}\left\langle\mathbf{c}_{t_i},\boldsymbol{e}_\alpha^{(i)},\boldsymbol{e}_j^{(i)}\right\rangle\right] \tag{41}$$

$$- \frac{4c}{n(m-1)^3}\sum_{i=1}^{n}\left[(1-p_{t_i}^{(i)})\sum_{\alpha=1}^{m-1}\sum_{\beta=1}^{m-1}\left\langle\mathbf{c}_{t_i},\boldsymbol{e}_\alpha^{(i)},\boldsymbol{e}_\beta^{(i)}\right\rangle\right] \tag{42}$$

$$- \frac{8c}{n(m-1)} \sum_{i=1}^{n} \left[(1 - p_{t_i}^{(i)}) \left\langle \mathbf{c}_{t_i}, (\hat{\boldsymbol{z}}_m^{(i)})^{\odot 3}, \boldsymbol{e}_j^{(i)} \right\rangle \right] \tag{43}$$

$$+ \frac{8c}{n(m-1)^2} \sum_{i=1}^{n} \left[(1 - p_{t_i}^{(i)}) \sum_{\beta=1}^{m-1} \left\langle \mathbf{c}_{t_i}, (\hat{\boldsymbol{z}}_m^{(i)})^{\odot 3}, \boldsymbol{e}_\beta^{(i)} \right\rangle \right]. \tag{44}$$

Here, the outer summation $\frac{1}{n}\sum_{i=1}^{n}$ averages the contribution of each sample. For every sample i, the gradient is weighted by the model's uncertainty $(1 - p_{t_i}^{(i)})$, ensuring that well-classified samples contribute less to the parameter update, while the structural terms (interactions between target prototypes $\mathbf{c}_{t_i}$, history $\boldsymbol{e}_\alpha^{(i)}$, and current token $\boldsymbol{e}_j^{(i)}$) drive the learning process.

Controlling the sample-dependent output factor. Let $a(\boldsymbol{x}) = \boldsymbol{u}^\top \phi(\hat{\boldsymbol{z}}_m)$ and $D_j a = \partial a / \partial w_{j,m}$. The symmetric output logits are $(a, -a)$, so $p_t = (1 + y \tanh(a))/2$. The single-example gradient has the exact decomposition

$$\frac{\partial \ell}{\partial w_{j,m}} = -2(1 - p_t) y D_j a = -y D_j a + R_j(\boldsymbol{x}), \qquad R_j(\boldsymbol{x}) := \tanh(a) D_j a. \tag{45}$$

The prediction term R_j is controlled jointly, rather than dropping the output factor from individual contractions.

Lemma 3 (Prediction-term centering at initialization)**.** *Condition on fixed initialized vectors $\boldsymbol{u}, \boldsymbol{v}$, independent of the evaluation samples. Suppose the visible bits $x_1, \ldots, x_{m-1}$ are i.i.d. uniform over $\{\pm 1\}$, the embeddings are $\boldsymbol{e}_j = x_j \boldsymbol{v}$, and $\boldsymbol{w}_m = \mathbf{0}$, so that $\sigma_j(\boldsymbol{w}_m) = 1/(m-1)$ for every $j < m$ in each example. Assume $\|\boldsymbol{u}\|_1$, $\|\boldsymbol{v}\|_\infty$, and ϕ' on the attained pre-activation range are bounded independently of m, n. Then $\mathbb{E}[R_j] = 0$ for every $j < m$. For n independent examples, with probability at least $1 - p$,*

$$\max_{j<m} \left| \frac{1}{n} \sum_{i=1}^{n} R_j(\boldsymbol{x}^{(i)}) \right| \le \frac{C_{\mathrm{pred}}}{m-1} \sqrt{\frac{2\log(2(m-1)/p)}{n}} = O\Big(\frac{\kappa}{m}\Big), \tag{46}$$

where C_{pred} depends only on these bounds and $\kappa = O(\sqrt{\log(m/p)/n})$.

Proof. At uniform attention, a is invariant under permutations of the visible bits, and the derivatives $D_j a$ permute with their indices. Exchangeability therefore makes $\mathbb{E}[R_j]$ identical for all j. Softmax centering gives, for every input,

$$\sum_{j<m} D_j a = \boldsymbol{u}^\top \operatorname{diag}(\phi'(\hat{\boldsymbol{z}}_m)) \sum_{j<m} \sigma_j(\boldsymbol{e}_j - \hat{\boldsymbol{z}}_m) = 0. \tag{47}$$

Thus $\sum_{j<m} R_j = 0$, and each expectation must be zero. Moreover, $|R_j| \le |D_j a| \le C_{\mathrm{pred}}/(m-1)$, with $C_{\mathrm{pred}} = 2\|\boldsymbol{u}\|_1 \|\boldsymbol{v}\|_\infty \sup|\phi'|$. Hoeffding's inequality and a union bound over $m-1$ positions yield Eq. (46). □

Consequently, the empirical gradient equals the empirical target-dependent term $-n^{-1}\sum_i y^{(i)} D_j a^{(i)}$ plus $O(\kappa/m)$. Lemma 2 is applied only to this unweighted target-dependent term. The two concentration events are combined by assigning failure probability $p/2$ to each, which does not change the stated rate. This argument does not require constant output probabilities or zero means for individual weighted contractions.

Condition for extensions. For general output logits $s_k(X)$, the exact cross-entropy gradient is $-\partial s_y/\partial w_{j,m} + R_j(X)$, where $R_j(X) = \sum_k p_k(X) \partial s_k(X)/\partial w_{j,m}$. To retain the same decomposition beyond the exchangeable setting, we explicitly require, at the analyzed initialization,

$$\mathbb{E}[R_j(X)] = 0, \qquad |R_j(X)| \le \frac{C_{\mathrm{pred}}}{m-1} \quad \text{for every analyzed attention logit,} \tag{48}$$

with C_{pred} independent of m, n (but possibly architecture-dependent). Lemma 3 verifies this condition for the parity setup above. Correlated intermediate tokens, nonuniform attention, and architectural

extensions do not automatically satisfy it. If the mean is nonzero, the decomposition must retain the additional population term $\mathbb{E}[R_j(X)]$; increasing n controls fluctuations around this mean, not the mean itself.

Computing interaction strengths.

We now analyze the quadratic and quartic components of the target-dependent gradient $-n^{-1}\sum_i y^{(i)} D_j a^{(i)}$, denoted by $\widehat{G}_j^{(2)}$ and $\widehat{G}_j^{(4)}$. Lemma 3 controls the remaining prediction term. We use the exact normalization $1/(m-1)$ in coefficients and retain m as the asymptotic context-length scale.

Based on Lemma 2, the sample average of the contraction for any index tuple (a, b) concentrates based on its membership in these sets:

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle \mathbf{c}_{t_i}, \boldsymbol{e}_a^{(i)}, \boldsymbol{e}_b^{(i)}\right\rangle = \begin{cases} C_{\text{signal}}^{(2)} & \text{if } (a,b)\in\mathcal{S}_{2,m}, \\ \text{noise}\in[-\kappa,\kappa] & \text{if } (a,b)\in\mathcal{I}_{2,m}, \end{cases} \tag{49}$$

where $\kappa = O(\sqrt{\frac{\log m}{n}})$ represents the noise floor.

1. Analyzing the current-token quadratic contraction in (41): Consider the term $\sum_{\alpha=1}^{m-1}\left\langle \mathbf{c}_{t_i}, \boldsymbol{e}_\alpha^{(i)}, \boldsymbol{e}_j^{(i)}\right\rangle$. We classify the contribution based on whether the fixed index j can form a valid signal tuple with any α.

- **Case: j is a relevant feature.** If j is part of the true parity support, there exists exactly one index α' such that the pair $(\alpha', j) \in \mathcal{S}_{2,m}$ (representing the complementary feature in the parity pair). For all other $\alpha \neq \alpha'$, the pair $(\alpha, j) \in \mathcal{I}_{2,m}$.
- **Case: j is an irrelevant feature.** If j is not part of the support, then for all $\alpha \in [m-1]$, the pair (α, j) never matches the target parity. Thus, $(\alpha, j) \in \mathcal{I}_{2,m}$ for the entire summation.

By Eq. (45), these unweighted contractions occur in the target-dependent gradient; the sample-dependent output factor has not been taken outside the sample average. Their empirical sum satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\alpha=1}^{m-1}\left\langle \mathbf{c}_{t_i}, \boldsymbol{e}_\alpha^{(i)}, \boldsymbol{e}_j^{(i)}\right\rangle = \begin{cases} C_{\text{signal}}^{(2)} + O(m\kappa) & \text{if } \exists\alpha' : (\alpha', j)\in\mathcal{S}_{2,m}, \\ O(m\kappa) & \text{otherwise.} \end{cases} \tag{50}$$

2. Analyzing the inner sum of (42): This term involves a double summation over all pairs $(\alpha, \beta) \in [m-1]\times[m-1]$. This global term is independent of j.

- The sum accumulates $C_{\text{signal}}^{(2)}$ for every tuple (α, β) that belongs to $\mathcal{S}_{2,m}$.
- Since the order of indices matters in the tensor contraction but parity is symmetric, $\mathcal{S}_{2,m}$ contains all permutations of the support indices. For a degree-2 parity, $|\mathcal{S}_{2,m}| = 2$.
- All other $(m-1)^2 - 2$ tuples belong to $\mathcal{I}_{2,m}$.

Thus, the bound is:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{\alpha,\beta}\left\langle \mathbf{c}_{t_i}, \boldsymbol{e}_\alpha^{(i)}, \boldsymbol{e}_\beta^{(i)}\right\rangle &= |\mathcal{S}_{2,m}|\cdot C_{\text{signal}}^{(2)} + O(m^2\kappa) \\ &= 2C_{\text{signal}}^{(2)} + O(m^2\kappa).\end{aligned} \tag{51}$$

3. Combining the terms:

Substituting (50) and (51) into the target-dependent gradient, we define a generalized membership indicator dependent on the interaction order r.

Let $\mathbb{I}_j^{(r)}$ denote whether feature j is an active component in the ground-truth signal set of order r, denoted $\mathcal{S}_{r,m}$:

$$\mathbb{I}_j^{(r)} := \mathbb{I}\left(\exists\,(i_1,\ldots,i_r)\in\mathcal{S}_{r,m} \text{ such that } j\in\{i_1,\ldots,i_r\}\right). \tag{52}$$

This indicator effectively acts as a selector: $\mathbb{I}_j^{(r)} = 1$ if feature j contributes to an r-th order interaction, and 0 otherwise.

For our specific case where the leading interaction is second-order ($r = 2$), the combined gradient expression becomes:

$$\widehat{G}_j^{(2)} = \frac{2c}{(m-1)^2}\left(C_{\text{signal}}^{(2)}\mathbb{I}_j^{(2)} + O(m\kappa)\right) - \frac{2c}{(m-1)^3}\left(2C_{\text{signal}}^{(2)} + O(m^2\kappa)\right) \tag{53}$$

$$= \frac{2c}{(m-1)^2}C_{\text{signal}}^{(2)}\mathbb{I}_j^{(2)} - \frac{4c}{(m-1)^3}C_{\text{signal}}^{(2)} + O\Big(\frac{\kappa}{m}\Big). \tag{54}$$

4. Analyzing the inner sum of (43): This term involves a triple summation over indices $(k_1, k_2, k_3) \in [m]^3$ with one index j fixed. We examine the tuple formed by (k_1, k_2, k_3, j).

- Case A (Relevant Feature): If j is part of the ground-truth support set (i.e., $\mathbb{I}_j^{(4)} = 1$), then for the tuple to belong to $\mathcal{S}_{4,m}$, the indices (k_1, k_2, k_3) must be a permutation of the remaining 3 support indices. The number of such permutations is $(4-1)! = 3! = 6$.
- Case B (Irrelevant Feature): If j is not in the support set (i.e., $\mathbb{I}_j^{(4)} = 0$), no choice of (k_1, k_2, k_3) can form a valid support tuple. The signal contribution is 0.
- The total number of terms in the summation is m^3.

The aggregated contribution is:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k_1,k_2,k_3}\left\langle \mathbf{c}_{t_i}, \boldsymbol{e}_{k_1}^{(i)}, \boldsymbol{e}_{k_2}^{(i)}, \boldsymbol{e}_{k_3}^{(i)}, \boldsymbol{e}_{j}^{(i)}\right\rangle = \begin{cases} 6C_{\text{signal}}^{(4)} + O(m^3\kappa) & \text{if } \mathbb{I}_j^{(4)} = 1, \\ O(m^3\kappa) & \text{if } \mathbb{I}_j^{(4)} = 0. \end{cases} \tag{55}$$

5. Analyzing the inner sum of (44): This term involves a quadruple summation over all indices (β, k_1, k_2, k_3). This global term is independent of j.

- The sum accumulates $C_{\text{signal}}^{(4)}$ for every tuple (k_1, k_2, k_3, β) that belongs to $\mathcal{S}_{4,m}$.
- Since the sum covers all possible combinations, it includes all valid permutations of the support set. For a degree-4 parity, $|\mathcal{S}_{4,m}| = 4! = 24$.
- The total number of terms in the summation is approximately m^4.

Thus, the bound is:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\beta,k_1,k_2,k_3}\left\langle \mathbf{c}_{t_i}, \boldsymbol{e}_{k_1}^{(i)}, \boldsymbol{e}_{k_2}^{(i)}, \boldsymbol{e}_{k_3}^{(i)}, \boldsymbol{e}_{\beta}^{(i)}\right\rangle = 24C_{\text{signal}}^{(4)} + O(m^4\kappa). \tag{56}$$

6. Combining the 4th-order terms:

Substituting (55) and (56) into the target-dependent quartic gradient gives

$$\widehat{G}_j^{(4)} = -\frac{4c}{(m-1)^4}\left(6C_{\text{signal}}^{(4)}\mathbb{I}_j^{(4)} + O(m^3\kappa)\right) + \frac{4c}{(m-1)^5}\left(24C_{\text{signal}}^{(4)} + O(m^4\kappa)\right) \tag{57}$$

$$= -\frac{24c}{(m-1)^4}C_{\text{signal}}^{(4)}\mathbb{I}_j^{(4)} + \frac{96c}{(m-1)^5}C_{\text{signal}}^{(4)} + O\Big(\frac{\kappa}{m}\Big). \tag{58}$$

Combined Gradient Analysis and Learning Dynamics

Combining the target-dependent expansion with Lemma 3 recovers the empirical cross-entropy gradient. The prediction term changes only the finite-sample remainder. For the quartic activation, $\gamma_2 = -2c$ and $\gamma_4 = 4c$; retaining these signed coefficients gives the same convention as the main theorem.

Total Gradient Approximation

For any feature index j, the leading support-dependent contributions can therefore be summarized as

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = \sum_{r\in\{2,4\}} \gamma_r \left[-\Theta(m^{-r})\mathbb{I}_j^{(r)} C_{\text{signal}}^{(r)} + \Theta(m^{-(r+1)}) C_{\text{signal}}^{(r)}\right] + \underbrace{O\Big(\frac{\kappa}{m}\Big)}_{\text{Noise}}, \tag{59}$$

where $\mathbb{I}_j^{(r)} \in \{0,1\}$ is the indicator function, which equals 1 if and only if feature j belongs to the ground-truth support set $\mathcal{S}_{r,m}$ of order r.

A.4 Generalization to Arbitrary Activation Functions

While the previous analysis utilized a specific quartic polynomial to demonstrate the emergence of second and fourth-order interaction signals, specific functional forms restrict the generality of the findings. To provide a comprehensive understanding of the Transformer's learning dynamics, we extend this framework to general non-linearities. In this section, we generalize our derivation to arbitrary smooth activation functions using Taylor series expansions. This allows us to strictly quantify the signal intensity for interaction terms of any order r. **Expansion Point Justification**: The validity of the perturbative analysis depends on expanding the function around the operating point of the hidden states at initialization. Given the symmetric initialization of embeddings $\mathbf{e} \sim \mathcal{U}(-1,1)$ and the linearity of the attention aggregation, the pre-activation feature $\hat{\boldsymbol{z}}_m$ satisfies $\mathbb{E}[\hat{\boldsymbol{z}}_m] = \mathbf{0}$. Consequently, the statistical behavior of the gradient is dictated by the local geometry of the activation function around zero. We therefore employ the Maclaurin series (Taylor expansion at $z = 0$). **Formal Definition**: Let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth activation function. We express it as:

$$\phi(z) = \sum_{k=0}^{\infty} \frac{\phi^{(k)}(0)}{k!} z^k. \tag{60}$$

Recall that the gradient flow is modulated by the derivative $\phi'(\hat{\boldsymbol{z}}_m)$. Differentiating the series, we define the gradient scaling function as:

$$\phi'(z) = \sum_{k=1}^{\infty} \frac{\phi^{(k)}(0)}{(k-1)!} z^{k-1} = \sum_{r=1}^{\infty} \gamma_r z^{r-1}, \tag{61}$$

where $\gamma_r \triangleq \frac{\phi^{(r)}(0)}{(r-1)!}$ is the *Interaction Coefficient* for order r.

Generalized Gradient Decomposition: First apply Eq. (45) and Lemma 3, which hold for any activation satisfying the stated derivative bound. Substituting the series for ϕ' into the target-dependent term then gives

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = -\sum_{r=1}^{\infty} \gamma_r \cdot \Bigg\{ \underbrace{\frac{1}{n(m-1)} \sum_{i=1}^{n} \Big\langle \mathbf{c}_{t_i}, (\hat{\boldsymbol{z}}_m^{(i)})^{\odot(r-1)}, \boldsymbol{e}_j^{(i)} \Big\rangle}_{\text{Term A: Current-token interaction}} - \underbrace{\frac{1}{n(m-1)^2} \sum_{i=1}^{n} \sum_{\beta=1}^{m-1} \Big\langle \mathbf{c}_{t_i}, (\hat{\boldsymbol{z}}_m^{(i)})^{\odot(r-1)}, \boldsymbol{e}_\beta^{(i)} \Big\rangle}_{\text{Term B: Context centering}} \Bigg\} + O\Big(\frac{\kappa}{m}\Big). \tag{62}$$

For an irreducible order-r support, the leading population signal in Term A scales as $\Theta(m^{-r})$, while its centering contribution in Term B has one additional inverse-context factor. The prediction term and the empirical fluctuations of the unweighted contractions are both included in $O(\kappa/m)$; no zero-mean assumption is made for individual loss-weighted contractions.

Generalizing the asymptotic behavior observed in the quadratic and quartic cases, we can express the total gradient as a sum of signal and bias terms for each order r. The final generalized scaling law is:

$$\frac{\partial \mathcal{L}}{\partial w_{j,m}} = \sum_{r=1}^{\infty} \gamma_r \cdot \left[\underbrace{-\Theta(m^{-r}) \cdot \mathbb{I}_j^{(r)} \cdot C_{\text{signal}}^{(r)}}_{\text{Order-}r\text{ Signal}} + \underbrace{\Theta(m^{-(r+1)}) \cdot C_{\text{signal}}^{(r)}}_{\text{Order-}r\text{ Bias}} \right] + \underbrace{O\left(\frac{\kappa}{m}\right)}_{\text{Noise}}. \tag{63}$$

where $\mathbb{I}_j^{(r)}$ represents the structured interaction tensor of order r specific to token j. This confirms that lower-order terms (small r) dominate the learning dynamics, while higher-order terms decay exponentially with the complexity of the interaction.

Conclusion and Implications

The analytical result in (59) reveals the intrinsic learning dynamics of the model. We summarize the key findings as follows:

First, the model demonstrates a discriminative learning mechanism that effectively distinguishes between relevant and irrelevant features. For features within the support set (where $\mathbb{I}_j^{(r)} = 1$), the gradient is dominated by a signal term scaling as $-\Theta(m^{-r})$, which provides a deterministic drive for the weights to capture the interaction. Conversely, for irrelevant features, this signal vanishes, leaving the gradient driven solely by higher-order bias terms and stochastic noise. This distinction allows the model to selectively amplify features in the support set while suppressing non-informative inputs.

Second, there exists a clear hierarchy in the learning difficulty of different interaction orders, governed by the sequence length m. The effective signal strength decays exponentially with the interaction order, dropping from $\Theta(m^{-2})$ for second-order terms to $\Theta(m^{-4})$ for fourth-order terms. Although higher-order terms are theoretically learnable, their significantly smaller gradient magnitudes compared to lower-order terms imply that the model exhibits a spectral bias, prioritizing the learning of lower-order interactions before fitting the more complex, high-order dependencies.

Third, the sample size n plays a critical role in noise suppression, acting as the fundamental bottleneck for learning high-order interactions. Since the fourth-order signal intensity $\Theta(m^{-4})$ is extremely weak, the noise term $O(\kappa/m)$—where κ is typically inversely proportional to $\sqrt{n}$—must be suppressed to a level lower than the signal. Consequently, learning high-order logic imposes a strict requirement for a sufficiently large sample size to reduce statistical variance; otherwise, the faint high-order signal will be overwhelmed by the noise derived from lower-order couplings, preventing the model from converging to the correct solution.

B Experimental Validation of Gradient Scaling

Theorem 1 predicts that the support-selective gradient weakens with interaction order and context length, whereas finite-sample fluctuations decrease with sample size. We test these three dependencies directly in a minimal attention model at symmetric initialization, using exact enumeration for short contexts and paired gradient sampling for longer ones.

Experimental setup. Let m denote the target position, with $m-1$ visible input positions. Each input is a binary vector $(x_1, \ldots, x_{m-1}) \in \{-1,+1\}^{m-1}$, with target $y = \prod_{j=1}^{r} x_j$; thus, the first r positions form the irreducible support and the remaining positions are irrelevant. To avoid cancellation of the interaction signal under the symmetric output projection, we choose unequal, nonzero embedding coordinates. A single causal attention head reads two-dimensional embeddings $\boldsymbol{e}_j = x_j \boldsymbol{v}$, where $\boldsymbol{v} = (1, \frac{1}{2})^\top$. All attention logits $w_{j,m}$ are zero, giving exact uniform weights $1/(m-1)$. With the value map set to the identity, the attended state is $\hat{\boldsymbol{z}}_m = (m-1)^{-1} \sum_{j=1}^{m-1} \boldsymbol{e}_j$. We use the elementwise activation $\phi(t) = e^t$, whose nonzero higher-order derivatives allow us to probe $r = 2, \ldots, 6$ with the same model. The symmetric binary-class logits are $(s, -s)$, where $s = \boldsymbol{u}^\top \phi(\hat{\boldsymbol{z}}_m)$ and $\boldsymbol{u} = (1,-1)^\top$. For each example $\boldsymbol{x}$, we denote the cross-entropy gradient with respect to the *pre-softmax* attention logit at uniform attention by $g_j(\boldsymbol{x}) = \frac{\partial \mathcal{L}}{\partial w_{j,m}}\Big|_{\boldsymbol{w}_m = \boldsymbol{0}}$.

Signal and noise measurement. Let $G_j = 2^{-(m-1)} \sum_{\boldsymbol{x} \in \{-1,+1\}^{m-1}} g_j(\boldsymbol{x})$ be the population gradient, computed by exhaustive enumeration where feasible, and let $\widehat{G}_j = n^{-1} \sum_{i=1}^{n} g_j(\boldsymbol{x}^{(i)})$ be

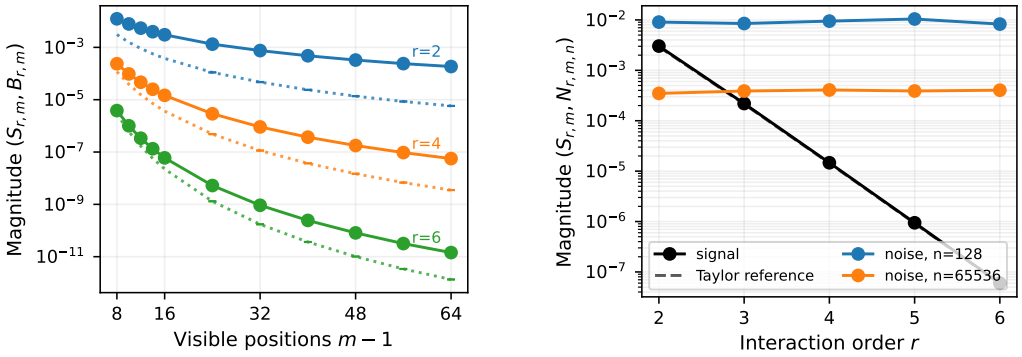
its estimate from an independently and uniformly sampled batch. We use the following measures:

$$S_{r,m} = \left| \frac{1}{r} \sum_{j=1}^{r} G_j - \frac{1}{m-1-r} \sum_{j=r+1}^{m-1} G_j \right|,$$

$$B_{r,m} = \left| \frac{1}{m-1-r} \sum_{j=r+1}^{m-1} G_j \right|,$$

$$N_{r,m,n} = \widehat{Q}_{0.95}\left(\max_{1 \le j \le m-1} |\widehat{G}_j - G_j| \right), \qquad \mathrm{SNR}_{r,m,n} = S_{r,m}/N_{r,m,n}, \tag{64}$$

The difference $S_{r,m}$ compares mean gradients on support and irrelevant positions, isolating the support-selective signal; $B_{r,m}$ measures the background gradient on irrelevant positions. $N_{r,m,n}$ is the empirical 95th percentile, over independent batches, of the largest deviation $|\widehat{G}_j - G_j|$ across positions. Finally, $\mathrm{SNR}_{r,m,n}$ compares the support signal with this finite-sample noise floor.



(a) Support contrast (solid) and centering bias (dotted) versus visible positions $m-1$.

(b) Population signal, Taylor reference, and noise floors across orders at $m = 17$.

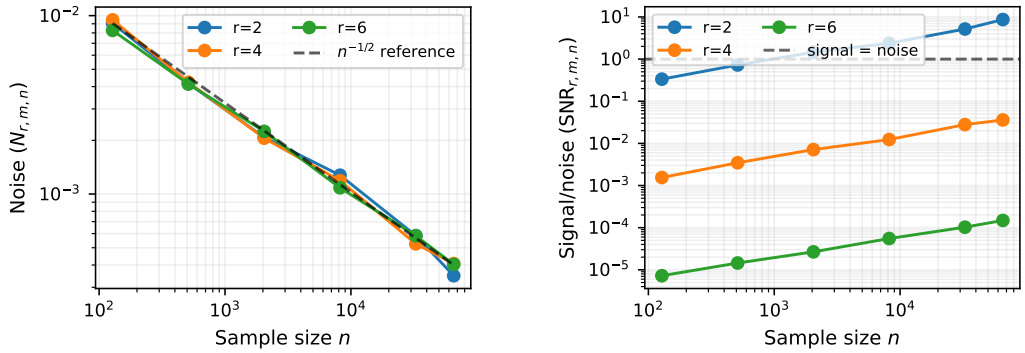
Figure 6: Gradient scaling on parity interactions. Signal and bias are exact for $m \le 17$ and estimated by paired gradient sampling for larger m; 95% confidence intervals are smaller than the markers. Noise floors in (b) use independent batches at $m = 17$. The context-length axis is linear; vertical axes are logarithmic.

Results and interpretation. Figure 6a shows that the support contrast weakens as the context length grows, especially at higher interaction orders, while the centering bias remains smaller. This matches the theorem's $\Theta(m^{-r})$ support signal and weaker bias. At fixed context length, Figure 6b shows that higher-order signals rapidly fall toward or below the noise floor, in line with the predicted order-dependent hierarchy.

Figure 7a shows that sampling noise decreases with n, following the $n^{-1/2}$ reference predicted by the theorem. Accordingly, the signal-to-noise ratios in Figure 7b rise with n, but only the second-order signal exceeds the noise level in the tested range. Together, these curves support the theorem's prediction that higher-order interactions require more samples for signal separation in this controlled parity setting. The compression experiments in Section 6.2 examine the consequences for training.

C Discussion on two latent token paradigms

Existing approaches for implementing implicit CoT typically employ special control tokens—such as <bot> (Begin of Thought) and <eot> (End of Thought)—to delimit the latent space. Under this common framework, the reasoning process is structured as a sequence $[\boldsymbol{x}_{\text{context}}, \texttt{<bot>}, \boldsymbol{c}_s, \texttt{<eot>}]$. However, as shown in Figure 8, the derivation of the latent token $\boldsymbol{c}_s$ varies significantly across different studies, generally falling into two paradigms:



(a) Finite-sample noise and the $n^{-1/2}$ reference.

(b) Signal-to-noise ratio; the dashed line marks equality.

Figure 7: Finite-sample detectability versus sample size n at $m = 17$. Noise is the empirical 95th percentile of the maximum gradient error over 32 independent batches.

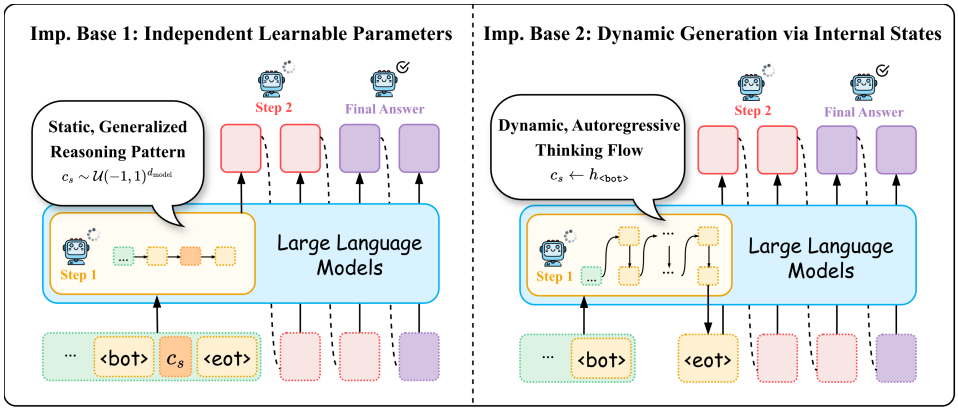


Figure 8: Two different implementation paradigms for implicit CoT: Independent Learnable Parameters and Dynamic Generation via Internal States.

•**Imp.Base-1: Independent Learnable Parameters.** One category of methods treats the latent token $\boldsymbol{c}_s$ as a standalone, global parameter Zhang et al. (2025a); Goyal et al. (2024). In this setting, $\boldsymbol{c}_s$ is decoupled from the model's immediate runtime activations and functions similarly to a "soft prompt" or a continuous instruction embedding. The vector is typically initialized by sampling from a high-dimensional uniform distribution:

$$\boldsymbol{c}_s \sim \mathcal{U}(-1, 1)^{d_{\text{model}}}. \tag{65}$$

During training, this vector is optimized via backpropagation to encode a static, generalized reasoning pattern that bridges the gap between the context and the answer, remaining constant across different inference instances.

•**Imp.Base-2: Dynamic Generation via Internal States.** The other category generates $\boldsymbol{c}_s$ dynamically from the model's internal processingHao et al. (2024), enabling an instance-specific "thinking" flow. This process is autoregressive:

(1) **State Extraction:** Given the context ending with the start marker $[\ldots, \boldsymbol{x}_t, \texttt{<bot>}]$, the model performs a forward pass. The output hidden state at the <bot> position is captured to serve as the latent token:

$$\boldsymbol{c}_s \leftarrow \boldsymbol{h}_{\texttt{<bot>}}. \tag{66}$$

(2) **Recurrent Inference:** This computed state $\boldsymbol{c}_s$ is then fed back into the sequence followed by <eot>. The model processes $[\ldots, \texttt{<bot>}, \boldsymbol{c}_s, \texttt{<eot>}]$ to predict the subsequent explicit token.

Following the proof in Section A, we analyze target-dependent contractions; their full cross-entropy gradient interpretation assumes Eq. (48), since adding latent tokens need not preserve exchangeability. We use the same Order-r Interaction support $\mathcal{S}^{(r)}$ as in Definition 1. The relevant question is not merely whether $\boldsymbol{c}_s$ can appear in an algebraic product with context tokens. The key question is whether it provides a *support-selective*, sample-dependent feature that has non-zero population correlation with the target

$$y = \prod_{j \in \mathcal{S}^{(r)}} x_j, \qquad |\mathcal{S}^{(r)}| = r. \tag{67}$$

Under independent Rademacher inputs, for any context subset B,

$$\mathbb{E}\left[y \prod_{j \in B} x_j \right] = 0 \quad \text{unless } B = \mathcal{S}^{(r)}. \tag{68}$$

Thus an omitted reasoning step can be replaced by a latent token only if the latent state contributes an input-dependent feature that covers part of the missing support. Otherwise, the downstream model must still collect the entire raw support $\mathcal{S}^{(r)}$, and the effective interaction order is not reduced.

C.1 Analysis of Implicit CoT: Independent Learnable Parameters

No sample-dependent support information. An independent latent token is a global learnable vector, fixed across samples. It can certainly participate in downstream products, but it enters them only through sample-independent coefficients. For example, for a scalar projection or tensor contraction involving a context subset B,

$$T_{B,\boldsymbol{c}_s} = \langle \{x_j \boldsymbol{v}\}_{j \in B}, \boldsymbol{c}_s \rangle = \left(\prod_{j \in B} x_j \right) \underbrace{\langle \{\boldsymbol{v}\}_{j \in B}, \boldsymbol{c}_s \rangle}_{C_{\boldsymbol{c},B}}, \tag{69}$$

where $C_{\boldsymbol{c},B}$ is a constant determined by $\boldsymbol{c}_s$ and the fixed vectors in the contraction, but not by the input instance. The target correlation of this term is therefore

$$\mathbb{E}\left[y T_{B,\boldsymbol{c}_s} \right] = C_{\boldsymbol{c},B} \mathbb{E}\left[y \prod_{j \in B} x_j \right]. \tag{70}$$

By Eq. (68), this expectation is zero for every proper support subset $B \subsetneq \mathcal{S}^{(r)}$. Hence a static latent parameter cannot replace any omitted support token or intermediate reasoning step. The only case where Eq. (70) can be non-zero is $B = \mathcal{S}^{(r)}$, but then the context has already collected the entire raw support; the latent token merely rescales an already complete order-r interaction and does not reduce the effective order.

Finite-sample behavior. For $B \neq \mathcal{S}^{(r)}$, the corresponding terms are zero-mean variables. They may appear in the empirical gradient because of finite-sample fluctuations, but they do not form a stable population signal. By invoking the same concentration argument used in Lemma 2 (Hoeffding's inequality applied to bounded zero-mean variables), we can bound their maximum empirical contribution.

Lemma 4 (Concentration of Latent Interactions)**.** *Consider the set of context subsets $\mathcal{J}$ whose latent interaction terms involve the independent token $\boldsymbol{c}_s$ but do not contain the complete raw support $\mathcal{S}^{(r)}$. Under i.i.d. uniform inputs over $\{\pm 1\}$, for any confidence level $1-p$, the maximum empirical magnitude is bounded by*

$$\max_{B \in \mathcal{J}} \left| \frac{1}{n} \sum_{i=1}^{n} y^{(i)} T^{(i)}_{B,\boldsymbol{c}_s} \right| \leq \kappa_s, \tag{71}$$

where $\kappa_s := C_{\boldsymbol{c}} \sqrt{\frac{2}{n} \log \frac{2|\mathcal{J}|}{p}}$ and $C_{\boldsymbol{c}}$ uniformly bounds the magnitude of the contractions involving $\boldsymbol{c}_s$.

Implication. Independent latent parameters can behave as soft prompts or global control vectors, but under the task loss alone they cannot create support-selective interactions with proper context subsets. Unlike explicit intermediate tokens, which are sample-dependent functions of their parents and can provide low-order supports for later reasoning, an independent $\boldsymbol{c}_s$ contributes only constants to products of input signs. Consequently, it cannot reduce the effective interaction order; all apparent proper-subset interactions involving $\boldsymbol{c}_s$ are finite-sample noise. This limitation applies to unsupervised independent latent tokens. ALiCoT avoids relying on the raw input parameter itself by supervising the contextual latent hidden state to be distinguishable with respect to reasoning-step identities.

C.2 Analysis of Implicit CoT: Dynamic Generation via Internal States

Dynamic latent states have a different nature. In this paradigm, the latent token is no longer a global parameter, but an input-dependent representation

$$\boldsymbol{c}_s = \boldsymbol{h}_{\texttt{<bot>}} = F_\theta(X). \tag{72}$$

Thus it can contain nontrivial functions of the current context. To make the signal-noise tradeoff explicit, consider a scalar projection of the generated latent state,

$$z_s(X) = \mathbf{a}^\top \boldsymbol{c}_s. \tag{73}$$

Since the parity monomials form an orthogonal basis over $\{\pm 1\}^m$, this scalar feature can be expanded as

$$z_s(X) = \sum_{A \subseteq [m]} \beta_A \prod_{j \in A} x_j. \tag{74}$$

The coefficients β_A summarize which interaction components are encoded in the generated latent state. If a downstream term also selects the context product $\prod_{j \in B} x_j$, then its target-aligned population signal is

$$\mathbb{E}\left[y z_s(X) \prod_{j \in B} x_j \right] = \sum_{A \subseteq [m]} \beta_A \mathbb{E}\left[y \left(\prod_{j \in A} x_j \right) \left(\prod_{j \in B} x_j \right) \right]. \tag{75}$$

Because $x_j^2 = 1$, a term in Eq. (75) has non-zero expectation only when the latent component $\prod_{j \in A} x_j$ and the downstream context product $\prod_{j \in B} x_j$ jointly reduce to the target product $\prod_{j \in \mathcal{S}^{(r)}} x_j$ after repeated variables cancel. This gives the precise sense in which dynamic latent states can help: if $\boldsymbol{c}_s$ already contains a component aligned with part of the missing support, then the downstream context only needs to provide the complementary factors. For example, if $\mathcal{S}^{(4)} = \{2, 4, 6, 8\}$ and the latent state contains the component $x_2 x_4$, then a downstream second-order factor $x_6 x_8$ is sufficient because $(x_2 x_4)(x_6 x_8) = \prod_{j \in \mathcal{S}^{(4)}} x_j$. In this way, dynamic latent generation can raise the representational ceiling: a later low-order computation can realize a higher-order raw interaction if the generated latent state has already encoded part of the support.

Noise from dense interaction mixtures. This benefit comes with a cost. A dynamically generated latent state is not a clean reasoning step like an explicit CoT token; it is a dense summary of the whole prefix. Therefore, it may contain a useful component such as x_2x_4, but it also contains many other components involving irrelevant indices. Only the components that can combine with downstream context factors to recover $\prod_{j\in\mathcal{S}^{(r)}} x_j$ provide coherent signal. The other components have zero population correlation with the target, but on a finite training set they do not cancel perfectly and appear as empirical noise.

Schematically, if the generated state contains N_{dyn} non-negligible interaction components whose magnitudes are bounded by B_{dyn}, the corresponding fluctuation level scales as

$$\kappa_{\mathrm{dyn}} = O\left(B_{\mathrm{dyn}}\sqrt{\frac{\log(N_{\mathrm{dyn}}/p)}{n}}\right). \tag{76}$$

The point is qualitative: compared with an independent latent parameter, a dynamic latent state can contain many more input-dependent components, so there are more irrelevant terms that must be averaged out.

Moreover, the useful high-order components are themselves weak near initialization. For instance, a Taylor expansion of an attention-aggregated hidden state has the schematic form

$$\boldsymbol{h}_{<\mathtt{bot}>} \approx \sum_{q\geq 0} c_q \left(\frac{1}{m}\sum_{\alpha<m} x_\alpha \boldsymbol{v}\right)^{\odot q}, \tag{77}$$

so a fully specified order-q component still carries q attention-averaging factors and is suppressed at scale $\Theta(m^{-q})$. Thus dynamic generation raises the expressivity ceiling by creating more possible high-order signal paths, but it also makes signal isolation harder: the correct component is weak and surrounded by many irrelevant zero-mean components.

Implication. Latent token generation via internal states is therefore high-capacity but low signal-to-noise. It can, in principle, encode rich sample-dependent features and increase the maximum raw interaction order that downstream layers can realize. However, without explicit alignment, the useful support-aligned components are embedded in a combinatorially large pool of irrelevant components. This raises the representational upper bound, but also increases the noise floor and makes the correct reasoning signal harder to isolate during training.

D Architectural Extensions under Uniformly Random Support Locations

Theorem 1 gives a single-layer gradient decomposition for a fixed support. Here we prove a different multi-layer statement: under uniformly random support locations, the *expected ℓ_1 norm of the target-correlated attention-logit gradient* is at most $O_{L,r}(m^{-r})$ for fixed depth L and order r. The proof controls the combined contribution of all input interactions, including those formed at early causal positions, and then averages over support locations.

D.1 Model and Random-Support Experiment

Write $\mathcal{V}(t)$ for the positions visible to receiver t and $d(t) = |\mathcal{V}(t)|$. The strict-prefix convention of Appendix A.1 gives $d(t) = t-1$; including the current input position gives $d(t) = t$, with the hidden state predicting the next token. Put $M = d(m)$ and index the visible raw inputs by $[M]$. Thus $M = m-1$ or $M = m$, respectively.

Let $X = (x_1, \ldots, x_M)$ have independent uniform $\{\pm 1\}$ coordinates. Independently of the initialization, choose

$$S \sim \mathrm{Unif}\{T \subseteq [M] : |T| = r\}, \qquad y_S(X) = \chi_S(X) := \prod_{i\in S} x_i, \qquad 1 \leq r \leq M. \tag{78}$$

The support is chosen once and held fixed across the examples of a task. The initialized input-to-score function is the same for all supports: S determines the label, but is not supplied as an additional input or encoded in the initialized parameters. Expectations over X describe a fixed task; expectations

over S describe the ensemble of support locations. These are distinct from averaging over random initializations or resampling an unobserved support for every example.

We retain the position-only softmax attention logits of the base model. Condition on a fixed initialization with uniform attention:

$$\hat{\boldsymbol{z}}_t^{(\ell,h)} = \sum_{\alpha \in \mathcal{V}(t)} \sigma_{\alpha,t}^{(\ell,h)} \mathbf{V}_{\ell,h} \boldsymbol{h}_\alpha^{(\ell-1)}, \qquad \sigma_{\alpha,t}^{(\ell,h)} = \frac{1}{d(t)}. \tag{79}$$

For an empty prefix we set the attention output to zero; differentiated rows always have $d(t) > 0$. The attention weights are independent of X at every layer. Gradients below are with respect to individual attention logits, holding the other logits fixed, as in the main analysis.

A layer with H heads and optional attention and FFN residuals is

$$\begin{aligned} \boldsymbol{u}_t^{(\ell)} &= \rho_{\ell,A} \boldsymbol{h}_t^{(\ell-1)} + \sum_{h=1}^{H} \mathbf{W}_{O,\ell,h} \hat{\boldsymbol{z}}_t^{(\ell,h)}, \\ \boldsymbol{h}_t^{(\ell)} &= \rho_{\ell,F} \boldsymbol{u}_t^{(\ell)} + \mathbf{W}_{2,\ell} \phi_\ell\big(\mathbf{W}_{1,\ell} \boldsymbol{u}_t^{(\ell)} + \boldsymbol{b}_{1,\ell}\big) + \boldsymbol{b}_{2,\ell}, \qquad \rho_{\ell,A}, \rho_{\ell,F} \in \{0,1\}. \end{aligned} \tag{80}$$

The activation acts coordinatewise. The scalar output score is $f(X) = \mathbf{a}^\top \boldsymbol{h}_m^{(L)} + b_{\mathrm{out}}$; the binary logits are $(f(X), -f(X))$. Raw embeddings are $\boldsymbol{h}_j^{(0)} = x_j \boldsymbol{v}$; any additional query state is constant in X. As in the base theoretical model, the stack has no normalization sublayers. Content-dependent query/key derivatives and gradients of shared query/key parameters are not part of the logit-coordinate result.

Proof roadmap. For a logit coordinate $q = (\ell, h, j, t)$, write $D_q(X) := \partial f(X)/\partial w_{j,t}^{(\ell,h)}$ and $G_q(S) := \mathbb{E}_X[\chi_S(X) D_q(X)]$. Here D_q measures how the score changes when one attention logit changes, and $G_q(S)$ measures the part correlated with the target. We proceed in two steps. First, under the bounds stated below, we bound the sum of the absolute Fourier coefficients of all D_q by a constant independent of M. Second, the signal for support S is exactly the coefficient indexed by S. Averaging over the $\binom{M}{r}$ possible supports therefore divides this total bound by $\binom{M}{r}$, giving $O(m^{-r})$ for fixed r. The next two subsections establish these steps.

D.2 Controlling the Total Fourier Coefficients

On uniform Boolean inputs, the products $\chi_T(X)$ form an orthonormal basis: each product describes an interaction among the inputs in T. Every scalar function of X therefore has the unique multilinear expansion

$$u(X) = \sum_{T \subseteq [M]} \widehat{u}(T) \chi_T(X), \qquad \widehat{u}(T) = \mathbb{E}_X[u(X)\chi_T(X)]. \tag{81}$$

Define $\|u\|_{\mathcal{A}} := \sum_T |\widehat{u}(T)|$ and, for a vector, $\|\boldsymbol{u}\|_{\mathcal{A},\infty} := \max_a \|u_a\|_{\mathcal{A}}$. For example, $u(X) = \frac{1}{2}x_1 + \frac{1}{4}x_1x_2$ has $\widehat{u}(\{1,2\}) = \frac{1}{4}$ and $\|u\|_{\mathcal{A}} = \frac{3}{4}$. The coefficient selects one particular interaction; the norm bounds the total absolute contribution across all interactions. The matrix norm $\|W\|_\infty$ below is the induced maximum-row-sum norm. Parity orthogonality and $\chi_T \chi_U = \chi_{T \triangle U}$ imply

$$\begin{aligned} \|uv\|_{\mathcal{A}} &\le \|u\|_{\mathcal{A}} \|v\|_{\mathcal{A}}, & |u(X)| &\le \|u\|_{\mathcal{A}}, \\ \|W\boldsymbol{u}\|_{\mathcal{A},\infty} &\le \|W\|_\infty \|\boldsymbol{u}\|_{\mathcal{A},\infty}, & \left\| \sum_j \lambda_j \boldsymbol{u}_j \right\|_{\mathcal{A},\infty} &\le \max_j \|\boldsymbol{u}_j\|_{\mathcal{A},\infty} \quad (\lambda_j \ge 0,\ \textstyle\sum_j \lambda_j = 1). \end{aligned} \tag{82}$$

For the product inequality, expand uv, group terms with the same symmetric difference, and apply the triangle inequality. The symmetric difference appears because $x_i^2 = 1$: an input occurring in both factors cancels from their product. The other bounds follow directly from the same coefficient norm. In particular, averaging a short causal prefix does not increase the total absolute coefficients, even though its individual attention weights can be much larger than $1/M$.

Write $\phi_\ell(z) = \sum_{k\geq 0} c_{\ell,k} z^k$ and define the following power series by taking the absolute value of each coefficient:

$$\Phi_\ell(s) = \sum_{k\geq 0} |c_{\ell,k}| s^k, \qquad \Phi'_\ell(s) = \sum_{k\geq 1} k|c_{\ell,k}| s^{k-1}. \tag{83}$$

We assume the power series for ϕ_ℓ and ϕ'_ℓ represent these functions on the required domain, and that Φ_ℓ and Φ'_ℓ are finite at the coefficient bounds specified below. Polynomial activations satisfy this requirement automatically. For analytic activations this is an absolute-convergence condition at the *coefficient-norm bound*, which is stronger than convergence only on the range of realized scalar inputs. Under this condition,

$$\|\phi_\ell(u)\|_{\mathcal{A}} \leq \Phi_\ell(\|u\|_{\mathcal{A}}), \qquad \|\phi'_\ell(u)\|_{\mathcal{A}} \leq \Phi'_\ell(\|u\|_{\mathcal{A}}). \tag{84}$$

These inequalities follow by submultiplicativity and absolute convergence, so no Taylor remainder is discarded.

The next lemma tracks forward representations and backward derivatives separately. The bounds U_ℓ, τ_ℓ, and B_ℓ control the attention output with its residual, the input to the FFN activation, and the layer output, respectively. The factors F_ℓ and J_ℓ bound derivative propagation through the FFN with its residual and through the whole layer. Their role is to keep all architecture-dependent factors explicit and independent of M.

Lemma 5 (Forward and derivative coefficient bounds)**.** *Suppose* $\max_t \|\boldsymbol{h}_t^{(0)}\|_{\mathcal{A},\infty} \leq B_0$. *Define the following deterministic bounds recursively:*

$$\begin{aligned}
\alpha_\ell &:= \rho_{\ell,A} + \sum_{h=1}^{H} \|\mathbf{W}_{O,\ell,h}\|_\infty \|\mathbf{V}_{\ell,h}\|_\infty, & U_\ell &:= \alpha_\ell B_{\ell-1}, \\
\tau_\ell &:= \|\mathbf{W}_{1,\ell}\|_\infty U_\ell + \|\boldsymbol{b}_{1,\ell}\|_\infty, & B_\ell &:= \rho_{\ell,F} U_\ell + \|\mathbf{W}_{2,\ell}\|_\infty \Phi_\ell(\tau_\ell) + \|\boldsymbol{b}_{2,\ell}\|_\infty, \\
F_\ell &:= \rho_{\ell,F} + \|\mathbf{W}_{2,\ell}\|_\infty \Phi'_\ell(\tau_\ell) \|\mathbf{W}_{1,\ell}\|_\infty, & J_\ell &:= \alpha_\ell F_\ell.
\end{aligned} \tag{85}$$

Assume L, H, B_0*, the displayed weight and bias bounds, and the majorants at all* τ_ℓ *are bounded independently of* M*. Then* $\max_t \|\boldsymbol{h}_t^{(\ell)}\|_{\mathcal{A},\infty} \leq B_\ell$*. For* $q = (\ell, h, j, t)$ *with* $j \in \mathcal{V}(t)$*, the score derivative defined above satisfies*

$$\|D_q\|_{\mathcal{A}} \leq \frac{C_{\ell,h}}{d(t)}, \qquad C_{\ell,h} := 2\|\mathbf{a}\|_1 B_{\ell-1} \|\mathbf{W}_{O,\ell,h}\|_\infty \|\mathbf{V}_{\ell,h}\|_\infty F_\ell \prod_{k=\ell+1}^{L} J_k. \tag{86}$$

Here the empty product is one, and $\|\mathbf{a}\|_1$ *is also assumed bounded independently of* M*. Consequently,* $C_L := \max_{\ell,h} C_{\ell,h}$ *is independent of* M *under these conditions, although it can depend strongly on depth and width. In addition, the derivatives of a fixed layer and head satisfy the joint bound*

$$\sum_{t:d(t)>0} \sum_{j\in\mathcal{V}(t)} \|D_{\ell,h,j,t}\|_{\mathcal{A}} \leq C_{\ell,h}. \tag{87}$$

Thus, over all attention logits $\mathcal{Q}_{\mathrm{all}}$*,*

$$\sum_{q\in\mathcal{Q}_{\mathrm{all}}} \|D_q\|_{\mathcal{A}} \leq C_{\mathrm{attn}}, \qquad C_{\mathrm{attn}} := \sum_{\ell,h} C_{\ell,h} \leq LHC_L. \tag{88}$$

Proof. Apply Eq. (82) to each attention average and projection, obtaining $\|\boldsymbol{u}_t^{(\ell)}\|_{\mathcal{A},\infty} \leq U_\ell$. Equation (84), including the constant Fourier coefficients introduced by the biases, then gives B_ℓ by induction.

For a perturbation $\dot{\mathbf{H}}^{(\ell-1)}$, measure its size by $\max_t \|\dot{\boldsymbol{h}}_t^{(\ell-1)}\|_{\mathcal{A},\infty}$. With the layer logits fixed, the derivative of the attention and residual map has norm at most α_ℓ. The derivative of the FFN and its residual has norm at most F_ℓ: each diagonal activation derivative has coefficient norm at most $\Phi'_\ell(\tau_\ell)$. Thus the whole layer propagates the perturbation with factor at most J_ℓ.

Differentiating a single logit introduces a perturbation only in head h at receiver t of layer ℓ:

$$\frac{\partial \hat{\boldsymbol{z}}_t^{(\ell,h)}}{\partial w_{j,t}^{(\ell,h)}} = \frac{1}{d(t)} \left(\mathbf{V}_{\ell,h} \boldsymbol{h}_j^{(\ell-1)} - \hat{\boldsymbol{z}}_t^{(\ell,h)} \right). \tag{89}$$

Its coefficient norm is at most $2\|\mathbf{V}_{\ell,h}\|_\infty B_{\ell-1}/d(t)$. Apply the head output projection, the current FFN derivative bound F_ℓ, the subsequent layer bounds J_k, and the scalar readout bound $\|\mathbf{a}\|_1$. This gives Eq. (86).

To bound all logits together, we track the sum of backward derivatives over positions, rather than multiplying a per-logit bound by the number of logits. Define $\Delta_t^{(\ell,h)} := \partial f/\partial \hat{\boldsymbol{z}}_t^{(\ell,h)}$ and $\|\mathbf{v}\|_{\mathcal{A},1} := \sum_a \|v_a\|_{\mathcal{A}}$. For a function-valued matrix K, define $\|K\|_{\mathcal{A},\mathrm{row}} := \max_i \sum_j \|K_{ij}\|_{\mathcal{A}}$. Expanding the matrix-vector product gives

$$\|K^\top \boldsymbol{\eta}\|_{\mathcal{A},1} \le \|K\|_{\mathcal{A},\mathrm{row}} \|\boldsymbol{\eta}\|_{\mathcal{A},1}. \tag{90}$$

Viewing all positions as one stacked vector, the attention-and-residual Jacobian has row coefficient norm at most α_ℓ, since each attention row has unit mass. The FFN-and-residual Jacobian is block diagonal across positions. Its block at position t is

$$K_{\ell,F,t} = \rho_{\ell,F} I + \mathbf{W}_{2,\ell}\,\mathrm{diag}\Big[\phi'_\ell\Big(\mathbf{W}_{1,\ell}\boldsymbol{u}_t^{(\ell)} + \boldsymbol{b}_{1,\ell}\Big)\Big]\,\mathbf{W}_{1,\ell}, \qquad \|K_{\ell,F,t}\|_{\mathcal{A},\mathrm{row}} \le F_\ell.$$

The product inequality in Eq. (82) also gives $\|AB\|_{\mathcal{A},\mathrm{row}} \le \|A\|_{\mathcal{A},\mathrm{row}}\|B\|_{\mathcal{A},\mathrm{row}}$, so the whole layer has bound $J_\ell = \alpha_\ell F_\ell$. The readout derivative has total coefficient norm $\|\mathbf{a}\|_1$ and is supported at receiver m. Propagating it to the outputs of head (ℓ,h) gives

$$\sum_t \|\Delta_t^{(\ell,h)}\|_{\mathcal{A},1} \le \|\mathbf{W}_{O,\ell,h}\|_\infty F_\ell \|\mathbf{a}\|_1 \prod_{k=\ell+1}^{L} J_k. \tag{91}$$

The local chain rule and Eq. (89) give

$$\|D_{\ell,h,j,t}\|_{\mathcal{A}} \le 2\sigma_{j,t}^{(\ell,h)} \|\mathbf{V}_{\ell,h}\|_\infty B_{\ell-1} \|\Delta_t^{(\ell,h)}\|_{\mathcal{A},1}. \tag{92}$$

Sum over j, using $\sum_j \sigma_{j,t}^{(\ell,h)} = 1$, and then over t. Equation (91) gives Eq. (87); summing over layers and heads proves Eq. (88). The backward bound controls the sum over receivers; normalized attention controls the sum within each row. Neither step introduces an extra factor of M. □

This argument accounts for all terms of the composed network and their possible cancellations through a bound on absolute coefficients. Multi-head mixing enters through α_ℓ, residual routes through $\rho_{\ell,A}$ and $\rho_{\ell,F}$, and biases through τ_ℓ and B_ℓ. No separate assumption of coherent head contributions is needed. Earlier layers use only nonnegative, normalized, input-independent attention in this proof; uniformity supplies the factor $1/d(t)$ for the differentiated row.

D.3 Average Signal under Uniformly Random Support Locations

For the parity target in Eq. (78), orthogonality identifies the target-correlated score derivative with one Fourier coefficient:

$$G_q(S) := \mathbb{E}_X[\chi_S(X) D_q(X)] = \widehat{D}_q(S). \tag{93}$$

The target-dependent part of the population cross-entropy gradient is $-G_q(S)$. Thus the result below concerns a learning signal rather than only a coefficient of the forward representation.

Theorem 6 (Average initialization signal in a deep causal stack)**.** *Under Eq.* (78) *and Lemma 5, write* $\mathbf{G}(S) := (G_q(S))_{q\in\mathcal{Q}_{\mathrm{all}}}$ *for the target-correlated gradient over all attention logits. Then*

$$\mathbb{E}_S\|\mathbf{G}(S)\|_1 \le \frac{C_{\mathrm{attn}}}{\binom{M}{r}} = O_{L,r}(m^{-r}) \quad \textit{for fixed } L, H, r. \tag{94}$$

This controls all receivers, layers, and heads jointly. For an individual logit $q = (\ell,h,j,m)$ at the final receiver, conditioning on its input position being relevant gives

$$\mathbb{E}_{S|j\in S}|G_q(S)| \le \frac{C_{\ell,h}}{M\binom{M-1}{r-1}} \le \frac{C_L}{M\binom{M-1}{r-1}}. \tag{95}$$

For fixed L, H, r and the stated coefficient bounds, the right-hand side is $O_{L,r}(m^{-r})$, with architectural constants included. More generally, for a logit $q = (\ell,h,j,t)$ at an earlier receiver,

$$\mathbb{E}_{S|j\in S}|G_q(S)| \le \frac{C_{\ell,h}}{d(t)\binom{M-1}{r-1}}. \tag{96}$$

The earlier-receiver statement retains $d(t)$; it is not a uniform m^{-r} conditional bound independent of the receiver.

Proof. The initialized functions D_q are the same for every support choice. Average their target coefficients over the $\binom{M}{r}$ equally likely supports and apply the total coefficient bound:

$$\begin{aligned}\mathbb{E}_S\|\mathbf{G}(S)\|_1 &= \frac{1}{\binom{M}{r}}\sum_{|S|=r}\sum_{q\in\mathcal{Q}_{\mathrm{all}}}|\widehat{D}_q(S)| \\ &\le \frac{1}{\binom{M}{r}}\sum_{q\in\mathcal{Q}_{\mathrm{all}}}\|D_q\|_{\mathcal{A}} \le \frac{C_{\mathrm{attn}}}{\binom{M}{r}}.\end{aligned} \tag{97}$$

The last step uses Eq. (88). For fixed r, $\binom{M}{r}^{-1} \sim r!M^{-r}$ and $M = \Theta(m)$, giving Eq. (94).

Conditional on $j \in S$, there are exactly $\binom{M-1}{r-1}$ equally likely supports. The same argument for one derivative, using Lemma 5, gives

$$\begin{aligned}\mathbb{E}_{S|j\in S}|G_q(S)| &= \frac{1}{\binom{M-1}{r-1}}\sum_{\substack{S\subseteq[M],\ |S|=r\\ j\in S}}|\widehat{D}_q(S)| \\ &\le \frac{\|D_q\|_{\mathcal{A}}}{\binom{M-1}{r-1}} \le \frac{C_{\ell,h}}{d(t)\binom{M-1}{r-1}}.\end{aligned} \tag{98}$$

Set $t = m$ to obtain Eq. (95). For fixed r, $[M\binom{M-1}{r-1}]^{-1} \sim (r-1)!M^{-r}$ and $M = \Theta(m)$, proving the stated asymptotic upper bound. □

The conditional expectation measures the gradient at a position known to belong to the support. Averaging this bound over relevant positions gives another useful interpretation. For a fixed layer and head, define their mean absolute signal by

$$A_{\ell,h}(S) := \frac{1}{r}\sum_{j\in S}|G_{\ell,h,j,m}(S)|. \tag{99}$$

Using $\Pr(j \in S) = r/M$ and Eq. (95),

$$\mathbb{E}_S A_{\ell,h}(S) \le \frac{C_L}{M\binom{M-1}{r-1}}. \tag{100}$$

For $0 < \delta < 1$, Markov's inequality applied to Eq. (94) gives, with probability at least $1-\delta$ over S,

$$\max_{q\in\mathcal{Q}_{\mathrm{all}}}|G_q(S)| \le \|\mathbf{G}(S)\|_1 \le \frac{C_{\mathrm{attn}}}{\delta\binom{M}{r}}. \tag{101}$$

For fixed L, H, r, δ, this is an $O(m^{-r})$ upper bound simultaneously for all attention logits, including early receivers. This averages over S without conditioning on a particular early position being relevant. Its failure probability concerns support locations, not finite-sample noise. If r, depth, or δ^{-1} grows with m, their dependence must be retained in the displayed formulas.

Why early causal features do not invalidate this average bound. Under strict-prefix attention, a quadratic FFN at position 3 can form $((x_1 + x_2)/2)^2 = (1 + x_1x_2)/2$. A later read can therefore carry a two-variable interaction with only one final-position averaging factor. Such fixed layouts need not satisfy a pointwise m^{-2} bound. Under Eq. (78), however,

$$\Pr(S \subseteq [k]) = \frac{\binom{k}{r}}{\binom{M}{r}}, \qquad r \le k \le M. \tag{102}$$

The probability suppresses the contribution of supports confined to a fixed early prefix. Theorem 6 controls all support layouts together through the coefficient bound, without requiring a separate path count for each layout.

D.4 Cross-Entropy Gradients and Finite-Sample Fluctuations

Fix the sampled support S and draw n independent examples of X. With binary logits $(f, -f)$,

$$\mathcal{L}_S(X) = \log(2\cosh f(X)) - \chi_S(X) f(X),$$
$$g_q(X;S) := \frac{\partial \mathcal{L}_S}{\partial w_q} = \big(\tanh f(X) - \chi_S(X)\big) D_q(X). \tag{103}$$

Define the prediction contribution $R_q := \mathbb{E}_X[\tanh f(X) D_q(X)]$, which is independent of S under the random-support experiment. The empirical gradient is

$$\widehat{g}_q = R_q - G_q(S) + \xi_q. \tag{104}$$

Here $\xi_q := \widehat{g}_q - \mathbb{E}_X[g_q(X;S)]$ is the deviation of the sample average from its population value. The decomposition separates the prediction contribution, the target signal, and finite-sample fluctuations. The prediction-centering condition in Eq. (48), when verified, sets $R_q = 0$. The coefficient argument bounds $G_q(S)$ without assuming this centering; it does not establish centering itself.

Since $|D_q(X)| \le \|D_q\|_{\mathcal{A}}$ and $|\tanh f - \chi_S| \le 2$, Lemma 5 gives $|g_q(X;S)| \le 2C_L/d(t)$. For any finite set $\mathcal{Q}$ of analyzed logits, put $N = |\mathcal{Q}|$. For $0 < p < 1$, Hoeffding's inequality and a union bound yield, conditional on every fixed S, with probability at least $1 - p$ over the n examples,

$$|\xi_q| \le \frac{2C_L}{d(t)} \sqrt{\frac{2\log(2N/p)}{n}} \qquad \text{for all } q = (\ell, h, j, t) \in \mathcal{Q}. \tag{105}$$

For all final-receiver logits, $N = LHM$ and $d(t) = M$, recovering an $O(M^{-1}\sqrt{\log(M/p)/n})$ fluctuation bound at fixed architecture. The support event in Eq. (101) and this sampling event hold jointly with probability at least $1 - \delta - p$; independence between the two events is unnecessary.

An upper bound on $\mathbb{E}_S|G_q(S)|$ does not provide a nonzero signal for every task, nor a lower bound on stochastic gradient noise. Therefore it does not establish either a universal training lower bound or a sufficient sample size depending only on r. If prediction centering holds and a particular support additionally satisfies $|G_q(S)| \ge \mu_q > 0$, then

$$n \ge \frac{32 C_L^2}{d(t)^2 \mu_q^2} \log \frac{2N}{p} \tag{106}$$

is sufficient to make the sampling error at most $\mu_q/2$. At the final receiver, a separately established $\mu_q = \Theta(m^{-r})$ recovers the single-layer sample scale $m^{2(r-1)}$ up to logarithms. The random-support upper bound alone does not supply that additional signal lower bound.

D.5 Scope of the Multi-Layer Extension

The main theorem and this extension have complementary scopes. The single-layer result gives a support-selective decomposition for a fixed support under its stated centering and non-degeneracy conditions. The multi-layer result bounds the expected ℓ_1 norm of the target-correlated gradient over all attention logits by $C_{\mathrm{attn}}/\binom{M}{r}$, and the conditional mean absolute signal at a relevant final-receiver logit by $C_L/[M\binom{M-1}{r-1}]$. Both are $O(m^{-r})$ for fixed r, under uniformly random support locations and explicit coefficient control. Depth, multiple heads, residuals, and biases can change the constants, but do not remove this average context-length suppression within the analyzed model class.

The exact binomial expression is retained when r varies; the asymptotic notation does not hide a constant uniform in r. The result permits exceptional supports with larger signals and signals that vanish by cancellation. It does not assert a $\Theta(m^{-r})$ lower bound, a guarantee for every fixed support, a bound on shared query/key parameter gradients, or a description of training after attention departs from the reference initialization.

E Discussion on General Natural Language Reasoning Tasks

The parity analysis is a clean setting in which high-order dependencies and zero-mean irrelevant tuples can be computed exactly. Natural language reasoning is less discrete: a relevant "variable" may

be a token, an entity, a number, a relation, or a short proposition. This section states the conditions under which the same gradient hierarchy applies: irreducible semantic support, the correlation bounds below, and the prediction-centering condition in Eq. (48). Semantic irreducibility alone does not imply prediction centering.

From parity variables to semantic units. Let

$$X = (x_1, \dots, x_m), \qquad y \in \mathcal{Y}, \tag{107}$$

denote a natural-language context and its target. In the token-level analysis, each position t has embedding $\boldsymbol{e}_t$. For natural language, we allow a reasoning unit to be a bounded-length span rather than a single token. Let

$$U_j \subseteq [m] \tag{108}$$

denote the j-th semantic unit. If U_j is a single token, then $U_j = \{j\}$. If it is a contiguous span, write

$$U_j = \{a_j, a_j + 1, \dots, b_j\}, \qquad |U_j| = b_j - a_j + 1, \qquad |U_j| \le s_{\max} = O(1). \tag{109}$$

Let $\boldsymbol{e}_{U_j}$ be a centered bounded representation of this span, for example

$$\boldsymbol{e}_{U_j} = \mathrm{Pool}\left(\boldsymbol{e}_{a_j}, \dots, \boldsymbol{e}_{b_j}\right) - \mathbb{E}\left[\mathrm{Pool}\left(\boldsymbol{e}_{a_j}, \dots, \boldsymbol{e}_{b_j}\right)\right], \qquad \|\boldsymbol{e}_{U_j}\| \le B. \tag{110}$$

The original token-level setting is recovered by taking $U_j = \{j\}$ and $\boldsymbol{e}_{U_j} = \boldsymbol{e}_j$.

This span abstraction does not change the attention scaling. Attention is normalized over token positions, not over semantic units. Near the symmetric initialization,

$$\sum_{\alpha \in U_j} \sigma_{\alpha,m} = |U_j| \Theta(m^{-1}) = \Theta(m^{-1}), \tag{111}$$

because $|U_j| = O(1)$. Therefore introducing r bounded semantic units still contributes r attention-averaging factors:

$$\prod_{q=1}^{r} \left(\sum_{\alpha \in U_{j_q}} \sigma_{\alpha,m} \right) = \Theta(m^{-r}). \tag{112}$$

Irreducible semantic dependencies. We say that a natural-language task contains an irreducible semantic order-r dependency if there exists a support set of semantic units

$$\mathcal{S}^{(r)} = \{U_{j_1}, \dots, U_{j_r}\} \tag{113}$$

whose joint representation has a stable target correlation, whereas every proper subset has zero or negligible target correlation. Define the joint representation by a tensor product

$$\boldsymbol{e}_{\mathcal{S}^{(r)}} = \boldsymbol{e}_{U_{j_1}} \otimes \cdots \otimes \boldsymbol{e}_{U_{j_r}}, \tag{114}$$

or, equivalently for coordinatewise FFN expansions, by a Hadamard product. Let $\mathcal{C}_y$ denote a bounded contraction tensor from the target-logit derivative, absorbing the target readout direction and fixed projection factors. The sample-dependent prediction term is controlled separately by Eq. (48). The relevant support signal is

$$Z_{\mathcal{S}^{(r)}}(X, y) = \left\langle \mathcal{C}_y, \boldsymbol{e}_{\mathcal{S}^{(r)}} \right\rangle. \tag{115}$$

The irreducibility condition is

$$\left| \mathbb{E}\left[Z_{\mathcal{S}^{(r)}}(X, y) \right] \right| = \left| C^{(r)}_{\mathrm{signal}} \right| \ge c_r > 0, \tag{116}$$

while for every proper subset $T \subsetneq \mathcal{S}^{(r)}$,

$$\left| \mathbb{E}\left[\langle \mathcal{C}_y, \boldsymbol{e}_T \rangle \right] \right| \le \varepsilon_{\mathrm{shortcut}}. \tag{117}$$

Here $\boldsymbol{e}_T$ denotes the joint representation of the semantic units in T. The term $\varepsilon_{\mathrm{shortcut}}$ accounts for weak low-order semantic shortcuts in real corpora. In a strictly irreducible task, $\varepsilon_{\mathrm{shortcut}} = 0$; in natural language, we only require it to be smaller than the order-r signal scale.

Relevant supports provide non-zero population signal. For a semantic tuple

$$J = (U_{j_1}, \dots, U_{j_r}), \tag{118}$$

define

$$Z_J(X, y) = \left\langle \mathcal{C}_y, \boldsymbol{e}_{U_{j_1}} \otimes \cdots \otimes \boldsymbol{e}_{U_{j_r}} \right\rangle. \tag{119}$$

If J covers the true support $\mathcal{S}^{(r)}$, then Eq. (116) gives

$$\mathbb{E}\left[Z_J(X, y)\right] = C^{(r)}_{\text{signal}}, \qquad |C^{(r)}_{\text{signal}}| \geq c_r. \tag{120}$$

Consequently, the Taylor-expanded attention gradient contains a support-selective population component of the same form as in Theorem 1:

$$\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial w_{j,m}}\right]_{\text{signal}} = -\Theta(m^{-r})\mathbb{I}^{(r)}_j C^{(r)}_{\text{signal}}. \tag{121}$$

Here j may index either a token or a semantic unit. If U_j is a span, one may differentiate an anchor token inside the span or aggregate the attention mass over the span; since $|U_j| = O(1)$, this changes only constants. The key point is that the non-zero signal comes from the population-level irreducible semantic correlation, not from assuming that attention has already aligned to the support.

Irrelevant semantic tuples concentrate as noise. Let $\mathcal{I}_{r,m}$ denote the set of irrelevant order-r semantic tuples. For $J \in \mathcal{I}_{r,m}$, define $Z_J(X, y)$ as above. We additionally assume their population correlations are zero in the strict case and bounded by the shortcut level in the approximate case:

$$|\mathbb{E}\left[Z_J(X, y)\right]| \leq \varepsilon_{\text{shortcut}}. \tag{122}$$

Because semantic-unit representations and contraction tensors are bounded, there exists a constant B_r such that

$$|Z_J(X, y)| \leq B_r. \tag{123}$$

For n independent samples, let

$$\bar{Z}_J = \frac{1}{n}\sum_{s=1}^{n} Z_J(X^{(s)}, y^{(s)}). \tag{124}$$

Hoeffding's inequality yields

$$\Pr\left(\left|\bar{Z}_J - \mathbb{E}[Z_J]\right| \geq \kappa\right) \leq 2\exp\left(-\frac{2n\kappa^2}{B_r^2}\right). \tag{125}$$

A union bound over $\mathcal{I}_{r,m}$ gives, with probability at least $1 - p$,

$$\kappa = O\left(B_r\sqrt{\frac{\log(|\mathcal{I}_{r,m}|/p)}{n}}\right). \tag{126}$$

Thus irrelevant semantic tuples contribute at most

$$O\left(\frac{\kappa + \varepsilon_{\text{shortcut}}}{m}\right) \tag{127}$$

to the attention-gradient background. If $\varepsilon_{\text{shortcut}}$ is large, the task has a strong low-order shortcut and its effective interaction order is smaller; if it is small, the order-r hierarchy remains visible.

Zipf frequencies and effective sample size. Zipfian token frequencies do not invalidate the bounded concentration argument above: the Hoeffding bound depends on the range of Z_J, not directly on the marginal frequency of a token. However, frequency affects how often a relevant semantic configuration appears. Let

$$p_{\mathcal{S}^{(r)}} \tag{128}$$

denote the probability that the support configuration $\mathcal{S}^{(r)}$ appears in the required semantic form. The effective number of samples that expose this configuration is

$$n_{\text{effective}}\left(\mathcal{S}^{(r)}\right) = n \cdot p_{\mathcal{S}^{(r)}}. \tag{129}$$

The signal-noise separation requires

$$\Theta(m^{-r})|C^{(r)}_{\text{signal}}| \gg O\left(\frac{\kappa + \varepsilon_{\text{shortcut}}}{m}\right). \tag{130}$$

Ignoring constants and the shortcut term, this requires

$$\kappa \ll m^{-(r-1)}|C^{(r)}_{\text{signal}}|. \tag{131}$$

Using the effective sample size in Eq. (126), we obtain

$$n_{\text{effective}}\left(\mathcal{S}^{(r)}\right) = \Omega\left(\frac{B_r^2}{\left(C^{(r)}_{\text{signal}}\right)^2} m^{2(r-1)} \log\frac{|\mathcal{I}_{r,m}|}{p}\right). \tag{132}$$

Equivalently, in terms of the total sample size,

$$n = \Omega\left(\frac{B_r^2}{\left(C^{(r)}_{\text{signal}}\right)^2 p_{\mathcal{S}^{(r)}}} m^{2(r-1)} \log\frac{|\mathcal{I}_{r,m}|}{p}\right). \tag{133}$$

Thus Zipfian frequency does not change the exponent in m, but rare high-order semantic supports require more total samples because $p_{\mathcal{S}^{(r)}}$ is small.

Proposition E.1 (Natural-Language Irreducible Dependencies)**.** *Suppose a natural-language task contains an irreducible semantic support $\mathcal{S}^{(r)}$ satisfying Eq.* (116)*, irrelevant semantic tuples satisfy Eq.* (117) *with bounded contractions $|Z_J| \le B_r$, and the prediction-centering condition in Eq.* (48) *holds. Then, with probability at least $1 - p$, the attention-gradient decomposition contains a support-selective term of size*

$$-\Theta(m^{-r})\mathbb{I}^{(r)}_j C^{(r)}_{\text{signal}}, \tag{134}$$

while irrelevant semantic tuples contribute at most

$$O\left(\frac{\kappa + \varepsilon_{\text{shortcut}}}{m}\right), \qquad \kappa = O\left(B_r\sqrt{\frac{\log(|\mathcal{I}_{r,m}|/p)}{n}}\right). \tag{135}$$

Consequently, when the shortcut level is dominated by the order-r signal, the same gradient hierarchy and the sample-complexity scaling $n = \Omega(m^{2(r-1)})$ hold up to bounded semantic constants and effective-sample-size factors.

Examples. The distinction is easiest to see from two simple cases. In a mathematical word problem, the model may need to identify two quantity spans and the operation that relates them. Denote these semantic units by U_a, U_b, and U_{op}. The answer is determined only after the three pieces are combined:

$$\mathcal{S}^{(3)} = \{U_a, U_b, U_{\text{op}}\}, \qquad y = \text{op}(a, b). \tag{136}$$

If the model sees only U_a, only U_b, or only U_{op}, it cannot determine the answer. Even two of the three pieces are insufficient: the quantities without the operation are ambiguous, and the operation without both quantities is incomplete. Thus the useful gradient signal appears only when the interaction term contains all three semantic units. Since each unit must first be brought in through attention, the leading support-selective signal contains three attention-averaging factors and scales as $\Theta(m^{-3})$. When intermediate reasoning steps are omitted, the final answer may depend on more raw semantic units, increasing the effective order from 3 to a larger r and attenuating the signal as $\Theta(m^{-r})$.

Commonsense questions often behave differently. In "It is raining. What should one bring?", the answer "umbrella" is already strongly associated with the single semantic cue "raining". A written explanation might mention intermediate notions such as getting wet or staying dry, but the answer can often be predicted from one dominant cue. Such a task has a low effective order and therefore does not suffer the same high-order signal attenuation. This explains why implicit reasoning compression can be relatively harmless for tasks with stable low-order shortcuts, while it is much harder for mathematical or logical tasks whose answers depend on irreducible combinations of several semantic units.

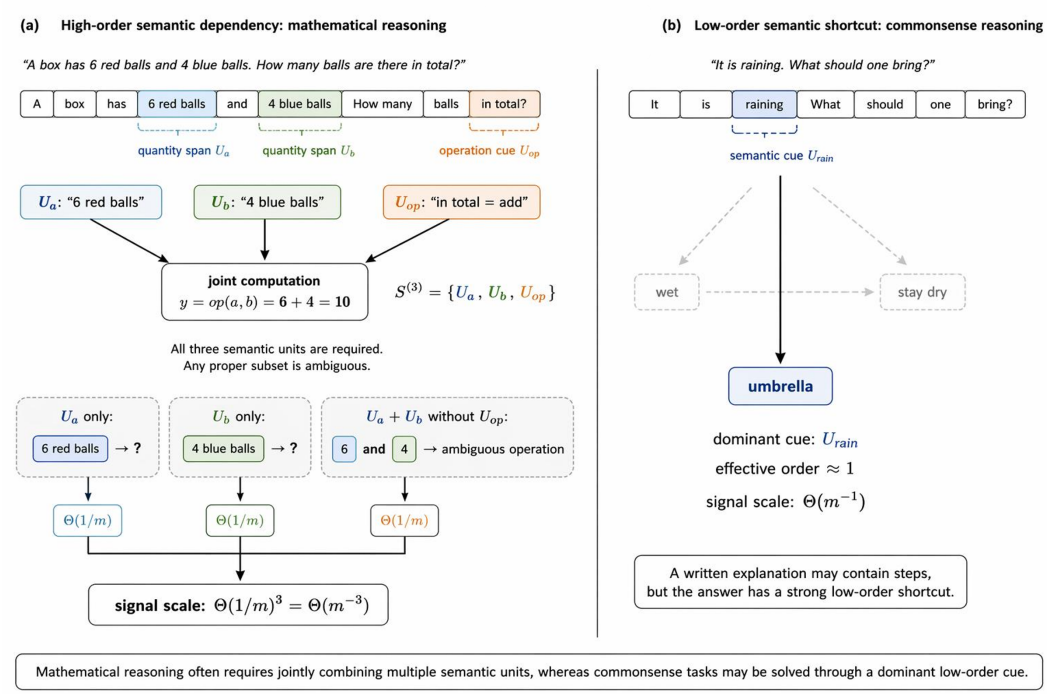


Figure 9: **Irreducible semantic support versus low-order shortcut.** Left: a mathematical word problem requires jointly combining two quantity spans U_a, U_b and an operation cue U_{op}, so the relevant support is $\mathcal{S}^{(3)} = \{U_a, U_b, U_{\mathrm{op}}\}$ and the leading support-selective signal scales as $\Theta(m^{-3})$. Right: a commonsense question can often be answered from one dominant cue, such as U_{rain}, yielding a much lower effective interaction order.

Conclusion. The parity problem instantiates a broader mechanism governed by semantic interaction order. Under the stated correlation and prediction-centering conditions, natural-language tasks with low-order shortcuts have smaller effective order, whereas irreducible order-r dependencies exhibit the same $\Theta(m^{-r})$ gradient attenuation and signal-separation sample scale, up to bounded semantic constants, shortcut bias, and effective sample size.

F Data Sample

Table 4 presents a complete sample from the **NatBool-DAG** dataset within the *Cyber Warfare* theme. Its recorded depth is **3**, counting the input layer and two layers of logical operations. The target depends on the XOR of four distinct observed states, computed through three XOR operations. The box below reproduces the input and the step-by-step derivation used for supervision.

G Implementation Details

Hardware and Code. All experiments are conducted on a server with 4 NVIDIA RTX 4090 GPUs. We implement Imp. Base-1 ourselves using independent learnable latent-token parameters, and adapt the official Coconut open-source implementation[2] as Imp. Base-2. All code will be released after publication.

Existing Assets and Licenses. We use publicly available assets under their stated terms: GPT-2 (Radford et al., 2019) under the Modified MIT License, Llama-3.2 (Meta AI, 2024) under the Llama 3.2 Community License, Qwen3 (Yang et al., 2025) under Apache-2.0, GSM8K (Cobbe et al., 2021) under the MIT License, and the Coconut codebase (Hao et al., 2024) under the MIT License.

[2] https://github.com/facebookresearch/coconut

Table 4: NatBool-DAG Sample Instance (Theme: Cyber Warfare, Depth: 3)

NatBool-DAG Sample ID: 1e32b423-2769-442d-b6df-92822c0843a5

[METADATA]
Theme: Cyber Warfare | **Reasoning Steps:** 2 | **Total Nodes:** 10

[INSTRUCTION]
Based on the observed data and system rules, determine the status of 'Root Privilege'. Is it True or False?

[INPUT CONTEXT]
=== SCENARIO: Cyber Warfare ===

— OBSERVED DATA —
[SYSTEM LOG] Service 'Proxy Server Beta' is OFFLINE/SILENT.
[SYSTEM LOG] Service 'Secure SSH Key' is OFFLINE/SILENT.
[SYSTEM LOG] Service 'Encrypted SSH Key' is OFFLINE/SILENT.
[SYSTEM LOG] Service 'Encrypted Proxy Server' is OFFLINE/SILENT.
[SYSTEM LOG] Service 'Ghost Admin Token' is currently ONLINE/DETECTED.
[SYSTEM LOG] Service 'Redundant Biometric Scanner' is currently ONLINE/DETECTED.

— SYSTEM RULES —
- Conflict Detection: 'Root Privilege' flag is raised when 'Encrypted Kernel Process' and 'Secure Packet Header' have conflicting states.
- Conflict Detection: 'Encrypted Kernel Process' flag is raised when 'Proxy Server Beta' and 'Ghost Admin Token' have conflicting states.
- Conflict Detection: 'Secure Packet Header' flag is raised when 'Secure SSH Key' and 'Encrypted Proxy Server' have conflicting states.
- Kill Switch: 'Encrypted Memory Buffer' remains stable unless both 'Encrypted Proxy Server' and 'Secure SSH Key' fail simultaneously.
- Conflict Detection: 'Memory Buffer Cluster-A' flag is raised when 'Encrypted SSH Key' and 'Secure SSH Key' have conflicting states.

[GROUND TRUTH CHAIN-OF-THOUGHT]
1. Check status of 'Proxy Server Beta': The logs indicate it is False.
1. Check status of 'Ghost Admin Token': The logs indicate it is True.
Analyzing 'Encrypted Kernel Process':
    - Requires: Proxy Server Beta (False) XOR Ghost Admin Token (True)
    - Logic: XOR logic evaluates to True.
    - Conclusion: 'Encrypted Kernel Process' is True.
1. Check status of 'Secure SSH Key': The logs indicate it is False.
1. Check status of 'Encrypted Proxy Server': The logs indicate it is False.
Analyzing 'Secure Packet Header':
    - Requires: Secure SSH Key (False) XOR Encrypted Proxy Server (False)
    - Logic: XOR logic evaluates to False.
    - Conclusion: 'Secure Packet Header' is False.
Analyzing 'Root Privilege':
    - Requires: Encrypted Kernel Process (True) XOR Secure Packet Header (False)
    - Logic: XOR logic evaluates to True.
    - Conclusion: 'Root Privilege' is True.

Final Answer: True

[OUTPUT]
True

We use these assets only for research experiments and comply with their stated terms. NatBool-DAG is a synthetic benchmark generated by us and will be released with the code after publication.

Training Setup. Unless otherwise stated, all methods use the same training hyperparameters: $\texttt{EPOCHS} = 10$ and $\texttt{LR} = 1 \times 10^{-5}$. We use FP16 mixed-precision optimization with maximum sequence length 2048, per-device batch size 1, and 8 gradient accumulation steps. We perform full-parameter fine-tuning for models up to the 4B scale; for the Qwen3-8B experiments, we use LoRA fine-tuning under the same epoch and learning-rate setting due to memory constraints.

Method Hyperparameters. For our proposed ALiCoT, the latent tokens are initialized as independent learnable parameters. We set the sequence length of these latent tokens to 20. The coefficient for the alignment loss (λ in Eq. 6) is set to 0.5, balancing the primary generation task with the semantic alignment of the latent space.

H Main Experimental Results

The aggregate accuracy table has been moved to the main text as Table 2. Here we provide additional interpretation of the same full-compression setting. We evaluate ALiCoT across diverse model architectures and scales, including GPT2, Llama-3.2, and Qwen3. For all evaluated methods, the entire chain of intermediate reasoning steps is suppressed to test the limit of implicit generation. The comparison primarily focuses on Imp. Base-1; Imp. Base-2 is reported only where evaluated due to computational constraints. All methods use identical training datasets and iterations for a controlled comparison.

Superior Accuracy under Extreme Compression. ALiCoT consistently outperforms the baselines across all model architectures and scales, demonstrating remarkable resilience to the extreme compression of intermediate reasoning steps. While *Imp. Base-1* struggles to maintain accuracy—particularly evident in the Qwen3-0.6B and Llama-3.2-3B settings where it lags significantly—ALiCoT achieves substantial gains. For instance, on Qwen3 (4B), ALiCoT improves the average accuracy by over **17%** compared to Imp. Base-1 (95.01

Robustness in Long-Reasoning. ALiCoT exhibits superior stability as the reasoning depth increases (from 3 to 10 hops). In the challenging 10-hop scenarios, where baseline performance typically degrades due to error accumulation in the implicit state, ALiCoT maintains high proficiency. Notably, in the Llama-3.2 (1B) setting, although the baseline performs competitively on shorter hops, ALiCoT surpasses it on longer chains (e.g., +10.17% on Hop 10), indicating that our method effectively internalizes complex multi-step logic without suffering from "reasoning collapse."

Impact of Latent Noise and Context Length. As demonstrated in Section C, autoregressive latent tokens (Imp. Base-2) introduce significantly higher noise levels compared to independent learnable parameters (Imp. Base-1 and ALiCoT). Given identical latent lengths, we observe divergent behaviors driven by this noise:

Imp. Base-1 (Low Noise): Due to the minimal interference from independent initialization, the model is not easily distracted. Consequently, it follows the standard complexity trend where performance is higher on simpler tasks (Short Hops) and naturally decays as reasoning complexity increases.

Imp. Base-2 (High Noise): The autoregressive noise creates a dependency on input context length for robustness.

Vulnerability in Short Contexts: In simple tasks (e.g., Hop 3), the input context is sparse. Lacking sufficient information to "anchor" its attention, the model becomes highly susceptible to the noise generated by the latent tokens, leading to performance collapse (33.48%). *Robustness in Long Contexts:* Conversely, as the reasoning chain lengthens (e.g., Hop 10), the richer context provides stronger supervision signals. This extended information helps the model filter out latent noise, resulting in a counter-intuitive performance recovery (53.80%). This contrast highlights that ALiCoT's independent initialization is crucial for stabilizing inference, especially for tasks with varying context lengths.