TEST-TIME SYMMETRIZATION: SCALING LAWS FOR MODEL-AGNOSTIC LEARNING UNDER STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time compute has emerged as a powerful way to improve the accuracy of language models at inference time, with scaling laws quantifying how accuracy improves with additional forward passes. Existing scaling laws, however, focus on repeated sampling from a fixed prompt, even though many tasks admit multiple equivalent representations of the same underlying problem. A graph does not depend on how its nodes are numbered, and a program does not depend on the names of its variables, yet language models can change their answers under such rewritings. In this paper, we study *test-time symmetrization*, which uses forward passes on rewritings of the input and aggregates the resulting answers over the group G of rewritings. We prove that no amount of sampling and repeated decoding from a single representation can replace rewriting, while only $O(\log |G|)$ random rewritings suffice to recover the information provided by all $|G|$ rewritings. The same set of $O(\log |G|)$ rewritings provides simultaneous approximation across models, decoding rules, and sampled reasoning traces, and the logarithmic scaling is optimal in general. Finally, exact computations confirm the rate and its logarithmic dependence on $|G|$, and graph-reasoning experiments with a frozen open-weight model (Qwen2.5-1.5B) show that a single set of rewritings attains it across six tasks and six decoding rules. Our results provide a principled way to enforce symmetries at inference time when they cannot be built into the model, which we call model-agnostic symmetrization.

1 INTRODUCTION

Test-time compute is now a standard way to improve a language model whose weights are fixed. A longer reasoning trace (Wei et al., 2022), a majority vote over several traces (Wang et al., 2023c), or a verifier that selects among candidates (Cobbe et al., 2021; Lightman et al., 2024) each convert extra forward passes into accuracy, and scaling laws predict the return on each pass (Brown et al., 2024; Snell et al., 2025). All of these methods query the model again on the same prompt: the randomness they exploit is that of *decoding*.

A second source of variation is *rewriting*. Relabeling the nodes of a graph, permuting the options of a multiple-choice question, renaming the variables of a program, or reordering in-context examples changes the prompt but not its meaning. These rewritings move the input along an *orbit* of equivalent representations, and a forward pass spent on a rewriting asks the same question in a different form. The two sources are not interchangeable. Repeated decoding explores the model at one representation and reveals nothing about its behavior elsewhere on the orbit (Figure 1).

Indeed, frozen models are far from constant along the orbit. Reordering answer options or few-shot examples changes benchmark scores and model rankings (Zheng et al., 2024; Lu et al., 2022; Sclar et al., 2024; Mizrahi et al., 2024), and the same instability appears for graph serializations and identifier names (Herbst et al., 2025; Miceli Barone et al., 2023). Querying several rewritings and combining the answers helps: permutation-based calibration reduces option-order bias (Zheng et al., 2024), prompt ensembles improve on single prompts (Jiang et al., 2020; Arora et al., 2023), and at matched compute, paraphrasing the input can beat sampling more traces (Kozodoi et al., 2026). These remedies are empirical, and three basic questions remain open: can more decoding replace rewriting, how many rewritings are enough, and how should rewriting be combined with decoding?

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.

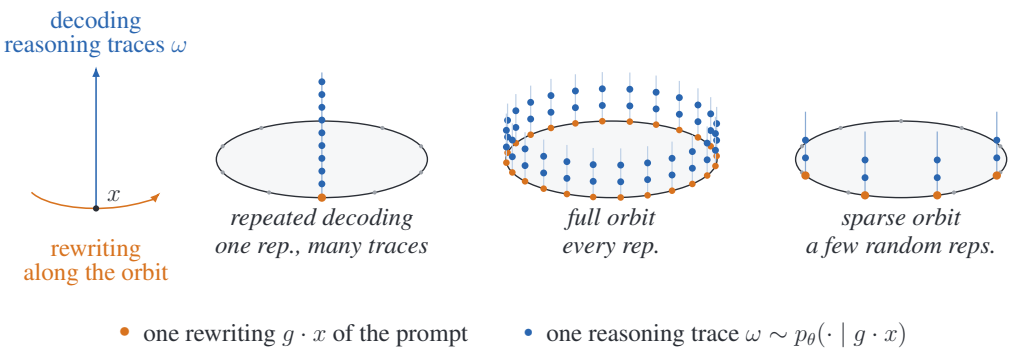


Figure 1: Rewriting by G explores the orbit of equivalent representations, while the frozen model p_θ samples reasoning traces at each representation. Full orbit averaging uses $|G|$ rewrites; test-time symmetrization uses only $O(\log |G|)$ random rewrites with a few traces each.

Kozodoi et al. (2026) point to the last one, the balance between input-side and output-side diversity, as a direction for future work. This paper takes up these questions.

Rewriting is necessary. We construct two models whose responses agree on every trace at the original prompt but lead to different decisions once averaged over the orbit. No decoding procedure, adaptive or not, can tell them apart, however many traces it draws (Theorem 1).

Logarithmically many rewritings suffice. Note that a full pass over the group is out of reach: a graph on 20 nodes has $20! \approx 2 \times 10^{18}$ relabelings. We prove that a random set of rewritings whose size grows with $\log |G|$ approximates the full group average in mean square over starting representations, and that the logarithm cannot be improved in general (Theorem 2). The set is drawn once and serves every frozen model at the same time. This uniformity is what makes the guarantee *model-agnostic*. Indeed, even though equivariant architectures build symmetry into the network during training, a frozen model that allows no such modification can still exploit task symmetries imposed externally through pre- and post-processing, at a logarithmic cost.

Any decoding rule can be used. The same set of rewritings also serves every decoding rule. The traces at one rewriting may be combined by majority vote, expected logits, verifier weighting, or any other rule, at any generation temperature, and the result is again a function on the group. The scaling law holds for all such functions at once, including every realized randomized decoder, so the decoding rule can be chosen after the rewritings are drawn (Corollary 3). An error decomposition separates the accuracy of the fully symmetrized model from the cost of approximating it (Corollary 4), and the number of rewritings and the number of traces per rewriting can be set independently (Section C).

In short, in this paper, we make the following contributions:

- We formulate test-time symmetrization, in which decoding and rewriting are two separate sources of test-time randomness, and prove that decoding alone cannot replace rewriting (Theorem 1).
- We prove a tight scaling law: $O(\log |G|)$ random rewritings approximate the full group average, uniformly over models, decoding rules, and realized decoders (Theorem 2 and Corollary 3).
- We verify the law exactly on symmetric and hypercube groups and on graph reasoning with a frozen language model (Qwen2.5-1.5B), where rewriting improves accuracy beyond any decoding at one representation on the tasks where the fully symmetrized model is more accurate (Section 5).

2 Related Work

Test-time compute. Spending additional forward passes at inference is a standard way to improve a frozen language model: lengthening a single reasoning trace (Wei et al., 2022; Muennighoff et al., 2025), sampling several traces and voting (Wang et al., 2023c), reranking them with a verifier (Cobbe

et al., 2021; Lightman et al., 2024), or searching over intermediate steps (Yao et al., 2023; Snell et al., 2025). Empirical scaling laws for these procedures relate coverage or accuracy to the number of samples (Brown et al., 2024; Wu et al., 2025; Chen et al., 2024a; Li et al., 2024), and several works explain their functional form (Schaeffer et al., 2025; Levi, 2025; Liu et al., 2025) or their limitations under imperfect verification (Stroebl et al., 2026); see Welleck et al. (2024) for a survey and Kaplan et al. (2020); Hoffmann et al. (2022) for the training-time counterparts. These procedures query the model using a fixed representation of the prompt, leaving the role of rewritings comparatively less studied. Our results show that this is a limitation in principle and that rewriting obeys its own scaling law, distinct from that of the decoding stage.

Sensitivity to rewritings. It is well documented that language models are not invariant to semantically neutral rewritings: the order of in-context examples (Lu et al., 2022; Zhao et al., 2021), prompt formatting (Sclar et al., 2024; Mizrahi et al., 2024), the order of multiple-choice options (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Zong et al., 2024; Gupta et al., 2024), the order of premises (Chen et al., 2024b), the names and numbers in a math problem (Mirzadeh et al., 2025), the position of information in a long context (Liu et al., 2024), the serialization of a graph (Wang et al., 2023a; Fatemi et al., 2024; Herbst et al., 2025), and identifier renaming in code (Miceli Barone et al., 2023; Wang et al., 2023b) can all change the model's prediction; consistency across paraphrases has also been studied on its own (Elazar et al., 2021). Existing remedies use small collections of rewritings to estimate and remove option-position bias (Zheng et al., 2024), aggregate predictions from several prompts (Jiang et al., 2020; Schick & Schütze, 2021; Arora et al., 2023), modify input representations to make outputs order-independent (McIlroy-Young et al., 2024), or train for robustness to permutations (Chen et al., 2025). Closest to our setting, Kozodoi et al. (2026) show empirically that allocating forward passes to input paraphrases can outperform self-consistency at matched compute, motivating the allocation problem between input-side and output-side diversity. Our framework separates this decoding choice from the cost of enforcing symmetry. Related work on model-agnostic symmetrization and on the theory of invariance is discussed in Section A.

3 Test-Time Symmetrization

Rewriting orbits and decoding fibers. Let $\mathcal{X}$ denote the space of prompts and let $\mathcal{C}$ denote the set of answer classes, with $C = |\mathcal{C}|$. The target function $y^\star : \mathcal{X} \to \mathcal{C}$ assigns each prompt its correct answer. Let a finite group G act on $\mathcal{X}$.[1] For a fixed semantic instance $x \in \mathcal{X}$, its *orbit* $G \cdot x = \{g \cdot x : g \in G\}$ contains all rewritings generated by the group action. A frozen model queried on $g \cdot x$ defines a conditional distribution $p_{\theta,\tau}(\cdot \mid g \cdot x)$ over reasoning traces. We write

$$\omega \sim p_{\theta,\tau}(\cdot \mid g \cdot x), \qquad V_x(g, \omega) \in \mathcal{Z}, \tag{1}$$

where τ is the token generation temperature, ω is a sampled trace, and $V_x(g, \omega)$ is its response representation in a measurable space $\mathcal{Z}$. It may contain an answer, logits, a verifier score, or the entire trace. Without loss of generality, we assume that the task is invariant, $y^\star(g \cdot x) = y^\star(x)$. If G acts equivariantly on the answer coordinates, we first align them using the inverse output action; the same results then apply. Task invariance does not require $V_x(g, \omega)$ to be invariant. Its variation across g is precisely what test time symmetrization aims to remove.

We use *rewriting* and *serialization* for the same representations, with a small distinction in emphasis: g denotes the rewriting operation, while $g \cdot x$ is the resulting serialization of the prompt. The response field has two distinct coordinates. The group element g selects a representation of the prompt, while ω selects a decoder outcome from $p_{\theta,\tau}(\cdot \mid g \cdot x)$. We call the conditional probability space

$$\big(\Omega, p_{\theta,\tau}(\cdot \mid g \cdot x)\big)$$

the *decoding fiber* at $g \cdot x$. Thus, an orbit contains the equivalent representations of one semantic instance, and each representation has its own conditional distribution over decoder outcomes. We refer to V_x as the *response field* over these two coordinates. Formal definitions of the group action, the conditional decoder model, and the response field are collected in Section B.

[1] Most rewritings of interest are exact group actions once they are applied to the structured object behind the prompt (node labels, answer options, in-context examples, or identifiers) and the result is serialized in a fixed way, as in Section H.1. Theorem 2 and Corollary 3 use only functions on G, so they apply to any family of rewritings indexed by G.

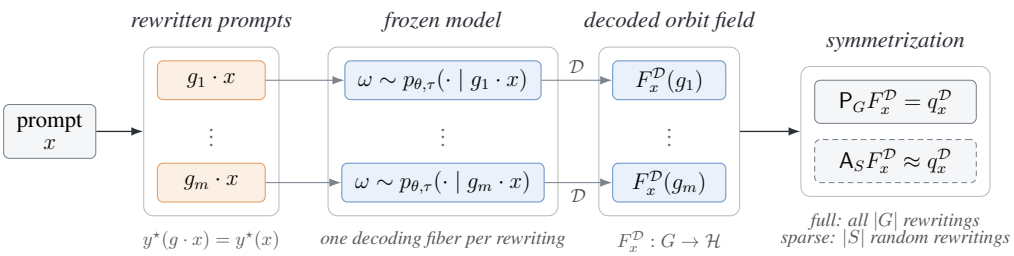


Figure 2: Decoding each rewriting produces the orbit field $F_x^{\mathcal{D}}$. Full group averaging gives the invariant target $q_x^{\mathcal{D}}$, while sparse averaging uses only the rewritings in S to approximate it.

Decoding before symmetrization. A *decoding rule* $\mathcal{D}$ maps the conditional response distribution at one rewriting to an element of a real Hilbert space $\mathcal{H}$. We first describe this rule at the distribution level to isolate the cost of rewriting from the number of traces used to estimate it. The rule induces the *decoded orbit field*

$$F_x^{\mathcal{D}}(g) := \mathcal{D}(p_{\theta,\tau}(\cdot \mid g \cdot x), V_x(g, \cdot)) \in \mathcal{H}. \tag{2}$$

For class prediction, one usually takes $\mathcal{H} = \mathbb{R}^C$. Three examples are

$$F_x^{\mathrm{SC}}(g)_c := \mathbb{P}_{\omega|g}\big(a_x(g,\omega) = c\big), \tag{3}$$

$$F_x^{\mathrm{logit}}(g) := \mathbb{E}_{\omega|g}\big[V_x(g,\omega)\big], \tag{4}$$

$$F_x^{\beta_D}(g)_c := \frac{\mathbb{E}_{\omega|g}\left[e^{\beta_D r_x(g,\omega)} \mathbf{1}\{a_x(g,\omega) = c\}\right]}{\mathbb{E}_{\omega|g}\left[e^{\beta_D r_x(g,\omega)}\right]}. \tag{5}$$

Here $a_x(g,\omega)$ is the answer extracted from a trace and $r_x(g,\omega)$ is a verifier score. The first rule gives self consistency probabilities, the second averages logits, and the third is a verifier weighted vote. In Equation (5), $\beta_D = 0$ recovers self consistency and larger β_D gives more weight to highly verified traces. This decoder inverse temperature is distinct from the generation temperature τ inside $p_{\theta,\tau}$. Our results hold simultaneously for all of these choices and for any other decoding rule whose output lies in a Hilbert space.

The *fully symmetrized response* is

$$q_x^{\mathcal{D}} := \mathsf{P}_G F_x^{\mathcal{D}} = \frac{1}{|G|} \sum_{g \in G} F_x^{\mathcal{D}}(g). \tag{6}$$

It is invariant to the starting rewriting, but it need not be constant across answer classes. For $\mathcal{H} = \mathbb{R}^C$, it remains a class score vector and can select a unique answer. Invariance is therefore imposed only in the rewriting coordinate, not in the answer coordinate. This is the core idea of model agnostic symmetrization: a frozen model is left unchanged, and task structure is imposed through external averaging.

Let $e \in G$ denote the identity rewriting, so that $e \cdot x = x$. Self consistency on the original prompt estimates $F_x^{\mathrm{SC}}(e)$, whereas full test time symmetrization returns $\mathsf{P}_G F_x^{\mathrm{SC}}$. Repeated decoding at $e \cdot x$ provides information only about the original representation. Querying $g \cdot x$ reveals the decoded response at another point on the orbit. Describing the decoding rule at the distribution level removes trace estimation error from the rewriting scaling law, but it does not erase this information difference.

A reusable symmetrization operator. A full average over the orbit requires $|G|$ forward passes. To reduce this cost, let $S = \{\{s_1, \ldots, s_{|S|}\}\}$ be a multiset of group elements and define the sparse rewriting operator

$$(\mathsf{A}_S f)(h) := \frac{1}{|S|} \sum_{i=1}^{|S|} f(s_i h), \qquad h \in G. \tag{7}$$

The argument h indexes the starting rewriting $h \cdot x$ from which the same inference procedure is invoked. The full group averaging operator is

$$(\mathsf{P}_G f)(h) := \frac{1}{|G|} \sum_{g \in G} f(gh), \tag{8}$$

which is constant over h and is the orthogonal projection onto the subspace of functions that are constant on G, as proved in Lemma 10. It is important to distinguish the group coordinate from the answer coordinate. The invariant subspace consists of score fields that are constant as the starting representation h changes. It does not require the scores of the C answer classes to be equal. Indeed, if $f(h) = (f_1(h), \dots, f_C(h))$, then

$$(\mathsf{P}_G f)_c(h) = \frac{1}{|G|} \sum_{g \in G} f_c(gh), \qquad c \in \mathcal{C}, \tag{9}$$

is independent of h but can still vary with c. Consequently, the invariant score vector $\mathsf{P}_G f$ can select a unique answer through $\arg\max_c (\mathsf{P}_G f)_c$. The projection therefore removes sensitivity to how the prompt is represented without removing the information needed to distinguish answer classes. If the original score field is already invariant, then $\mathsf{P}_G f = f$.

Our goal is to derive a scaling law for the number $|S|$ of forward passes spent exploring G. The sampled multiset S is chosen once and remains valid uniformly over frozen models, decoding rules, generation and verifier temperatures, decoder realizations, and output spaces, with error measured in mean square over starting representations. This separates the forward pass cost of rewriting from the statistical cost of estimating a chosen decoder and makes S a reusable model agnostic symmetrization rule.

3.1 Why decoder resampling alone is insufficient

We show that drawing more reasoning traces from the original prompt cannot, in general, recover the decision obtained by averaging over rewritings. Two possible model behaviors can be identical on the original prompt but have different average behavior over its orbit. No decoding rule that observes only the original prompt can distinguish between them.

Theorem 1 (Decoder resampling cannot replace rewriting)**.** *Let G contain at least two elements and let $C \geq 2$. Consider any possibly randomized inference algorithm that is allowed to draw an arbitrary number of independent traces*

$$\omega_t \sim p_{\theta,\tau}(\cdot \mid x)$$

from the original prompt. The algorithm may use internal randomness together with all previous traces and scores to decide whether to draw another trace or to stop and output a class, but it never queries $g \cdot x$ for $g \neq e$. There exist two bounded response fields $V_x^{(0)}$ and $V_x^{(1)}$, sharing the same decoder distribution at x, such that

$$V_x^{(0)}(e, \omega) = V_x^{(1)}(e, \omega) \qquad \textit{for every } \omega \in \Omega.$$

Thus, the complete observation sequence of the algorithm has the same distribution in the two worlds, regardless of how many traces it draws. Nevertheless, the two worlds have different unique invariant decisions:

$$\arg\max_{c \in \mathcal{C}} (\mathsf{P}_G F_x^{(0)})_c \neq \arg\max_{c \in \mathcal{C}} (\mathsf{P}_G F_x^{(1)})_c,$$

where $F_x^{(j)}(g) := \mathbb{E}_{\omega|g}[V_x^{(j)}(g, \omega)]$. Consequently, if the two worlds are equally likely, every such algorithm has error at least $1/2$, regardless of its decoder budget.

The reason is simple. The algorithm receives exactly the same information in both worlds and must therefore behave in the same way, but the correct invariant decisions are different. It must fail in at least one world. Thus, additional decoding can reduce randomness at the original prompt, but it cannot reveal how the model behaves at unseen rewritings. The construction and proof are given in Section D.

4 Scaling Laws for Model-Agnostic Symmetrization

We now quantify how many rewritings are needed to approximate the full group average. The result is stated briefly here; its stronger operator and representation theoretic form appears in Theorem 17.

Theorem 2 (Sparse rewriting scaling law, informal)**.** *Draw the rewritings in S independently and uniformly from G. With probability at least $1-\delta$, the same set S satisfies, for every real Hilbert space $\mathcal{H}$ and every $F \in L^2(G;\mathcal{H})$,*

$$\|\mathsf{A}_S F - \mathsf{P}_G F\|_{L^2(G;\mathcal{H})} \le \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}\,\|F - \mathsf{P}_G F\|_{L^2(G;\mathcal{H})}. \tag{10}$$

This guarantee is simultaneous over all such fields, not one field fixed in advance. Moreover, the $\log|G|$ dependence is optimal in general: for some finite groups, fewer than $\log_2|G|$ rewritings cannot achieve any uniform error below one.

Thus, $|S| = O(\varepsilon^{-2}(\log|G| + \log(1/\delta)))$ rewritings give relative L^2 error at most ε. The event depends only on S, which is why a single sampled set works for every model, decoder, trace realization, and answer space. In particular, there is no dependence on the number of answer classes. The complete theorem also shows that the same sample approximates averaging in every unitary representation. Its proof and the matching lower bound are given in Section E. The argument uses random Cayley expansion and adapts the operator and averaging complexity ideas of Alon & Roichman (1994); Tahmasebi et al. (2026); Tahmasebi & Weber (2026).

4.1 A DECODER AGNOSTIC SCALING LAW

We now apply the same event after decoding. This order keeps the decoding rule arbitrary while leaving the rewriting rate unchanged.

Corollary 3 (Decoder agnostic symmetrization)**.** *On the event in Theorem 2, simultaneously for every prompt, frozen model, generation temperature, decoding rule, and Hilbert valued output,*

$$\|\mathsf{A}_S F_x^{\mathcal{D}} - q_x^{\mathcal{D}}\|_{L^2(G;\mathcal{H})} \le \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}\,\|F_x^{\mathcal{D}} - q_x^{\mathcal{D}}\|_{L^2(G;\mathcal{H})}. \tag{11}$$

The same statement holds for every realized randomized decoder.

There is no dependence on C, on the size of the trace space, or on a decoder covering number. A single draw of S therefore works for self consistency, expected logits, every verifier temperature in Equation (5), and any other decoding rule selected after S is drawn. The guarantee is also pathwise: a finite randomized decoder produces a function on G, and the same event controls every seed and trace realization without a union bound over the trace space. It compares sparse and full group averaging of the same realized decoder field. Note that estimating a distribution level decoder from finitely many traces is a separate statistical question and does not alter the rewriting scaling law. The complete statement and proof are given in Section F; Section C explains why directly sampling joint pairs (g,ω) gives a different guarantee.

This universal control is what separates the two sources of diversity. The rewriting set is chosen without reference to the reasoning traces or to how they will be decoded. One may therefore use any trace sampling and decoding procedure at each rewriting, while the same event controls the resulting field. The rewriting budget $|S|$ and the trace budget can then be calibrated independently: our scaling law controls the former, while a decoder specific analysis controls the latter. No uniform estimation of every individual trace is required. An explicit decomposition of these two errors is given in Section C.

This distinction also explains why we average over rewritings. If the goal were instead to find the single best rewriting, an isolated maximizer could require $\Theta(|G|)$ random queries to discover. The logarithmic law concerns approximate invariance under group averaging, not worst case search over the orbit.

4.2 FROM SYMMETRIZATION TO PREDICTION ERROR

The scaling law controls scores. We next translate it into prediction error. Let X be drawn from the task distribution and take $\mathcal{H} = \mathbb{R}^C$. Fix any decoding rule and abbreviate $F_X = F_X^{\mathcal{D}}$ and $q_X = \mathsf{P}_G F_X$. Let

$$c_G(X) := \arg\max_{c\in\mathcal{C}} (q_X)_c, \qquad \widehat{c}_S(h\cdot X) := \arg\max_{c\in\mathcal{C}} (\mathsf{A}_S F_X)(h)_c,$$

using a fixed tie breaking rule. Write Δ_X for the gap between the largest and second largest coordinates of q_X, with $\Delta_X = 0$ at a tie, and define

$$R_G := \mathbb{P}_X\big(c_G(X) \neq y^\star(X)\big), \qquad R_S := \mathbb{P}_{X,h}\big(\widehat{c}_S(h \cdot X) \neq y^\star(X)\big),$$

where $h \sim \mathrm{Unif}(G)$ independently of X.

Corollary 4 (Prediction error decomposition)**.** *On the event in Theorem 2, for every $t > 0$,*

$$R_S \leq R_G + \mathbb{P}_X(\Delta_X \leq t) + \frac{128}{9t^2}\frac{\log(|G|/\delta)}{|S|}\mathbb{E}_X\left[\|F_X - \mathsf{P}_G F_X\|^2_{L^2(G;\mathbb{R}^C)}\right]. \tag{12}$$

The three terms play different roles. The first, R_G, is the error of the fully symmetrized model, which uses all $|G|$ rewritings; it is a property of the model and the task, and no number of rewritings can reduce it. The second counts examples whose symmetrized scores have a small margin Δ_X, where a small approximation error can change the decision. The third is the finite rewriting error, which decays as $\log|G|/|S|$ with no explicit dependence on the number of classes C; the threshold t trades it against the second term. The bound thus separates three requirements. A large $|S|$ controls only the third term, and the decision can remain unstable when the margin is small. A large margin controls the second term, and even then a task can be hard for the model regardless of how it is written, in which case R_G stays large. In particular, the result does not assume that full symmetrization is optimal: it separates this modeling error from the compute used to approximate it. The proof is given in Section G.

5 Experiments

We test the theory in three steps, with details in Section H. First, we compute the constant of Theorem 2 exactly and check that a frozen language model stays below it. Second, we check that one set of rewritings works for every task and decoding rule (Corollary 3). Third, we ask when rewriting improves accuracy over decoding alone (Theorem 1 and Corollary 4).

Setup. We generate Erdős–Rényi graphs (Erdős & Rényi, 1959) and serialize them as edge lists with integer node labels (Herbst et al., 2025). Throughout, n denotes the number of nodes of a graph. The group $G = S_n$ of node relabelings, of order $|G| = n!$, acts by relabeling the nodes and re-sorting the edge list, so that $g \mapsto g \cdot x$ is a group action on prompts; questions that name nodes are relabeled with the graph. We use six tasks whose answers are invariant under relabeling, similar to those of standard graph-reasoning benchmarks for language models (Fatemi et al., 2024; Wang et al., 2023a): edge count, triangle count, the degree of a named node, the number of isolated nodes, whether two named nodes are adjacent, and the number of connected components. The model is Qwen2.5-1.5B-Instruct (Qwen Team et al., 2024), frozen and prompted to answer with a single token. Reading its next-token distribution over the answer classes gives the distribution-level decoded field $F_x^{\mathcal{D}}(g)$ exactly, for six decoding rules: expected logits Equation (4), self consistency probabilities Equation (3) at generation temperatures $\tau \in \{0.5, 1, 2\}$, greedy decoding, and one realized sampled answer per rewriting.

The uniform constant, computed exactly. For a finite group, the smallest constant that makes Equation (10) hold simultaneously for all fields is the operator norm of $\mathsf{A}_S - \mathsf{P}_G$ on $L^2(G)$, which we compute exactly (Section H.2). Figure 3(a) shows it for graphs with up to $n = 7$ nodes ($|G| = 5040$). It decays at the rate $|S|^{-1/2}$ and lies about four times below the bound: at $n = 7$ and $|S| = 256$ it equals 0.128, against a bound of 0.548. Figure 3(b) isolates the dependence on the group. For S_n, the rescaled constant $\sqrt{|S|}\,\|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}}$ levels off at about 2.05 for $n \geq 6$ and shows no measurable growth with $\log n!$. For the hypercube $\mathbb{Z}_2^d$ it grows from 2.1 at $d = 4$ to 4.9 at $d = 20$ and tracks $\sqrt{2\log|G|}$. The logarithmic factor of Theorem 2 is therefore attained on some groups, as its lower bound states, while node relabeling is cheaper than the worst case at these sizes.

One set of rewritings for every task and decoder. For graphs with $n \in \{4, 5\}$ nodes we query the model on every relabeling of every graph ($|G| = 24$ and 120 prompts per graph), so that $\mathsf{P}_G F$ is exact rather than estimated, and obtain 376 non-constant fields from six graphs per task and six decoding rules. Each set S is drawn once and applied to all of them. Figure 3(c) shows the relative

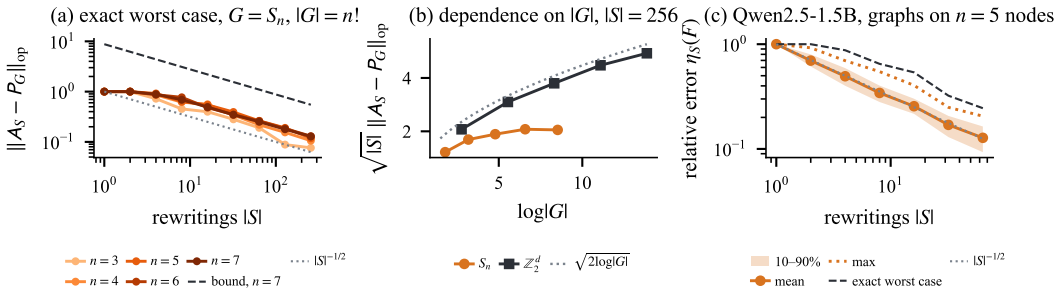


Figure 3: **The sparse rewriting law, exactly and on a frozen model.** (a) Exact uniform constant $\|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}}$ of Theorem 2 for node relabeling of graphs on $n = 3, \ldots, 7$ nodes, $G = S_n$ with $|G| = n!$ (mean of five draws of S), against the upper bound of Theorem 2, Equation (10) with $\delta = 0.1$. (b) The same constant rescaled by $\sqrt{|S|}$ as a function of $\log|G|$, for S_n and for $G = \mathbb{Z}_2^d$ ($|G| = 2^d$, $d = 4, \ldots, 20$); the $\log|G|$ factor of Theorem 2 is attained on $\mathbb{Z}_2^d$ and not visible on S_n at these sizes. (c) Relative error $\eta_S(F) = \|\mathsf{A}_S F - \mathsf{P}_G F\|/\|F - \mathsf{P}_G F\|$ of Corollary 3 for Qwen2.5-1.5B on graphs with $n = 5$ nodes ($|G| = 120$), pooled over all six tasks and six decoding rules (192 fields, one shared S per size and draw), with the exact worst case of the same draws; Figure 6 shows each task and decoding rule separately.

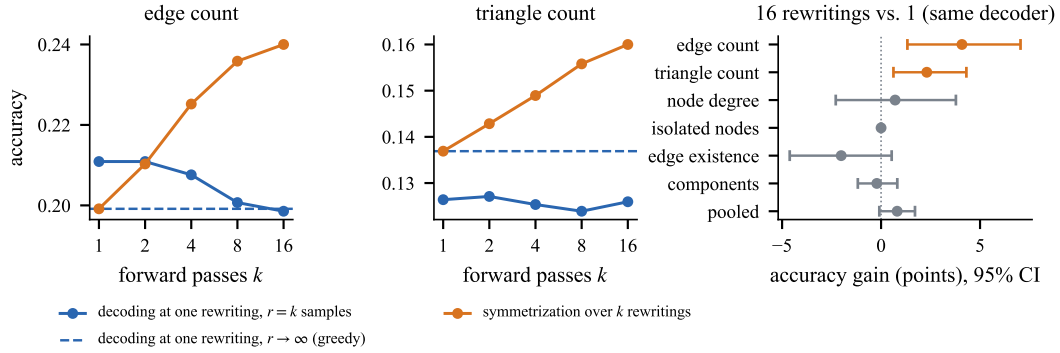


Figure 4: **Rewriting versus decoding at matched compute** (Theorem 1 and Corollary 4), Qwen2.5-1.5B, graphs with $n = 7$ nodes, 100 graphs per task. Left and middle: accuracy with k forward passes when all k are spent at one uniformly random rewriting (blue; $r = k$ sampled answers with a majority vote, and the $r \to \infty$ limit, greedy decoding) or on k random rewritings with the answer distributions averaged (orange). Right: accuracy gain of 16 rewritings over one rewriting with the same decoding rule, with 95% paired bootstrap intervals over graphs, for all six tasks.

error $\eta_S(F) := \|\mathsf{A}_S F - \mathsf{P}_G F\|/\|F - \mathsf{P}_G F\|$, the ratio bounded in Equation (10), at $n = 5$. Its mean is 0.128 at $|S| = 64$, matching $|S|^{-1/2} = 0.125$, and the per-task and per-decoder means lie in $[0.125, 0.128]$ and $[0.123, 0.129]$ (Figure 6). No field exceeds the exact worst-case constant of its draw for $|S| > 1$, as Corollary 3 requires. This holds on all six tasks, since the corollary concerns the approximation of $\mathsf{P}_G F$, not its accuracy.

When does rewriting change the answer? We now fix a budget of k forward passes on graphs with $n = 7$ nodes (100 graphs per task) and draw 16 uniformly random serializations of each graph from its orbit under relabeling and edge reordering. As in the definition of R_S, the starting serialization is itself uniformly random. Decoding at one rewriting can use the k passes for k sampled answers, and in the limit $r \to \infty$ it returns the greedy answer; by Theorem 1, no such procedure can see how the model behaves elsewhere on the orbit. Symmetrization instead averages the answer distributions of k rewritings. Figure 4 shows the outcome. On edge and triangle counting, decoding at one rewriting stays flat (0.199–0.211 and 0.124–0.137 across all k and $r \to \infty$), while symmetrization rises steadily with k to 0.240 and 0.160. With the decoding rule held fixed, 16 rewritings

improve on one by $+4.1$ points (95% interval $[+1.3, +7.0]$) and $+2.3$ points ($[+0.6, +4.3]$). On the other four tasks the change is not significant (Figure 4, right).

Corollary 4 explains when rewriting helps. Symmetrization converges to the decision c_G of the fully symmetrized model, so with many rewritings it improves on a single rewriting exactly when that decision is more accurate than a typical representation. Averaging recovers competence that the model shows on part of the orbit; it cannot create competence that the model lacks on all of it. The six tasks exhibit three cases (Table 1). On isolated-node and component counting, the model is at or below the majority-class baseline at every serialization, so there is nothing to recover. On node degree and edge existence, it is well above the baseline, but its errors do not cancel over the orbit, and symmetrization converges faithfully to an orbit average that is no better than a typical rewriting. On the two counting tasks, the model is right on some serializations and wrong on others, the orbit average is more accurate than a typical representation, and by Theorem 1 decoding at one representation cannot reach it. The gain requires a good decoder at each rewriting: with a single sampled answer per rewriting, as in joint sampling, the decoding error in Equation (19) dominates at these budgets and more rewritings do not help (Section H.5).

6 Conclusion

We studied test-time symmetrization, which spends forward passes on rewritings of the input rather than on repeated decoding. Decoding alone cannot replace rewriting, whatever its budget, while one random set of $O(\log |G|)$ rewritings approximates the full group average simultaneously for every frozen model, decoding rule, realized decoder, and Hilbert valued output, with no dependence on the number of answer classes or the size of the reasoning space. An error decomposition separates the accuracy of the fully symmetrized model, its margin, and the finite rewriting cost. Experiments confirm the law exactly and on a frozen language model, where one set of rewritings serves every task and decoding rule, and rewriting improves accuracy wherever the fully symmetrized model is more accurate than a typical representation.

Our guarantees control invariance in mean square over starting representations rather than at every representation, and they concern the approximation of the fully symmetrized model rather than its accuracy, which depends on the model and the task. They also require the symmetry to be known and exact, and our experiments cover one small model on graph tasks. Natural next steps are approximate and semantic symmetries such as paraphrasing, where rewritings are produced by a model rather than a group, symmetrization inside reasoning traces rather than only at the answer, and experiments with larger models and further exact symmetries such as identifier renaming and in-context example order.

AI use statement

Large language model tools (ChatGPT 5.6 Sol and Claude Fable 5.1) were used primarily for copy-editing, including improvements to grammar, wording, clarity, and LaTeX presentation. The research direction, core ideas, theoretical results, and experimental design are original contributions of the authors. The authors carefully reviewed and verified all technical content and take full responsibility for the final manuscript.

References

Ahmed M. Adly. Measuring the symmetry–data exchange rate. *arXiv preprint arXiv:2606.01090*, 2026. 16

Noga Alon and Yuval Roichman. Random Cayley graphs and expanders. *Random Structures & Algorithms*, 5(2):271–284, 1994. 6, 16

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. Ask me anything: A simple strategy for prompting language models. In *International Conference on Learning Representations*, 2023. arXiv:2210.02441. 1, 3

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020. arXiv:2002.06470. 15

Sourya Basu, Prasanna Sattigeri, Karthikeyan Natesan Ramamurthy, Vijil Chenthamarakshan, Kush R. Varshney, Lav R. Varshney, and Payel Das. Equi-tuning: Group equivariant fine-tuning of pretrained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6788–6796, 2023. arXiv:2210.06475. 15

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks from training data. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 15

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under invariance and geometric stability. In *Advances in Neural Information Processing Systems*, volume 34, 2021. 16

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 15

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. 15

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. 1, 3

Liang Chen, Li Shen, Yang Deng, Xiaoyan Zhao, Bin Liang, and Kam-Fai Wong. PEARL: Towards permutation-resilient LLMs. In *International Conference on Learning Representations*, 2025. arXiv:2502.14628. 3

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. 3

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. 16

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *International Conference on Machine Learning*, 2024b. arXiv:2402.08939. 3

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 1, 2

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, 2016. 15

Persi Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 1988. 25

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993. 25

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. arXiv:2102.01017. 3

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969, 2021. 16

Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959. 7, 25

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations*, 2024. arXiv:2310.04560. 3, 7, 25

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The Lie derivative for measuring learned equivariance. In *International Conference on Learning Representations*, 2023. 15

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. Changing answer order can decrease MMLU accuracy. *arXiv preprint arXiv:2406.19470*, 2024. 3

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pp. 11–15, 2008. 25

Daniel Herbst, Lea Karbevska, Divyanshu Kumar, Akanksha Ahuja, Fatemeh Gholamzadeh Nasrabadi, and Fabrizio Frasca. Lost in serialization: Invariance and generalization of LLM graph reasoners. *arXiv preprint arXiv:2511.10234*, 2025. 1, 3, 7, 25

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. 3

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. 1, 3

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15546–15566, 2023. 15

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

Jinwoo Kim, Tien Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2306.02866. 15

Masanari Kimura. Understanding test-time augmentation. *arXiv preprint arXiv:2402.06892*, 2024. 15

Nikita Kozodoi, Zainab Afolabi, and Jack Butler. Test-time augmentation for LLMs: When input diversity beats output diversity at matched compute. *arXiv preprint arXiv:2608.09351*, 2026. 1, 2, 3

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. 15

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. arXiv:1612.01474. 15

Zeph Landau and Alexander Russell. Random Cayley graphs are expanders: a simple proof of the Alon–Roichman theorem. *The Electronic Journal of Combinatorics*, 11(1):R62, 2004. 16

Richard B. Lehoucq, Danny C. Sorensen, and Chao Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, 1998. 25

Noam Itzhak Levi. A simple model of inference scaling laws. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33984–33998, 2025. arXiv:2410.16377. 3

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024. arXiv:2402.05120. 3

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. arXiv:2305.20050. 1, 3

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. arXiv:2307.03172. 3

Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, and Tieniu Tan. Rethinking the role of prompting strategies in LLM test-time scaling: A perspective of probability theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27962–27994, 2025. arXiv:2505.10981. 3

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. arXiv:2104.08786. 1, 3

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020. 16

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. 15

Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. Order-independence without fine tuning. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.06581. 3

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3351–3418, 2021. 16

Antonio Valerio Miceli Barone, Fazl Barez, Shay B. Cohen, and Ioannis Konstas. The larger they are, the harder they fail: Language models do not recognize identifier swaps in Python. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 272–292, 2023. arXiv:2305.15507. 1, 3

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025. arXiv:2410.05229. 3

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. arXiv:2401.00595. 1, 3

Arnab Kumar Mondal, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Sai Rajeswar, and Siamak Ravanbakhsh. Equivariant adaptation of large pretrained models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2310.01647. 15

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. 2

Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *International Conference on Learning Representations*, 2019. 15

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 24

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2305.17592. 16

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. arXiv:2308.11483. 3

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. 15

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 7, 24

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53132–53176, 2025. arXiv:2502.17578. 3

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, 2021. arXiv:2001.07676. 3

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. arXiv:2310.11324. 1, 3

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *IEEE/CVF International Conference on Computer Vision*, 2021. arXiv:2011.11156. 15

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. arXiv:1409.1556. 15

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. arXiv:2408.03314. 1, 3

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The limits of inference scaling through resampling. In *International Conference on Learning Representations*, 2026. arXiv:2411.17501. 3

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. 16

Behrooz Tahmasebi and Melanie Weber. Achieving approximate symmetry is exponentially easier than exact symmetry. In *International Conference on Learning Representations*, 2026. arXiv:2512.11855. 6, 16, 22

Behrooz Tahmasebi, Melanie Weber, and Stefanie Jegelka. Data augmentation: A Fourier analysis perspective. In *Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6114–6155, 2026. arXiv:2606.24418. 6, 16, 22

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. 16, 18

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015. arXiv:1501.01571. 16

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. 25

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? In *Advances in Neural Information Processing Systems*, 2023a. arXiv:2305.10037. 3, 7

Shiqi Wang, Zheng Li, Haifeng Qian, Chenghao Yang, Zijian Wang, Mingyue Shang, Varun Kumar, Samson Tan, Baishakhi Ray, Parminder Bhatia, Ramesh Nallapati, Murali Krishna Ramanathan, Dan Roth, and Bing Xiang. ReCode: Robustness evaluation of code generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13818–13843, 2023b. arXiv:2212.10264. 3

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023c. arXiv:2203.11171. 1, 2, 24

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2201.11903. 1, 2

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. From decoding to meta-generation: Inference-time algorithms for large language models. *Transactions on Machine Learning Research*, 2024. arXiv:2406.16838. 3

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. 24

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations*, 2025. arXiv:2408.00724. 3

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.10601. 3

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 15

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021. arXiv:2102.09690. 3

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. arXiv:2309.03882. 1, 3

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. In *International Conference on Machine Learning*, 2024. arXiv:2310.01651. 3

Appendix Contents

A Additional Related Work **15**

B Mathematical Preliminaries **16**

B.1 Finite groups, actions, and orbits 16

B.2 Conditional decoding and response fields 16

B.3 Functions on the group and invariant projection 17

B.4 Translation operators and matrix concentration 17

B.5 Hilbert valued averaging and unitary representations 18

B.6 Margins and decision disagreement 20

C On Joint Rewriting and Decoding Sampling **20**

D Proof of Theorem 1 **21**

E Proof of Theorem 2 **22**

F Proof of Corollary 3 **24**

G Proof of Corollary 4 **24**

H Experimental Details and Additional Results **24**

H.1 Setup 24

H.2 Computing the uniform constant 25

H.3 Relative error of language-model fields 25

H.4 Accuracy on all tasks 25

H.5 Allocating rewritings and samples 27

I Notation **28**

A Additional Related Work

Test-time augmentation. Averaging predictions over transformed inputs at inference is an old idea in vision, going back to multi-crop evaluation (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015). Later work studies how to aggregate the augmented predictions (Shanmugam et al., 2021), when averaging helps (Kimura, 2024), and how it compares with ensembles (Ashukha et al., 2020; Breiman, 1996; Lakshminarayanan et al., 2017). Our setting differs in that the transformations form a group, so the cost can be measured against $|G|$, and the predictor is itself random.

Model-agnostic symmetrization. Symmetry is usually built into the architecture (Cohen & Welling, 2016; Bronstein et al., 2021; Zaheer et al., 2017; Maron et al., 2019). When this is not possible, it can be imposed from outside by averaging over the group: Janossy pooling (Murphy et al., 2019), frame averaging (Puny et al., 2022), canonicalization (Kaba et al., 2023; Mondal et al., 2023), probabilistic symmetrization (Kim et al., 2023), and equi-tuning of pretrained models (Basu et al., 2023) are examples. Related work measures how equivariant a trained model is (Gruver et al., 2023), learns invariances from data (Benton et al., 2020), or studies approximate symmetries

(Petrache & Trivedi, 2023). These methods generally alter model construction, training, or internal representations. Our model is frozen and accessed only through forward passes, so the cost of symmetrization is counted in queries.

Theory of invariance. Invariance is known to help at training time: data augmentation reduces variance (Chen et al., 2020; Lyle et al., 2020), equivariant models can generalize strictly better (Elesedy & Zaidi, 2021), and the gain has been quantified for kernel models (Bietti et al., 2021; Mei et al., 2021; Tahmasebi & Jegelka, 2023) and measured experimentally (Adly, 2026); see also Tahmasebi et al. (2026) for a Fourier view of augmentation. Most relevant to us, Tahmasebi & Weber (2026) show in their averaging-complexity framework that exact invariance can require all $|G|$ group elements while approximate invariance needs only $\Theta(\log |G|)$ random ones. Our Theorem 2 is of the same kind and relies on random Cayley graphs being expanders (Alon & Roichman, 1994; Landau & Russell, 2004) and on matrix concentration (Tropp, 2012; 2015). We extend this from deterministic functions to stochastic frozen models, make the guarantee uniform, combine it with decoding, and connect it to decisions through the margin.

B Mathematical Preliminaries

This appendix collects the definitions, notation, and elementary facts used throughout the paper. All expectations of vector valued functions are understood coordinatewise.

B.1 Finite groups, actions, and orbits

Definition 5 (Finite group)**.** A *finite group* G is a finite set equipped with an associative product $(g, h) \mapsto gh$, an identity element $e \in G$, and an inverse $g^{-1} \in G$ for every $g \in G$. Thus, $eg = ge = g$ and $g^{-1}g = gg^{-1} = e$.

Definition 6 (Group action and orbit)**.** A left *action* of G on a set $\mathcal{X}$ is a map $(g, x) \mapsto g \cdot x$ satisfying

$$e \cdot x = x, \qquad (gh) \cdot x = g \cdot (h \cdot x)$$

for all $g, h \in G$ and $x \in \mathcal{X}$. The *orbit* of x is

$$G \cdot x = \{g \cdot x : g \in G\}.$$

Elements of the same orbit are the rewritings that the task regards as semantically equivalent.

Definition 7 (Invariant and equivariant targets)**.** Let $y^\star : \mathcal{X} \to \mathcal{C}$ be the target function. It is *invariant* if

$$y^\star(g \cdot x) = y^\star(x)$$

for every $g \in G$ and $x \in \mathcal{X}$. More generally, suppose that each $g \in G$ induces a permutation $\rho(g) : \mathcal{C} \to \mathcal{C}$ of the answer classes, with $\rho(gh) = \rho(g)\rho(h)$. The target is *equivariant* if

$$y^\star(g \cdot x) = \rho(g) y^\star(x).$$

In the equivariant case, applying $\rho(g)^{-1}$ to the output scores aligns them with the original answer coordinates and reduces the problem to the invariant setting.

B.2 Conditional decoding and response fields

Equip $\mathcal{X}$ with a sigma algebra $\mathcal{B}_\mathcal{X}$, and let $(\Omega, \mathcal{B}_\Omega)$ be the measurable space of reasoning traces. A frozen model specifies, for every prompt $x \in \mathcal{X}$, a probability measure

$$p_{\theta,\tau}(\cdot \mid x)$$

on Ω. Equivalently, $p_{\theta,\tau}$ is a conditional probability kernel from prompts to reasoning traces, where τ may specify a generation temperature. Querying the model on $g \cdot x$ means drawing

$$\omega \sim p_{\theta,\tau}(\cdot \mid g \cdot x).$$

We assume throughout that the maps below are measurable and square integrable whenever an L^2 norm is used.

Definition 8 (Response field and decoding fiber)**.** For a fixed semantic instance x, a *response field* is a measurable function

$$V_x : G \times \Omega \to \mathcal{Z}.$$

The value $V_x(g, \omega)$ is the response representation obtained from trace ω at rewriting $g \cdot x$. It may contain an answer, logits, a verifier score, or the trace itself. The pair

$$\big(\Omega, p_{\theta,\tau}(\cdot \mid g \cdot x)\big)$$

is the *decoding fiber* at $g \cdot x$.

The joint rewriting and decoding experiment is

$$g \sim \mathrm{Unif}(G), \qquad \omega \mid g \sim p_{\theta,\tau}(\cdot \mid g \cdot x).$$

When V_x is vector valued and integrable, its joint expectation is

$$\mathbb{E}_{g,\omega|g}[V_x(g, \omega)] = \frac{1}{|G|} \sum_{g \in G} \int_\Omega V_x(g, \omega) p_{\theta,\tau}(d\omega \mid g \cdot x).$$

Definition 9 (Distribution level decoding rule)**.** Let $\mathcal{H}$ be a real Hilbert space. A distribution level decoding rule is any measurable map $\mathcal{D}$ from the conditional response distribution at a rewriting to an element of $\mathcal{H}$. Its decoded orbit field is

$$F_x^{\mathcal{D}}(g) = \mathcal{D}(p_{\theta,\tau}(\cdot \mid g \cdot x), V_x(g, \cdot)) .$$

No linearity or finite dimensionality is assumed. A randomized finite decoder with seed ξ similarly induces a realized field $\widehat{F}_{x,\xi} : G \to \mathcal{H}$.

B.3 Functions on the group and invariant projection

For a real Hilbert space $\mathcal{H}$, equip $L^2(G; \mathcal{H})$ with the uniform inner product

$$\langle f, q \rangle_{L^2(G)} := \frac{1}{|G|} \sum_{h \in G} \langle f(h), q(h) \rangle_{\mathcal{H}}.$$

Lemma 10 (Full group averaging)**.** *For $f \in L^2(G; \mathcal{H})$, define*

$$(\mathsf{P}_G f)(h) := \frac{1}{|G|} \sum_{g \in G} f(gh).$$

Then $\mathsf{P}_G f$ is constant in h, and P_G is the orthogonal projection onto the subspace of functions that are constant on G.

Proof. For fixed $h \in G$, the map $g \mapsto gh$ is a bijection of G. Hence

$$(\mathsf{P}_G f)(h) = \frac{1}{|G|} \sum_{u \in G} f(u),$$

which does not depend on h. Thus, the image of P_G consists of constant functions, and $\mathsf{P}_G q = q$ for every constant function q. Moreover, $f - \mathsf{P}_G f$ has mean zero and is therefore orthogonal to every constant function. This proves the claim. □

B.4 Translation operators and matrix concentration

For a scalar function $f \in L^2(G)$ and $s \in G$, define the translation operator

$$(\mathsf{T}_s f)(h) := f(sh), \qquad h \in G. \tag{13}$$

Because $h \mapsto sh$ is a permutation of G, T_s is unitary under the uniform measure. In this notation,

$$\mathsf{A}_S = \frac{1}{|S|} \sum_{i=1}^{|S|} \mathsf{T}_{s_i}, \qquad \mathsf{P}_G = \mathbb{E}_{s \sim \mathrm{Unif}(G)}[\mathsf{T}_s]. \tag{14}$$

The operator norm used in Theorem 17 is

$$\|\mathsf{T}\|_{L^2(G)\to L^2(G)} := \sup_{f\neq 0} \frac{\|\mathsf{T}f\|_{L^2(G)}}{\|f\|_{L^2(G)}}. \tag{15}$$

Thus, an operator norm bound is one event that controls the approximation error for every function $f \in L^2(G)$ simultaneously.

We use the following standard matrix Bernstein inequality (Tropp, 2012). It is stated for operators so that it applies directly to translations on $L^2(G)$.

Lemma 11 (Matrix Bernstein inequality)**.** *Let $X_1, \dots, X_m$ be independent, mean zero linear operators on a d dimensional Hilbert space. Suppose that $\|X_i\|_{\mathrm{op}} \le R$ almost surely, and define*

$$\sigma^2 := \max\left\{ \left\| \sum_{i=1}^m \mathbb{E}[X_i X_i^*] \right\|_{\mathrm{op}}, \left\| \sum_{i=1}^m \mathbb{E}[X_i^* X_i] \right\|_{\mathrm{op}} \right\}.$$

Then, for every $t \ge 0$,

$$\mathbb{P}\left(\left\| \sum_{i=1}^m X_i \right\|_{\mathrm{op}} \ge t \right) \le 2d \exp\left(-\frac{t^2}{2\sigma^2 + 2Rt/3} \right).$$

The nonsymmetric form above follows from the self-adjoint matrix Bernstein inequality by applying it to the self-adjoint dilation of each X_i.

B.5 Hilbert valued averaging and unitary representations

We collect the representation theoretic facts behind the universality in Theorem 17. We state them over complex Hilbert spaces, where the standard Fourier decomposition is most transparent. A real representation is covered by passing to its complexification, which preserves operator norms.

Definition 12 (Unitary representation and invariant subspace)**.** A *unitary representation* of G on a Hilbert space $\mathcal{H}$ is a map $\pi : G \to \mathcal{U}(\mathcal{H})$ satisfying

$$\pi(gh) = \pi(g)\pi(h), \qquad \pi(e) = I_{\mathcal{H}}.$$

Its invariant subspace is

$$\mathcal{H}^G := \{v \in \mathcal{H} : \pi(g)v = v \text{ for every } g \in G\}.$$

A nonzero representation is *irreducible* if its only closed invariant subspaces are $\{0\}$ and the whole space.

The *left regular representation* λ acts on $L^2(G;\mathbb{C})$ by

$$(\lambda(g)f)(h) := f(g^{-1}h).$$

Our translation $\mathsf{T}_g f(h) = f(gh)$ equals $\lambda(g^{-1})$. This inverse changes neither the relevant operator norm nor the distribution of a uniform group sample.

Let $\widehat{G}$ denote a complete set of inequivalent irreducible unitary representations, and let d_ρ be the dimension of $\rho \in \widehat{G}$. The finite group Fourier decomposition gives

$$L^2(G;\mathbb{C}) \cong \bigoplus_{\rho\in\widehat{G}} \mathbb{C}^{d_\rho} \otimes \mathbb{C}^{d_\rho}, \qquad \lambda(g) \cong \bigoplus_{\rho\in\widehat{G}} \rho(g) \otimes I_{d_\rho}. \tag{16}$$

Thus, every irreducible representation occurs inside the regular representation, with multiplicity equal to its dimension. More generally, every unitary representation of a finite group is an orthogonal direct sum of irreducible representations. These are the finite group forms of Maschke's theorem and the Peter Weyl decomposition.

The Hilbert valued space $L^2(G;\mathcal{H})$ is naturally the Hilbert tensor product $L^2(G) \otimes \mathcal{H}$. A scalar operator T on $L^2(G)$ acts on it as $T \otimes I_{\mathcal{H}}$.

Lemma 13 (Hilbert valued tensorization)**.** *For every bounded operator $T : L^2(G) \to L^2(G)$ and every real Hilbert space $\mathcal{H}$,*

$$\|T \otimes I_{\mathcal{H}}\|_{L^2(G;\mathcal{H})\to L^2(G;\mathcal{H})} = \|T\|_{L^2(G)\to L^2(G)}.$$

Consequently, if $T\mathsf{P}_G = \mathsf{P}_G T = 0$, then

$$\|TF\|_{L^2(G;\mathcal{H})} \le \|T\|_{\mathrm{op}} \|F - \mathsf{P}_G F\|_{L^2(G;\mathcal{H})}.$$

Proof. Choose an orthonormal basis $(u_j)_j$ for the closed span of the finite set $\{F(g) : g \in G\}$. Writing $F = \sum_j f_j u_j$ and using Parseval's identity gives

$$\|(T \otimes I_{\mathcal{H}})F\|^2 = \sum_j \|Tf_j\|^2_{L^2(G)} \le \|T\|^2_{\mathrm{op}} \sum_j \|f_j\|^2_{L^2(G)}.$$

This proves the upper bound, and testing on $F = fu$ for a unit vector $u \in \mathcal{H}$ proves equality. The final claim follows from $TF = T(F - \mathsf{P}_G F)$. □

Lemma 14 (Group averaging is an invariant projection)**.** *For a unitary representation $\pi : G \to \mathcal{U}(\mathcal{H})$, the operator*

$$\mathsf{P}_\pi := \frac{1}{|G|} \sum_{g\in G} \pi(g)$$

is the orthogonal projection onto $\mathcal{H}^G$.

Proof. For any $a \in G$, a change of variables gives

$$\pi(a)\mathsf{P}_\pi = \frac{1}{|G|} \sum_{g\in G} \pi(ag) = \mathsf{P}_\pi.$$

Hence the image of P_π lies in $\mathcal{H}^G$. Conversely, $\mathsf{P}_\pi v = v$ for every $v \in \mathcal{H}^G$. Moreover,

$$\mathsf{P}_\pi^* = \frac{1}{|G|} \sum_{g\in G} \pi(g^{-1}) = \mathsf{P}_\pi,$$

and the same change of variables gives $\mathsf{P}_\pi^2 = \mathsf{P}_\pi$. It is therefore the orthogonal projection onto $\mathcal{H}^G$. □

Lemma 15 (The regular representation controls every representation)**.** *Let $\pi : G \to \mathcal{U}(\mathcal{H})$ be a unitary representation of a finite group, and let*

$$\mathsf{P}_\pi := \frac{1}{|G|} \sum_{g\in G} \pi(g).$$

For every multiset $S \subseteq G$,

$$\left\| \frac{1}{|S|} \sum_{s\in S} \pi(s) - \mathsf{P}_\pi \right\|_{\mathrm{op}} \le \|\mathsf{A}_S - \mathsf{P}_G\|_{L^2(G)\to L^2(G)}.$$

Moreover, P_π is the orthogonal projection onto the vectors fixed by every $\pi(g)$.

Proof. Write $m = |S|$ and define

$$B_\pi := \frac{1}{m} \sum_{i=1}^{m} \pi(s_i) - \mathsf{P}_\pi.$$

Decompose π into irreducibles. On an isotypic component associated with $\rho \in \widehat{G}$, the operator B_π acts as

$$\left(\frac{1}{m} \sum_{i=1}^{m} \rho(s_i) - \mathsf{P}_\rho \right) \otimes I,$$

so $\|B_\pi\|_{\mathrm{op}}$ is the supremum of the norms of these irreducible blocks. By Equation (16), every such block occurs in the regular representation.

Our convention uses $\mathsf{T}_s = \lambda(s^{-1})$. The corresponding regular block contains

$$\frac{1}{m}\sum_{i=1}^{m}\rho(s_i^{-1}) - \mathsf{P}_\rho = \left(\frac{1}{m}\sum_{i=1}^{m}\rho(s_i) - \mathsf{P}_\rho\right)^*.$$

Taking an adjoint does not change the operator norm. Therefore, each irreducible block of B_π has norm at most $\|\mathsf{A}_S - \mathsf{P}_G\|_{L^2(G)\to L^2(G)}$, and taking the supremum over the blocks proves the inequality. The final statement follows from Lemma 14. □

B.6 Margins and decision disagreement

For a score vector $q \in \mathbb{R}^C$ with a unique maximizer c^*, define its margin

$$\Delta(q) := q_{c^*} - \max_{c \neq c^*} q_c.$$

We use any fixed rule to break ties in approximate score vectors.

Lemma 16 (A changed decision requires score error)**.** *Let $q, v \in \mathbb{R}^C$, suppose that q has unique maximizer c^*, and let $\Delta(q) > 0$ be its margin. If $\arg\max_c v_c \neq c^*$, then*

$$\|v - q\|_2^2 \geq \frac{\Delta(q)^2}{2}.$$

Proof. Let $c \neq c^*$ be the class selected by v. Since $v_c \geq v_{c^*}$,

$$\Delta(q) \leq q_{c^*} - q_c \leq |v_{c^*} - q_{c^*}| + |v_c - q_c| \leq \sqrt{2}\,\|v - q\|_2.$$

Squaring both sides proves the claim. □

C On Joint Rewriting and Decoding Sampling

A natural alternative is to sample pairs

$$g_i \sim \mathrm{Unif}(G), \qquad \omega_i \sim p_{\theta,\tau}(\cdot \mid g_i \cdot x),$$

and aggregate the resulting responses directly. This is a valid inference procedure, but it answers a different statistical question from the operator approximation studied in this paper. We explain the distinction here.

Joint sampling is exact for linear decoding rules. Suppose that a distribution level decoding rule has the linear form

$$F_x^{\mathcal{D}}(g) = \mathbb{E}_{\omega|g}\big[\psi_x(g,\omega)\big]$$

for some Hilbert valued response ψ_x. Then the fully symmetrized response satisfies

$$q_x^{\mathcal{D}} = \mathbb{E}_{g\sim\mathrm{Unif}(G)}\mathbb{E}_{\omega|g}[\psi_x(g,\omega)] = \mathbb{E}_{g,\omega|g}[\psi_x(g,\omega)]. \tag{17}$$

Consequently, independent joint samples give an unbiased Monte Carlo estimator of $q_x^{\mathcal{D}}$. Self consistency with one hot vote vectors and expected logits are examples. For one fixed instance, one fixed decoder, and one fixed starting representation, standard concentration may therefore be the simplest analysis.

Nonlinear decoding changes the target quantity. The equality in Equation (17) does not hold for a general decoder. The quantity studied in this paper is

$$q_x^{\mathcal{D}} = \mathbb{E}_g\big[\mathcal{D}(P_{x,g})\big], \qquad P_{x,g} := \mathrm{Law}\big(V_x(g,\omega) \mid g\big), \tag{18}$$

which first decodes each conditional trace distribution and then removes rewriting sensitivity. Pooling joint samples generally corresponds instead to applying a decoder after mixing the conditional distributions. These operations need not commute:

$$\mathbb{E}_g\big[\mathcal{D}(P_{x,g})\big] \neq \mathcal{D}(\mathbb{E}_g P_{x,g})\,.$$

For example, the verifier weighted decoder in Equation (5) averages the per rewriting ratios

$$\mathbb{E}_g\left[\frac{\mathbb{E}_{\omega|g}[e^{\beta_D r_x(g,\omega)}\mathbf{1}\{a_x(g,\omega)=c\}]}{\mathbb{E}_{\omega|g}[e^{\beta_D r_x(g,\omega)}]}\right].$$

Pooling all pairs before normalization instead produces a ratio of joint expectations. The two weight rewritings differently and can select different answers.

Pointwise estimation is not universal operator approximation. Even in the linear case, joint Monte Carlo estimates one chosen quantity $q_x^{\mathcal{D}}$ for one response field. Our goal is stronger: draw S once so that the induced group operator approximates P_G for every response field simultaneously. This operator event remains valid after the model, decoder, generation temperature, verifier temperature, output dimension, and decoder realization are chosen. It also controls mean square sensitivity to the starting rewriting. A concentration bound for one joint expectation does not provide this reusable guarantee. Indeed, the lower bound in Theorem 17 shows that uniform approximation over all orbit functions requires the $\log|G|$ dependence that can be absent from a pointwise Monte Carlo bound.

What happens with finite trace sampling. The distribution level formulation isolates the cost of exploring the orbit. In practice, a finite decoder is run at each queried rewriting, producing a realized field $\widehat{F}_{x,\xi}$. The bound in Equation (25) then compares sparse and full group averaging of that same realized field, simultaneously for every seed and trace configuration. Any difference between $\widehat{F}_{x,\xi}$ and the desired field $F_x^{\mathcal{D}}$ is a decoder estimation problem and depends on the chosen decoding algorithm. Keeping this term separate makes clear which error is reduced by more rewritings and which is reduced by more traces per rewriting.

The separation has a direct error decomposition. Since A_S is an average of unitary translation operators, $\|\mathsf{A}_S\|_{\mathrm{op}} \le 1$. On the event in Theorem 17,

$$\|\mathsf{A}_S\widehat{F}_{x,\xi} - q_x^{\mathcal{D}}\|_{L^2(G;\mathcal{H})} \le \|\widehat{F}_{x,\xi} - F_x^{\mathcal{D}}\|_{L^2(G;\mathcal{H})} + \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}\|F_x^{\mathcal{D}} - \mathsf{P}_G F_x^{\mathcal{D}}\|_{L^2(G;\mathcal{H})}. \tag{19}$$

The first term is the finite decoding error and the second is the rewriting approximation error. They can be controlled independently and then combined through Equation (19). In particular, the first term requires control only of the aggregate decoded field in $L^2(G;\mathcal{H})$, not uniform approximation over the trace space Ω.

For example, consider self consistency with n_D independent traces at each rewriting. If $F_x^{\mathrm{SC}}(g)$ is the class probability vector and $\widehat{F}_x^{\mathrm{SC}}(g)$ is its empirical frequency vector, then

$$\mathbb{E}\Big[\|\widehat{F}_x^{\mathrm{SC}} - F_x^{\mathrm{SC}}\|^2_{L^2(G;\mathbb{R}^C)}\Big] = \frac{1}{n_D}\,\mathbb{E}_{g\sim\mathrm{Unif}(G)}\left[1 - \|F_x^{\mathrm{SC}}(g)\|_2^2\right] \le \frac{1}{n_D}. \tag{20}$$

Thus, even this decoder term need not carry an explicit dependence on the number of classes. Averaged logits admit the usual Hilbert valued concentration bounds under a norm or tail assumption. Verifier weighted decoding additionally requires control of the weights and of its normalization denominator. These assumptions and rates belong to the selected decoder, whereas the rewriting rate is unchanged.

Thus, joint sampling is not ruled out, and it is fully appropriate for estimating a fixed linear joint expectation. It is not the object of our main scaling law because it can change the target for nonlinear decoders and, even when it does not, it does not yield the universal model agnostic symmetrization operator that is the focus of this paper.

D Proof of Theorem 1

Proof of Theorem 1. Choose two distinct classes $c_0, c_1 \in \mathcal{C}$, and let $u_{c_0}, u_{c_1} \in \mathbb{R}^C$ be their standard basis vectors. Use the same decoder distribution in both worlds and define response fields that do

not depend on ω:

$$V_x^{(j)}(g,\omega) := \begin{cases} 0, & g = e, \\ u_{c_j}, & g \neq e, \end{cases} \qquad j \in \{0,1\}.$$

Both response fields take values in $[0,1]^C$. At the original prompt,

$$V_x^{(0)}(e,\omega) = V_x^{(1)}(e,\omega) = 0$$

for every ω. Since the decoder distribution at x is also the same, every trace and score observed by an algorithm that queries only x has the same distribution in both worlds. This remains true under adaptive sampling because, conditional on any previous transcript, the next observation has the same conditional distribution in the two worlds. For a randomized algorithm, couple its internal random seed to be identical in both worlds. The complete transcript, including the internal randomness, then has the same distribution in both worlds.

The response fields are independent of ω, so decoder averaging leaves them unchanged. Therefore,

$$\mathsf{P}_G F_x^{(j)} = \frac{|G|-1}{|G|} u_{c_j}.$$

The unique invariant decision is c_j in world j, and these decisions are different.

Let Q be the common output distribution of the algorithm in the two worlds. Under a uniform prior on the worlds, its success probability is

$$\frac{1}{2}Q(c_0) + \frac{1}{2}Q(c_1) \leq \frac{1}{2},$$

because $c_0 \neq c_1$. Hence its error probability is at least $1/2$, regardless of the number of decoder samples. □

E Proof of Theorem 2

The proof adapts the operator concentration viewpoint used for partial group averaging in Tahmasebi et al. (2026) and the averaging complexity ideas of Tahmasebi & Weber (2026). Here these ideas are applied to test time rewriting operators acting on frozen model response fields.

We first state the complete version of Theorem 2. In addition to the Hilbert valued bound used in the main text, it records the underlying operator estimate and its consequence for every unitary representation of G.

Theorem 17 (Universal sparse group averaging)**.** *Let S contain $|S|$ independent samples from $\mathrm{Unif}(G)$ and let $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\|\mathsf{A}_S - \mathsf{P}_G\|_{L^2(G)\to L^2(G)} \leq \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}. \tag{21}$$

On the same event, for every real Hilbert space $\mathcal{H}$ and every $F \in L^2(G;\mathcal{H})$,

$$\|\mathsf{A}_S F - \mathsf{P}_G F\|_{L^2(G;\mathcal{H})} \leq \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}\|F - \mathsf{P}_G F\|_{L^2(G;\mathcal{H})}. \tag{22}$$

Moreover, for every unitary representation $\pi : G \to \mathcal{U}(\mathcal{H})$, if P_π denotes the orthogonal projection onto its invariant vectors, then

$$\left\| \frac{1}{|S|}\sum_{i=1}^{|S|} \pi(s_i) - \mathsf{P}_\pi \right\|_{\mathrm{op}} \leq \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}. \tag{23}$$

Finally, the logarithmic dependence is necessary in general: for $G = (\mathbb{Z}/2\mathbb{Z})^n$, every multiset with fewer than $n = \log_2|G|$ elements has operator error equal to one.

Proof of Theorem 17. Write $m = |S|$ and $N = |G|$. The case $N = 1$ is immediate, so assume $N \geq 2$. For each sampled element s_i, let

$$X_i := \mathsf{T}_{s_i} - \mathsf{P}_G.$$

By Equation (14), the operators X_i are independent and have mean zero. Translation preserves constant functions and preserves the uniform mean, so

$$\mathsf{T}_s \mathsf{P}_G = \mathsf{P}_G \mathsf{T}_s = \mathsf{P}_G \qquad \text{for every } s \in G.$$

Since T_s is unitary and P_G is an orthogonal projection, it follows that

$$X_i X_i^* = X_i^* X_i = I - \mathsf{P}_G, \qquad \|X_i\|_{\mathrm{op}} = 1.$$

Consequently, the variance parameter in Lemma 11 satisfies

$$\sigma^2 = m\|I - \mathsf{P}_G\|_{\mathrm{op}} = m.$$

Applying Lemma 11 on the N dimensional space $L^2(G)$ gives, for every $u \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{m} X_i\right\|_{\mathrm{op}} \geq u\right) \leq 2N \exp\left(-\frac{u^2}{2m + 2u/3}\right).$$

Equivalently, for every $t \geq 0$,

$$\mathbb{P}\left(\|\mathsf{A}_S - \mathsf{P}_G\|_{L^2(G)\to L^2(G)} \geq t\right) \leq 2N \exp\left(-\frac{mt^2}{2 + 2t/3}\right). \tag{24}$$

Moreover, $\|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}} \leq 1$, since on constant functions the difference vanishes and on their orthogonal complement it is an average of unitary operators. If

$$\frac{8}{3}\sqrt{\frac{\log(N/\delta)}{m}} \geq 1,$$

the claimed bound is therefore deterministic. Otherwise, set

$$t = \frac{8}{3}\sqrt{\frac{\log(N/\delta)}{m}} < 1.$$

Since $2 + 2t/3 \leq 8/3$, Equation (24) gives

$$\mathbb{P}\left(\|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}} \geq t\right) \leq 2N \left(\frac{N}{\delta}\right)^{-8/3} \leq \delta,$$

where the last inequality uses $N/\delta \geq 2$. This proves Equation (21) with the explicit constant $8/3$.

Set $T = \mathsf{A}_S - \mathsf{P}_G$. Since $T\mathsf{P}_G = \mathsf{P}_G T = 0$, Lemma 13 gives Equation (22) for every Hilbert space and every Hilbert valued field on the same event. Applying Lemma 15 gives Equation (23) for every unitary representation, again on the same event. No further probability bound or union bound is required.

It remains to prove the worst case necessity of $\log|G|$. Let $G = (\mathbb{Z}/2\mathbb{Z})^n$, so that $|G| = 2^n$, and suppose that $m < n$. The sampled elements $s_1, \ldots, s_m$ span a proper subspace of G over $\mathbb{Z}/2\mathbb{Z}$. Hence there is a nonzero $a \in G$ such that

$$\langle a, s_i \rangle = 0 \pmod 2 \qquad \text{for every } i.$$

Define the nonconstant character $\chi_a(h) = (-1)^{\langle a,h\rangle}$. It has uniform mean zero, so $\mathsf{P}_G\chi_a = 0$. On the other hand,

$$(\mathsf{T}_{s_i}\chi_a)(h) = \chi_a(s_i + h) = \chi_a(s_i)\chi_a(h) = \chi_a(h),$$

and therefore $\mathsf{A}_S\chi_a = \chi_a$. Thus,

$$\|\mathsf{A}_S - \mathsf{P}_G\|_{L^2(G)\to L^2(G)} \geq 1.$$

The reverse inequality was established above, so the error equals one. Since $n = \log_2|G|$, fewer than $\log_2|G|$ rewritings cannot achieve any operator error strictly below one for this family of groups. This proves the final claim. □

F Proof of Corollary 3

For completeness, the pathwise statement used in the main text is

$$\|\mathsf{A}_S\widehat{F}_{x,\xi} - \mathsf{P}_G\widehat{F}_{x,\xi}\|_{L^2(G;\mathcal{H})} \le \frac{8}{3}\sqrt{\frac{\log(|G|/\delta)}{|S|}}\|\widehat{F}_{x,\xi} - \mathsf{P}_G\widehat{F}_{x,\xi}\|_{L^2(G;\mathcal{H})}. \tag{25}$$

Proof of Corollary 3. Fix the operator event in Theorem 17. For any choices of x, θ, τ, $\mathcal{D}$, and $\mathcal{H}$, the decoded response $F_x^{\mathcal{D}}$ belongs to $L^2(G;\mathcal{H})$. Applying Equation (22) to this function and using $q_x^{\mathcal{D}} = \mathsf{P}_G F_x^{\mathcal{D}}$ proves Equation (11). The event depends only on S, so the conclusion is simultaneous over all of these choices.

For a randomized decoder, condition on any seed ξ and all sampled traces. Its output is then a deterministic field $\widehat{F}_{x,\xi} \in L^2(G;\mathcal{H})$. A second application of Equation (22) proves Equation (25). Because the operator event already holds for every function, the conclusion is simultaneous over all seeds and trace realizations and requires no union bound over Ω. □

G Proof of Corollary 4

Proof of Corollary 4. If $\widehat{c}_S(h \cdot X) \neq y^\star(X)$, then either $c_G(X) \neq y^\star(X)$ or $\widehat{c}_S(h \cdot X) \neq c_G(X)$. Hence

$$R_S \le R_G + \mathbb{P}_{X,h}\big(\widehat{c}_S(h \cdot X) \neq c_G(X)\big).$$

Split the second probability according to whether $\Delta_X \le t$. On $\{\Delta_X > t\}$, Lemma 16 shows that a changed decision requires

$$\|(\mathsf{A}_S F_X)(h) - q_X\|_2^2 \ge \frac{t^2}{2}.$$

For each X with $\Delta_X > t$, Markov's inequality over $h \sim \mathrm{Unif}(G)$ therefore gives

$$\begin{aligned}\mathbb{P}_h\big(\widehat{c}_S(h \cdot X) \neq c_G(X),\ \Delta_X > t \mid X\big) &\le \frac{2}{t^2}\|\mathsf{A}_S F_X - \mathsf{P}_G F_X\|^2_{L^2(G;\mathbb{R}^C)} \\ &\le \frac{128}{9t^2}\frac{\log(|G|/\delta)}{|S|}\|F_X - \mathsf{P}_G F_X\|^2_{L^2(G;\mathbb{R}^C)},\end{aligned}$$

where the second line uses Equation (22). Taking expectation over X and adding $\mathbb{P}_X(\Delta_X \le t)$ proves Equation (12). □

H Experimental Details and Additional Results

This appendix gives the full setup of Section 5 and additional results.

H.1 Setup

Software and model. All experiments use PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020). The model is Qwen2.5-1.5B-Instruct (Qwen Team et al., 2024) with its chat template and default system prompt. We run it in 32-bit floating point. The model does not generate text: for each prompt we run a single forward pass and read the next-token log-probabilities of the answer tokens (`Yes`/`No`, or the digits `0`–`9`), all of which are single tokens with distinct identifiers. These tokens receive 99.7% of the next-token probability mass on average and at least 96% on every prompt. The decoded fields are computed from these log-probabilities: expected logits (centered log-probabilities), self consistency (Wang et al., 2023c) probabilities $\mathrm{softmax}(\log p/\tau)$ restricted to the answer classes for $\tau \in \{0.5, 1, 2\}$, the one-hot greedy answer, and one realized answer sampled at $\tau = 1$ with a fixed seed.

Graphs and prompts. Graphs are drawn from the Erdős–Rényi model $G(n, p)$ (Erdős & Rényi, 1959), generated with NetworkX (Hagberg et al., 2008), with p uniform in $[0.15, 0.6]$ for $n \leq 5$ and in $[0.15, 0.4]$ for $n = 7$, rejecting graphs without edges, and are balanced across the answer classes that occur. For edge counting, graphs with more than nine edges are rejected so that the answer is a single digit. The degree and adjacency questions name nodes drawn uniformly at random once per graph. A prompt reads

```
The graph has nodes 0, 1, 2, 3, 4, 5, 6 and the following
undirected edges:  (0, 5), (1, 5), (5, 6), (2, 3).
Question:  How many neighbors does node 1 have?  Answer with
only a single digit.
```

Prompts use the edge-list encoding with integer labels (Herbst et al., 2025; Fatemi et al., 2024). A relabeling $g \in S_n$ maps node v to $g \cdot v$ in the edge list and in the question; each edge is written with its smaller label first, and the edges are sorted, so that the serialization depends only on the relabeled graph and $g \mapsto g \cdot x$ is a group action. Graphs are seeded by task and size, so every analysis uses the same instances.

Rewritings in the accuracy experiment. For the accuracy experiment of Figure 4 (graphs with $n = 7$ nodes), each rewriting also reorders the edge list uniformly at random. The set of serializations is then the orbit of the canonical edge sequence under the product group $S_n \times S_{|E|}$, acting by relabeling node names and permuting positions in the sequence, and our rewritings are independent uniform draws from this orbit. We draw 16 rewritings per graph and evaluate every budget $k \leq 16$ on subsets of them, averaging over 200 random subsets and starting rewritings.

H.2 Computing the uniform constant

For a multiset S, the operator $\mathsf{A}_S - \mathsf{P}_G$ acts on $L^2(G)$ with uniform measure, and $\mathsf{A}_S - \mathsf{P}_G = (\mathsf{A}_S - \mathsf{P}_G)(I - \mathsf{P}_G)$ since $\mathsf{A}_S\mathsf{P}_G = \mathsf{P}_G\mathsf{A}_S = \mathsf{P}_G$. Hence

$$\sup_{F \notin \operatorname{im} \mathsf{P}_G} \frac{\|\mathsf{A}_S F - \mathsf{P}_G F\|}{\|F - \mathsf{P}_G F\|} = \|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}}$$

for scalar fields, and the same value holds for Hilbert-valued fields by applying the scalar bound coordinatewise in an orthonormal basis. For $G = S_n$ we represent functions on G as vectors indexed by the lexicographic rank of permutations, implement A_S and its adjoint by index maps $h \mapsto s_i h$, and compute the largest singular value with ARPACK (Lehoucq et al., 1998) through SciPy (Virtanen et al., 2020). For $n \leq 5$ we checked every draw against a dense singular value decomposition; all 70 draws agree to 10^{-6}. For $G = \mathbb{Z}_2^d$ every irreducible representation is a character (Diaconis, 1988) $\chi_T(s) = (-1)^{\langle T, s \rangle}$, so $\|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}} = \max_{T \neq 0} \left| \frac{1}{|S|} \sum_i \chi_T(s_i) \right|$, which we read off the Walsh–Hadamard transform of the empirical distribution of S in $O(d\, 2^d)$ time. Figure 5 shows the full curves with their spread over five draws.

H.3 Relative error of language-model fields

Figure 6 breaks down the relative error of Figure 3(c) by group size, task, and decoding rule. Fields that are constant over the orbit have $F = \mathsf{P}_G F$ and no defined relative error; they are excluded, which removes the greedy and realized decoders on graphs whose answer never changes. Across all $|S| > 1$, all draws, and both group sizes, no field exceeds the exact worst-case constant of its draw. At $|S| = 1$ both equal one, and the ratio for fields differs from one only by floating-point rounding (at most 2.4×10^{-7}).

H.4 Accuracy on all tasks

Table 1 reports, for each task, the accuracy of the decoding and rewriting procedures of Figure 4 at $k = 16$, together with two paired comparisons: rewriting against greedy decoding at one rewriting (the same decoding rule, as in the main text), and rewriting against self consistency with 16 samples at one rewriting. Intervals are 95% paired bootstrap intervals (Efron & Tibshirani, 1993) over graphs (20,000 resamples for the first comparison and 10,000 for the second). The column "Base" is the

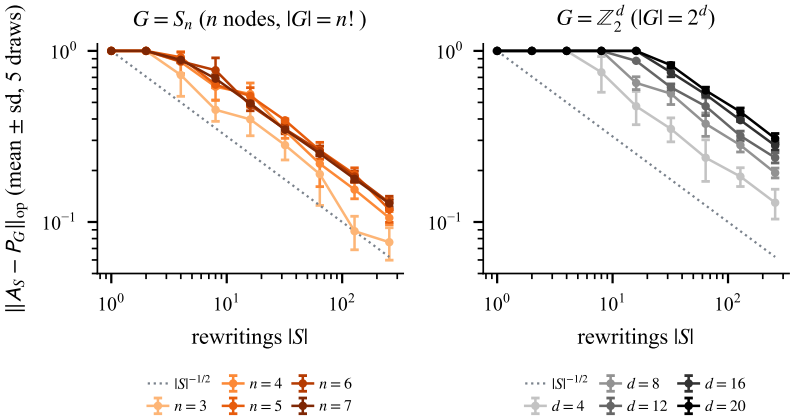


Figure 5: Exact uniform constant $\|\mathsf{A}_S - \mathsf{P}_G\|_{\mathrm{op}}$ of Theorem 2 against $|S|$ for $G = S_n$ (graphs on $n = 3, \ldots, 7$ nodes, $|G| = n!$) and $G = \mathbb{Z}_2^d$ ($|G| = 2^d$, $d = 4, \ldots, 20$); mean and standard deviation over five draws of S. For $\mathbb{Z}_2^d$ the constant stays at one until $|S|$ is of order d, and larger groups need proportionally more rewritings; for S_n the curves nearly coincide.

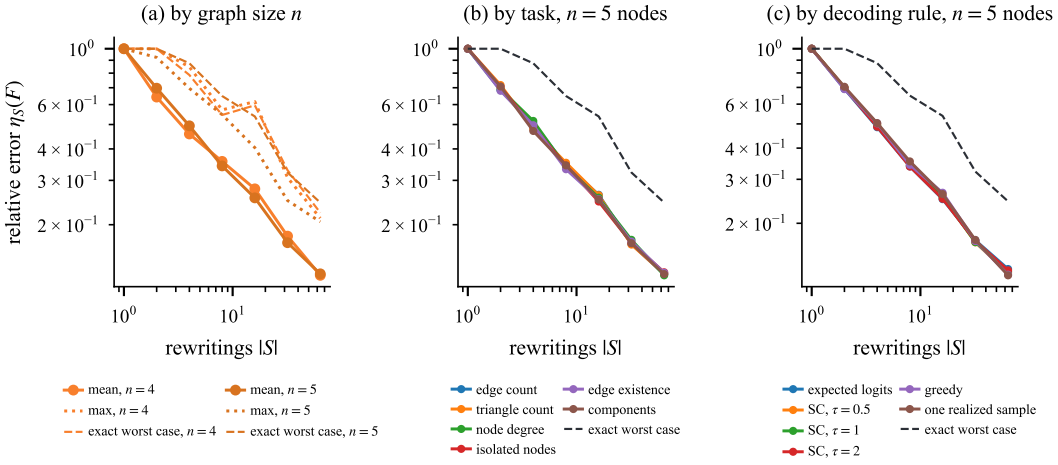


Figure 6: Relative error $\eta_S(F)$ for Qwen2.5-1.5B fields with one shared S per size and draw. (a) Mean and maximum over fields, and the exact worst case of Theorem 2, for graphs with $n = 4$ nodes ($|G| = 24$) and $n = 5$ nodes ($|G| = 120$). (b) Mean by task and (c) by decoding rule at $n = 5$. The curves are nearly indistinguishable, as Corollary 3 predicts.

majority-class baseline, the accuracy of always giving the most frequent answer, and the last column reports the fraction of graphs whose greedy answer changes across the 16 rewritings. Figure 7 shows the accuracy and the gain against self consistency at every budget.

The table groups the tasks by the three cases discussed in Section 5. Whether the answer changes across the orbit is necessary for a gain but does not determine its size or sign: node degree changes as often as edge count, yet its errors do not cancel under averaging, and on edge existence the small loss is concentrated on exactly the graphs whose answer changes. What matters is whether the orbit-averaged decision is more accurate than a typical representation, the term R_G of Corollary 4. On isolated-node counting the model outputs a single answer for all 1,600 prompts, so its accuracy is the majority-class baseline; on component counting it is below the baseline at every serialization. No aggregation over rewritings can help when no serialization carries the right answer.

The comparisons use a uniformly random starting serialization, as in the definition of R_S. A fixed canonical serialization (identity labels, sorted edges) is a different, non-uniform choice outside our analysis; on these 100-graph sets its greedy accuracy is 0.16 (edge count), 0.14 (triangle count), 0.27

Table 1: Accuracy at $k = 16$ forward passes (graphs with $n = 7$ nodes, 100 graphs per task). "Greedy" and "SC" decode at one uniformly random rewriting; "Sym." averages the answer distributions of 16 rewritings. Gains in accuracy points with 95% bootstrap intervals over graphs.

Task	Base	Greedy	SC	Sym.	Sym. − Greedy	Sym. − SC	Changes
Partial competence that varies across the orbit							
Edge count	0.12	0.199	0.199	**0.240**	+4.1 [+1.3, +7.0]	+4.1 [+0.9, +7.5]	70%
Triangle count	0.10	0.137	0.126	**0.160**	+2.3 [+0.6, +4.3]	+3.4 [−0.5, +7.5]	44%
Competent, errors do not cancel over the orbit							
Node degree	0.15	0.283	0.287	0.290	+0.7 [−2.3, +3.8]	+0.3 [−3.1, +3.7]	72%
Edge existence	0.50	0.710	0.706	0.690	−2.0 [−4.6, +0.5]	−1.6 [−4.0, +0.8]	31%
No competence at any serialization							
Isolated nodes	0.17	0.170	0.170	0.170	0.0	0.0	0%
Components	0.17	0.062	0.084	0.060	−0.2 [−1.2, +0.8]	−2.4 [−4.8, +0.1]	17%
Pooled		0.260	0.262	0.268	+0.8 [−0.1, +1.7]	+0.7 [−0.5, +1.8]	

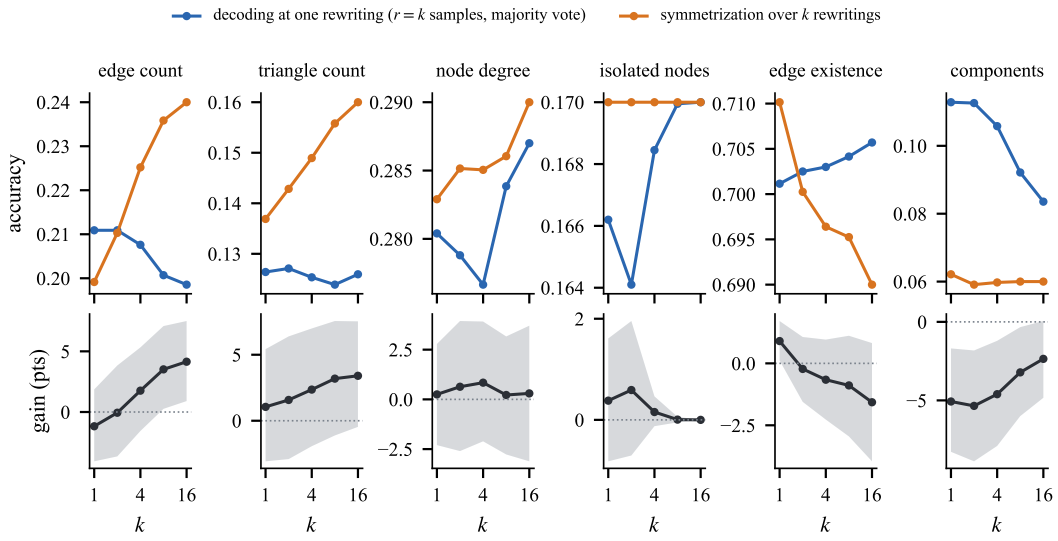


Figure 7: Top: accuracy against the number of forward passes k for self consistency at one random rewriting ($r = k$, blue) and symmetrization over k rewritings (orange), for all six tasks. Bottom: paired gain with 95% bootstrap intervals over graphs.

(node degree), 0.77 (edge existence), 0.17 (isolated nodes), and 0.05 (components), so it is better than a random serialization only on edge existence.

H.5 Allocating rewritings and samples

Figure 8 spends a budget of $k = m \cdot r$ forward passes on m distinct random rewritings with r sampled answers each, pooled by majority vote, and reports the accuracy relative to $m = r = 1$. The column "p" uses the exact answer distribution at each rewriting, which our setup obtains from a single forward pass. The column $r = 1$ is joint sampling of (rewriting, answer) pairs (Section C). On the two counting tasks, adding rewritings at $r = 1$ does not improve accuracy at these budgets, whereas averaging the exact answer distribution of each rewriting (column "p") does, by up to +3.0 and +3.7 points; on edge count, increasing r at $m = 16$ recovers most of that gain (+2.4 points at $r = 16$). This is the behavior predicted by the decomposition Equation (19): the finite-decoder term decreases with the number of samples per rewriting and not with $|S|$, while the rewriting term decreases with $|S|$. Because majority voting is linear, joint sampling is unbiased here (Equation (17)), so the gap is a variance effect at small budgets rather than a change of target. On edge existence and components, adding rewritings lowers accuracy, consistent with the orbit-averaged decision being less accurate

there. Each cell averages 50 random draws of rewritings and samples on 100 graphs, so differences of about one point are within simulation noise.

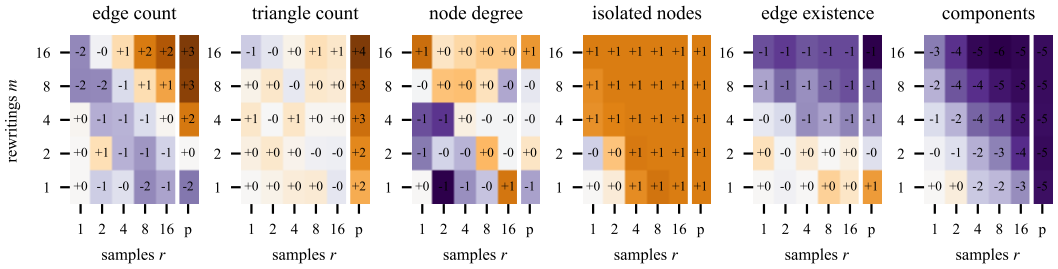


Figure 8: Accuracy change (points) relative to one rewriting with one sample, for m rewritings and r sampled answers per rewriting; "p" is the exact answer distribution per rewriting (cost m). Color scales are per task.

I NOTATION

This table collects the notation used in the paper and the place where each symbol is introduced.

Symbol	Meaning	Introduced
$\mathcal{X}$	space of prompts	Section 3
$\mathcal{C}, C = \lvert\mathcal{C}\rvert$	set of answer classes and its size	Section 3
$y^\star : \mathcal{X} \to \mathcal{C}$	target function, invariant: $y^\star(g \cdot x) = y^\star(x)$	Section 3
$G, \lvert G\rvert, e$	finite group of rewritings, its order, and its identity	Section 3
$g \cdot x$	rewriting (serialization) of prompt x by $g \in G$	Section 3
$G \cdot x$	orbit of x: all rewritings of the same instance	Section 3
h	starting rewriting from which inference is invoked	Equation (7)
$\rho(g)$	permutation of answer classes induced by g (equivariant targets)	Section B
X	prompt drawn from the task distribution	Section 4.2
$p_{\theta,\tau}(\cdot \mid x)$	frozen model: distribution of reasoning traces given prompt x	Equation (1)
τ	generation temperature	Equation (1)
$\omega \in \Omega$	a sampled reasoning trace and the trace space	Equation (1)
$V_x(g, \omega) \in \mathcal{Z}$	response field: response of trace ω at rewriting $g \cdot x$	Equation (1)
$a_x(g, \omega), r_x(g, \omega)$	answer extracted from a trace and its verifier score	Equation (5)
$\mathcal{D}, \mathcal{H}$	decoding rule and the real Hilbert space of its outputs	Equation (2)
$F_x^{\mathcal{D}}(g)$	decoded orbit field: output of $\mathcal{D}$ at rewriting g	Equation (2)
$F_x^{\mathrm{SC}}, F_x^{\mathrm{logit}}, F_x^{\beta_D}$	self consistency, expected logits, verifier weighted decoding	Section 3
β_D	decoder (verifier) inverse temperature	Equation (5)
$q_x^{\mathcal{D}} = \mathsf{P}_G F_x^{\mathcal{D}}$	fully symmetrized response	Equation (6)
$\widehat{F}_{x,\xi}, \xi$	field of a realized randomized decoder and its seed	Section B
n_D	number of traces per rewriting in finite self consistency	Section C
$L^2(G; \mathcal{H})$	$\mathcal{H}$-valued functions on G with the uniform inner product	Section B
$S = \{\{s_1, \ldots, s_{\lvert S\rvert}\}\}$	multiset of rewritings, drawn i.i.d. from $\mathrm{Unif}(G)$	Equation (7)
A_S	sparse rewriting operator $(\mathsf{A}_S f)(h) = \frac{1}{\lvert S\rvert} \sum_i f(s_i h)$	Equation (7)
P_G	full group average, the orthogonal projection onto constants	Equation (8)
T_s	translation operator $(\mathsf{T}_s f)(h) = f(sh)$	Equation (13)
$\lVert\cdot\rVert_{\mathrm{op}}$	operator norm on $L^2(G)$	Equation (15)
δ	failure probability of the random draw of S	Theorem 2
$\pi, \mathcal{U}(\mathcal{H})$	unitary representation of G and the unitary group of $\mathcal{H}$	Section B
$\mathcal{H}^G, \mathsf{P}_\pi$	invariant vectors of π and the projection onto them	Section B
λ	left regular representation of G	Section B

Symbol	Meaning	Introduced
$\widehat{G}$, d_ρ	irreducible unitary representations and the dimension of $\rho \in \widehat{G}$	Section B
$c_G(X)$	decision of the fully symmetrized model, $\arg\max_c (q_X)_c$	Section 4.2
$\widehat{c}_S(h \cdot X)$	decision of sparse symmetrization from starting rewriting h	Section 4.2
Δ_X	margin between the two largest coordinates of q_X	Section 4.2
R_G, R_S	error of full and of sparse symmetrization, $h \sim \mathrm{Unif}(G)$	Section 4.2
t	margin threshold in the error decomposition	Corollary 4
n	number of nodes of a graph; node relabelings form $G = S_n$, $\lvert G\rvert = n!$	Section 5
$\mathbb{Z}_2^d$	group of d binary flips, $\lvert G\rvert = 2^d$ (the family of the lower bound)	Section 5
$\eta_S(F)$	relative error $\lVert \mathsf{A}_S F - \mathsf{P}_G F\rVert / \lVert F - \mathsf{P}_G F\rVert$	Section 5
k	budget of forward passes per question	Section 5
m, r	number of rewritings and number of sampled answers per rewriting, $k = m \cdot r$	Section H.5