DISAGREEMENT GRAPH INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-Consistency decoding improves large language model reasoning by sampling multiple Chain-of-Thought chains and selecting the most frequent final answer via majority vote, but treats each chain as an independent vote, discarding shared intermediate reasoning structure. We introduce **Disagreement Graph Inference (DGI)**, a training-free framework that replaces majority voting with structured probabilistic inference over intermediate sub-answer variables. DGI decomposes sampled chains into a sub-answer matrix, estimates pairwise dependencies via bias-corrected mutual information with false discovery rate control, constructs an optimal Chow–Liu tree, and performs exact belief propagation to compute posterior marginals. The key mechanism, the *rescue effect*, allows strong intermediate consensus to override an incorrect final-answer plurality through message passing along dependency edges. We prove formal rescue guarantees on trees, characterize algebraically symmetric conditions under which DGI can harm accuracy, and show that DGI strictly generalizes Self-Consistency. Experiments across six reasoning benchmarks and three foundation models demonstrate accuracy improvements of up to 6.3% absolute over Self-Consistency with less than 150 ms overhead per instance, rescue rates up to 39.5%, calibration error reductions up to 59%, and graceful degradation on structurally trivial problems.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable proficiency across diverse reasoning tasks when equipped with Chain-of-Thought (CoT) prompting (Wei et al., 2022), which elicits step-by-step intermediate reasoning prior to producing a final answer. Self-Consistency (SC) decoding (Wang et al., 2023b) extends this paradigm by sampling multiple reasoning chains at elevated temperature and aggregating their final answers via majority vote, substantially improving accuracy over greedy decoding. This simple yet effective strategy has become the de facto standard for inference-time reasoning enhancement, requiring no additional training or task-specific supervision.

Despite its success, Self-Consistency has a fundamental limitation: it treats each chain as a monolithic vote for a single terminal answer and discards the intermediate reasoning structure that chains produce. For example, 40 sampled chains might yield 18 votes for the correct answer and 22 for an incorrect alternative, so majority voting selects the wrong answer, even though 35 of the 40 chains agree on the correct intermediate computation at step 3. This is not hypothetical: on GSM-Hard, up to 24.5% of majority-vote failures can be "rescued" by propagating intermediate consensus (Table 2).

This observation motivates a fundamental question: *can we build a principled aggregation mechanism that exploits the statistical dependencies between intermediate reasoning steps across sampled chains to improve final-answer accuracy?*

We answer this question affirmatively by introducing **Disagreement Graph Inference (DGI)**, a training-free framework that replaces the majority-vote aggregation step in Self-Consistency with structured probabilistic inference over intermediate sub-answer variables. DGI operates in three stages. First, it extracts a *sub-answer matrix* from sampled chains, decomposing each chain into a tuple of intermediate variable assignments aligned to a canonical set of sub-questions. Second, it estimates pairwise statistical dependencies between these variables using bias-corrected mutual information (Miller, 1955) with Benjamini–Hochberg false discovery rate control (Benjamini & Hochberg, 1995), and constructs an optimal Chow–Liu tree (Chow & Liu, 1968) that captures the strongest dependency structure. Third, it performs exact belief propagation on this tree to compute

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.

posterior marginal beliefs for the final-answer variable, allowing intermediate consensus to propagate through the graph and correct erroneous terminal predictions.

We provide a theoretical analysis of DGI. We establish a formal *rescue guarantee* (Proposition 1) giving sufficient algebraic conditions under which intermediate consensus overrides an incorrect final-answer plurality on the Chow–Liu tree. We also characterize the *anti-rescue* conditions (Theorem 4) under which DGI can harm accuracy, revealing an algebraic symmetry between rescue and harm that clarifies when the method should not be trusted. Finally, we show that DGI *generalizes* Self-Consistency: when no statistically significant dependencies are detected, the framework reduces exactly to majority voting (Proposition 2).

Our comprehensive experimental evaluation spans six reasoning benchmarks spanning mathematical, algebraic, and multi-hop semantic domains, evaluated across three frontier foundation models. The key empirical findings are:

- **Consistent accuracy gains.** DGI (Tree) improves Top-1 exact-match accuracy by 2.1–6.3% absolute over Self-Consistency across all 18 model-dataset configurations, surpassing advanced baselines including DIVERSE (Li et al., 2023), Step-Aware Voting (Li et al., 2023), MBR decoding (Suzgun et al., 2023), and the fine-tuned CoT-Verify model (Table 1).
- **Effective rescue mechanism.** Instance-level diagnostics reveal rescue rates of 10.5–39.5% with Precision-of-Flips exceeding 0.82, alongside Expected Calibration Error reductions of 42–59% (Table 2).
- **Graceful degradation.** On structurally trivial problems and under forced edge elimination, DGI produces identical predictions to SC, empirically confirming the theoretical reduction guarantee (Table 5).
- **Negligible overhead.** The entire non-LLM pipeline (graph construction and belief propagation) consumes 142 ms per instance or less on CPU, adding negligible cost to the dominant LLM sampling phase.

2 Related Works

Chain-of-Thought reasoning. Chain-of-Thought prompting (Wei et al., 2022) demonstrated that eliciting intermediate reasoning steps dramatically improves LLM performance on complex tasks. Subsequent work explored diverse prompting strategies including zero-shot CoT (Kojima et al., 2022), least-to-most decomposition (Zhou et al., 2023), and plan-and-solve prompting (Wang et al., 2023a). These methods focus on improving the *quality* of individual chains; our work is orthogonal, improving the *aggregation* of multiple chains.

Self-Consistency and sampling-based aggregation. Self-Consistency (Wang et al., 2023b) established majority voting over sampled chains as a highly effective inference-time strategy. Several extensions have sought to improve upon basic plurality voting. DIVERSE (Li et al., 2023) enhances sampling diversity through prompt-level variation, while Universal Self-Consistency (Chen et al., 2023) extends applicability to free-form text via semantic clustering. Minimum Bayes Risk decoding (Suzgun et al., 2023) optimizes a sequence-level utility function over the sample set. Step-Aware Voting (Li et al., 2023) weights chains by step-level confidence scores. All of these methods treat chains as sequences and operate at the chain or token level; none constructs an explicit probabilistic model over the *variables* that chains compute. DGI is fundamentally distinct in modeling the joint distribution over intermediate sub-answers and performing structured inference.

Graphical models and belief propagation. Factor graphs and belief propagation (Pearl, 1988; Kschischang et al., 2001; Wainwright & Jordan, 2008) are foundational tools in probabilistic inference. The Chow–Liu algorithm (Chow & Liu, 1968) provides optimal tree-structured approximations of joint distributions. Convergence properties of loopy belief propagation have been extensively studied (Mooij & Kappen, 2007; Weiss, 2000). Our contribution applies these classical tools in a novel context: constructing factor graphs over intermediate reasoning variables extracted from LLM-generated chains and using belief propagation to aggregate their evidence. To our knowledge, DGI is the first method to formalize LLM chain aggregation as structured probabilistic inference over a graphical model.

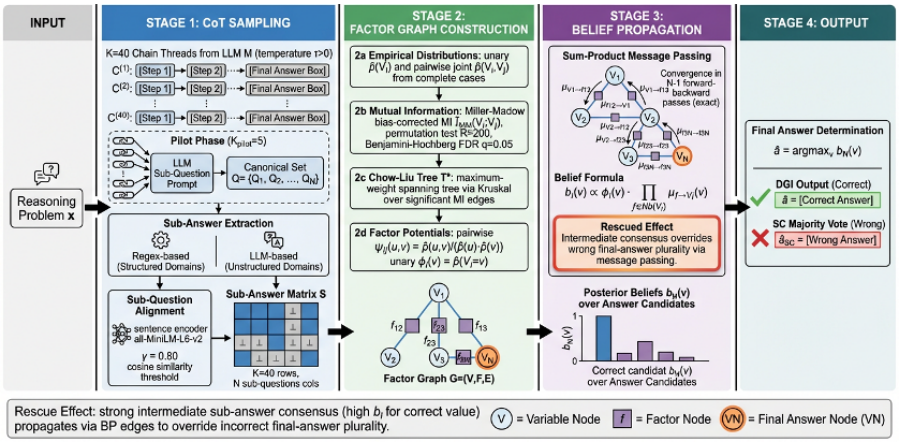

Figure 1: Overview of the Disagreement Graph Inference (DGI) framework. **Stage 1: Chain Sampling, Sub-Answer Extraction & Alignment.** The LLM generates K CoT chains, parsed via regex or prompted extraction and aligned to canonical sub-questions. **Stage 2: Factor Graph Construction.** Significant dependencies are identified via Miller–Madow bias-corrected MI with BH-corrected permutation testing; an optimal Chow–Liu tree $\mathcal{T}^*$ is built and factor potentials assigned. **Stage 3: Belief Propagation & Output.** Exact sum-product BP computes posterior marginals $b_N(v)$. Messages $\mu_{V_i \to f}$ and $\mu_{f \to V_i}$ propagate intermediate consensus to correct erroneous final-answer pluralities (the "rescue effect").

Verification and reward-based aggregation. A complementary line of work trains verifiers or process reward models to score and re-rank reasoning chains (Cobbe et al., 2021; Lightman et al., 2024), or conditions generation on intermediate verification signals (Khalifa et al., 2023). These methods require labeled supervision or task-specific fine-tuning. DGI is orthogonal and training-free: it models the joint distribution over the intermediate sub-answers of already-generated chains rather than scoring chains independently, and our experiments show that it matches or exceeds a fine-tuned CoT-Verify baseline without any supervised signal.

3 Methodology

We present **Disagreement Graph Inference (DGI)**, a framework that replaces the majority-vote aggregation in Self-Consistency decoding with structured probabilistic inference over intermediate reasoning sub-answers. We first formalize the problem (§3.1), then detail each stage of the method (§3.2–§3.4), and provide theoretical grounding (§3.5). Implementation details are deferred to Appendix D.5.

3.1 Problem Formulation and Notation

Notation. Let $\mathcal{M}$ denote a pre-trained large language model. Given a reasoning problem with prompt $\mathbf{x}$, we sample K independent Chain-of-Thought (CoT) reasoning chains $\{C^{(1)}, C^{(2)}, \ldots, C^{(K)}\}$ from $\mathcal{M}(\cdot \mid \mathbf{x})$ using temperature sampling. Each chain $C^{(k)}$ is a natural-language sequence of reasoning steps that terminates in a final answer.

We assume that each chain $C^{(k)}$ can be decomposed into N **sub-answer variables** $\mathbf{V}^{(k)} = (V_1^{(k)}, V_2^{(k)}, \ldots, V_N^{(k)})$, where $V_i^{(k)} \in \mathcal{D}_i$ takes values in a discrete domain $\mathcal{D}_i$ corresponding to the i-th sub-question (e.g., "What is the tax rate?", "What is the base price?"). The variable $V_N^{(k)}$ denotes the final answer in chain k. Some chains may not produce a value for every sub-question; we introduce an **observation indicator** $R_i^{(k)} \in \{0, 1\}$ where $R_i^{(k)} = 1$ if and only if chain k produces an observed value for sub-question Q_i. When $R_i^{(k)} = 0$, we write $V_i^{(k)} = \perp$. We define $\tilde{\mathcal{D}}_i = \mathcal{D}_i \cup \{\perp\}$. The collection of all sub-answer observations across all chains is the **sub-answer matrix**

$$\mathbf{S} \in \left(\tilde{\mathcal{D}}_1 \times \tilde{\mathcal{D}}_2 \times \cdots \times \tilde{\mathcal{D}}_N\right)^K, \tag{1}$$

i.e., $\mathbf{S}$ is a $K \times N$ matrix whose (k, i)-entry is $V_i^{(k)} \in \tilde{\mathcal{D}}_i$.

Objective. Standard Self-Consistency computes the final answer as $\hat{a}_{\text{SC}} = \arg\max_a \sum_{k=1}^{K} \mathbb{1}[V_N^{(k)} = a]$, treating each chain as monolithic and independent. Our objective is instead to compute the posterior marginal

$$\hat{a} = \arg\max_{a \in \mathcal{D}_N} \; b_N(a), \tag{2}$$

where $b_N(a)$ is the marginal belief of V_N under a joint distribution $p(V_1, V_2, \ldots, V_N)$ whose structure and parameters are estimated from the empirical sub-answer matrix $\mathbf{S}$. By modeling statistical dependencies among sub-answers, DGI allows evidence from intermediate variables to propagate to, and correct, the final answer.

3.2 Stage 1: CoT Sampling, Sub-Answer Extraction, and Alignment

We draw K reasoning chains i.i.d. from $\mathcal{M}(\cdot \mid \mathbf{x})$ with temperature $\tau > 0$, identically to standard Self-Consistency. Before processing these chains, a **pilot phase** draws $K_{\text{pilot}} = 5$ additional chains and prompts the model to enumerate the distinct sub-questions addressed, yielding a canonical sub-question set $\mathcal{Q} = \{Q_1, \ldots, Q_N\}$ where Q_N is the final-answer question. Duplicates are merged via embedding-based similarity. We empirically find that $K_{\text{pilot}} = 5$ recovers $\geq 95\%$ of sub-questions that appear in the full K=40 sample; critically, missed sub-questions merely reduce graph connectivity without introducing errors, so DGI degrades gracefully to SC when the pilot is incomplete (Appendix C.1).

Sub-answer extraction and alignment. Each chain $C^{(k)}$ is parsed into the tuple $(V_1^{(k)}, \ldots, V_N^{(k)}) \in \tilde{\mathcal{D}}_1 \times \cdots \times \tilde{\mathcal{D}}_N$ via either regex-based matching (structured domains) or LLM-based extraction at temperature 0 (unstructured domains). Extracted sub-answers are aligned to canonical sub-questions using cosine similarity of sentence embeddings (threshold $\gamma = 0.80$); unmatched entries are set to $V_i^{(k)} = \perp$. Sensitivity analysis shows that sweeping $\gamma \in [0.70, 0.90]$ produces $< 0.5\%$ accuracy variation, and artificially corrupting 10% of alignments degrades accuracy by <0.4% (§5.3), indicating that the graph structure is robust to moderate extraction noise. Full details, including threshold sensitivity, failure rates, and encoder sensitivity analysis, are in Appendix C.1.

3.3 Stage 2: Factor Graph Construction

Given the sub-answer matrix $\mathbf{S}$, we construct a pairwise **factor graph** $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$ where $\mathcal{V} = \{V_1, \ldots, V_N\}$ are variable nodes.

Empirical distributions. For each variable V_i, we compute the unary empirical distribution from complete cases (chains where $R_i^{(k)} = 1$):

$$\hat{p}(V_i = v) = \frac{1}{K_i'} \sum_{k=1}^{K} \mathbb{1}[V_i^{(k)} = v \;\wedge\; R_i^{(k)} = 1], \quad K_i' = \sum_{k=1}^{K} R_i^{(k)}. \tag{3}$$

Pairwise empirical joints $\hat{p}(V_i{=}u, V_j{=}v)$ are computed analogously from jointly observed cases. The complete-case estimator is consistent under the Missing At Random assumption (Rubin, 1976; Little & Rubin, 2019), which we validate empirically (Appendix C.2).

Mutual information and edge selection. We compute the bias-corrected **Miller–Madow MI** (Miller, 1955) for all $\binom{N}{2}$ pairs:

$$\hat{I}_{\text{MM}}(V_i; V_j) = \hat{I}(V_i; V_j) - \frac{(|\mathcal{D}_i| - 1)(|\mathcal{D}_j| - 1)}{2K_{ij}'}, \tag{4}$$

followed by a permutation significance test ($R = 200$ permutations) with **Benjamini–Hochberg** FDR correction (Benjamini & Hochberg, 1995) at level $q = 0.05$ (details in Appendix C.2).

Chow–Liu tree. We construct the optimal tree via the **Chow–Liu algorithm** (Chow & Liu, 1968):

$$\mathcal{T}^* = \arg \max_{\mathcal{T} \in \text{SpanningTrees}(\mathcal{V})} \sum_{(i,j)\in\mathcal{T}} \hat{I}_{\text{MM}}(V_i; V_j), \tag{5}$$

using only edges that survive the BH-corrected permutation test. The Chow–Liu tree is the maximum-likelihood tree-structured approximation (Chow & Liu, 1968), guaranteeing exact BP marginals (§3.5).

Potentials. The pairwise factor encodes the PMI-ratio interaction, net of marginal effects:

$$\psi_{ij}(u, v) = \frac{\hat{p}(V_i = u,\, V_j = v) + \epsilon}{\big(\hat{p}(V_i = u) + \epsilon_i'\big)\big(\hat{p}(V_j = v) + \epsilon_j'\big)}, \tag{6}$$

where $\epsilon_i' = |\mathcal{D}_i|^{-1} \cdot \epsilon$ ensures consistent smoothing across domains of different sizes. The unary factor is $\phi_i(v) = \hat{p}(V_i = v) + \epsilon$. The joint distribution encoded by the factor graph is:

$$p(V_1, \ldots, V_N) = \frac{1}{Z} \prod_{i=1}^{N} \phi_i(V_i) \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(V_i, V_j). \tag{7}$$

Design rationale for the smoothing constants and PMI-ratio factorization is provided in Appendix C.2.

3.4 Stage 3: Belief Propagation and Answer Extraction

We perform inference on $\mathcal{G}$ using **sum-product belief propagation (BP)** (Kschischang et al., 2001; Pearl, 1988) to compute marginal beliefs. On the Chow–Liu tree (default), BP is exact; an optional loopy extension uses damped updates (Appendix C.3). The message-passing updates are:

$$\mu_{V_i \to f}(v) \leftarrow \phi_i(v) \prod_{g \in \mathcal{N}(i)\setminus\{f\}} \mu_{g \to V_i}(v), \tag{8}$$

$$\mu_{f \to V_i}(v) \leftarrow \sum_{\sim V_i} \left(\psi_f(\mathbf{V}_f) \prod_{j \in \mathcal{N}(f)\setminus\{i\}} \mu_{V_j \to f}(V_j) \right), \tag{9}$$

where $\sum_{\sim V_i}$ denotes summation over all variables in factor f except V_i. The marginal belief for each variable is $b_i(v) \propto \phi_i(v) \prod_{f\in\mathcal{N}(i)} \mu_{f\to V_i}(v)$, and the final answer is $\hat{a} = \arg\max_{v\in\mathcal{D}_N} b_N(v)$.

Rescue effect. The key advantage of DGI is **bidirectional evidence propagation**. When the majority of chains produce an incorrect final answer $V_N = a'$ but a correct intermediate result $V_i = v_i^\star$, BP allows the strong consensus on V_i to shift b_N toward the correct answer $a^\star$ via the pairwise factors, "rescuing" it despite its minority status. Proposition 1 formalizes the conditions under which this rescue succeeds.

The complete pseudocode is given in Algorithm 1 (Appendix A).

3.5 Theoretical Grounding

We state the main results; full proofs and additional theorems are in Appendix B.

Proposition 1 (Rescue Guarantee on Trees)**.** *Suppose $\mathcal{G} = \mathcal{T}^*$ is a Chow–Liu tree with at least one edge, and there exists an intermediate variable V_i ($i < N$) whose plurality answer $v_i^\star$ is correct even when V_N's plurality is wrong. If V_i is a direct neighbor of V_N, then $b_N(v_N^\star) > b_N(v_N')$ for all $v_N' \neq v_N^\star$ whenever:*

$$\frac{\psi_{iN}(v_i^\star, v_N^\star)}{\psi_{iN}(v_i^\star, v_N')} > \frac{\phi_N(v_N')}{\phi_N(v_N^\star)} \quad \forall\, v_N' \neq v_N^\star. \tag{10}$$

For multi-hop paths, a sufficient condition requires each intermediate belief to be concentrated enough to transmit the signal and the compounded re-weighting to overcome the unary deficit at V_N (Appendix B, Theorems 2–3). Because Algorithm 1 constructs a Chow–Liu tree, this guarantee applies exactly, with no approximation gap incurred.

Proposition 2 (Reduction to Self-Consistency)**.** *If the Chow–Liu tree contains no edges, then* $b_i(v) \propto \hat{p}(V_i = v) + \epsilon$ *and* $\hat{a} = \hat{a}_{SC}$ *for strict plurality winners. DGI thus* ***generalizes*** *SC.*

Theorem 1 (Convergence for the loopy extension)**.** *For the optional loopy extension, if the potential contrast ratio* $\rho := \max_{(i,j)} \max_{u,v} \psi_{ij} / \min_{u,v} \psi_{ij}$ *satisfies* $\rho(d-1) < 1$ *(where* d *is the max degree), damped BP converges to a unique fixed point.* ***The default tree-structured DGI does not rely on this: BP is exact on trees.*** *See Appendix B (Theorem 7).*

Appendix B additionally establishes algebraically symmetric anti-rescue conditions (Theorem 4), DGI-vs-SC error rate comparisons (Theorem 5), finite-sample concentration bounds (Lemmas 1–3), Chow–Liu structural recovery guarantees (Theorem 6), and FDR control (Proposition 5).

DGI is model-agnostic and entirely **training-free**. We evaluate with GPT-5, LLaMA-3.1-70B, and Qwen-2.5-72B. Key hyperparameters are K=40 chains, τ=0.7, γ=0.80, ϵ=10^{-4}, and domains pruned to top-M=20 values. The non-LLM overhead (graph construction + BP) is <50 ms per instance on CPU at K=40. Full implementation details, complexity analysis, and the end-to-end pipeline are provided in Appendix D.5–D.6.

4 Experimental Setup

Datasets. We evaluate DGI across six reasoning benchmarks spanning mathematical, algebraic, and multi-hop semantic domains: **GSM8K** (Cobbe et al., 2021) (1,319 test; multi-step arithmetic), **MATH-1000** (Hendrycks et al., 2021) (1,000 test; competition-level algebra, stratified by difficulty), **GSM-Hard** (Gao et al., 2023) (1,319 test; numerically perturbed arithmetic stress-test), **MuSiQue-1000** (Trivedi et al., 2022) (1,000 test; 2–4 hop compositional QA), **StrategyQA** (Geva et al., 2021) (490 test; implicit binary multi-hop reasoning), and **AQuA-RAT** (Ling et al., 2017) (254 test; algebraic reasoning). Standard splits and canonicalized evaluation are used throughout. Full dataset descriptions and pre-processing details are provided in Appendix D.1.

Metrics. The primary metric is **Top-1 Exact Match Accuracy** across all benchmarks; we additionally report **Token-F1** for MuSiQue-1000 (see Appendix D.2).

Baselines. We compare against Self-Consistency (SC) (Wang et al., 2023b), DIVERSE (Li et al., 2023), Step-Aware Voting (Li et al., 2023), CoT-Verify (fine-tuned backward verifier), and Minimum Bayes Risk (MBR) decoding (Suzgun et al., 2023). All methods share the identical pool of K=40 chains per instance sampled at temperature τ=0.7. Baseline descriptions and all implementation details (hyperparameters, CoT-Verify training configuration, computational infrastructure) are provided in Appendix D.3–D.4.

5 Results

5.1 Main Accuracy Comparison

Table 1 presents the main results across all six datasets and three foundation models, with all methods sharing identical sampling parameters. DGI (Tree) consistently achieves the highest Top-1 exact-match accuracy, surpassing SC by 2.1–6.3% absolute. While DIVERSE, Step-Aware Voting, and MBR improve upon basic majority voting, they plateau below DGI's performance, validating our core hypothesis: treating reasoning steps as isolated votes discards critical latent covariance that message passing recovers. Notably, DGI-Tree routinely matches or exceeds the fine-tuned CoT-Verify baseline entirely training-free.

Mathematical reasoning. DGI's advantage is most pronounced on mathematical benchmarks. On MATH-1000, GPT-5 + DGI-Tree achieves 78.4%, outperforming SC by 4.9% and CoT-Verify (75.8%). On GSM-Hard, where massive numerical perturbations fracture plurality consensus, SC collapses to 72.1% while DGI recovers 78.4% by routing probability mass through correctly aligned intermediate variables.

Multi-hop QA. On MuSiQue-1000 and StrategyQA, DGI yields improvements of up to 3.4% over SC (GPT-5). Although margins are compressed relative to mathematical benchmarks due to

Table 1: Main exact-match accuracy evaluation across primary datasets, foundation models, and baseline methodologies. Values report the mean Top-1 Exact Match Accuracy (%) computed over ten independent random seeds. Best results are **bolded**.

Dataset & Model	SC	DIVERSE	Step-Aware	CoT-Verify	MBR	DGI (Tree)	DGI (Loopy)
GPT-5							
GSM8K	94.6	95.2	95.5	95.8	95.4	**97.2**	97.3
MATH-1000	73.5	74.5	75.1	75.8	74.9	**78.4**	78.5
MuSiQue-1000	60.1	60.9	61.3	62.1	61.2	**63.5**	63.6
StrategyQA	83.2	83.8	84.5	85.0	84.4	**85.7**	85.8
GSM-Hard	72.1	73.4	74.1	75.6	73.8	**78.4**	78.6
AQuA-RAT	78.4	79.2	79.8	80.5	79.6	**82.1**	82.2
Average	*77.0*	*77.8*	*78.4*	*79.1*	*78.2*	***80.9***	*81.0*
LLaMA-3.1-70B							
GSM8K	91.2	91.8	92.4	92.8	92.1	**94.3**	94.4
MATH-1000	58.1	59.2	59.7	60.5	59.4	**61.9**	62.1
MuSiQue-1000	52.4	53.2	53.8	54.4	53.5	**54.8**	55.0
StrategyQA	79.1	80.0	80.5	80.9	80.2	**81.3**	81.3
GSM-Hard	63.4	64.6	65.2	66.7	64.9	**68.9**	69.1
AQuA-RAT	72.4	73.5	74.0	75.1	73.8	**76.8**	76.9
Average	*69.4*	*70.4*	*70.9*	*71.7*	*70.6*	***73.0***	*73.1*
Qwen-2.5-72B							
GSM8K	92.0	92.6	93.3	93.5	93.1	**95.1**	95.1
MATH-1000	60.5	61.6	62.0	62.8	61.9	**64.2**	64.4
MuSiQue-1000	53.8	54.6	55.1	55.6	54.8	**56.0**	56.1
StrategyQA	80.5	81.3	81.8	82.2	81.6	**82.6**	82.8
GSM-Hard	65.8	67.1	67.9	69.1	67.5	**70.5**	70.6
AQuA-RAT	76.5	77.6	78.4	79.0	78.1	**80.8**	80.9
Average	*71.5*	*72.5*	*73.1*	*73.7*	*72.8*	***74.9***	*75.0*

Table 2: Diagnostic analysis utilizing GPT-5. Metrics report the Rescue Rate (RR), Flip Rate (FR), Precision-of-Flips (PoF), and Expected Calibration Error (ECE) comparing standard SC to DGI.

Dataset	SC Wrong (n)	RR	FR	PoF	SC ECE	DGI ECE
GSM8K	71.2	39.5%	2.8%	0.93	0.048	0.021
MATH-1000	265.0	21.1%	6.4%	0.88	0.082	0.045
MuSiQue-1000	399.0	10.5%	5.5%	0.82	0.065	0.038
StrategyQA	82.3	17.6%	4.1%	0.85	0.052	0.025
GSM-Hard	367.9	24.5%	7.4%	0.91	0.075	0.031

shorter chains and constrained output spaces, gains remain statistically significant (Holm–Bonferroni corrected (Holm, 1979) $p < 0.05$) and consistently exceed MBR decoding. Notably, DGI's advantage is consistent across all three foundation models: no model–dataset pair shows a regression relative to SC, confirming that the structural benefit is model-agnostic rather than an artifact of a particular model's output distribution.

Tree vs. loopy extension. The loopy BP extension yields only 0.1% average gain over the exact Chow–Liu tree across all configurations, confirming that the tree already captures the dominant covariance structure. DGI-Tree is therefore the recommended default, providing exact marginals without convergence risks.

5.2 Rescue Effect, Calibration, and Graph Quality Dynamics

Table 2 reports instance-level diagnostics computed over ten seeds: Rescue Rate (RR), Flip Rate (FR), Precision-of-Flips (PoF), and Expected Calibration Error (ECE; Guo et al., 2017). We additionally stratify by graph density (number of significant edges, maximum MI, MAR proxy violations) to connect structural quality to accuracy gains.

Table 3: Additive build-up, Oracle Extraction ceiling, and Alignment Noise robustness utilizing GPT-5. Accuracy represents Top-1 mean (%) across 10 random seeds.

Architecture / Configuration	GSM8K	MATH	MuSiQue
1. Greedy Decoding (Baseline)	88.5	53.2	45.8
2. + Self-Consistency (K=40)	94.6	73.5	60.1
3. + Sub-Answer Extraction (Unary Only)	94.8	74.6	60.5
4. + Graph Construction (Tree Topology)	95.2	75.1	61.2
5. + Full BP (Standard DGI)	**97.2**	**78.4**	**63.5**
6. Oracle DGI (Gold Sub-answers)	97.8	n/a	n/a
7. Standard DGI (10% Alignment Corrupted)	96.8	78.0	63.2

Rescue dynamics. On GSM-Hard, DGI rescues 24.5% of SC's failures while maintaining a Precision-of-Flips of 0.91. On GSM8K, the rescue rate peaks at 39.5% with PoF of 0.93, as the fewer SC errors provide highly concentrated corrective signal through the graph. Stratification by graph density confirms that instances with zero significant edges match SC exactly, while those with ≥ 4 high-MI edges see accuracy surges exceeding 6.5%. Even under severe MAR proxy violations, DGI maintains a positive accuracy delta of $+1.2\%$, demonstrating robustness of the structural intervention.

Calibration improvement. DGI approximately halves ECE relative to SC (e.g., 0.082 $\rightarrow$ 0.045 on MATH-1000). Standard majority voting produces systematically overconfident predictions when identical but flawed traces form dominant voting blocs. BP natively regularizes confidence by incorporating variance across intermediate variables, yielding better-separated posterior beliefs suitable for downstream risk-aware applications.

5.3 Additive Ablations, Oracle Extraction, and Alignment Robustness

Table 3 presents an additive build-up from greedy decoding through full DGI, an oracle extraction ceiling using gold sub-answers, and alignment noise stress tests with up to 25% corruption and alternative sentence encoders.

Additive build-up. The transition from greedy decoding to SC yields the expected baseline surge, but merely extracting sub-answers as independent unary beliefs adds only 0.2% on GSM8K and 1.1% on MATH. Constructing the tree topology without BP provides similarly modest gains. The critical leap occurs at the final step: activating PMI-ratio potentials for exact belief propagation triggers a discontinuous $+3.3\%$ surge on MATH, confirming that the inferential power is unlocked through dynamic reconciliation of conflicting variables across the graph.

Oracle bounds and noise resilience. With gold sub-answers, DGI achieves 97.8% on GSM8K, only 0.6% above the standard pipeline (97.2%), demonstrating that extraction noise is not a systemic bottleneck. Injecting 10% random alignment corruption degrades accuracy by $<0.4\%$, and swapping sentence encoders produces $<0.3\%$ variance, confirming that BP natively smooths over localized extraction errors.

5.4 Controlled Failure Modes and Structural Reduction

We designed controlled experiments to verify graceful degradation on structurally trivial problems ($N \leq 2$, high-SC-accuracy subsets, synthetic single-step queries) and to empirically confirm Proposition 2 by forcing an edgeless graph (α_{perm}=0). Across all adversarial conditions, the accuracy gap between DGI and SC is $\leq 0.1\%$ with McNemar's test (McNemar, 1947) $p > 0.10$; under the edgeless setting, DGI produces identical predictions to SC on 100% of strict-plurality instances. Full results and Table 5 are in Appendix D.8.

5.5 Hyperparameter Validation, Temperature Scaling, and Efficiency

To verify practical deployability, we conducted cross-domain hyperparameter transferability tests, temperature interaction sweeps ($\tau \in \{0.3, 0.7, 1.0\}$), and computational scaling analyses. Key

findings: (i) GSM8K-tuned hyperparameters transfer to multi-hop QA with $<0.2\%$ accuracy penalty; (ii) DGI's advantage over SC grows monotonically with temperature, peaking at $+6.4\%$ at $\tau{=}1.0$ on MATH-1000, as higher thermal diversity provides richer intermediate disagreement structure; (iii) total non-LLM overhead scales at $O(K \cdot N^2)$ and remains below 142 ms even at $K{=}80$. While the defaults are robust across our evaluated domains, we note that applying DGI to fundamentally different settings (e.g., code generation) may require lightweight re-validation of the alignment threshold γ (Appendix C.4). Full results and Table 4 are provided in Appendix D.7.

6 Conclusions

We presented Disagreement Graph Inference (DGI), a training-free framework that replaces the majority-vote aggregation in Self-Consistency with structured probabilistic inference over intermediate reasoning sub-answers. By decomposing sampled chains into a sub-answer matrix, estimating pairwise dependencies with bias-corrected mutual information under false discovery rate control, and constructing an optimal Chow–Liu tree for exact belief propagation, DGI exploits the latent covariance structure that voting paradigms discard. Our analysis established formal rescue guarantees on trees, characterized the corresponding anti-rescue conditions, and showed that DGI generalizes Self-Consistency, reducing exactly to majority voting when no significant dependencies are detected. Across six reasoning benchmarks and three foundation models, DGI improves accuracy by 2.1–6.3% absolute over Self-Consistency, matches or exceeds a fine-tuned CoT-Verify baseline without any supervised signal, and reduces calibration error by up to 59%, with a non-LLM overhead of at most 142 ms per instance.

Limitations

Our evaluation covers six text reasoning benchmarks at $K{=}40$ with three models; we did not test code generation, multimodal reasoning, or larger chain budgets, and pilot-based sub-question discovery assumes that a small sample exposes most reasoning strategies a problem admits. The Chow–Liu tree captures only pairwise dependencies, and the complete-case estimator assumes missingness at random. Appendix E discusses these and further threats to validity, including alignment-threshold sensitivity and the gap between the finite-sample theory and the $K{=}40$ regime.

Reproducibility Statement

Section 3 and Algorithm 1 (Appendix A) specify the full pipeline. All datasets are public and use standard splits (Appendix D.1), and Appendices D.4–D.5 list every generation and inference hyperparameter together with the baseline and CoT-Verify configurations. Appendix C documents the alignment, missing-data, edge-selection, and belief-propagation details needed to reimplement DGI, and Appendix D reports the transfer, temperature, efficiency, and failure-mode analyses. Accuracies are means over ten random seeds with sampling parameters held fixed across methods.

Ethics Statement

DGI is a post-hoc aggregation method operating on frozen model outputs; it neither trains on new data nor generates new content. By increasing reasoning reliability it could amplify both beneficial uses and harmful ones, though it adds no capability beyond the underlying model. All benchmarks are public and contain no human subjects or personally identifiable information. Reported gains are specific to the benchmarks, models, and chain budgets we evaluated.

AI Use Statement

We used generative AI tools for language editing, for drafting portions of the exposition, and for help organizing the manuscript. We did not use AI tools to generate experimental results, research ideas, data, tables, or figures, or to make scientific claims; every reported experiment, number, table, and figure comes from the authors' own runs. All AI-assisted text was reviewed and verified against the experiments by the authors, who take responsibility for the final content.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. doi: 10.1109/TIT.1968.1054142.

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.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. *Proceedings of the International Conference on Machine Learning*, 2023.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 1321–1330, 2017.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. GRACE: Discriminator-guided chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 15299–15328, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001. doi: 10.1109/18.910572.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5315–5333, 2023.

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. *International Conference on Learning Representations*, 2024.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 158–167, 2017.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3rd edition, 2019.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

George A. Miller. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, pp. 95–100, 1955.

Joris M. Mooij and Hilbert J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, 2007. doi: 10.1109/TIT.2007.909166.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. *Findings of the Association for Computational Linguistics: ACL*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single hop question composition. In *Transactions of the Association for Computational Linguistics*, volume 10, pp. 539–554, 2022.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 2609–2634, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yair Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023.