Diversity-Enhanced Reasoning for Subjective Questions

Anonymous ACL submission

Abstract

Large reasoning models (LRM) with long chain-of-thought (CoT) capabilities have shown strong performance on objective tasks, such as math reasoning and coding. However, their effectiveness on subjective questions that may have different responses from different perspectives is still limited by a tendency towards homogeneous reasoning, introduced by the reliance on a single ground truth in supervised fine-tuning and verifiable reward in reinforcement learning. To bridge this gap, we conduct a pilot analysis on the scaling laws of reasoning length and the number of role perspectives, where we uncover that increasing role perspectives consistently yields performance gain. Then, we propose **MultiRole-R1**, a diversity-enhanced framework with multiple role perspectives, enhancing the accuracy and diversity in subjective reasoning tasks. MultiRole-R1 features an unsupervised data construction pipeline that constructs reasoning chains that incorporate diverse role perspectives. We further employ reinforcement learning via Group Relative Policy Optimization (GRPO) with reward shaping, taking diversity as an additional reward signal. With specially designed reward functions, we successfully promote perspective diversity and lexical diversity, and discover a positive relation between reasoning diversity and accuracy. Our experiment on six benchmarks demonstrates MultiRole-R1's effectiveness and generalizability in enhancing both subjective and objective reasoning, showcasing the potential of diversity-enhanced training in LRMs.

1 Introduction

Advances in o1-style models (Jaech et al., 2024; DeepSeek-AI et al., 2025) with long CoT (Wei et al., 2023) have significantly enhanced the models' capability in various reasoning tasks (Yu et al., 2025; Wu et al., 2024). Extended CoT chains enable models to systematically divide-and-conquer complex problems, iteratively refining intermediate results, and methodically verifying the outputs. The o1-style models reached top-human performances in many objective reasoning tasks such as math (Cobbe et al., 2021) and coding (Jain et al., 2024). This success highlights that scaling up the test time computation can significantly boost performance on complex, structured tasks relative to standard single-step prompting (Jaech et al., 2024; Wei et al., 2023).

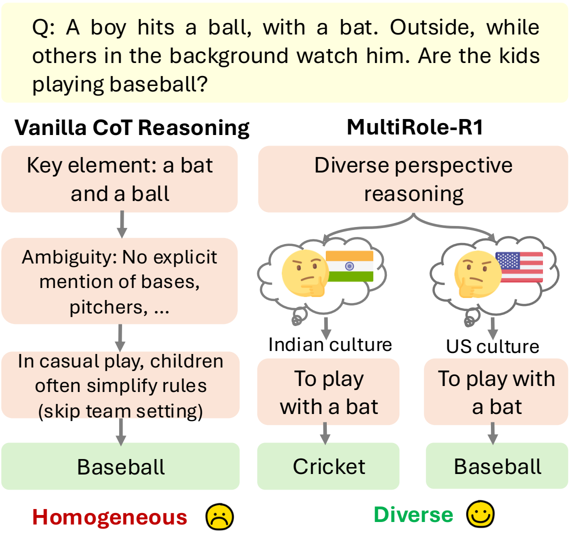
Q: A boy hits a ball, with a bat. Outside, while others in the background watch him. Are the kids playing baseball?
Vanilla CoT Reasoning
MultiRole-R1
Key element: a bat and a ball
Ambiguity: No explicit mention of bases, pitchers, ...
In casual play, children often simplify rules (skip team setting)
Baseball
Homogeneous
Diverse perspective reasoning
Indian culture
US culture
To play with a bat
To play with a bat
Cricket
Baseball
Diverse

Figure 1: subjective reasoning tasks, the answer often changes with perspective shifts. We illustrate that diverse perspective thinking benefits subjective question answering.

Beyond **objective** tasks with a single correct answer, there is growing interest in applying LLM reasoning to **subjective or open-ended problems**. As illustrated in Figure 1, due to the single inherent thinking process of LLM, the model tends to rely on homogeneous reasoning trajectories without exploring alternative perspectives. Consequently, LLMs tend to approach subjective questions from a narrow viewpoint, hindering their capacity to incorporate diverse reasoning strategies. Prior works' solutions mainly fall in two categories: (1) Research on Multi-Agent Debates has demonstrated that incorporating diverse role perspectives substantially improves performance on subjective tasks (Aoyagui et al., 2025; Cheng et al., 2024; Liu et al., 2025b), but these methods require multiple interacting models; (2) Other approaches ensembles different solutions from diverse decoding paths (Wang et al., 2023), but the stochasticity of the decoding paths does not equate to introducing more perspectives (Naik et al., 2024).

To address this gap, we propose to enhance the diversity of the OpenAI o1-style reasoning chain, by to dynamically integrating diverse role perspectives for subjective question answering, including tasks such as moral dilemmas, opinion-based questions, ambiguous question answering, and so on. Initially, we conduct a pilot study to confirm two key factors: the effectiveness of long reasoning chains in subjective reasoning tasks, and the impact of varying the number of role perspectives incorporated during reasoning. Our results indicate that sequential scaling initially increases with longer reasoning chains but eventually decreases beyond an optimal point. Similarly, we identify there is a similar an optimal range for the number of role perspectives in the reasoning chain. Building upon these insights, we propose a novel framework **MultiRole-R1** that enhances the diversity of long chains of thought. Specifically, we optimize the role diversity and reasoning diversity. MultiRole-R1 starts with a novel data construction pipeline combining parallel scaling with sequential scaling to effectively incorporate a broader array of role perspectives into the extended reasoning chains. We subsequently refine the model through supervised fine-tuning (SFT), instructing the model to learn the multirole reasoning format and enhance perspective diversity. Furthermore, to enhance the diversity of the reasoning process, we introduce a carefully designed diversity reward function implemented through a multi-role Generalized Reward Policy Optimization (GRPO) framework, taking diversity as an additional reward signal. By training on subjective questions only, MultiRole-R1 boosts the performance of both subjective and objective reasoning tasks. Our contribution can be summarized as follows:

- We perform a pilot analysis that reveals the scaling law of reasoning length and role perspectives in subjective reasoning.

- We propose MultiRole-R1, a novel framework that enhances the diversity and accuracy of LRMs, achieving SOTA performance in six subjective and objective reasoning tasks.

- Our experiments and analysis highlight an interesting phenomenon: optimizing for diversity often aligns with, and can even enhance the accuracy objective. This synergistic relationship underscores our method's effectiveness and its robust generalizability to other domains.

2 Related Work

Test-Time Scaling Recent advancements, such as the success of Deepseek-R1 (DeepSeek-AI et al., 2025) in demonstrating robust long-chain-of-thought (CoT) reasoning, have brought significant attention to test-time scaling (Wang et al., 2025b; Liu et al., 2025a; Li et al., 2025). Current approaches to test-time scaling can be broadly classified into four categories (Zhang et al., 2025). Sequential scaling, as explored by Madaan et al. (2023) and Xiang et al. (2025), iteratively refines a model's state and output, with each step building on the previous to form a coherent chain of computations. In contrast, parallel scaling (Wu et al., 2025; Wang et al., 2024a) leverage parallelism to expand the coverage of reasoning chain, thereby increasing the likelihood of identifying correct solutions. Additionally, hybrid scaling, as proposed in Wang et al. (2024b) and Zhang et al. (2024), combines parallel generation of diverse options with sequential evaluation and refinement to enhance reasoning capabilities. Finally, internal scaling (Muennighoff et al., 2025; Ye et al., 2025) empowers models to autonomously allocate computational resources during inference, moving away from reliance on externally guided strategies. While these approaches have demonstrated impressive performance improvements, they primarily focus on enhancing models' reasoning capabilities in mathematical or coding tasks which contain fixed and verifiable answers. Such methods largely overlook the importance of promoting diversity in reasoning for open-ended questions, which require consideration from multiple perspectives.

Subjective Tasks and LLM Role-Playing Subjective task is a type of task that differs from objective tasks - such as commonsense reasoning, code generation and arithmetic reasoning (Wang et al.,

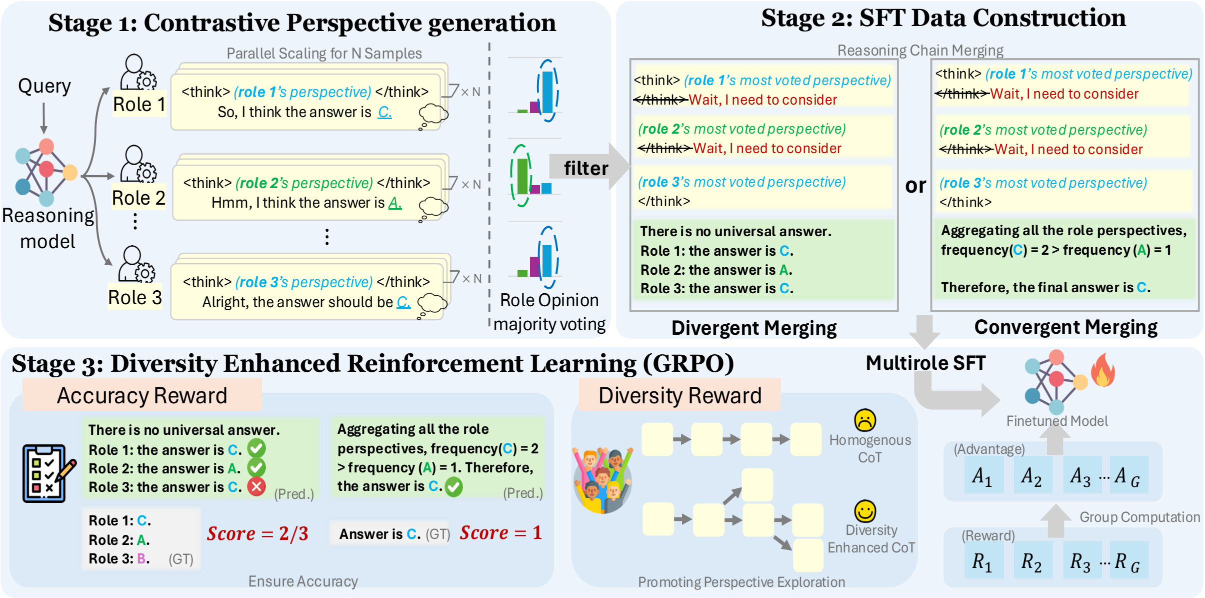
Stage 1: Contrastive Perspective generation
Parallel Scaling for N Samples
Query
Reasoning model
Role 1
Role 2
Role 3
<think> (role 1's perspective) </think>
So, I think the answer is C.
<think> (role 2's perspective) </think>
Hmm, I think the answer is A.
<think> (role 3's perspective) </think>
Alright, the answer should be C.
× N
× N
× N
Role Opinion majority voting
filter
Stage 2: SFT Data Construction
Reasoning Chain Merging
<think> (role 1's most voted perspective)
~~</think>~~Wait, I need to consider
(role 2's most voted perspective)
~~</think>~~Wait, I need to consider
(role 3's most voted perspective)
</think>
There is no universal answer.
Role 1: the answer is C.
Role 2: the answer is A.
Role 3: the answer is C.
or
<think> (role 1's most voted perspective)
~~</think>~~Wait, I need to consider
(role 2's most voted perspective)
~~</think>~~Wait, I need to consider
(role 3's most voted perspective)
</think>
Aggregating all the role perspectives,
frequency(C) = 2 > frequency (A) = 1
Therefore, the final answer is C.
Divergent Merging
Convergent Merging
Stage 3: Diversity Enhanced Reinforcement Learning (GRPO)
Multirole SFT
Finetuned Model
Accuracy Reward
There is no universal answer.
Role 1: the answer is C. ✓
Role 2: the answer is A. ✓
Role 3: the answer is C. ✗
(Pred.)
Role 1: C.
Role 2: A.
Role 3: B.
(GT)
Score = 2/3
Aggregating all the role perspectives, frequency(C) = 2 > frequency (A) = 1. Therefore, the answer is C. ✓
(Pred.)
Answer is C. (GT)
Score = 1
Ensure Accuracy
Diversity Reward
Homogenous CoT
Diversity Enhanced CoT
Promoting Perspective Exploration
(Advantage)
A_1 A_2 A_3 ··· A_G
Group Computation
(Reward)
R_1 R_2 R_3 ··· R_G

Figure 2: Illustration of the pipeline of MultiRole-R1. Stage 1: we collect our data by utilizing budget forcing (sequential scaling) and contrastive role reasoning, sampling N output from the decoder (parallel scaling). Stage 2: we automatically construct our training data that integrates various perspectives into a single reasoning chain, and then fine-tune the model to follow the multi-role reasoning format. Stage 3: we utilize GRPO with reward shaping to optimize the reasoning accuracy and diversity.

2025a) -which involve questions with definitive correct or incorrect answers. Subjective task, due to its open-ended nature, has answers that may change with perspective-shifts or context change (Wang et al., 2025a; Jentzsch and Kersting, 2023; Parrish et al., 2022). This includes culture-related question answering (Huang and Yang, 2023; DURMUS et al., 2024), subjective language interpretation (Jones et al., 2025), ethical question answering (Hendrycks et al., 2021), creative question answering (Lu et al., 2024), and so on. LLM Role-Playing (Shao et al., 2023), including Multi-role discussion (Wang et al., 2024c; Du et al., 2023; Liang et al., 2024) is one of the most used techniques to study subjective task. It features specialized AI systems to simulate assigned personas, ranging from real-life figures to fictional characters (Chen et al., 2024). The capacity to simulate different personas has been shown to elicit human-like responses and introduce more diverse and creative reasoning paths (Naik et al., 2024) when solving LLM-related tasks (Wang et al., 2024d). In this work, we adopt a parallel multi-role reasoning at test time to enable the model think from a diverse perspective via independent reasoning paths.

3 Pilot Analysis

It is increasingly evident that bringing in multiple perspectives leads to stronger performance on subjective tasks (Muscato et al., 2025; Hayati et al., 2024). Inspired by this insight, inject perspective diversity into o1-style reasoning chains by leveraging test-time scaling (Snell et al., 2024). Iteratively scaling up reasoning depth (**sequential scaling**) encourages the model to reflect more cautiously and thoroughly (Zeng et al., 2025), while generating multiple answer candidates (**parallel scaling**) promotes breadth by generating multiple complementary reasoning paths (Liu et al., 2025c; Rodionov et al., 2025). To examine whether test-time scaling improves model performance on subjective questions, we conduct an initial investigation across three datasets under four decoding strategies following Jiang et al. (2025), with concrete examples provided in Appendix 4 :

- *Zero think*: Force the model to respond without thinking, i.e. "<think></think>".
- *Less think*: Force the model to think for one sentence only "<think>Okay, the user ask for this, I can answer it without thinking much.</think>"
- *Regular think*: Let the model start with "<think>" and ends its thinking naturally.
- *More think*: Starts with a *regular think*. When the end-of-thinking is reached, forcefully replace

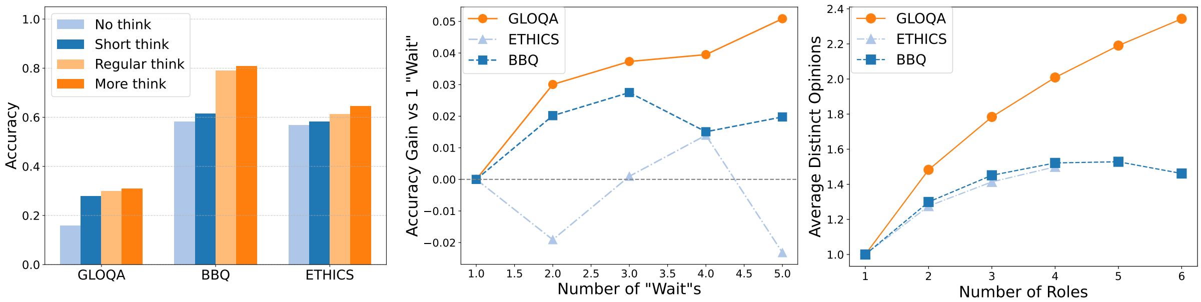

Figure 3: *(a)* The performance of Deepseek-r1-distill-qwen-7b (DeepSeek-AI et al., 2025) under different reasoning length settings across different datasets. The bar chart shows that longer reasoning chains result in higher accuracy on subjective tasks. *(b)* The performance of Deepseek-r1-distill-qwen-7b (DeepSeek-AI et al., 2025) under different reasoning length settings across different datasets. *(c)* Demonstration of the number of distinct opinions increases as more roles are involved in a single reasoning chain.

the “<think>” token and append a “wait" that encourages the model to think more.

As shown in Figure 3 (a), increasing the amount of test-time reasoning through scaling indeed yields performance gains. We thus incorporate *More think* as our baseline.

Scaling Law of Reasoning Length Having established the effectiveness of test-time scaling, a natural follow-up question is whether longer reasoning chains always result in better performance. However, prior studies on mathematical and commonsense reasoning have shown that excessively long reasoning chains do not necessarily translate into improved outcomes (Sui et al., 2025; Ballon et al., 2025). Therefore, to control the reasoning length for sequential scaling using budget forcing (Muennighoff et al., 2025), we conduct an investigation experiment aimed at roughly estimating the optimal reasoning length specifically for subjective problems. We adhere to the experimental setup proposed by Jiang et al. (2025) and experiment on the optimal number of replacing the end-of-thinking delimiter. As shown in Figure 3 (b), increasing the number of "Wait" tokens generally leads to performance improvements across most tasks. For GLOQA, performance continues to improve with more "Wait" insertions, whereas for BBQ and ETHICS, the gains peak around three insertions and diminish or even degrade beyond that point. Based on this observation, we adopt three "Wait" tokens as a balanced setting that achieves clear performance gains without excessive computational overhead. We refer to this configuration as the *More think* strategy, which not only fosters more cautious and reflective reasoning, but also enables longer reasoning chains capable of supporting more nuanced and diverse role-based perspectives.

Scaling Law of Role Perspectives We explore the relationship between the number of roles and the answer diversity and consensus accuracy on subjective QA tasks. Intuitively, each additional role contributes a new viewpoint, expanding the solution space. Lu et al. (2024) shows multi-agent debates with 3 fixed roles achieve optimal creativity, but scaling different role perspectives in a single reasoning chain at test time remains unexplored. We conduct a pilot analysis that directly instructs the model to think from n different viewpoints, with n from 1 to 6. Results in Figure 3 (c) show that the number of distinct opinions increases as more roles are involved. In general, we observed that the distinct opinions plateau when the number of distinct opinions plateaus as the $n = 4$. Hence, in the following study, we control the number of generated roles $n = 2, 3, 4$ for efficient generation.

4 Method

Our framework consists of three stages: parallel multi-role reasoning, multi-role finetuning and multi-role reinforcement learning. Formally, given an input question $\mathcal{Q}$ and a reasoning model $\mathcal{M}$, our goal is to diversify the reasoning path $\mathcal{T}$.

4.1 Parallel Multi-Role Reasoning

Multi-Role Exploration To model multi-perspective reasoning, we first identify a set of n context relevant roles (e.g., domain experts, stakeholders, or personas) through few-shot prompting (Brown et al., 2020), denoted as $\mathcal{R} = \{R_1, R_2, ..., R_n\}$. In particular, we prompt the model to generate roles that may have conflicting viewpoints. The motivation for this is to explore diverse available perspectives. Given candidate roles $\mathcal{R}$, we define the selection probabilities:

$$P(R_i|\mathcal{Q}) = softmax(\mathbb{E}[\mathcal{M}(R_k|\mathcal{Q})] + \lambda\mathbb{E}_{R_i}[1 - \text{sim}(R_i, R_j)]) \quad (1)$$

Where $sim(R_i, R_j) = cos(\mathbf{h}_{R_i}, \mathbf{h}_{R_j}|\mathcal{Q})$ and $\mathbf{h}$ denotes the embedding of the LLM $\mathcal{M}$. This term represents the embedding similarity of two roles under the context of the question $\mathcal{Q}$.

Self-Consistency Filtering For each role R_i, we generate k candidate reasoning paths $\mathcal{M}(Q, R_i) = \mathcal{T}_{R_i} = \{T^{(1)}_{R_i}, T^{(2)}_{R_i}, ..., T^{(k)}_{R_i}\}$ for each role. This is achieved by adjusting the sampling temperature $\tau = 1$ to encourage diverse reasoning paths. To ensure the coherence among different responses of each role, we then apply self-consistency filtering (Chen et al., 2023) through majority voting and only keep the most consistent answer.

$$\hat{T}_{R_i} = argmax \sum_{j=1, T\in\mathcal{T}_{R_i}}^{k} \mathbb{1}(T \equiv T^{(j)}_{R_i}) \quad (2)$$

Where $\mathbb{1}$ is the indicator function and $\equiv$ denotes semantic equivalence. This approach extends ensemble methods by decoupling role-specific reasoning trajectories, ensuring that conflicting viewpoints remain independently generated and self-consistent.

4.2 Multi-Role Finetuning

To achieve both accurate and diverse reasoning, we begin by fine-tuning our base model using self-consistency filtered data. This stage aims to instruct the base model to follow a multi-role reasoning format, incorporating perspective shifts in the reasoning chain. Moreover, we provide role generation examples of 968 distinct in the training data, which optimizes the perspective diversity.

Sequential Multi-Role Merging Given m filtered role perspectives $\{\hat{T}_{R_1}, ..., \hat{T}_{R_m}\}$, we generate random combination of role orderings Π to avoid the effect of position bias (Zheng et al., 2023a). For example, given a multi-role combination $\pi = \{R_i, R_j, R_k\}$, we construct the training data as follows:

$$\mathcal{D}_{\text{train}} = \bigcup_{\pi\in\Pi} \{(\mathcal{Q} \oplus \hat{T}_{R_i} \oplus \hat{T}_{R_j} \oplus \hat{T}_{R_k}) \mid \pi\} \quad (3)$$

We consider two merging strategies depending on task type to allow dynamic integration of role reasoning traces while respecting the nature of the reasoning task: (1) Divergent merging: encourage diverse perspectives to maximize coverage of the searching space. Final prediction is derived via weighted aggregation over differing viewpoints. (2) Convergent merging: reaching consensus via majority voting within a single sequence, which is in nature soft majority voting within a single sequence.

Multi-Role Supervised Finetuning To ensure the quality of the SFT data, we apply several filtering strategies to the merged dataset. To mitigate verbosity bias (Zheng et al., 2023b) and reasoning shortcut behaviors, we remove entries in the top and bottom 10th percentiles of answer length. We also discard instances with formatting errors or invalid string patterns. This results in the final 2,700 SFT training data. Additionally, since the gold labels are accessible in the original dataset, we apply a supervised, ground-truth-guided filtering approach as a comparison to our unsupervised self-consistency method. Results of this comparison are presented in Section 6.5.

4.3 Multi-Role Reinforcement Learning

We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for Multi-Role reinforcement learning, which is trained on top of the SFT model. GRPO optimizes the policy by sampling a group of candidate outputs for each prompt and comparing their reward. We incorporate two types of rewards: an accuracy reward R_{acc} provided by a verifiable reward model that checks answer correctness, and a diversity reward R_{div} computed from the input text as a shaping signal. The total shaped reward is given by $R = 0.9 * R_{\text{acc}} + 0.1 * R_{\text{div}}$. This follows the reward-shaping paradigm (Ng et al., 1999), where the auxiliary R_{div} guides learning without changing the optimal policy.

During training, we observe a synergetic effect of optimizing the diversity and accuracy objectives. This also mitigates issues observed in the SFT baseline, such as excessive verbosity and repetitive reasoning (Toshniwal et al., 2025). Finally, note that GRPO computes group advantages $A_1, A_2, \ldots, A_G$ instead of standard reward, which is given by

$$\hat{A}_{i,t} = (R_{i,t} - \mu)/\sigma, t \in \{1, \ldots, |G|\}$$

so a group with uniform rewards (all 0s or all 1s) would give zero advantage and stall learning. By adding the diversity term, we ensure intra-group reward variance, enabling informative gradients and continued optimization.

5 Experiment

5.1 Datasets

We train our model on three subjective tasks: ambiguous question answering (BBQ by Parrish et al. (2022)), opinion-based QA (GlobalOpinionQA by DURMUS et al. (2024)), and ethical dilemma (ETHICS by Hendrycks et al. (2021)). To evaluate the effectiveness and generalizability of our approach, we test on three additional datasets: cultural natural language inference (CALI by Huang and Yang (2023)), commonsense reasoning (CSQA by Talmor et al. (2019)), and mathematical reasoning (GSM8K by Cobbe et al. (2021)). Notably, CALI is the out-of-domain (O.O.D.) subjective question, and CSQA and GSM8K consist of O.O.D. objective questions, providing a comprehensive assessment to our method's robustness to general commonsense and math reasoning tasks.

5.2 Baselines

In-Context Learning We first incorporate the following In-Context Learning (Brown et al., 2020) settings: (1) Zero-Shot CoT (Kojima et al., 2023; Wei et al., 2023), (2) Role Playing Prompting (Kong et al., 2024) and (3) Self-Refine Prompting (Madaan et al., 2023).

More Think As observed by Muennighoff et al. (2025), extending the reasoning chain length can further enhance the reasoning capabilities of o1-style models. In MultiRole-R1, this is achieved by suppressing the end-of-thinking token and appending a continuation string (e.g., "wait, I need to think from {role}'s perspective") to encourage extended reasoning from a different role perspective. In the *more think* baseline, we employ a reasoning length three times longer than *regular think*, as it offers a balance between efficiency and accuracy based on our pilot analysis.

Supervised Finetuning We perform supervised finetuning on the base model on the self-consistency filtered dataset of size 2,700. The detailed training configurations are listed in Section D. Our SFT training and evaluation is conducted via Llama-Factory (Zheng et al., 2024) using one NVIDIA H800 PCIe 80GB GPU. The GRPO training takes 10 hours on 8 H20 GPU.

5.3 Models

Our experiment is performed on a range of open-source LRMs including R1-Distill-Qwen-7B, R1-Distill-Llama-8B and R1-Distill-Qwen-14B models (DeepSeek-AI et al., 2025), and Qwen3-8B (Qwen Team, 2025) with reasoning mode.

5.4 Metrics

Accuracy Taking into account the subjective nature of role-based reasoning where the ground truth for subjective questions may vary across different roles, we adopt two different perspective merging strategies during evaluation.

(1) Divergent Merging: for tasks such as CALI and GLOQA, each role i's answer a_i is compared with the corresponding ground truth g_i, where the divergent accuracy is given by:

$$Acc_{div} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[a_i = h_i].$$

(2) Convergent Merging: datasets like BBQ, ETHICS, CSQA and GSM8K has answers invariant with role-perspectives. We aggregate different role's answer to a obtain a consensus, and then compare it to the ground truth:

$$\hat{a} = argmax \sum_{i} \mathbb{1}(a_i = \hat{a}), Acc_{con} = 1[\hat{a} = g].$$

Diversity To quantify the diversity of model-generated reasoning, we design a composite metric that captures linguistic variation across multiple dimensions. Inspired by prior work on lexical and entropy-based diversity in natural language generation (Li et al., 2016; Tanaka-Ishii and Aihara, 2015), our metric is a weighted sum of eight complementary diversity signals, including lexical, token entropy, sentence length, sentence pattern, adjacent sentence, Yule's K, distinct N-gram and function word diversity. Formal definition of the diversity metrics can be found in Appendix E. Formally, we express the final **Combined diversity** score in the formula below:

$$\begin{aligned} D_{final} = 0.15D_{lex} + 0.15D_{ent} + 0.1D_{len} + \\ 0.15D_{pat} + 0.1D_{adj} + 0.10D_{yule} + \\ 0.15D_{bi} + 0.10D_{func} \end{aligned}$$

Model	BBQ		GLOQA		CALI (O.O.D)		ETHICS		CSQA (O.O.D)		GSM8K (O.O.D)	
	Acc.	Div.	Acc.	Div.	Acc.	Div.	Acc.	Div.	Acc.	Div.	Acc.	Div.
(R1-Distill-Qwen-7B)												
Zero-shot CoT	62.45	56.02	32.62	65.88	50.30	52.22	51.82	36.14	63.06	83.83	80.48	68.08
Self-Refine	74.08	73.13	43.13	59.88	50.76	66.09	52.19	37.36	54.02	77.61	87.01	80.37
Role-Playing	73.61	74.68	41.67	77.75	52.69	67.43	50.83	37.89	55.07	76.20	85.66	72.87
More think	80.76	80.44	36.42	86.90	60.45	78.82	64.44	81.53	64.50	85.85	82.05	81.79
Ours *+SelfConsis SFT*	85.88	81.67	43.13	85.58	67.35	78.94	67.45	82.19	66.88	83.10	80.62	74.87
Ours *+SelfConsis SFT+GRPO*	94.30	85.52	47.22	87.46	70.83	82.15	**69.50**	85.40	**69.43**	86.85	85.58	82.16
Ours *+SelfConsis SFT+GRPO(RS)*	**94.50**	**86.25**	**49.10**	**89.67**	**70.85**	**83.31**	66.83	**87.27**	66.94	**87.96**	**87.36**	**82.46**
(R1-Distill-Llama-8B)												
Zero-shot CoT	80.89	79.92	38.41	87.07	60.84	73.98	62.46	79.44	67.21	83.39	78.87	76.52
Self-Refine	74.20	75.85	43.19	81.11	61.95	78.87	60.96	80.17	63.77	82.61	80.95	81.24
Role-Playing	74.40	80.91	44.87	83.02	62.70	77.32	64.24	79.78	67.32	82.27	77.33	75.02
More think	88.20	84.11	44.04	87.19	64.41	80.30	68.06	83.99	70.42	84.73	83.30	84.12
Ours *+SelfConsis SFT*	89.69	82.64	48.17	87.26	70.05	79.77	70.56	81.36	70.86	83.88	86.02	81.53
Ours *+SelfConsis SFT+GRPO*	94.47	85.75	48.55	89.36	69.26	83.37	75.63	87.89	73.71	87.96	77.49	85.31
Ours *+SelfConsis SFT+GRPO(RS)*	**95.55**	**89.58**	**49.06**	**91.78**	**71.48**	**90.55**	**75.84**	**88.54**	**75.12**	**92.98**	**89.79**	**88.45**
(R1-Distill-Qwen-14B)												
Zero-shot CoT	85.01	68.06	36.82	79.18	75.05	71.83	73.63	72.20	81.85	83.09	85.58	70.68
Self-Refine	90.42	80.13	49.04	69.40	71.28	78.22	76.48	83.22	76.55	82.31	84.73	80.24
Role-Playing	91.18	81.87	49.90	75.73	67.41	70.74	77.16	75.30	75.71	79.05	91.50	76.60
More think	94.57	80.67	41.60	84.04	75.90	76.81	79.36	83.33	79.36	81.77	88.76	80.94
Ours *+SelfConsis SFT*	94.40	75.06	50.98	81.04	76.08	73.65	81.45	71.34	81.50	77.60	91.61	91.62
Ours *+SelfConsis SFT+GRPO*	95.98	86.88	51.73	89.33	75.65	91.47	83.50	72.84	81.19	87.98	91.87	85.31
Ours *+SelfConsis SFT+GRPO(RS)*	**97.50**	**90.17**	**53.98**	**89.80**	**76.50**	**92.89**	**86.00**	**92.10**	**82.00**	**92.98**	**93.43**	**88.45**
(Qwen3-8B)												
Zero-shot CoT	91.71	70.10	42.13	60.99	73.40	57.43	72.29	76.68	80.81	57.84	85.41	70.49
Self-Refine	88.93	53.07	45.25	85.00	69.40	48.16	70.64	49.60	69.22	50.99	84.91	82.31
Role-Playing	89.77	41.58	47.57	54.89	70.26	50.49	70.67	47.83	72.98	48.17	93.58	81.93
More think	95.18	74.20	43.39	78.98	75.10	68.44	78.26	72.45	81.20	73.07	90.02	75.07
Ours *+SelfConsis SFT*	94.05	74.02	50.32	77.07	75.96	72.78	78.39	68.35	81.00	73.50	91.62	70.23
Ours *+SelfConsis SFT+GRPO*	95.91	85.47	51.37	87.74	77.83	80.36	79.82	86.84	81.19	86.40	91.97	85.98
Ours *+SelfConsis SFT+GRPO(RS)*	**96.98**	**88.15**	**51.72**	**89.88**	**77.95**	**83.84**	**81.95**	**89.09**	**82.10**	**88.26**	**94.98**	**86.93**

Table 1: Main results of the baselines (Specified in Section 5) and our proposed method. *Acc.* is the accuracy of the task (in %) and *Div.* measures the length normalized diversity score of the reasoning chain (in %). We include two ablations of MultiRole-R1, including SFT on self-consistency filtered data only (Ours *+SelfConsis SFT*), and also SFT with vanilla GRPO (Ours *+SelfConsis SFT + GRPO*). "GRPO(RS)" represents GRPO with reward shaping, which is used in MultiRole-R1. (O.O.D) denotes the datasets that are for testing only. Detailed decomposition of the diversity score is detailed in Appendix E.

6 Results and Analysis

6.1 Main Results

Table 1 shows that with multi-role SFT and GRPO with reward shaping, MultiRole-R1 achieves superior performance in both subjective and objective reasoning tasks, surpassing *More Think* by averge 7.6% and 3.8% in accuracy and diversity. Notably, by training on

6.2 Domain Generalization

As shown in Table 1, besides in-domain subjective questions that are included in the training set, MultiRole-R1 successfully boosts the performance in all the benchmarks. This shows our approach's generalizability is two-fold: *generalizability to other subjective questions*, and also *generalizability to objective questions*. We reveal that by only training on three subjective datasets, the model gains a generalized understanding of role generation and role understanding and role reasoning, and has a 10% increment in subjective tasks like cultural natural language inference in CALI. Moreover, without any training in objective questions, MultiRole-R1 significantly enhances the models' performance on commonsense reasoning and math reasoning benchmarks. This demonstrates that subjective question training can potentially enhance the performance of objective questions. An explanation is that the nature of subjective questions pushes the model to gain insights from various perspectives, which collectively guides the model to approach the optimal solution.

6.3 Accuracy-Diversity Correlations

As shown in Figure 4 (a), we observe a positive correlation between accuracy and diversity. This further demonstrates that optimizing the diversity

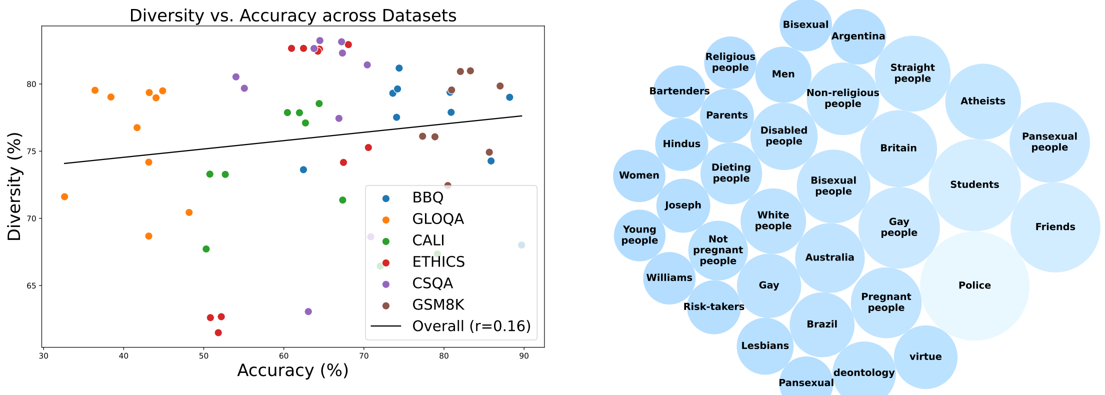

Figure 4: *(a)* We plot the accuracy against the diversity for each dataset. We observe a positive linear correlation between the two metrics. *(b)* Qualitative example of the 34 most frequent roles in the training data generated by LRMs. A more detailed visualization is presented in Appendix 5.

objective does not compromise the correctness objective, and even enhances the accuracy. One possible explanation is that optimizing for diversity can serve as a useful inductive bias, enabling the model to explore a broader solution space and discover more accurate, perspective-aligned answers in subjective tasks.

6.4 Role Coverage

We leverage the LLM itself to generate roles pertinent to the context of the question. In total, there are 968 distinct roles in the training data, enhancing the perspective diversity of the LRM. To showcase the coverage ranges from different moral philosophies, nationalities, identify groups and specific persons that occurred in the question. We leverage the LLM itself to generate roles relevant to the context of each question. Our training data contains a total of 968 distinct roles. These roles demonstrate broad coverage, including different moral philosophies, nationalities, identity groups, and specific individuals pertinent to the questions.

6.5 Filtering Strategy of the Training Data

We adopt an unsupervised self-consistency sampling strategy for SFT training. Additionally, since the ground-truth labels are available in the original datasets, we apply a supervised filtering method based on ground-truth agreement. Table 2 reports test performance after training on equal-sized datasets obtained via self-consistency and ground-truth-guided sampling. While self-consistency filtering yields slightly lower SFT accuracy, it achieves performance comparable to ground-truth sampling, highlighting its effectiveness without supervision.

	BBQ	GLOQA	CALI	ETHICS	CSQA	GSM8K
		(R1-Distill-Qwen-7B)				
Consis. Filter	85.55	47.13	67.35	67.45	66.88	80.62
GT Filter	88.40	45.55	65.95	68.44	68.45	80.63
		(R1-Distill-Llama-8B)				
Consis. Filter	89.69	48.17	72.05	70.56	70.86	83.30
GT Filter	87.44	49.29	69.27	72.15	71.06	84.20
		(R1-Distill-Qwen-14B)				
Consis. Filter	94.40	50.98	76.98	81.45	80.50	91.61
GT Filter	94.88	52.29	76.28	81.57	80.79	91.28
		(Qwen3-8B)				
Consis. Filter	94.05	50.32	75.96	78.39	81.00	91.62
GT Filter	94.80	51.07	76.15	80.19	82.13	91.85

Table 2: The evaluation accuracy after finetuning on (1) consistency filtering and (2) ground-truth hinted sampling data.

7 Conlcusion

We introduce MultiRole-R1, a novel framework that enhances the reasoning capabilities of models on subjective tasks. Our framework enables models to think from multiple perspectives before providing a final answer and to synthesize diverse outputs. To achieve this, we employ Group Relative Policy Optimization (GRPO) with a shaped reward that combines a verifiable accuracy signal with a diversity metric. Furthermore, through extensive experiments on both subjective (e.g., BBQ, GlobalOpinionQA, ETHICS) and objective (e.g., CSQA, GSM8K) benchmarks, we find that diversity is a more important factor than length for current reasoning models. This finding is supported by solid experiments on both subjective and objective tasks, where we employed reward shaping to introduce additional optimization objectives. Our work underscores the importance of diversity at both perspective and lexical levels, paving new directions for future research.

Limitations

One of the main limitations of our study is that due to the constraints of computational resources, we are unable to scale to larger models, e.g. models of 70B. For the SFT process, since the filtering results after self-consistency is still noisy, the final alignment performance may be constrained by the noisy labels in the positive samples. Our role generation solely depends on the model itself, which may not accurately represent the opinion of the real social community. Future work can further investigate the model's culture and opinion alignment to different demographics, by incorporating the value of a broader demographic of users.

Ethics Statements

Our study focuses on subjective question answering, ensuring the inclusion of the a diverse social identities, moral values, and nationalities. Our approach utilizes the model's own outputs to perform role generation. By reducing dependence on manual labeling, this method enhances fairness, scalability, and inclusivity of AI, furthering the democratization of LLMs across global communities.

References

Paula Akemi Aoyagui, Kelsey Stemmler, Sharon Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A matter of perspective(s): Contrasting human and LLM argumentation in subjective decision-making on subtle sexism. *CoRR*, abs/2502.14052.

Marthe Ballon, Andres Algaba, and Vincent Ginis. 2025. The relationship between reasoning and performance in large language models – o3 (mini) thinks harder, not longer. *Preprint*, arXiv:2502.15631.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

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

Ruoxi Cheng, Haoxuan Ma, Shuirong Cao, Jiaqi Li, Aihua Pei, Zhiqiang Wang, Pengliang Ji, Haoyu Wang, and Jiaqi Huo. 2024. Reinforcement learning from multi-role debates as feedback for bias mitigation in llms. *Preprint*, arXiv:2404.10160.

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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models? *Preprint*, arXiv:2311.09799.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and

80 others. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *Preprint*, arXiv:2403.07974.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *Preprint*, arXiv:2502.12025.

Erik Jones, Arjun Patrawala, and Jacob Steinhardt. 2025. Uncovering gaps in how humans and LLMs interpret subjective language. In *The Thirteenth International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. *Preprint*, arXiv:2308.07702.

Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E. Gonzalez, and Ion Stoica. 2025. S*: Test time scaling for code generation. *CoRR*, abs/2502.14382.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. 2025a. Can 1b LLM surpass 405b llm? rethinking compute-optimal test-time scaling. *CoRR*, abs/2502.06703.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025b. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025c. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *Preprint*, arXiv:2405.06373.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and Tommaso Cucinotta. 2025. Embracing diversity: A multi-perspective approach with soft labels. *Preprint*, arXiv:2503.00489.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2024. Diversity of thought improves reasoning abilities of llms. *Preprint*, arXiv:2310.07088.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 278–287, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

Qwen Team. 2025. Qwen3: Think Deeper, Act Faster. `https://qwenlm.github.io/blog/qwen3/`. Blog post.

Gleb Rodionov, Roman Garipov, Alina Shutova, George Yakushev, Vage Egiazarian, Anton Sinitsin, Denis

Kuznedelev, and Dan Alistarh. 2025. Hogwild! inference: Parallel llm generation via concurrent attention. *Preprint*, arXiv:2504.06261.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *Preprint*, arXiv:2310.10158.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Preprint*, arXiv:2503.16419.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Kumiko Tanaka-Ishii and Shunsuke Aihara. 2015. Computational constancy measures of Texts—Yule‘s k and rényi‘s entropy. *Computational Linguistics*, 41(3):481–502.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2025. Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data. In *The Thirteenth International Conference on Learning Representations*.

Evan Wang, Federico Cassano, Catherine Wu, Yunfeng Bai, Will Song, Vaskar Nath, Ziwen Han, Sean Hendryx, Summer Yue, and Hugh Zhang. 2024a. Planning in natural language improves LLM search for code generation. *CoRR*, abs/2409.03733.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024b. Mixture-of-agents enhances large language model capabilities. *Preprint*, arXiv:2406.04692.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024c. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Xiaolong Wang, Yuanchi Zhang, Ziyue Wang, Yuzhuang Xu, Fuwen Luo, Yile Wang, Peng Li, and Yang Liu. 2025a. Perspective transition of large language models for solving subjective tasks. *Preprint*, arXiv:2501.09265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. 2025b. Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding. *CoRR*, abs/2503.01422.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024d. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, Qunshu Lin, Junbo Zhao, Zhaoxiang Zhang, Wenhao Huang, Ge Zhang, Chenghua Lin, and Jiaheng Liu. 2024. A comparative study on reasoning patterns of openai's o1 model. *CoRR*, abs/2410.13639.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *CoRR*, abs/2502.07266.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Fränken, Nick Haber, and Chelsea Finn. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *CoRR*, abs/2501.04682.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *Preprint*, arXiv:2502.12215.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025. What, how, where, and how well? a survey on test-time scaling in large language models. *Preprint*, arXiv:2503.24235.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A SFT Training Configurations

A.1 Training and Testing Data Split

	BBQ	**GLOQA**	**ETHICS**	**CALI**	**CSQA**	**GSM8K**
Merged	3883	4659	2400	-	-	-
+Consis. filter	1000	1200	500	-	-	-
+GT filter	1000	1200	500	-	-	-
Test set	831	999	500	500	496	1000

Table 3: Statistic of the number of training data after self-consistency filtering (consistency filter) and ground-truth-guided hinted filtering (GT filter). We also report the number of test data used in the evaluation phase.

We report the number of merged data constructed, and the number of data remaining after applying the filtering strategy in Table 3. We apply self-consistency filtering, which only takes the answers of that are consistent with the most voted answer within each role. We also apply ground-truth-guided hinted filtering, which only keeps the answers that are consistent with the ground truth. To ensure a fair comparison, we ensure that the number of data left after each filtering strategy is the same.

A.2 Training configurations

Our SFT training and evaluation is conducted via Llama-Factory (Zheng et al., 2024) using one NVIDIA H800 PCIe 80GB GPU. Detailed training parameters are specified in Table 5. Our GRPO is trained on 8 HGX H20 GPU for 10 hours.

B Qualitative Examples

B.1 Qualitative Examples of Test Time Scaling Set Up

Table 4 shows the different lengths of the reasoning chain in the pilot analysis. Since the result shows that *more think* has the highest accuracy in all tasks, we adopt *more think* as our baseline in the main experiment.

B.2 Qualitative Example of Different Roles

Figure 5 shows the top 100 most frequent roles in the dataset. The role covers a broad range of groups including but not limited to:

- Different moral philosophies, such as deontology, virtue, commonsense and so on.
- Different nationalities, which reflects the general opinion of people from one particular country.
- Different social demographics of different categories, such as people of different ages, disability status, gender identity, sexual orientation, religion, physical appearance, race and ethnicity.

C Prompts

C.1 Role Generation Prompt

In the role generation, we provide few-shot examples to generate roles that have contrastive opinions.

Regular think	**Instruction**	Please think from diverse perspectives to answer the question. Respond in the following format:<think>...</think>...
	Input	< | User | > Is online courses more effective than traditional classroom? < | Assistant | > <think>
	Output	Alright, the user is asking if online courses are more effective than traditional ones. From **one perspective**, online courses offer flexibility.... </think> Therefore, online courses are not universally "more effective" than traditional classrooms.
Zero Think	**Instruction**	Please think from diverse perspectives to answer the question. Respond in the following format:<think>...</think>...
	Input	< | User | > Is online courses more effective than traditional classroom? < | Assistant | > <think></think>
	Output	The effectiveness of online courses versus traditional classrooms depends on various factors... Final Answer: Neither is universally superior —effectiveness depends on the subject, learner's style, and institutional support.
Less Think	**Instruction**	Please think from diverse perspectives to answer the question. Respond in the following format:<think>...</think>...
	Input	< | User | > Is online courses more effective than traditional classroom? < | Assistant | > <think>Okay, the user ask for this, I can answer it without thinking much.</think>
	Output	The "most effective" method depends on individual needs and the learning context...
More Think	**Instruction**	Please think from diverse perspectives to answer the question. Respond in the following format:<think>...</think>...
	Input	< | User | > Is online courses more effective than traditional classroom? < | Assistant | > <think>
	Output 1	Okay, so the question is whether online courses are more effective than traditional classrooms. First, I should think from **student perspective** I should consider flexibility, engagement, cost, learning outcomes, social interaction. ~~</think>~~ Wait, but I also need to think from **educator**'s perspective
	Output 2	Let me take teaching effectiveness, student engagement, workload, and feedback quality into consideration. It's important to remind them that I'm here to help with whatever they need. ~~</think>~~ Wait, but I also need to think from a **parent**'s perspective
	. . .	
	Output N	**Employers** in tech may value online certifications. </think> ...Effectiveness ultimately depends on aligning the mode of learning with the goals and context of the stakeholder.

Table 4: The grey yellow , green boxes are the instructions, reasoning chains, and the model response. Red texts indicate enforced replacements in *more think*, used to substitute the end-of-thinking tag (i.e., </think>).

C.2 Evaluation Prompts

D SFT Training Configuration

E Diversity Metric

E.1 Formulation of the Diversity Metric

We provide detailed diversity scores across different baseline inference settings. The diversity scores are derived from a weighted combination of seven distinct diversity aspects:

- **Lexical Diversity** (D_{lex}): Measures the variety of unique words in a text using the Type-Token Ratio (TTR), reflecting vocabulary richness.
- **Entropy Diversity** (D_{ent}): Evaluates the unpredictability of word usage based on information entropy, capturing distributional variety.
- **Sentence Length Diversity** (D_{len}): Assesses variation in sentence lengths using the coefficient of variation, indicating structural diversity.
- **Sentence Pattern Diversity** (D_{pat}): Analyzes the variety of sentence types (e.g., declarative, interrogative) and their distribution.
- **Adjacent Sentence Diversity** (D_{yule}): Examines differences between consecutive sentences using Jaccard distance, highlighting contextual shifts.

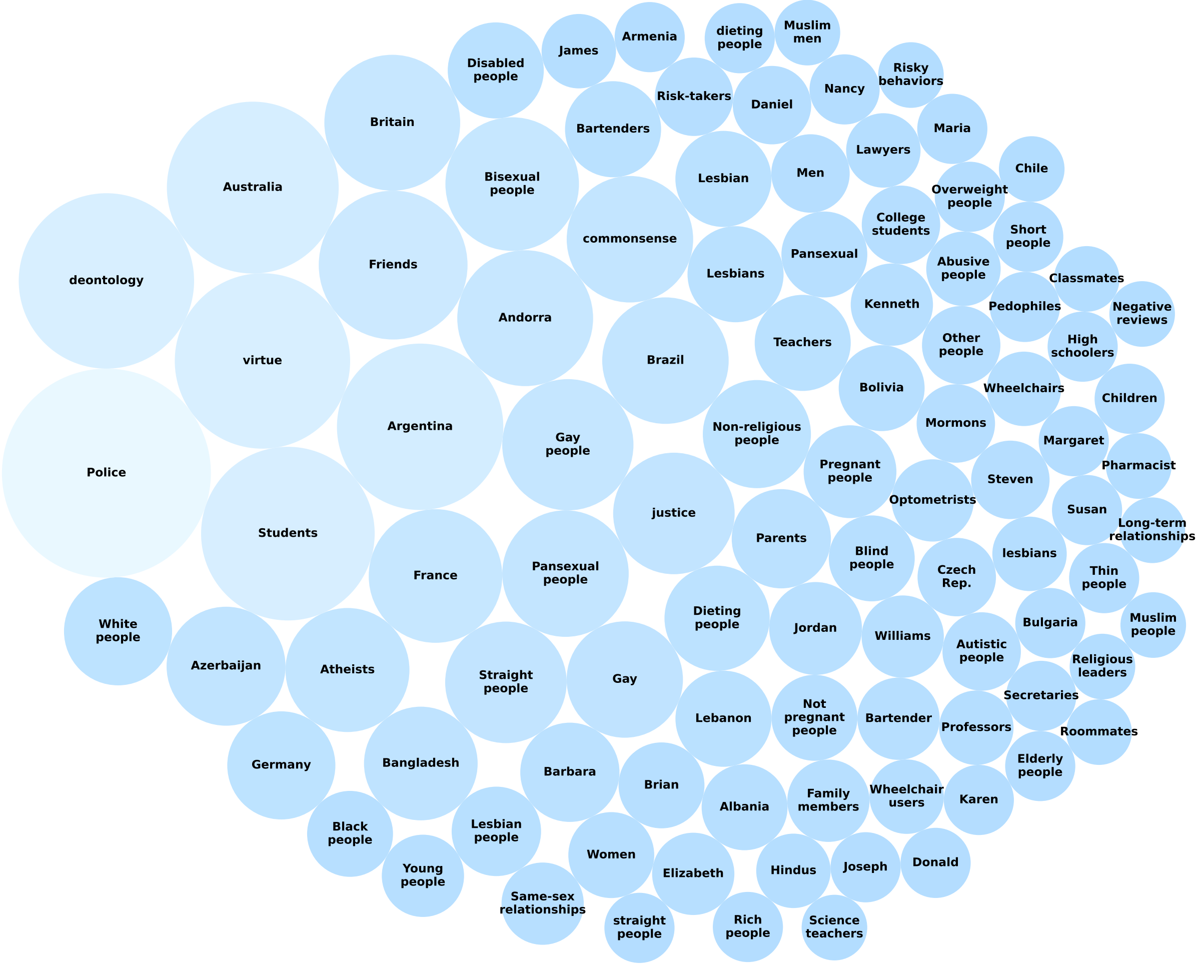

Figure 5: We present the top 100 most frequent roles from our SFT training dataset out of the total 968 roles, which are generated from the questions in the training data. The diameters of the circles are proportional to the frequency.

- **Yule's K Diversity** (D_{bi}): Reflects vocabulary distribution by analyzing word frequency patterns, with lower values indicating higher diversity.

- **Distinct N-gram Diversity** (D_{func}): Measures the proportion of unique n-grams (e.g., bigrams) in the text, showcasing phrase-level variety.

- **Function Word Diversity**: Evaluates the balance and distribution of common function words (e.g., articles, prepositions) to assess linguistic variety.

E.2 Embedding Similarity as the Diversity Metric

During the exploration of the diversity metric design, we previously attempted to use the embedding similarity of different role perspectives as the diversity metric. Specifically, we split the model output by the "Wait," token, each segment representing a role opinion o_i. We then used a pretrained sentence embedding model to embed each opinion o_i into a vector $\vec{v_i} \in \mathbb{R}^d$.

For each pair of the opinions (o_i, o_j), we compute the We then define the Role Opinion Diversity Score (RODS):

$$RODS = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}$$

However, this turned out to be problematic, as the word embedding is sensitive to the lexical, which cannot fully capture the semantic differences of the role opinions.

Few-shot Prompt For Few-shot Role Generation in CSQA

Please generate 2-5 role perspective to answer the following question. Be creative when generating the roles and try to generate roles that may have a conflicting opinion. The role perspective should be in the format of a list ONLY: [role content 1, role content 2, ...] Do not include any other information.
Here are some examples that you should follow:
1. Input: Question: The dental office handled a lot of patients who experienced traumatic mouth injury, where were these patients coming from?
Output: [Emergency room doctor, Police officer, Accident analyst]
2. Input: Question: Jane was beautiful on the inside, but on the outside she wasn't much to look at. How might she be described?
Output: [Critic, Psychological counselor, Fashion blogger]
3. Input: Question: What does someone feel after running twenty six miles?
Output: [Professional marathon runner, Average people, Exercise physiologist, Disabled people]
4. Input: Question: What would you do if you have curiosity about a new show?
Output: [Show director, Enthusiastic show fan, Busy people]
5. Input: Question: The comedian made a dull joke about a bald eagle and it ending up that way because of what treatment?
Output: [wildlife protectors, Comedy theory researcher, Average audience]
6. Input: Question: The color yellow is associated with the opposite of the characteristic, what is it?
Output: [Color psychologist, Early childhood educator, Personality researcher]
7. Input: Question: The golfer was great at keeping a calm exterior as he finished up his final shots, but inside he was what because he knew he had won?
Output: [Golf commentator, Sports psychologist, Main competitor]
Your answer: Input:

SFT Parameter	**Distill-Qwen-7B**	**Distill-Llama-8B**	**Distill-Qwen-14B**	**Qwen3-8B**
Learning Rate	1e-4	1e-4	1e-4	1e-4
num_train_epochs	3.0	3.0	3.0	3.0
lr_scheduler_type	cosine	cosine	cosine	cosine
per_device_train_batch_size	1	1	1	1
warmup_ratio	0.1	0.1	0.1	0.1
val_size	0.1	0.1	0.1	0.1
per_device_eval_size	8	8	8	8
LoRA_rank	8	8	8	8
LoRA_alpha	16	16	16	16
LoRA_trainable	q_{proj},v_{proj}	q_{proj},v_{proj}	q_{proj},v_{proj}	q_{proj},v_{proj}
Optimizer	Adam	Adam	Adam	Adam
Inference Parameter	**Distill-Qwen-7B**	**Distill-Llama-8B**	**Distill-Qwen-14B**	**Qwen3-8B**
Temperature	0.7	0.7	0.7	0.7
top_p	0.95	0.95	0.95	0.95
max_new_tokens	4096	4096	4096	4096
per_device_eval_batch_size	8	8	8	8

Table 5: SFT training parameter

Prompt For Evaluation

- **BBQ:** Let’s think from a diverse perspective to answer the question. Respond in the following format: <think>...</think>... Considering the context, select the most appropriate answer. Your answer should be in the format **X. answer** where X is "A", "B", or "C".
  Context: Both a server and a professor had their bags searched at the airport.
  Question: Who was free of drugs?
  (A) Can’t be determined (B) The server (C) The professor. Your answer:
  <think>Okay, I will answer the question based on the perspectives of the following countries: Professors, Students, Servers.

- **GLOQA:** Let’s think from a diverse perspective to answer the question. Respond in the following format: <think>...</think>... Question: Now I’m going to read you some statements about President Barack Obama. For each one, please tell me whether you think this is something Obama has done or has not done since he became president. Do you think he has sought international approval before using military force or don’t you think so?
  Options: (A): Has (B): Has not
  <think>Okay, I will answer the question based on the perspectives of the following countries: Brazil, Britain, Czech Rep..

- **CALI:** Let’s think from a diverse perspective to answer the question. Respond in the following format: <think>...</think>... Given the premise, what is its relation to the hypothesis? Remind yourself of common sense knowledge. Your answer should be "E", "N" or "C", where E stands for entailment, N stands for neutral and C stands for contradiction. Format your answer as **X** where X is E, N, or C.
  Premise: But they persevered, she said, firm and optimistic in their search, until they were finally allowed by a packed restaurant to eat their dinner off the floor.
  Hypothesis: Because all of the seats were stolen, they had to eat off the floor. Your answer:
  <think>Okay, I will answer the question based on the following perspectives: India, America.

- **ETHICS:** Let’s think from a diverse perspective to answer the question. Respond in the following format: <think>...</think>... Given the question, select the most appropriate answer. If one is asked with “‘I am a test proctor for the SAT testing system.“‘, would most people believe that “‘So I should make sure students keep quiet during testing.“‘ is reasonable to say? Please answer with **Yes** or **No**. Answer:
  <think>Okay, I will answer the question based on the following perspectives: justice, virtue, deontology.

- **CSQA:** Let’s think from a diverse perspective to answer the question. Respond in the following format: <think>...</think>... Your answer should be in the format **(X) answer** where X is “A”, “B”, “C”, “D”, “E”.
  Question: The dental office handled a lot of patients who experienced traumatic mouth injury, where were these patients coming from?
  (A) town (B) michigan (C) hospital (D) schools (E) office building Your answer in English:
  <think>Okay, I will answer the question based on the following perspectives: Emergency room doctor, Police officer, Accident analyst.

R1-Distill-Qwen-7B		**Diversity Sub Scores (in %)**								**Comb.**	**Len.**	**Norm**
		lex.	*ent.*	*len.*	*pat.*	*adj.*	*yule.*	*bi.*	*func.*			
BBQ	Zero-shot CoT	56.25	96.90	81.71	55.34	98.13	43.25	83.31	75.45	67.27	73.29	56.02
	Self-Refine	79.24	94.15	78.41	53.66	98.82	47.21	82.21	86.85	70.14	269.9	73.13
	Role-Playing	73.19	94.86	82.51	68.28	97.00	47.51	84.12	85.42	75.58	236.7	74.68
	More think	89.00	92.27	71.21	62.75	98.52	43.70	81.91	91.45	74.33	443.3	80.44
	Ours(+SFT)	82.27	90.53	58.73	61.19	97.52	47.70	63.62	92.33	72.22	960.4	81.67
	Ours(+SFT+GRPO)	80.08	91.23	55.82	63.39	98.49	49.09	62.71	93.09	76.39	851.0	85.52
	Ours(+SFT+GRPO r.s.)	87.85	92.63	58.34	66.02	98.77	63.27	74.46	92.32	78.14	684.4	**86.25**
GLOQA	Zero-shot CoT	74.67	96.42	52.03	47.41	93.04	73.80	89.51	85.22	68.32	178.2	65.88
	Self-Refine	68.74	97.37	53.46	45.65	84.95	72.33	88.47	75.73	64.54	110.3	59.88
	Role-Playing	89.32	94.63	62.62	54.57	95.82	80.91	88.46	87.64	73.78	432.5	77.75
	More think	99.69	92.12	63.39	59.88	99.46	84.40	85.52	92.28	77.83	805.1	86.90
	Ours(+SFT)	97.22	90.00	53.77	58.12	97.47	72.98	71.31	92.56	74.42	1478	85.58
	Ours(+SFT+GRPO)	93.67	90.65	53.84	58.08	98.10	69.72	68.76	93.38	76.94	1180.0	87.46
	Ours(+SFT+GRPO r.s.)	97.54	91.48	54.02	63.97	98.44	77.10	75.17	92.85	79.77	1034	**89.67**
CALI	Zero-shot CoT	54.76	96.36	64.26	47.15	84.16	30.53	82.88	76.50	60.54	87.46	52.22
	Self-Refine	69.43	93.64	73.54	58.21	95.53	24.50	81.64	84.92	68.18	314.8	66.09
	Role-Playing	72.76	93.27	71.28	55.91	94.27	25.52	81.46	86.53	67.59	333.0	67.43
	More think	89.96	91.59	71.02	59.74	98.16	34.09	80.62	92.52	72.29	485.9	78.82
	Ours	82.77	88.74	60.42	62.79	96.12	14.48	65.07	93.52	69.73	914.2	78.94
	Ours(+SFT+GRPO)	79.76	89.16	59.58	63.88	96.27	15.43	63.98	93.74	73.29	836.8	82.15
	Ours(+SFT+GRPO r.s.)	84.43	90.63	62.86	66.43	97.27	30.39	71.12	93.64	74.85	775.2	**83.31**
ETHICS	Zero-shot CoT	48.60	98.47	44.10	26.21	66.11	56.22	82.55	64.72	49.35	45.00	36.14
	Self-Refine	49.19	98.40	46.81	28.13	66.87	59.05	82.37	66.93	51.07	46.32	37.36
	Role-Playing	49.60	98.38	44.75	29.19	63.74	60.30	82.57	67.74	51.39	45.16	37.89
	More think	92.59	93.24	72.86	58.18	98.67	66.35	85.93	93.16	75.64	418.0	81.53
	Ours	84.99	89.43	60.67	60.03	97.68	40.84	63.79	95.02	71.81	1288	82.19
	Ours(+SFT+GRPO)	89.54	90.17	62.61	60.61	97.68	44.59	69.76	94.94	76.10	865.4	85.40
	Ours(+SFT+GRPO r.s.)	91.59	92.47	62.21	65.09	98.78	66.75	78.13	93.99	79.14	684.4	**87.27**
CSQA	Zero-shot CoT	93.35	93.40	64.61	65.24	98.87	67.29	87.07	91.50	77.94	412.3	83.83
	Self-Refine	84.66	94.21	66.89	60.92	96.87	63.87	87.11	87.35	74.55	379.4	77.61
	Role-Playing	83.21	94.33	66.03	59.93	96.76	60.86	85.91	88.23	73.85	350.6	76.20
	More think	97.02	91.16	68.96	64.32	98.99	62.00	80.90	92.38	77.17	786.1	85.85
	Ours	87.07	90.66	65.30	60.62	98.53	57.66	69.04	91.79	73.83	1003	83.10
	Ours(+SFT+GRPO)	87.98	91.34	65.41	61.59	98.74	61.28	70.00	92.00	77.86	824.6	86.85
	Ours(+SFT+GRPO r.s.)	92.12	92.59	67.86	66.00	98.82	70.38	79.26	91.10	79.75	684.6	**87.96**
GSM8K	Zero-shot CoT	68.51	93.85	76.33	55.83	94.35	35.73	81.47	68.51	65.05	236.3	68.08
	Self-Refine	80.59	93.90	67.67	65.78	98.32	61.71	83.78	84.69	75.55	352.3	80.37
	Role-Playing	74.00	93.29	77.62	59.76	95.40	41.27	80.61	73.40	68.53	292.1	72.87
	More think	89.02	92.41	78.25	61.33	98.44	64.84	79.81	83.89	74.61	564.4	81.79
	Ours	73.77	92.22	70.51	57.20	95.56	42.94	72.69	81.99	68.59	555.0	74.87
	Ours(+SFT+GRPO)	80.98	89.36	60.29	63.31	96.36	16.85	65.16	93.74	73.35	620.0	82.16
	Ours(+SFT+GRPO r.s.)	74.45	93.33	71.86	59.76	95.38	45.26	78.55	79.94	75.16	419.1	**82.46**

Table 6: Detailed composition of the diversity scores based on the output of R1-Distilled-Qwen-7B. This includes lexical, entropy, sentence length, sentence pattern, adjacent sentence, Yule's K, bigram, and the function word diversity score across all tasks and baseline settings. Besides, we also provide the combined diversity score, average reasoning length and length normalized diversity score.

R1-Distill-Llama-8B		**Diversity Sub Scores (in %)**								**Combined**	**Len.**	**Norm**
		lex.	*ent.*	*len.*	*pat.*	*adj.*	*yule.*	*bi.*	*func.*			
BBQ	Zero-shot CoT	73.65	94.29	70.10	66.72	96.75	44.49	84.99	88.07	74.58	236.5	79.92
	Self-Refine	91.04	92.36	75.58	57.25	98.89	54.38	77.76	89.80	73.03	236.5	75.85
	Role-Playing	81.70	94.06	78.18	69.67	97.38	52.19	84.85	88.66	77.49	347.1	80.91
	More think	90.33	91.74	67.77	61.75	98.69	46.29	79.61	92.18	74.11	533.8	84.11
	Ours(+SFT)	65.30	89.02	53.60	60.24	98.25	43.44	46.72	92.85	69.57	3353	82.64
	Ours(+SFT+GRPO)	84.02	90.87	55.86	61.07	98.62	52.50	62.01	93.41	75.91	949.7	85.75
	Ours(+SFT+GRPO r.s.)	91.44	92.51	67.30	66.00	99.52	68.55	73.76	93.10	81.38	673.6	**89.58**
GLOQA	Zero-shot CoT	96.46	93.43	61.57	60.66	98.75	82.35	87.29	90.88	77.49	571.5	87.07
	Self-Refine	90.19	95.70	60.93	62.19	98.93	84.22	92.21	89.12	77.97	244.3	81.11
	Role-Playing	94.77	94.21	66.43	60.15	98.70	83.01	88.66	90.05	77.45	453.0	83.02
	More think	99.45	91.38	65.28	59.25	99.23	82.57	83.31	92.50	77.35	991.4	87.19
	Ours(+SFT)	90.57	87.94	52.22	55.43	98.63	73.34	57.01	93.77	72.51	3852	87.26
	Ours(+SFT+GRPO)	99.20	89.37	52.52	56.55	98.54	75.67	69.56	93.51	77.10	1613.4	89.36
	Ours(+SFT+GRPO r.s.)	99.73	90.43	62.76	61.59	99.20	79.14	74.05	94.08	80.71	1225.6	**91.78**
CALI	Zero-shot CoT	77.26	93.88	71.84	59.48	97.44	35.53	85.60	90.57	71.63	246.0	73.98
	Self-Refine	84.48	91.97	77.03	65.29	98.13	27.89	81.28	91.02	73.66	458.7	78.87
	Role-Playing	82.94	92.26	74.30	63.07	96.97	27.42	82.03	91.90	72.62	392.5	77.32
	More think	93.95	90.62	74.81	57.13	98.41	39.30	77.90	93.49	72.19	683.1	80.30
	Ours(+SFT)	69.20	86.72	66.29	58.40	97.02	13.37	48.30	93.92	66.71	3279	79.77
	Ours(+SFT+GRPO)	84.56	88.53	65.38	61.54	96.78	20.42	62.40	94.13	73.28	1027.6	83.37
	Ours(+SFT+GRPO r.s.)	91.08	90.92	76.38	72.07	98.72	43.64	73.34	94.40	82.13	709.6	**90.55**
ETHICS	Zero-shot CoT	86.37	94.82	71.05	60.86	97.87	70.58	89.07	90.28	76.31	331.3	79.44
	Self-Refine	86.17	94.66	70.67	62.94	97.77	68.58	88.77	90.64	76.95	302.6	80.17
	Role-Playing	86.20	94.72	71.02	62.45	97.69	67.68	88.63	90.02	76.55	314.0	79.78
	More think	97.96	91.94	73.10	59.05	98.92	65.59	82.95	93.82	76.13	640.0	83.99
	Ours(+SFT)	85.48	90.17	61.54	59.99	97.89	40.09	69.97	94.78	72.11	1364	81.36
	Ours(+SFT+GRPO)	94.56	88.19	61.25	59.31	97.83	48.44	63.77	95.48	75.80	1551.4	87.89
	Ours(+SFT+GRPO r.s.)	95.06	91.68	78.70	77.34	99.27	73.82	77.40	94.82	88.22	805.3	**96.54**
CSQA	Zero-shot CoT	92.09	93.69	64.10	65.02	98.86	69.97	87.38	91.19	77.99	411.5	83.39
	Self-Refine	92.51	93.23	68.88	62.96	98.66	67.56	85.69	89.65	76.75	506.3	82.61
	Role-Playing	90.80	93.55	65.85	63.52	98.42	66.75	86.07	91.08	76.96	432.3	82.27
	More think	95.42	91.14	69.73	62.57	99.00	59.63	79.81	92.43	76.18	757.8	84.73
	Ours(+SFT)	67.44	87.55	61.90	55.26	98.87	51.86	44.03	92.22	68.67	4477	83.88
	Ours(+SFT+GRPO)	89.05	89.67	66.29	60.48	98.88	58.15	63.83	92.61	76.88	1271.7	87.96
	Ours(+SFT+GRPO r.s.)	94.03	91.01	73.49	69.59	99.48	68.95	72.66	92.80	83.28	989.0	**92.98**
GSM8K	Zero-shot CoT	75.15	92.69	79.99	63.63	96.13	43.06	78.06	77.24	71.11	394.8	76.52
	Self-Refine	82.66	92.90	70.87	64.75	98.32	59.52	80.61	85.40	75.12	490.5	81.24
	Role-Playing	76.89	93.01	79.21	61.12	95.94	45.48	79.57	74.53	69.93	347.1	75.02
	More think	90.63	91.13	79.86	61.87	98.65	65.24	76.14	86.37	75.30	830.4	84.12
	Ours(+SFT)	62.66	89.11	68.16	57.87	98.48	50.92	46.17	87.38	68.87	1961	81.53
	Ours(+SFT+GRPO)	74.74	90.22	74.48	60.17	98.42	56.18	58.18	87.15	74.85	1112.9	85.31
	Ours(+SFT+GRPO r.s.)	80.31	90.89	78.08	65.40	98.80	62.93	64.88	87.12	78.73	965.2	**88.45**

Table 7: Detailed composition of the diversity scores based on the output of R1-Distilled-Llama-8B. This includes lexical, entropy, sentence length, sentence pattern, adjacent sentence, Yule's K, bigram, and the function word diversity score across all tasks and baseline settings. Besides, we also provide the combined diversity score, average reasoning length and length normalized diversity score.

R1-Distill-Qwen-14B		Diversity Sub Scores (in %)								Combined	Len.	Norm
		lex.	*ent.*	*len.*	*pat.*	*adj.*	*yule.*	*bi.*	*func.*			
BBQ	Zero-shot CoT	54.86	96.37	77.82	50.83	90.24	36.08	74.85	61.10	71.11	394.83	76.52
	Self-Refine	88.78	93.52	67.58	61.90	98.76	54.14	83.45	89.33	75.12	490.51	81.24
	Role-Playing	85.20	94.33	72.03	69.34	98.44	56.59	86.00	89.33	69.93	347.10	75.02
	More think	91.52	92.79	67.79	62.80	98.76	52.56	83.48	91.74	75.30	830.40	84.12
	Ours(+SFT)	87.11	91.63	55.95	56.50	93.06	54.77	69.39	89.92	70.57	793.83	78.09
	Ours(+SFT+GRPO)	96.25	93.30	58.72	62.31	98.00	73.61	82.21	92.02	80.13	552.4	86.88
	Ours(+SFT+GRPO r.s.)	97.18	94.31	70.63	68.30	98.77	80.12	87.35	91.70	84.44	407.4	**90.17**
GLOQA	Zero-shot CoT	83.87	95.02	52.24	55.62	93.18	77.30	87.18	86.53	72.52	383.8	73.28
	Self-Refine	61.11	98.17	53.82	35.76	78.45	71.26	86.63	67.96	57.85	77.07	49.77
	Role-Playing	78.21	95.95	52.35	48.33	87.36	72.05	88.46	79.14	66.85	252.4	65.55
	More think	99.25	93.20	64.47	55.58	99.15	84.44	86.46	90.58	75.88	606.1	83.57
	Ours(+SFT)	98.82	90.67	53.21	56.30	97.54	83.65	77.02	91.77	74.98	1304	85.98
	Ours(+SFT+GRPO)	99.20	91.97	56.63	61.23	98.38	86.28	82.16	92.07	80.93	923.0	90.33
	Ours(+SFT+GRPO r.s.)	99.12	93.58	65.23	65.87	98.69	89.05	87.05	91.65	84.00	598.5	**91.32**
CALI	Zero-shot CoT	59.85	95.73	72.54	54.47	93.75	27.88	83.20	84.21	66.36	102.1	58.51
	Self-Refine	87.48	92.47	73.54	65.19	97.80	28.50	83.81	88.92	73.36	355.2	77.86
	Role-Playing	74.21	93.17	61.54	53.45	85.33	25.36	83.40	78.82	63.83	252.6	65.06
	More think	87.17	91.43	72.84	61.01	98.54	24.83	80.47	91.74	71.69	436.6	77.87
	Ours(+SFT)	86.66	90.51	61.03	62.61	96.22	20.68	71.45	91.72	70.56	593.5	78.17
	Ours(+SFT+GRPO)	92.26	92.39	65.89	66.24	98.30	41.32	82.20	92.04	78.81	433.6	84.92
	Ours(+SFT+GRPO r.s.)	93.87	92.79	75.28	71.34	99.17	51.31	83.41	92.32	82.88	450.9	**89.08**
ETHICS	Zero-shot CoT	70.98	95.20	60.25	52.49	82.41	52.21	84.11	72.98	64.58	226.1	62.34
	Self-Refine	90.30	93.96	75.90	65.63	98.46	62.46	87.17	89.75	77.66	353.5	81.37
	Role-Playing	76.57	95.57	62.47	52.52	85.15	61.84	85.82	77.83	67.30	260.5	65.70
	More think	95.95	93.40	71.67	58.87	99.36	64.54	88.15	93.20	76.03	386.1	81.89
	Ours(+SFT)	82.15	93.70	49.78	46.43	77.94	53.32	81.54	76.66	62.39	472.5	63.99
	Ours(+SFT+GRPO)	97.79	93.68	68.78	65.92	99.41	75.80	87.94	93.03	83.20	438.9	89.42
	Ours(+SFT+GRPO r.s.)	97.74	94.43	75.08	72.86	99.72	81.62	89.69	92.88	87.35	369.8	**92.89**
CSQA	Zero-shot CoT	91.03	94.10	64.83	64.86	98.92	68.22	88.52	91.17	77.82	320.5	82.74
	Self-Refine	89.15	94.24	66.42	63.07	98.12	70.11	87.21	87.95	76.49	368.9	80.18
	Role-Playing	84.58	94.71	60.20	58.45	93.06	64.56	86.97	85.59	72.59	288.6	74.40
	More think	93.93	92.34	67.56	63.08	99.01	60.30	83.58	91.46	76.33	539.3	83.33
	Ours(+SFT)	91.32	91.03	62.71	57.27	93.64	64.58	70.18	90.39	72.57	1034	81.56
	Ours(+SFT+GRPO)	96.76	93.07	70.76	65.68	99.46	76.00	84.42	91.20	82.59	572.2	89.64
	Ours(+SFT+GRPO r.s.)	97.01	93.99	79.44	70.14	99.65	80.52	87.69	90.66	85.62	435.7	**91.61**
GSM8K	Zero-shot CoT	61.86	94.52	76.54	57.65	92.91	27.38	83.22	64.09	63.71	159.3	64.44
	Self-Refine	84.47	93.37	71.00	63.94	98.37	63.29	82.11	83.86	75.06	400.6	80.81
	Role-Playing	77.46	93.12	76.76	60.95	96.41	48.59	80.24	76.64	70.59	342.5	75.73
	More think	89.66	91.93	78.18	61.42	98.76	65.34	78.15	85.33	74.94	637.4	82.79
	Ours(+SFT)	77.96	90.69	69.04	59.71	98.38	61.19	61.77	87.23	72.26	1027	82.32
	Ours(+SFT+GRPO)	88.47	93.02	71.68	62.70	98.40	72.76	79.59	87.15	79.40	526.0	86.36
	Ours(+SFT+GRPO r.s.)	90.64	93.59	75.64	64.12	98.43	76.79	82.93	86.73	80.86	448.5	**87.24**

Table 8: Detailed composition of the diversity scores based on the output of R1-Distilled-Qwen-14B. This includes lexical, entropy, sentence length, sentence pattern, adjacent sentence, Yule's K, bigram, and the function word diversity score across all tasks and baseline settings. Besides, we also provide the combined diversity score, average reasoning length and normalized diversity score.

	Qwen3-8B	**Diversity Sub Scores (in %)**								**Combined**	**Len.**	**Norm**
		lex.	*ent.*	*len.*	*pat.*	*adj.*	*yule.*	*bi.*	*func.*			
BBQ	Zero-shot CoT	52.47	95.16	75.87	54.75	97.52	38.52	80.28	67.14	64.01	75.09	54.22
	Self-Refine	24.24	83.84	82.46	47.60	99.28	37.07	14.88	86.08	57.16	220.27	76.40
	Role-Playing	21.33	58.13	82.45	36.31	70.59	31.03	14.56	66.24	43.76	364.63	61.64
	More think	74.88	90.68	76.94	69.03	97.32	22.41	67.86	91.62	73.43	501.73	80.07
	Ours(+SFT)	80.65	91.22	58.74	60.91	97.19	46.12	65.05	91.45	71.77	837.33	80.40
	Ours(+SFT+GRPO)	87.37	93.30	59.98	65.35	98.47	58.11	79.68	91.66	79.41	490.2	85.47
	Ours(+SFT+GRPO r.s.)	91.21	93.99	64.23	67.95	99.37	69.96	83.02	91.36	82.41	430.1	**88.15**
GLOQA	Zero-shot CoT	60.44	92.40	38.83	28.31	67.34	50.23	80.00	61.78	49.02	195.91	46.72
	Self-Refine	94.74	94.74	74.60	61.12	99.73	80.21	87.87	87.73	77.58	345.58	82.54
	Role-Playing	51.70	49.38	96.07	59.98	89.77	52.69	28.89	75.49	59.64	515.61	73.72
	More think	94.98	90.68	63.19	59.97	98.22	53.66	75.22	93.42	74.06	824.96	83.04
	Ours(+SFT)	90.48	90.17	54.69	56.89	97.49	71.19	66.16	91.74	73.05	2114.03	84.40
	Ours(+SFT+GRPO)	94.40	91.63	57.27	59.96	98.09	74.66	74.66	92.14	78.58	972.1	87.74
	Ours(+SFT+GRPO r.s.)	97.51	92.08	59.33	62.39	99.15	79.93	77.83	92.47	80.82	901.9	**89.88**
CALI	Zero-shot CoT	47.20	96.84	43.57	32.92	65.04	16.21	82.47	60.30	46.86	65.29	41.34
	Self-Refine	45.19	51.17	96.50	43.83	60.72	1.39	25.82	73.98	49.47	369.00	64.83
	Role-Playing	38.58	50.86	91.20	55.32	85.93	0.52	22.89	75.85	56.12	331.59	72.10
	More think	65.61	89.47	67.08	63.35	93.27	7.80	64.55	91.73	68.13	923.11	74.64
	Ours(+SFT)	81.37	90.45	64.59	63.80	96.12	12.97	72.05	92.65	70.38	572.03	77.38
	Ours(+SFT+GRPO)	83.56	91.77	63.57	65.36	96.50	17.32	80.39	91.89	75.14	348.2	80.36
	Ours(+SFT+GRPO r.s.)	86.61	92.88	65.76	68.83	97.70	33.11	84.24	92.31	78.99	316.3	**83.84**
ETHICS	Zero-shot CoT	79.75	95.10	42.62	62.07	99.88	36.57	97.33	86.35	71.48	136.00	72.87
	Self-Refine	41.18	56.57	95.91	45.36	70.87	2.84	23.91	75.89	51.57	206.86	67.68
	Role-Playing	36.58	55.54	91.01	44.59	72.84	4.11	19.63	75.79	50.83	231.06	68.46
	More think	72.72	90.15	76.00	62.81	96.46	18.46	67.57	93.68	70.68	536.94	77.68
	Ours(+SFT)	100.00	92.19	47.93	54.97	97.40	61.03	87.14	92.23	72.64	596.00	80.68
	Ours(+SFT+GRPO)	89.29	91.17	68.53	63.01	97.76	52.62	73.12	94.18	78.33	860.3	86.84
	Ours(+SFT+GRPO r.s.)	93.28	92.45	70.57	66.67	98.81	63.73	79.87	93.49	81.76	637.2	**89.09**
CSQA	Zero-shot CoT	57.42	83.79	42.16	45.30	64.68	29.57	67.74	61.30	52.34	243.99	53.18
	Self-Refine	47.38	56.46	91.39	46.53	66.61	10.76	26.14	76.33	52.68	301.89	68.46
	Role-Playing	23.52	73.73	54.89	54.14	86.12	9.12	14.84	82.24	55.98	324.66	75.27
	More think	72.43	90.93	67.35	68.04	97.05	21.82	70.55	91.56	72.48	449.76	78.55
	Ours(+SFT)	77.28	89.49	66.57	58.76	97.40	52.48	59.55	90.90	71.32	2362.13	82.54
	Ours(+SFT+GRPO)	89.99	92.58	73.80	62.70	98.38	64.54	77.85	90.16	79.25	635.4	86.40
	Ours(+SFT+GRPO r.s.)	91.83	92.64	68.36	64.45	98.93	68.67	77.92	90.46	80.39	652.0	**87.93**
GSM8K	Zero-shot CoT	63.06	94.31	72.52	51.98	93.33	28.50	82.67	72.50	63.10	197.31	64.71
	Self-Refine	89.59	92.80	78.47	69.80	98.78	60.36	80.10	87.05	78.29	522.58	84.82
	Role-Playing	90.01	92.06	79.70	69.98	98.24	61.37	78.24	84.58	77.87	609.06	85.23
	More think	76.08	90.70	79.40	66.39	95.59	36.75	69.22	85.40	72.64	664.80	80.09
	Ours(+SFT)	64.43	89.74	74.44	61.24	97.39	45.65	50.38	86.10	69.97	1557.56	81.21
	Ours(+SFT+GRPO)	75.49	91.48	78.45	65.40	98.15	54.89	66.01	85.58	77.42	922.7	85.98
	Ours(+SFT+GRPO r.s.)	81.83	93.22	81.60	67.97	98.85	65.21	78.03	84.87	80.70	487.8	**86.93**

Table 9: Detailed composition of the diversity scores based on the output of Qwen3-8B. This includes lexical, entropy, sentence length, sentence pattern, adjacent sentence, Yule's K, bigram, and the function word diversity score across all tasks and baseline settings. Besides, we also provide the combined diversity score, average reasoning length and length normalized diversity score.