电子科技大学
University of Electronic Science and Technology of China

LEARNING GLOBAL HYPOTHESIS SPACE FOR ENHANCING SYNERGISTIC REASONING CHAINT

Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Xudong Wang, Zhenzhen Huang, Pengcheng Zheng，Shuai Yuan, Sheng Zheng, Qigan Sun, Jie Zou, Lik-Hang Lee , Yang Yang

1. Introduction & Motivation

Computational / Reasoning Challenge

- CoT improves reasoning but relies on step-by-step autoregressive generation
- Early errors propagate and amplify across the chain
- Lack of global coordination → unstable reasoning

Core Limitations

- No mechanism for:
  - global consistency correction
  - redundancy pruning
  - structure-aware reasoning
- Existing methods (ToT / GoT / AoT):
  - expand reasoning space
  - but still rely on local heuristics

Core Question

How to build a globally consistent and structurally robust reasoning process instead of fragile step-by-step chains?

Key Insight

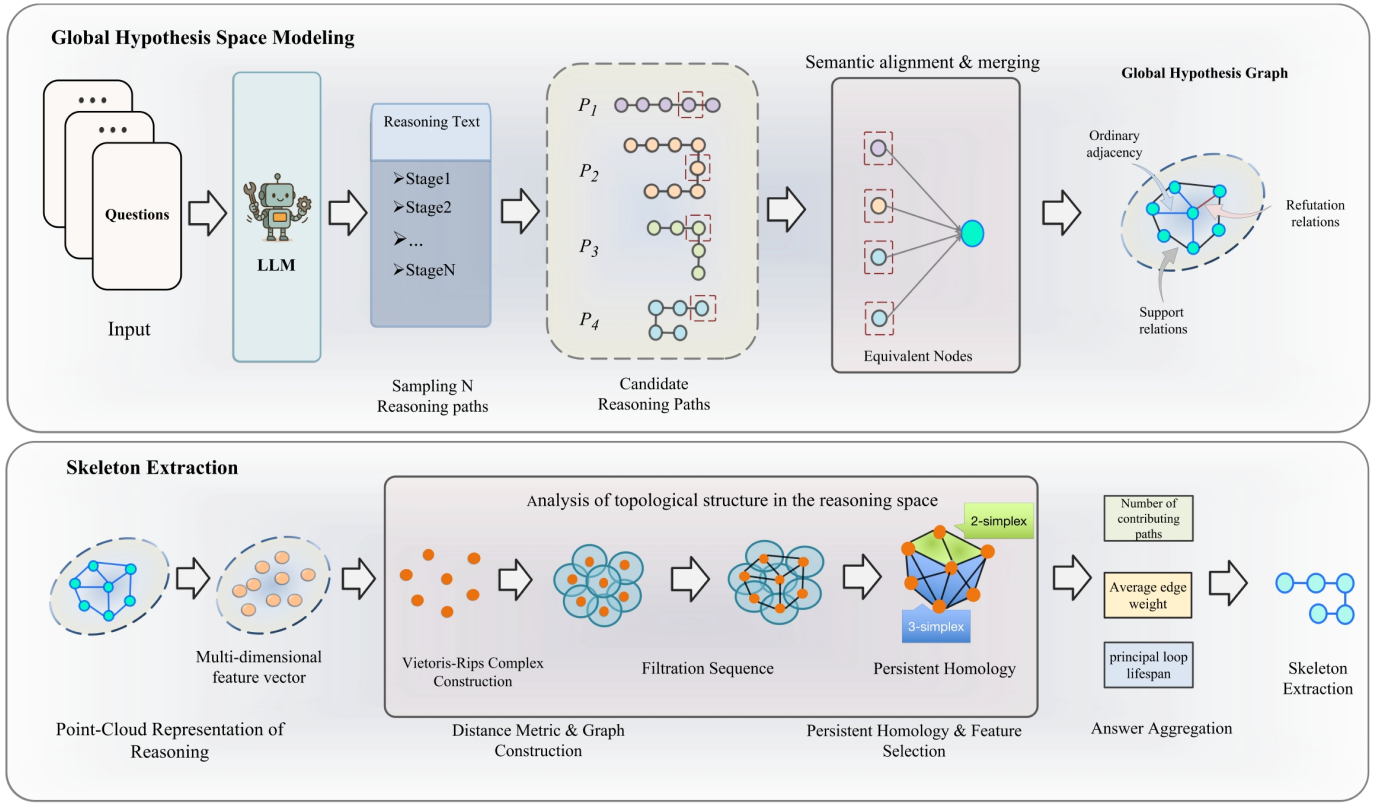
Reasoning quality depends not only on local correctness, but on global structural stability

2. Methodology

Framework Overview

We propose GHS-TDA, a two-stage framework:

- Construct global hypothesis space
- Extract stable reasoning skeleton via topology



2. Methodology

System Architecture

Stage 1: Global Hypothesis Graph (GHS)

- Merge multiple reasoning paths into one graph
- Nodes = reasoning steps
- Edges = logical relations (support / refute)

Stage 2: Topological Analysis (TDA)

Use persistent homology to extract:

- H0 → semantic clusters
- H1 → logical loops

2. Methodology

System Architecture

Stage 1: Global Hypothesis Graph (GHS)

- Merge multiple reasoning paths into one graph
- Nodes = reasoning steps
- Edges = logical relations (support / refute)

Stage 2: Topological Analysis (TDA)

Use persistent homology to extract:

- H0 → semantic clusters
- H1 → logical loops

Key Advantages

- Global hypothesis integration
- Topology-guided reasoning
- Multi-path coordination
- Self-consistent loop detection

2. Methodology

Core Mechanism

- Merge similar reasoning steps
- Build semantic-logical graph
- Extract persistent structures
- Generate final reasoning chain

Key Advantages

- Reduces error propagation
- Improves reasoning stability
- Enhances interpretability
- Enables global consistency

3. Experiments

Core Mechanism

Table 1: Performance comparison across datasets (EM %).

Method	MATH	OlympiadBench	gsm8k	BBH	MMLU-CF	LongBench	HotpotQA	MuSiQue	Avg
GPT-4o-mini									
CoT	78.3	9.3	90.9	78.3	69.6	57.6	67.2	34.1	60.7
CoT-SC (n=5)	81.8	10.2	92.0	83.4	71.1	58.6	66.2	33.8	62.1
Self-Refine	78.7	9.4	91.7	80.0	69.7	58.2	68.3	35.1	61.4
Analogical Prompting	65.4	6.5	87.2	72.5	65.8	52.9	64.7	32.8	56.0
AFlow	83.0	12.4	93.5	76.0	69.5	61.0	73.5	38.1	63.4
ToT	79.2	11.4	94.9	84.1	69.9	62.8	76.8	39.1	64.8
GoT	83.0	13.1	94.5	85.9	70.2	63.1	74.2	36.5	65.1
FoT (n=8)	82.5	12.5	94.0	82.4	70.6	59.1	66.7	35.8	63.0
AoT	83.6	12.1	95.0	86.0	70.9	68.5	80.6	38.4	66.9
GHS-TDA (Ours)	**83.9**	**14.5**	**95.2**	**88.4**	**71.6**	**69.5**	**81.4**	**39.8**	**68.0**
Qwen-Turbo									
CoT	78.1	8.9	90.7	78.1	69.4	57.3	66.8	33.6	60.4
CoT-SC (n=5)	81.4	9.9	91.5	83.2	70.8	58.4	65.9	33.5	61.8
Self-Refine	78.5	9.4	91.4	79.8	69.5	58.0	68.2	35.0	61.2
Analogical Prompting	65.2	6.2	87.0	72.2	65.2	52.7	64.5	32.6	55.7
AFlow	82.4	12.1	93.1	75.7	69.3	60.4	73.2	37.8	63.0
ToT	78.9	11.3	94.2	83.7	69.6	62.4	76.4	38.4	64.4
GoT	82.7	13.0	93.8	84.9	70.1	62.8	74.0	36.4	64.7
FoT (n=8)	82.2	12.3	93.9	82.3	70.4	59.0	66.4	35.8	62.8
AoT	83.5	12.6	94.7	85.4	70.5	68.1	80.0	39.2	66.8
GHS-TDA (Ours)	**83.7**	**14.4**	**94.8**	**87.9**	**71.2**	**68.6**	**80.3**	**39.6**	**67.6**
DeepSeek-V3									
CoT	78.5	9.5	91.3	78.5	69.9	57.7	67.4	34.2	60.9
CoT-SC (n=5)	82.0	10.4	92.1	83.6	71.5	58.9	66.6	34.0	62.4
Self-Refine	78.9	9.5	91.9	80.4	70.1	58.4	69.1	35.1	61.7
Analogical Prompting	65.6	6.7	87.6	72.8	66.1	53.4	64.9	33.1	56.3
AFlow	83.4	12.5	93.6	76.4	69.8	61.4	74.0	38.2	63.7
ToT	79.1	11.6	95.0	84.4	70.4	63.2	76.9	39.4	65.0
GoT	83.2	13.7	94.5	86.2	70.3	63.4	74.2	36.7	65.3
FoT (n=8)	82.7	12.6	94.2	82.6	70.5	59.3	66.8	36.2	63.1
AoT	84.0	13.1	95.1	86.1	70.8	68.7	80.6	39.6	67.3
GHS-TDA (Ours)	**84.5**	**14.7**	**95.2**	**88.7**	**71.6**	**69.9**	**81.7**	**40.1**	**68.3**

- GHS-TDA achieves best EM scores across benchmarks
- Outperforms CoT, ToT, GoT, AoT consistently

Table 2: Comparison of different path selection strategies within the Global Hypothesis Graph (GHS), combining quantitative evaluation and human-centered interpretability assessment.

Path Type	Accuracy %	Avg. Length	Avg. Conf.	Conf. Std ↓	Clarity	Coherence	Credibility	Conciseness
Shortest Path (GHS)	75.2	5.8	0.81	0.12	3.6	2.9	3.4	4.3
Max-Confidence Path (GHS)	82.1	11.5	0.93	0.21	4.1	4.2	4.3	3.9
Human-Selected Path (GHS)	83.6	9.2	0.88	0.07	4.5	4.6	4.7	4.4
TDA Skeleton (Ours)	**83.9**	**8.7**	**0.90**	**0.07**	**4.4**	**4.5**	**4.7**	**4.3**

3. Experiments

Reasoning Quality

- Comparable to human-selected reasoning paths
- Better balance between:
  accuracy/length/stability

Robustness

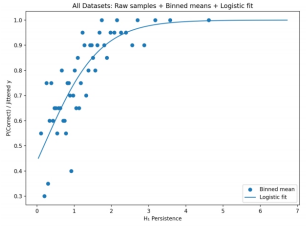
Table 3: Robustness under adversarial perturbations.

Strategy	Before (%)	After (%)	Change (%)
Max-Confidence	82.1	77.1	7.4
GHS-TDA (Ours)	**83.9**	**81.5**	**2.9**

Under perturbation:
GHS-TDA drop: 2.9% baseline: 7.4%

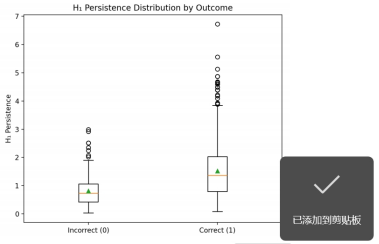
Interpretability & Theory

(a) Binned correct rate with logistic fit.

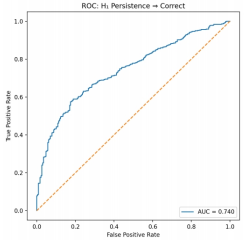


(b) Raw samples with binned means and logistic fit.

Figure 2: Global relation between H_1 persistence and reasoning correctness.



(a) Distribution of H_1 persistence by outcome.



(b) ROC curve using H_1 persistence.

Figure 3: Validation of the predictive role of topological persistence. Left: correct reasoning chains have consistently higher H_1 persistence values than incorrect ones. Right: ROC analysis shows persistence alone achieves an AUC of 0.74.

Table 4: Predictive power of H_1 persistence for reasoning correctness. Higher persistence consistently correlates with better performance.

Analysis Item	Value	Interpretation
Global Spearman ρ	0.349 ($p \approx 0$)	Moderate positive correlation
Logistic regression (std. H_1)	1.247 (OR $\approx$ 3.48)	Strong effect: +1 SD $\Rightarrow$ $\sim$3.5$\times$ odds
ROC–AUC (H_1 only)	0.74	Good discriminative ability
Per-dataset ROC–AUC		
GSM8K	0.748	Robust
MATH	0.704	Robust
OlympiadBench	0.703	Robust
BBH	0.729	Robust
MMLU-CF	0.733	Robust
LongBench	0.737	Robust
HotpotQA	0.778	Strongest
MuSiQue	0.709	Robust

- Higher H1 persistence → higher accuracy
- ROC AUC = 0.74

3. Experiments

Efficiency

- Fixed LLM calls (19 per task)
- ↓ 25–40% cost vs ToT / AoT
- ↓ ~30% token usage

Table 6: Comparison of LLM Call Computational Costs

Method	MATH	Olymp.	GSM8K	BBH	MMLU-CF	LongB.	HotpotQA	MuSiQue
CoT	1	1	1	1	1	1	1	1
CoT-SC ($n = 16$)	16	16	16	16	16	16	16	16
ToT	25.8	27.2	13.5	24.1	19.3	14.2	20.7	26.4
GoT	21.6	23.5	14.8	22.9	18.4	13.9	17.8	20.3
AoT	29.7	32.4	17.2	28.6	24.9	16.8	23.5	29.1
GHS-TDA (Ours)	**19**	**19**	**19**	**19**	**19**	**19**	**19**	**19**

Table 7: Comparison of token consumption.

Method	MATH	OlympiadBench	GSM8K	BBH	MMLU-CF	LongBench	HotpotQA	MuSiQue
CoT	2290	4590	278	642	1235	83	1305	485
CoT-SC	36640	73440	4448	10272	19760	1328	20880	7760
ToT	88623	187272	5630	23208	35753	1768	40520	19206
GoT	64303	140225	5349	19112	29541	1500	30198	12799
AoT	68013	148716	4782	18361	30752	1394	30668	14114
GHS-TDA	51411	83309	7709	15793	28862	1685	28358	10263

Ablation Study

- Removing semantic information → largest performance drop
- Removing structural components → significant performance decrease
- Removing uncertainty modeling → minor performance impact

Table 5: Ablation study of distance weights under the constraint $\alpha + \beta + \gamma = 1$.

Method	α	β	γ	Accuracy (Avg.)
GHS-TDA (Ours)	0.6	0.3	0.1	83.9%
Without β	0.9	0.0	0.1	81.2%
Without α	0.0	0.9	0.1	77.4%
Without γ	0.7	0.3	0.0	83.5%

4. Conclusions

- Propose GHS-TDA, a structure-aware reasoning framework
- Leverage topological structure to improve reasoning quality
- Achieve better accuracy, robustness, and efficiency
- Provide interpretable and predictive signals for reasoning