

A UNIFIED FRAMEWORK FOR EFFICIENCY AND EXPLAINABILITY

Yubo Dong, Hehe Fan, Linchao Zhu

1. Structured Reasoning Data & Tuning

(based on <IMAGE-0> - Data Collection)

Input: What is $\sqrt{53}$ in simplest radical form? Please reason step by step using structured reasoning tags.

Reasoning Steps:

<assumption>

To find the simplest radical form of $\sqrt{53}$, I need to check if 53 has any perfect square factors.

<decompose>

Left a prime number with factors 1 and 53.

<verify>

Since 53 is prime, no perfect squares greater than 1 divide it.

<inference>

Therefore, $\sqrt{53}$ cannot be simplified further, $\sqrt{53}$ is already in its simplest form.

<verify>

Double checking: If we had $\sqrt{53} = a/b$ where b is square free, then $a^2b = 53$.

<conclusion>

The simplest radical form is $\sqrt{53}$.

Output: $\sqrt{53}$

Output: $\sqrt{53}$

Data Pipeline:



```

Step 0: recognize 276: 3, 94
Find all positive perfect divisors of 166.

Step 1: list them: 1, 2, 83
So here to find we have factorization of 166.

Step 2: decompose 166: 2 * 83
166 = 2 * 83 = 2 * 83 = 2 * 83

Step 3: analyze 166: 2 * 83
Decomposition path: 2 * 83 = 2 * 83

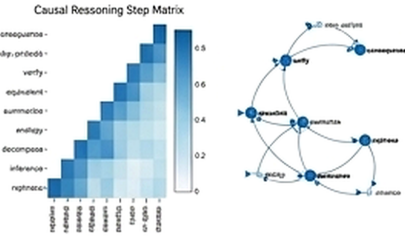
Step 4: summarize 166: 2 * 83, ... 2 * 83

Step 5: simplify 166: 2 * 83
Apply Euler Totient: (1-1)(83-1) = 82

Step 6: verify 166: 82, 166
Verify: exponent (1, 1) -> (1, 1) -> 0

Step 7: find project 276: 1, 161
List all subelements: 1, 2, 4, 8, 16, 20, 40, 80, 160

Step 8: recognize 166: 82, 167
Find all perfect divisors: 100 Test periods @ positive divisors
    
```



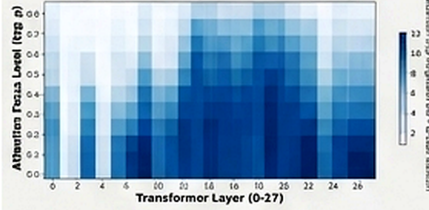
Structured Tuning Dataset:

- 500 high-quality samples from S1k, DeepScaleR.
- 23 Reasoning Tags.

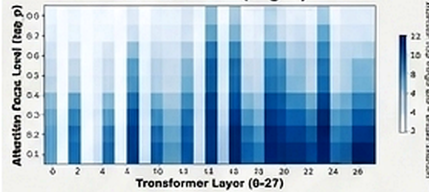
2. Methods - Traceability & Optimization

(based on <IMAGE-3>, <IMAGE-4>, and <IMAGE-2> Step dependency/Optimization

Layer-wise step Attention (Left)



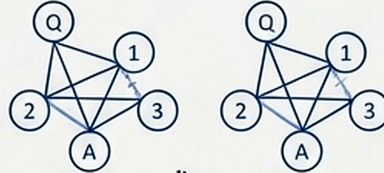
1.5B Models (Right)



local transition to global transition
Local at Layer 14.

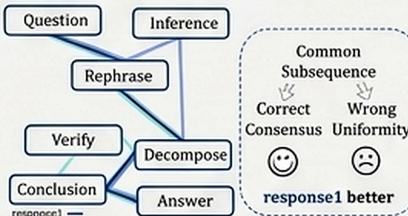
Middle layers integrate broader reasoning context.

A. MAX-Flow Algorithm: sparse graphs via ΔF_k (node importance)



$$\Delta F_k = \frac{1}{HT_i} \sum_{h=1}^n \sum_{acT_i} \max_{bcT_i} A_{h,a,b}$$

B. LCS Algorithm: optimal common subsequences from multiple generated responses

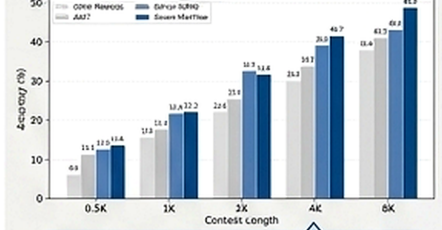


MAX-Flow ΔF_k outperforms other filters in EFE.

3. General Benchmarks

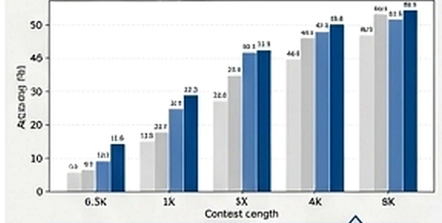
(based on <IMAGE-0>)

A) Efficiency (1.5B)



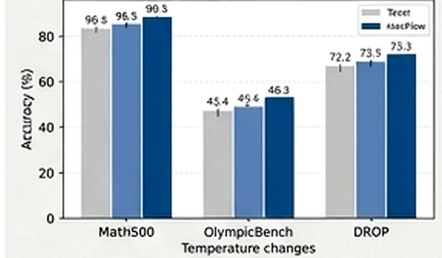
better performance with fewer steps at longer context lengths.

A) Efficiency (7B)

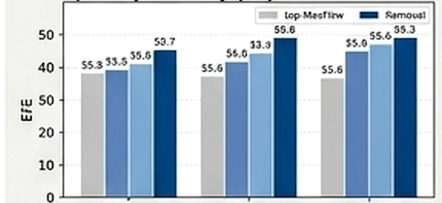


better performance lengths.
MATH500 7B MaxFlow: 92.67% accuracy

B) Stability (1.5B)



C) Interpretability (7B)



Interpretability Injection & Removal
top-matrix filter outperforms random and perplexity filter.

Conclusion

Structured Reasoning improves efficiency, stability, and interpretability.

Future: redundant layer pruning.



Project: <https://cnsdq-dyb.github.io/structured-reasoning/>