

Roads to Roam (R2R): Efficiently Navigating Divergent Reasoning Paths with Small-Large Model Token Routing

Tianyu Fu*, Yi Ge*, Yichen You, Enshu Liu,
Zhihang Yuan, Guohao Dai, Shengen Yan, Huazhong Yang, Yu Wang[†]

**equal contribution*

[†]corresponding author (yu-wang@tsinghua.edu.cn)



Tsinghua University

INFINIGENCE



SHANGHAI JIAO TONG
UNIVERSITY

Authors



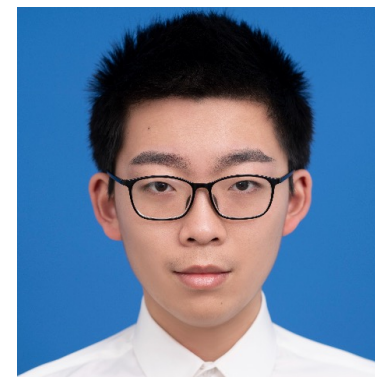
Tianyu Fu
Tsinghua University
Inference AI



Yi Ge
Tsinghua University



Yichen You
Tsinghua University



Enshu Liu
Tsinghua University



Zhihang Yuan
Inference AI



Guohao Dai
Shanghai Jiao Tong University
Inference AI



Shengen Yan
Inference AI



Huazhong Yang
Tsinghua University



Yu Wang
Tsinghua University

Motivation

Use small language model (SLM) and LLM for different reasoning steps

Motivation

Fast but weak SLM
slow but **strong** LLM

Tested results on AIME'24-25

| Type | Model | Accuracy | Latency (s / question) |
|------|---------|----------|---------------------------|
| SLM | R1-1.5B | ☹ 12% | 😊 199 |
| LLM | R1-32B | 😊 57% | ☹ 498 |

We should selectively use SLM and LLM for different generation steps, constructing a **fast and strong** mix-inference method

Insight

Given same context, SLM and LLM predictions are **often identical**

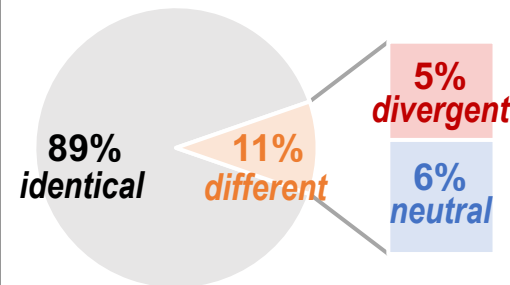
question: Compute $9999^2 - 9998 \times 1000$.

LLM: Okey, let's think step by step... 9999² is **hard**, rewrite it...

SLM: Okey, let us think step by step... 9999² is **999801...**

identical ✓ neutral ✓

divergent ✗



- 89% **identical** predictions
- 11% **different** predictions
 - some are **neutral**, like alternative expressions
 - only few **diverge** the meaning, logic, or conclusion of reasoning

Divergent Token Labeling

Label divergent token, then **train** a neural token-router,
utilizing LLMs only for path-divergent tokens during SLM generation

Method

Label divergent tokens
generate model preference training data

| | | | | | | | | | | | | | | | | | |
|------------------------|---------------------------------------|-----------------------------|-------|------|----|------|---|----|----|--------------|----|-----------------------------|---|----|-------|----|--|
| query | Compute $9999^2 - 9998 \times 1000$. | | | | | | | | | | | | | | | | |
| step0: LLM response | Let | 's | think | step | by | step | . | 99 | 99 | ² | is | hard | , | re | write | it | |
| step1: SLM prefill | Let | ① us different | think | step | by | step | . | 99 | 99 | ² | is | ② 99 different | , | re | write | it | |

Key idea:

Step1. **find** all predictions
where SLM-LLM **differ**

Step2. from the difference,
let LLM **continue generation**
until the end of current
sentence, to understand
difference's impact

Step3. ask **another LLM** to
verify if difference causes
divergence

Divergent Token Labeling

Label divergent token, then **train** a neural token-router, utilizing LLMs only for path-divergent tokens during SLM generation

Method

Label divergent tokens
generate model preference training data

| | | | | | | | | | | | | | | | | |
|----------------------------|---------------------------------------|-----------------------------|-----------|--------------------------------|------|------|------|----|----|--------------|--------------|-----------------------------|-----------|---------|-------|----|
| query | Compute $9999^2 - 9998 \times 1000$. | | | | | | | | | | | | | | | |
| step0: LLM response | Let | 's | think | step | by | step | . | 99 | 99 | ² | is | hard | , | re | write | it |
| step1: SLM prefill | Let | ① us different | think | step | by | step | . | 99 | 99 | ² | is | ② 99 different | , | re | write | it |
| step2: LLM continuation | ① | Let | us | → think about it step by step. | | | | | | | | | | | | |
| | ② | Let | 's | think | step | by | step | . | 99 | 99 | ² | is | 99 | → 9801. | | |

Key idea:

Step1. find all predictions where SLM-LLM **differ**

Step2. from the difference, let LLM **continue generation** until the end of current sentence, to understand difference's impact

Step3. ask another LLM to **verify** if difference causes **divergence**

Divergent Token Labeling

Label divergent token, then **train** a neural token-router, utilizing LLMs only for path-divergent tokens during SLM generation

Method

Label divergent tokens
generate model preference training data

| | | | | | | | | | | | | | | | | | |
|----------------------------|---------------------------------------|-----------------------------|---------------------------------------|--------------------------------|------|------|------|---|-----|--------------|--------------|-----------------------------|-----------|-----|---------------|-----|--|
| query | Compute $9999^2 - 9998 \times 1000$. | | | | | | | | | | | | | | | | |
| step0: LLM response | Let | 's | think | step | by | step | . | 99 | 99 | ² | is | hard | , | re | write | it | |
| step1: SLM prefill | Let | ① us different | think | step | by | step | . | 99 | 99 | ² | is | ② 99 different | , | re | write | it | |
| step2: LLM continuation | ① | Let | us | → think about it step by step. | | | | | | | | | | | | | |
| | ② | Let | 's | think | step | by | step | . | 99 | 99 | ² | is | 99 | → | 9801 . | | |
| step3: verify | ① | Verify | Let's think step by step | | | | and | Let us think about it step by step | | | | neutral | | | | | |
| | ② | Verify | 9999 ² is hard, rewrite it | | | | and | 9999 ² is 999801 | | | | divergent | | | | | |
| output label | | SLM | SLM | SLM | SLM | SLM | SLM | SLM | SLM | SLM | SLM | LLM | SLM | SLM | SLM | SLM | |

Key idea:

- Step1. find all predictions where SLM-LLM **differ**
- Step2. from the difference, let LLM **continue generation** until the end of current sentence, to understand difference's impact
- Step3. ask **another LLM to verify** if difference causes **divergence**

Tianyu Fu*, Yi Ge*, Yichen You, et al. "Efficiently Navigating Divergent Reasoning Paths with Small-Large Model Token Routing" NeurIPS'25.

Token-level Routing

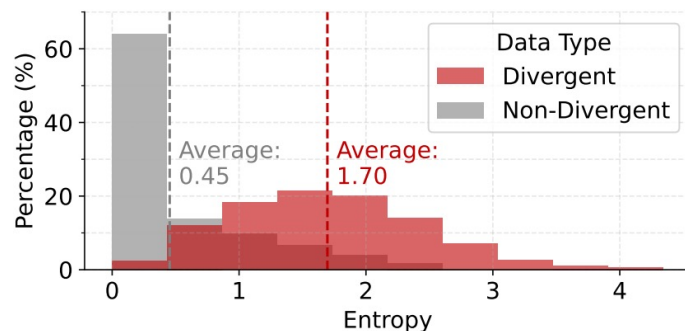
Label divergent token, then **train** a neural token-router, utilizing LLMs only for path-divergent tokens during SLM generation

Insight

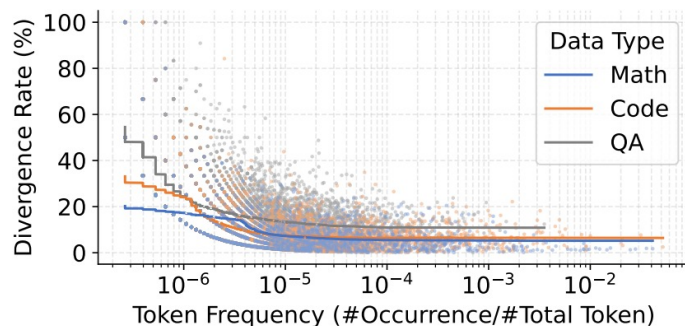
Predictive indicators for divergence

Key finding:

Divergent tokens exhibit **higher entropy** in SLM's output logits



Low-frequency tokens in the dataset are more likely to be divergent



Method

Train neural router, route to LLM for divergent SLM tokens

input: It's

SLM: It's 99

LLM: It's hard

It's hard, It's hard, re It's hard, rewrite

output: hard

,

re

write

Routing scheme:

We train a 56M neural router

Given SLM output token & its last-layer hidden states, it **classifies** whether this token is **divergent**, Immediately route to LLM if predicted as divergent

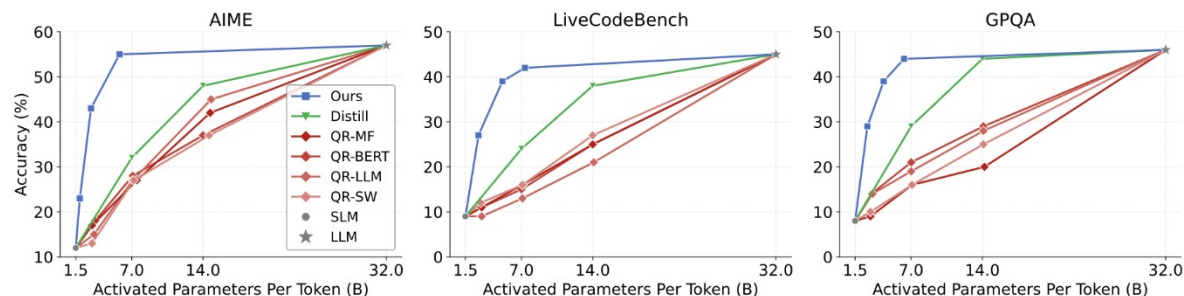
Experimental Results

Mixing R1-1.5B & 32B, using **R2R** with **5.6B** avg. activated param. per token achieve **performance exceeding R1-14B**

Result

Performance-Efficiency Pareto Frontier

same avg. parameter, better accuracy



R2R consistently outperforms both **distillation** models and **query-level routing** methods, setting a new state-of-the-art in performance-efficiency.

Ablation

Ablation results on **training objective** and **router input**

Tested results on AIME'24-25

| Objective | Router Input | Acc. | Param. | Latency (s / question) |
|------------------|---------------------|------------|-------------|---------------------------|
| Divergent | HS+Token +Logits | 55% | 5.6B | 218 |
| Different | HS+Token +Logits | 40% | 5.7B | 228 |
| Divergent | HS+Token | 47% | 5.8B | 253 |
| | HS | 42% | 6.0B | 245 |

Experimental Results

Mixing R1-1.5B & 32B, using R2R with 5.6B avg. activated param. per token achieve performance exceeding R1-14B

Result

Real-World Inference Latency

Tested results on AIME'24-25

| Model | Acc. | Speed (tok / s) | Latency (s / question) |
|----------|------|-----------------|------------------------|
| R1-14B | 48% | 52.1 | 328 |
| R1-32B | 57% | 30.5 | 498 |
| R2R-5.6B | 55% | 84.3 | 218 |

For R2R-5.6B (mix R1-1.5B & 32B)

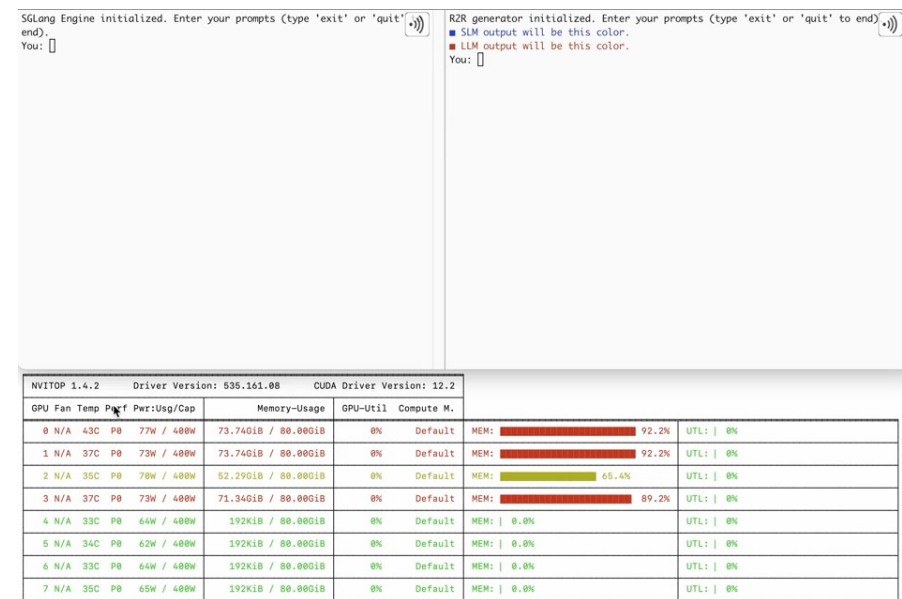
- Comparing R1-14B, 1.62x speedup, 1.15x accuracy
- Comparing R1-32B, 2.76x speedup, achieving 96% of its accuracy with only 11%-15% LLM usage

Demo




source code

Reaching 84.3 token/s on two A800-80GB GPUs



R1-32B
Finished in: 1min 12s

R2R
Finished in: 32s 

Analysis

R2R is trained to generalize across different tasks.
It learns **explainable routing patterns**

Training Data

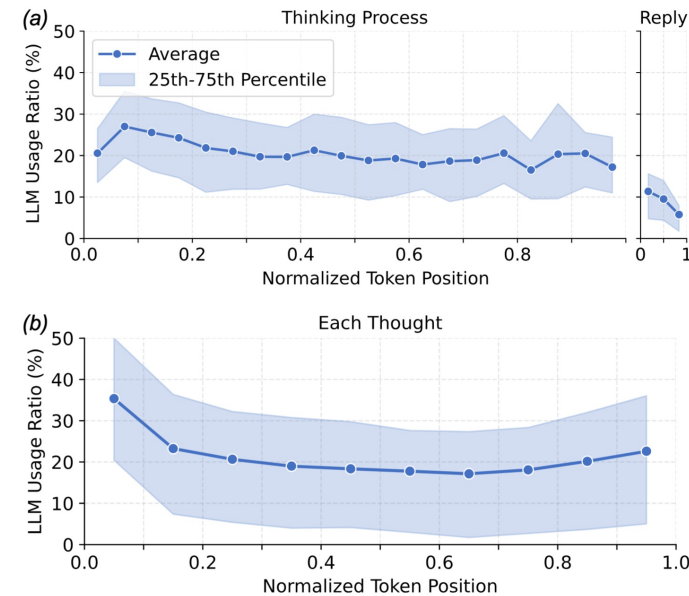
Statistics of tokens **difference** and **divergence** in the training dataset

| Type | #Query | #Token | Different Rate | Divergent Rate |
|---------|--------|--------|----------------|----------------|
| Math | 587 | 2.9M | 6.8% | 2.8% |
| Code | 698 | 3.2M | 10.3% | 4.7% |
| QA | 735 | 1.4M | 20.2% | 9.7% |
| Summary | 2094 | 7.6M | 10.8% | 4.9% |

In the training dataset, only **11%** of tokens **differ**, only **5%** are true **divergent** tokens that alter the reasoning path

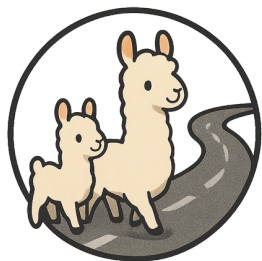
Routing Behavior

LLM usage rate at different positions during inference



During the **reply** phase, LLM usage **drops**.

At the **start** and **end** of a thought, LLM usage **peaks**.



Thank you



R2R: Efficiently Navigating Divergent Reasoning Paths with Small-Large Model Token Routing

Tianyu Fu*, Yi Ge*, Yichen You, Enshu Liu,
Zhihang Yuan, Guohao Dai, Shengen Yan, Huazhong Yang, Yu Wang[†]

**equal contribution [†]corresponding author (yu-wang@tsinghua.edu.cn)*



Tsinghua University

INFINIGENCE



SHANGHAI JIAO TONG
UNIVERSITY