ICML
International Conference
On Machine Learning

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

TRM

Characterizing, Evaluating, and Optimizing Complex Reasoning

Haoran Zhang[1,2], Yafu Li[2,3,*], Zhi Wang[4,2], Zhilin Wang[2,5], Shunkai Zhang[2,6], Xiaoye Qu[2], Yu Cheng[3,2,*]

[1]Shanghai Jiao Tong University [2]Shanghai Artificial Intelligence Laboratory [3]The Chinese University of Hong Kong
[4]Nanjing University [5]University of Science and Technology of China [6]Peking University

ICML 2026

01 INTRODUCTION

Reasoning traces are optimization targets

Large reasoning models are increasingly prominent due to their *strong logical reasoning abilities*

- Final responses
  - *concise* and *actionable*
- Reasoning traces
  - longer, *structurally more complex*
  - might be redundant, inconsistent, or incorrect
- Improving reasoning quality is therefore crucial

A reasoning trace from *Qwen3-8B*

.....

Therefore, I think my setup is correct.

Wait, but just to make sure, let me think ...

Alternatively, using the coordinates ...

So, that confirms the area is indeed 2.

But in our case, the integrand is ...

So the area is 2.

But let me check ...

Alternatively, maybe I can ...

Therefore, I think this is the correct ...

.....

01 INTRODUCTION

Three fundamental questions

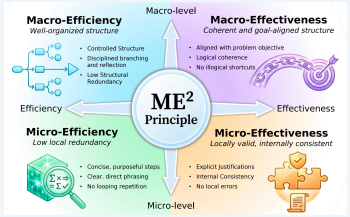
- Q1. *What constitutes a high-quality reasoning trace?*
  → Characterizing
- Q2. *How can we reliably evaluate long and implicitly structured traces?*
  → Evaluating
- Q3. *How can we leverage such evaluation signals to optimize reasoning at scale?*
  → Optimizing

01 INTRODUCTION

Our answers to the three questions

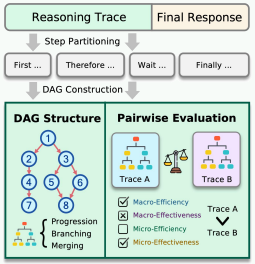
Define reasoning quality, evaluate it reliably, then turn it into a usable reward.

Q1 Characterizing



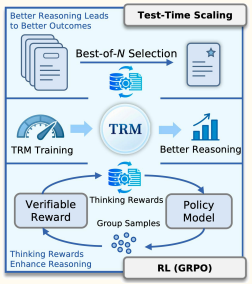
ME² principle name what good traces look like.

Q2 Evaluating



DAG-based pairwise evaluation produces preferences.

Q3 Optimizing

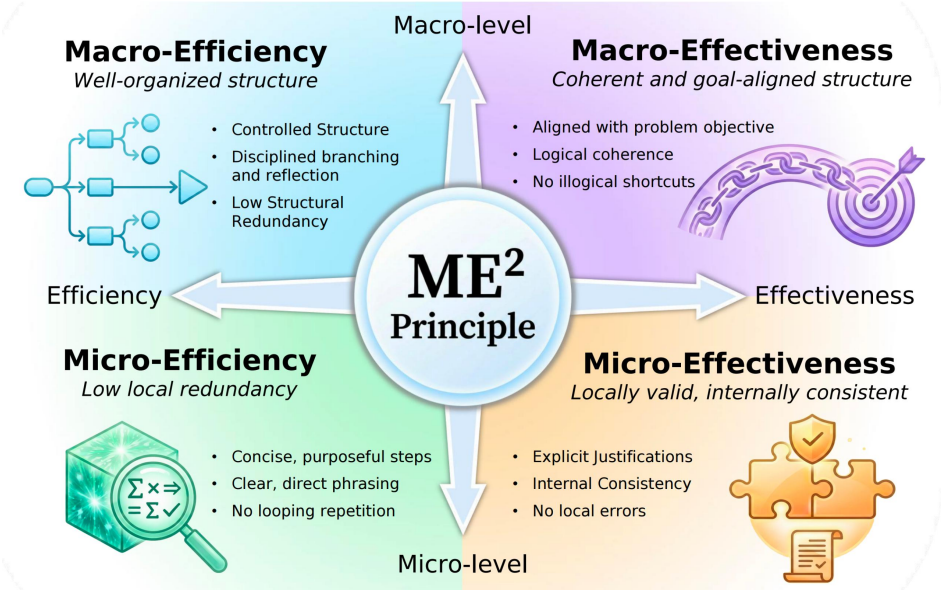


TRM turns preferences into a reward signal.

02 CHARACTERIZATION

ME² principle for reasoning quality

- Where quality manifests?
  - **Macro**: *global-level organization*
  - **Micro**: *step-level quality*
- What quality entails?
  - **Efficiency**: *compact reasoning without wasted effort*
  - **Effectiveness**: *coherent progress toward the goal*

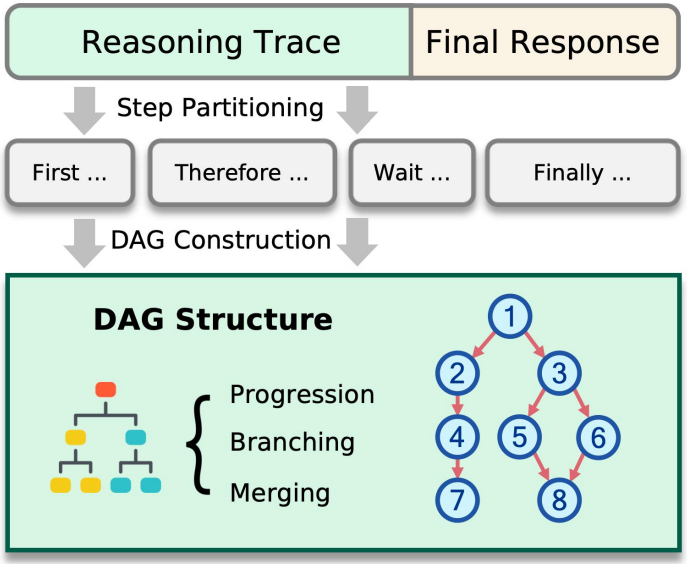


03 EVALUATION

How do we model complex reasoning structure?

- **We represent reasoning traces as DAGs (Directed Acyclic Graphs)**
  - Nodes are reasoning steps
  - Edges capture semantic dependencies
  - The DAG exposes how ideas branch, merge, and progress

- **Why DAGs?**
  - *Trees*: too rigid
  - *General graphs*: too unconstrained
  - *DAGs*: preserve step order while modeling complex dependencies



03 EVALUATION

Reasoning structuring

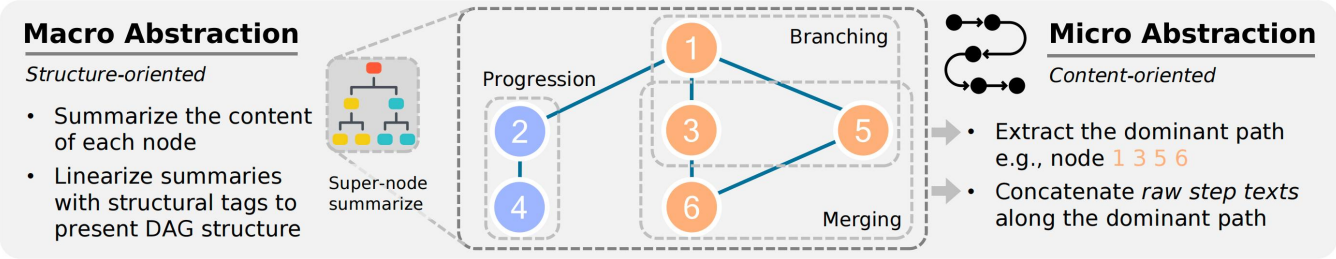
- **Step partitioning**
  - Start from coarse trace segments
  - Mine *high-frequency step prefixes*
  - Refine boundaries into atomic steps
- **DAG construction**
  - Traverse the steps in *generation order*
  - Select *parents* from earlier steps
  - Add *directed edges* for semantic dependencies

Algorithm 1 DAG Construction

Input: partitioned reasoning steps $(s_1, \ldots, s_n)$
Output: reasoning DAG $G = (V, E)$
Initialize $V \leftarrow \{v_1\}$, $E \leftarrow \emptyset$
for $i = 2$ **to** n **do**
  Create node v_i for step s_i and add it to V
  Construct attachment pool $\mathcal{P}_i \subseteq \{v_1, \ldots, v_{i-1}\}$
  Select parent nodes $\mathrm{Parent}(v_i) \subseteq \mathcal{P}_i$ using an LLM
  for each $v_j \in \mathrm{Parent}(v_i)$ **do**
    $E \leftarrow E \cup \{(v_j, v_i)\}$
  end for
end for
Collapse consecutive linear chains in G into super-nodes
Replace V with the resulting super-nodes

03 EVALUATION

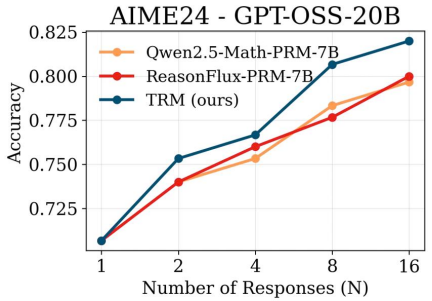
How do we compare reasoning quality?



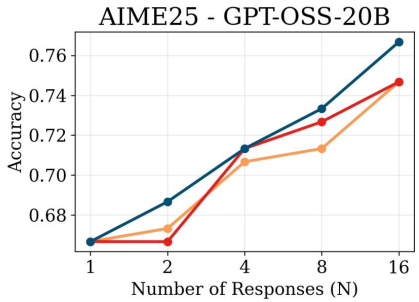
- **Dataset construction**
  - Compare pairs through *macro/micro abstractions*
  - Judge quality under ME² principle
  - Build ***TRM-Preference dataset*** (103K training pairs)
- **Thinking Reward Model (TRM)**
  - 8B reward model
  - Trained on the TRM-Preference dataset
  - *Decouple* reasoning quality from answer correctness

04 OPTIMIZATION

Better reasoning leads to better outcomes



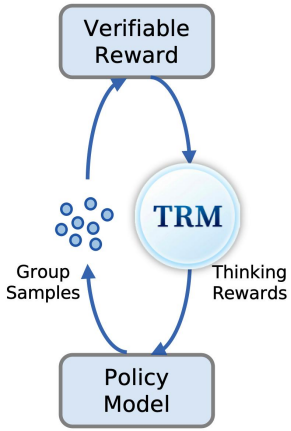
- **TRM-guided Test-Time Scaling (TTS)**
  - Trained only on verified-correct traces without explicit correctness supervision
  - At test-time, selecting higher-scored traces yields better final accuracy



Better reasoning, as measured by closer alignment with the ME² principle, leads to better final outcomes

04 OPTIMIZATION

TRM turns reasoning quality into training gains



Models	STEM Performance (%)						Math Performance (%)					
	GPQA	SuperGPQA	MMLU-Pro	BBEH	FS	Avg.	AMC	AIME24	AIME25	MATH500	Olympiad	Avg.
Qwen2.5-Math-1.5B												
BaseModel	10.1	13.2	19.9	1.3	7.7	10.4	22.2	2.7	0.6	32.0	15.4	14.6
Verifier	21.7	17.0	31.4	6.2	10.4	17.3$_{+6.9}$	43.8	7.9	4.3	69.0	**37.5**	32.5$_{+17.9}$
ReasonFlux	23.7	**17.6**	32.0	5.2	9.1	17.5$_{+7.1}$	45.0	10.0	4.9	71.4	35.0	33.3$_{+18.7}$
TRM (ours)	**27.3**	17.3	**33.0**	**6.8**	**12.6**	**19.4**$_{+9.0}$	**48.8**	**11.4**	**7.0**	**74.2**	37.3	**35.7**$_{+21.1}$
Qwen2.5-Math-7B												
BaseModel	22.7	19.8	38.6	6.1	14.4	20.3	50.8	12.8	7.7	69.0	32.4	34.5
Verifier	38.9	25.4	47.5	8.6	19.2	27.9$_{+7.6}$	62.0	16.8	13.7	81.6	46.1	44.0$_{+9.5}$
ReasonFlux	41.9	25.6	49.4	**9.5**	23.9	30.1$_{+9.8}$	61.9	17.1	13.2	**83.6**	46.5	44.5$_{+10.0}$
TRM (ours)	**42.4**	**26.7**	**50.5**	9.2	**26.5**	**31.1**$_{+10.8}$	**63.4**	**19.1**	**17.1**	83.0	**47.4**	**46.0**$_{+11.5}$
Llama-3.1-8B-Instruct												
BaseModel	16.2	22.1	43.6	8.6	20.3	22.2	19.4	3.9	0.2	47.6	10.4	16.3
Verifier	29.3	24.9	47.5	12.8	25.2	27.9$_{+5.7}$	24.7	5.1	1.2	52.6	18.4	20.4$_{+4.1}$
ReasonFlux	31.8	**27.3**	50.9	13.2	26.0	29.8$_{+7.6}$	29.3	**8.8**	1.4	**55.0**	21.5	23.2$_{+6.9}$
TRM (ours)	**36.4**	26.7	**52.6**	**14.0**	**29.1**	**31.8**$_{+9.6}$	**32.8**	8.1	**2.4**	53.8	**23.9**	**24.2**$_{+7.9}$

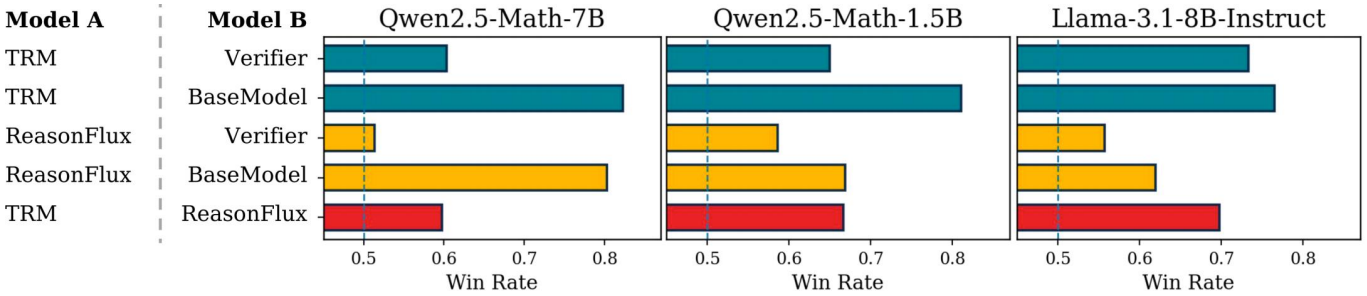
- **TRM-guided RL (GRPO)**
  - Gated reward shaping biases learning toward ME²-aligned reasoning

$$r = r_v \cdot (1 - \alpha + \alpha \cdot \mathrm{Sigmoid}(r_t))$$

TRM-guided RL improves over verifier-only training

04 OPTIMIZATION

Thinking rewards enhance reasoning



Trace quality win rate: Model A beats Model B; ties excluded.

- TRM-trained policies produce better reasoning traces
  - Verifiers enforce correctness, but cannot rank correct traces
  - TRM supplies the missing signal: which correct trace reasons better

Thinking rewards shift RL toward ME²-aligned reasoning

Thank you!

Contact: *zzzhr97@gmail.com*

Paper

Code

Our Team

https://github.com/Simplified-Reasoning