ESCAPING THE MODE: MULTI-ANSWER REINFORCEMENT LEARNING IN LMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LMs) are typically post-trained with reinforcement learning (RL) to produce a single best answer per query, implicitly optimizing for modal correctness. While effective for benchmark accuracy, this paradigm is poorly suited to applications such as medical diagnosis, where generating a *set* of plausible answers with associated uncertainty is essential. We propose *multi-answer reinforcement learning*, which modifies the RL objective to train LMs to generate multiple candidate answers in a single forward pass, internalizing aspects of inference-time search into the model's generative process. We instantiate this framework as *Multi-Answer Reinforcement Learning with Verifiable Rewards (Multi-RLVR)*, which generalizes standard RLVR to a set-level reward. We further introduce *Multi-Answer Reinforcement Learning with Calibrated Rewards (Multi-RLCR)*, which incorporates a set-level Brier score objective to produce calibrated uncertainty estimates for each answer. Multi-answer training encourages explicit representation of alternative hypotheses rather than repeated generation of a dominant mode. Across question-answering and medical diagnostic benchmarks, our approach improves diversity, recall, and set-level calibration relative to single-answer-trained models sampled multiple times. We additionally find that multi-answer RL is more token-efficient, requiring fewer tokens to produce multiple answers than inference-time scaling baselines. These results position multi-answer RL as an effective and efficient alternative to inference-time scaling.

1 INTRODUCTION

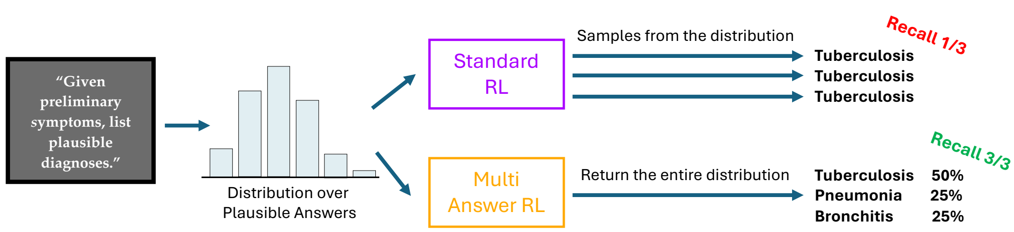
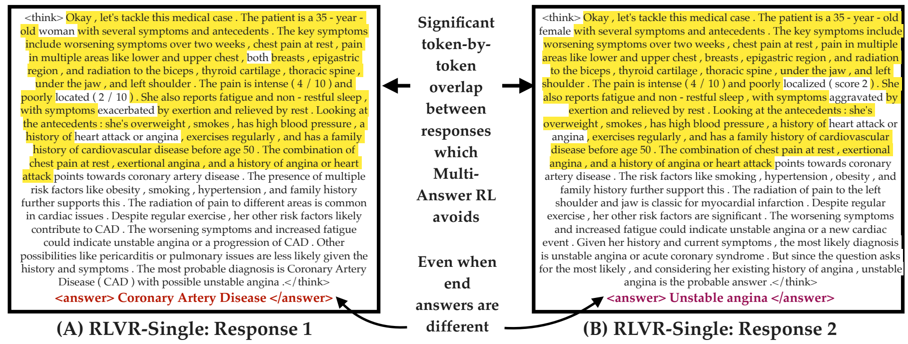


Figure 1: While standard RL trains LMs to consistently output the most likely answer to a question, **Multi-Answer Reinforcement Learning** trains models to output *distributions* of diverse answers.

Modern large language models are typically (post-)trained via RL to produce a single best answer per query, implicitly incentivizing the most likely correct answer Guo et al. (2025). While this objective is well aligned with standard benchmark evaluations, it is fundamentally mismatched to many real-world settings in which multiple distinct answers may be simultaneously correct, or where epistemic uncertainty arises due to incomplete or ambiguous information. As a motivating example, consider a clinical setting in which a patient comes in complaining of *right lower quadrant abdominal pain and fever*. A clinician in this setting might suspect a diagnosis of *acute appendicitis* or a *right-sided kidney stone*, but be unsure as to which due to incomplete information (i.e., epistemic uncertainty over a single true diagnosis). To check, they might order a *complete blood count* and a *urinalysis*, both of which would be clinically appropriate tests to order (i.e., multiple correct answers).

For such applications, models should ideally go beyond single answers and give a *set* of answers. However, single-answer training via RL can suppress alternative plausible hypotheses, leading models



Metric	RLVR Single
3-gram Overlap	0.199
4-gram Overlap	0.157
Pairwise Jaccard	0.152
ROUGE-L	0.424

Figure 2: **Significant subsequence overlap between independently sampled RLVR-Single responses, even those yielding different answers,** indicating that independent sampling largely re-instantiates the same reasoning tokens. Multi-Answer RL mitigates this by optimizing multiple generations jointly, reducing repeated token sequences and yield lower within-question overlap.

to repeatedly generate the same dominant answer even when other correct possibilities exist—indeed, entropy collapse is a commonly documented failure mode of models trained with RLVR (Lin et al., 2025; Yu et al., 2025; Jin et al., 2025; Wu & Choi, 2025). Beyond requiring a set of answers, many applications—especially in high-stakes settings—would further benefit from *uncertainty estimates* associated with each answer Kapoor et al. (2024). In other words, we ideally want a *distribution* P over plausible responses, including their estimated probabilities.

However, existing RL methods are trained to only provide a sample from that distribution $y \sim P$. As a result, models trained with them usually provide only the most likely answer. Inference-time techniques have been proposed to address some of these limitations: techniques for sampling multiple answers in parallel (Puri et al., 2025; Beirami et al., 2025) or sequentially Shinn et al. (2023); Xie et al. (2023) can produce an answer set for a given query; for uncertainty, models can be prompted or trained to verbalize uncertainty estimates for each answer (Xiong et al., 2024; Lin et al., 2022; Yang et al., 2025; Damani et al., 2025), or a separate model trained on top of model representations to output uncertainty scores (Azaria & Mitchell, 2023). However, since these post-hoc methods do not change the underlying training objective, they are fundamentally limited. As a result, such approaches may lead to poor performance in downstream decision-making scenarios where exploring and producing calibrated uncertainty estimates across multiple possibilities is crucial.

How can we move beyond single-answer RL objectives and instead train models to represent and generate sets of plausible answers directly? If diversity of responses and calibrated uncertainty estimates are desirable properties of language model outputs, then they should be explicitly optimized for during training rather than heuristically recovered at inference time. To this end, we propose **Multi-Answer RL**, an RL approach which explicitly optimizes language models to generate the distribution P over answers directly.

Concretely, Multi-Answer RL trains models to reason jointly over multiple plausible hypotheses within a single chain of thought and to produce structured sets of candidate answers in a single generation. Our approach supports both single-answer and multi-answer ground-truth settings and encourages explicit hypothesis exploration rather than repeated resampling. By introducing a reward function inspired by proper scoring rules that incentivizes calibrated answer distributions, we further show how this set generation approach can be extended to produce verbalized confidence scores for each answer in the set, thus making it into the full distribution P.

We empirically validate our approach on ambiguous and multi-label reasoning tasks (a general question answering dataset with incomplete information and a medical diagnostic dataset). We find that multi-answer RL scales better than baselines, producing substantial improvements in recall, answer diversity, and token efficiency. We further show that our set-level calibration reward enables models to produce better-calibrated uncertainty scores at the set level.

2 BACKGROUND

We start by describing the standard reinforcement learning setting for language models. A policy parameterized as an LM π_θ maps a prompt $x \in X$ to a distribution over textual outputs $y \in Y$. Given a dataset $D = \{(x_i, y_i^*)\}$ of prompt–answer pairs and a reward function $R : Y \times Y \to \mathbb{R}$, training aims to maximize expected reward: $\max_\theta \ \mathbb{E}_{(x,y^*)\sim D,\ y\sim\pi_\theta(\cdot|x)}\big[R(y, y^*)\big]$. This approach trains the LM to output samples which have high reward on average. In practice, this can result in policies that place much of its mass on single outputs.

Reinforcement Learning with Verifiable Rewards (RLVR): RLVR focuses on the class of reward functions where rewards are deterministically verifiable from model outputs. A standard choice is the binary correctness reward, $R_{correct}(y, y^*) = \mathbb{1}_{y\equiv y^*}$, where $\mathbb{1}_{y\equiv y^*} \in \{0, 1\}$ indicates whether the output y matches the ground-truth answer y^* (up to normalization).

Proper Scoring Rules: In many domains, users care not only about a model's answer y, but also about its reported confidence $q \in [0, 1]$ in that answer. In the case of binary outcomes, a scoring rule is a function $S : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ that maps a confidence estimate q and an outcome a to a scalar score. A scoring rule is said to be *proper* if it incentivizes truthful confidence reporting: the expected score is minimized when the reported confidence matches the true probability, i.e., $\mathbb{E}_{a\sim p(a)}\ S(p(a), a) \le \mathbb{E}_{a\sim p(a)}\ S(q, a)\ \ \forall q$.

Reinforcement Learning with Calibration Rewards (RLCR): A key limitation of RLVR is that models are incentivized to guess when uncertain. Building on the theory of proper scoring rules, RLCR (Damani et al., 2025) prompts models to produce reasoning chains that output both an answer y and an associated confidence estimate $q \in [0, 1]$. Models are then trained using a reward that jointly incentivizes correctness and calibration: $R_{\text{RLCR}}(y, q, y^*) \ = \ \mathbb{1}_{y\equiv y^*} \ - \ S(q, \mathbb{1}_{y\equiv y^*})$, where S is a *proper scoring rule*. Damani et al. (2025) prove that for a particular class of proper scoring rules, RLCR incentivizes predictions are both accurate and well-calibrated. In practice, they use the Brier score, which yields $S(q, \mathbb{1}_{y\equiv y^*}) = (q - \mathbb{1}_{y\equiv y^*})^2$.

3 MULTI-ANSWER REINFORCEMENT LEARNING

3.1 MULTI-ANSWER RLVR

We consider a generalized setting in which a prompt x is associated not with a single ground-truth answer, but with a *set* of valid answers $\mathcal{Y}^*(x) = \{y_1^*, y_2^*, \ldots, y_N^*\}$, where $N \ge 1$ may vary across instances. Our goal is to train a model to recover the full set—or a high-coverage subset—of these valid answers within a single generation.

We prompt the policy π_θ to produce a structured output consisting of a set of K *distinct* candidate answers, $A = \{a_1, a_2, \ldots, a_K\}$, within a single chain of thought. We then train the model using a *set-level* reward that checks how many of the generated answers belong to the ground-truth set: $R_{\text{RLVR}}^{\text{multi}}(A, \mathcal{Y}^*) \ = \ \sum_{i=1}^{K} \mathbf{1}[a_i \in \mathcal{Y}^*]$, This objective can be viewed as a natural generalization of RLVR from single-answer correctness to set-level correctness and subsumes several familiar training objectives:

1. $N = 1,\ K = 1$ **(Standard RLVR).** When there is a single ground-truth answer and the model produces a single output, the reward reduces exactly to the binary correctness signal used in vanilla RLVR.
2. $N = 1,\ K > 1$ **(Pass@K).** When there is a single correct answer but the model produces multiple candidates, the reward is equivalent to a pass@K objective.
3. $N > 1,\ K \le N$ **(Partial set recovery).** When multiple correct answers exist but the model is constrained to produce fewer candidates than the size of the ground-truth set, the objective encourages maximal coverage of distinct valid answers.
4. $N > 1,\ K \ge N$ **(Full set recovery).** When allowed to generate at least as many candidates as valid answers, the optimal policy recovers the entire ground-truth set.

> **Condensed RLCR-Multi Example Output**
>
> <think>Tuberculosis is a consideration, [...] then there's pneumonia, which can cause similar symptoms. [...] Another possibility is bronchitis, but again, hemoptysis is more common in other conditions. [...] Given the hemoptysis and fever, TB and pneumonia are higher. Since the patient is underweight, TB might be more likely as undernutrition can impair immune response. [...] Bronchitis is possible but less so. [...] ranking TB first, pneumonia second, bronchitis third[...] Assigning confidences: TB maybe 0.4, pneumonia 0.30, bronchitis 0.30.</think> <a1>Tuberculosis</a1><conf1>0.40</conf1> <a2>Pneumonia</a2><conf2>0.30</conf2> <a3>Bronchitis</a3><conf3>0.30</conf3>

Enforcing distinct outputs. Since we want to output distinct answers from the model, we use a *format reward* to enforce uniqueness among the generated answers. Full details in Appendix C.

3.2 Multi-Answer RLCR

Building on Multi-Answer RLVR, we introduce *Multi-Answer RLCR*, which additionally trains models to produce calibrated confidence estimates per answer. In a single chain-of-thought, the model is asked to output both a set of K distinct candidate answers $A = \{a_1, a_2, \dots, a_K\}$ and a corresponding set of confidence values $Q = \{q_1, q_2, \dots, q_K\}$ where each $q_i \in [0, 1]$ represents the model's reported confidence that the answer a_i is correct.

As in RLCR, training jointly optimizes answer correctness and confidence calibration by combining a correctness-based reward with a proper scoring rule. To measure calibration in the multi-answer setting, we introduce the *Multi-Brier score*, defined as the average squared error between reported confidences and answer-level correctness:

$$R_{\text{Brier}}^{\text{multi}}(A, Q, \mathcal{Y}^*) \;=\; \frac{1}{K}\sum_{i=1}^{K}\Big(q_i - \mathbf{1}[a_i \in \mathcal{Y}^*]\Big)^2.$$

Intuitively, this objective trains each confidence q_i to approximate the probability that the specific answer a_i is correct. The full Multi-Answer RLCR reward, which jointly incentivizes coverage of valid answers and calibrated confidence estimates at the answer level, is given by

$$R_{\text{RLCR}}^{\text{multi}}(A, Q, \mathcal{Y}^*) \;=\; R_{\text{RLVR}}^{\text{multi}}(A, \mathcal{Y}^*) \;-\; R_{\text{Brier}}^{\text{multi}}(A, Q, \mathcal{Y}^*),$$

Model output as a distribution. The model's output (A, Q) can be interpreted as defining a distribution over plausible answers. When $N = 1$, our format reward enforces that the K confidence scores sum to at most one, yielding a discrete probability distribution over the answer set. When $N > 1$, confidence scores are produced independently for each answer, and the output naturally corresponds to a multivariate Bernoulli distribution over correctness events.

4 Experiments

4.1 Experimental Setup

Training Details We use GRPO as the base RL algorithm with some modifications (App C). Specifically, we use the Qwen3-8B base models with a max response length of 1536. To enable easier verification and more structured outputs, we augment our reward with a simple format reward that encourages models to enclose CoTs within the right tags and keep answers produced in sets unique.

Datasets We evaluate our approach on two datasets (DDXPlus, HotPotQA-Modified) that require set-valued reasoning, but differ in underlying structure: DDXPlus admits multiple simultaneously correct answers, while HotPotQA-Modified has a single gold answer but significant ambiguity from incomplete or underspecified information.

1. DDXPlus Tchango et al. (2022) is a large-scale medical diagnostic dataset in which each example consists of basic patient demographics along with a brief description of symptoms and antecedents. The target output is a *differential diagnosis*—a *set* of medical conditions that are plausible given the available information. Because multiple diagnoses may be simultaneously correct ($N \geq 1$), performance is naturally evaluated using recall over the set of gold answers rather than top-1 accuracy. A representative example is shown in Figure 1. We train on 25,000 examples, evaluate correctness using exact string match, and prompt models to generate $K = 3$ diagnoses in a single output.

2. HotPotQA-Modified is a modified version of the HotPotQA distractor dataset Yang et al. (2018). Each example contains a multi-hop question and 10 paragraphs (2 relevant, 8 distractors), but we remove 0, 1, or both relevant paragraphs to create varying levels of information completeness. Although each question retains a single ground-truth answer ($N = 1$), incomplete information introduces ambiguity, making it beneficial for models to reason over and output multiple plausible candidates. This setting corresponds to the $N = 1$, $K > 1$ regime, where the Multi-Answer RL objective corresponds to a pass@K reward.

Methods We evaluate the following models:

1. **Base**: Pre-trained Qwen3-8B, prompted with single- and multi-answer RLVR/RLCR formats.
2. **RLVR Single**: Base trained w/ RLVR to produce one answer (confidence verbalized at eval).
3. **RLCR Single**: Base model trained with RLCR to produce one answer and confidence.
4. **Prompted Multi**: RLVR/RLCR Single models prompted at inference with multi-answer system prompts to generate multiple answers (and confidences for RLCR).
5. **RLVR Multi (ours)**: Base model trained with $R_{\text{RLVR\_Multi}}$ to generate multiple answers via `<answer{i}>` tags.
6. **RLCR Multi (ours)**: Base model trained with $R_{\text{RLCR\_Multi}}$ to generate answers and confidences via `<answer{i}>` and `<confidence{i}>` tags.

We compare Single-Answer and Multi-Answer methods by constructing sets of K candidate answers in both cases. For Single models, we sample K independent responses and treat them as a set, while Multi models naturally produce a single set of K answers in a single generation. This comparison allows us to isolate the effect of multi-answer training from inference-time sampling.

Metrics We evaluate correctness using Pass@1, Pass@K, Recall, average token count, and uniqueness, and evaluate calibration using Brier Score, ECE, and Set ECE. Detailed definitions provided in Appendix D.

Comparability Across Settings. Answer sets produced by single-answer and multi-answer methods differ in a fundamental way. Single-answer methods construct sets via repeated sampling, which often results in duplicate answers and variable effective set sizes. In contrast, multi-answer methods explicitly generate a set of distinct candidates. Because of this mismatch, **pooled** and **set-level** calibration metrics are not directly comparable across single and multi settings, as their values depend on set size. Accordingly, we use pooled and set-level calibration metrics only to compare *RLVR Multi* and *RLCR Multi*, where the answer set construction mechanism is shared. In contrast, **Top-1 Individual ECE** and **Top-1 Brier Score** evaluate calibration of the highest-ranked answer, and are directly comparable.

4.2 RESULTS

In this section, we compare *Single-Answer* to *Multi-Answer* RL, as well as *RLVR* to *RLCR*, along both correctness and calibration on DDXPLUS and HOTPOTQA-MODIFIED.

Correctness Table 1 reports correctness results across all settings. Across both RLVR and RLCR, Multi-Answer models substantially outperform their Single-Answer counterparts on set-level correctness metrics, including pass@K and recall. Moreover, simply prompting single-answer models to produce multiple answers performs substantially worse than trained multi-answer models, showing that generating answer sets requires explicit multi-answer training.

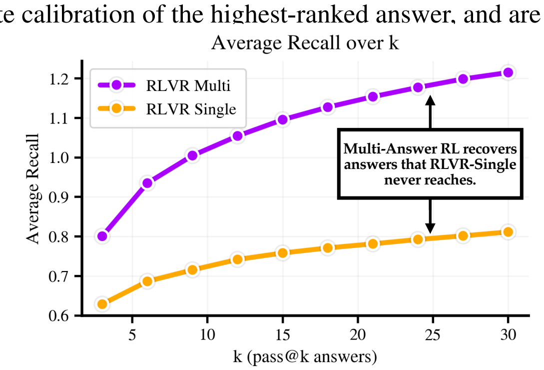


Figure 3: On DDXPlus, with equal total samples (30), RLVR-Multi yields 1.5× more unique correct answers than RLVR-Single, revealing reduced mode collapse.

(a) DDXPlus: Medical Differential Diagnoses

Method	Recall (↑)	Diversity (↑)	Efficiency (↓)	Pass@K (↑)	Top-1 Acc. (↑)
RLVR (Single Loss + Single Prompt)	0.62	0.62	1191	0.60	**0.50**
RLVR (Single Loss + Multi Prompt)	0.67	0.99	695	0.56	0.42
Zero-Shot (Multi Prompt)	0.31	0.93	1131	0.38	0.24
Multi-Answer RLVR (Multi Loss/ Prompt)	**0.79**	**1.00**	**677**	**0.65**	0.42
RLCR (Single Loss + Single Prompt)	0.65	0.50	1378	**0.63**	**0.55**
RLCR (Single Loss + Multi Prompt)	0.49	1.0	703	0.28	0.32
Zero-Shot (Multi Prompt)	0.41	0.89	1269	0.41	0.16
Multi-Answer RLCR (Multi Loss/ Prompt)	**0.77**	**1.00**	**510**	0.62	0.43

(b) HotPotQA-Modified

Method	Pass@K (↑)	Diversity (↑)	Efficiency (↓)	Top-1 Acc. (↑)
RLVR (Single Loss + Single Prompt)	0.17	0.59	1466	**0.19**
RLVR (Single Loss + Multi Prompt)	0.23	1.00	551	0.17
Zero-Shot (Multi Prompt)	0.20	0.75	1265	0.17
Multi-Answer RLVR (Multi Loss/ Prompt)	**0.27**	**1.00**	**544**	**0.19**
RLCR (Single Loss + Single Prompt)	0.22	0.62	1782	0.17
RLCR (Single Loss + Multi Prompt)	0.23	0.98	670	0.17
Zero-Shot (Multi Prompt)	0.20	0.73	1298	0.15
Multi-Answer RLCR (Multi Loss/ Prompt)	**0.27**	**1.00**	**622**	**0.19**

Table 1: **Correctness, diversity, and efficiency for Multi-Answer RL.** Multi-Answer RL (RLVR / RLCR) substantially improves recall, diversity, and efficiency over single-answer baselines in DDXPlus **(a)** and HotPotQA-Modified **(b)**. All results use $k = 3$.

(a) DDXPlus: Medical

Method	Set ECE (↓)	Top-1 ECE (↓)	Top-k ECE (↓)	Top-1 Brier (↓)	Top-k Brier (↓)
RLVR Single	–	0.34	–	0.35	–
RLCR Single	–	**0.02**	–	**0.24**	–
RLVR Multi	0.13	0.16	0.10	0.27	0.19
RLCR Multi (Ours)	**0.02**	**0.01**	**0.04**	**0.24**	**0.18**

(b) HotPotQA: Hard

Method	Set ECE (↓)	Top-1 ECE (↓)	Top-k ECE (↓)	Top-1 Brier (↓)	Top-k Brier (↓)
RLVR Single	–	0.48	–	0.38	–
RLCR Single	–	0.13	–	0.16	–
RLVR Multi	0.47	0.38	0.22	0.30	0.16
RLCR Multi (Ours)	0.44	0.31	0.21	0.22	0.12

Table 2: **Calibration metrics for Multi-Answer RL ($k = 3$).** RLCR Multi substantially improves set- and answer-level calibration over RLVR Multi on DDXPlus. On HotPotQA-Hard, RLCR Multi improves answer-level calibration but underperforms RLCR Single at the set level, consistent with a learned prior encouraging confidences to sum to one in extremely difficult single-label settings.

On DDXPLUS, where multiple diagnoses may be simultaneously correct, recall is the primary notion of correctness. Multi-Answer models recover a significantly larger fraction of the gold differential diagnoses per example, demonstrating that explicitly optimizing for set-valued outputs enables models to capture relevant alternatives that might be suppressed by single-answer RL. On HOTPOTQA-MODIFIED, each question has a single correct answer, but missing information induces substantial uncertainty. In this regime, Multi-Answer RL achieves large gains in pass@k, reflecting improved coverage of plausible answers. By reasoning over multiple plausible hypotheses within a single generation, Multi-Answer models avoid the strong answer collapse exhibited by Single-Answer models, which repeatedly regenerate a dominant answer and therefore underperform on pass@k.

Calibration. Table 2 reports calibration results across both settings. Across both datasets, RLCR consistently improves calibration relative to RLVR in both Single- and Multi-Answer settings.

On DDXPlus, RLCR-Multi achieves markedly better set-level calibration than RLVR-Multi across all metrics, demonstrating RLCR's ability to assign meaningful confidence to multiple plausible diagnoses. Calibration curves (Figure 4) provide a complementary view. RLCR-Multi closely tracks the identity line across confidence bins, while RLVR-Multi exhibits systematic overconfidence, particularly at higher confidence levels—effects that are less visible in aggregate metrics (i.e. ECE). Finally, of the calibration metrics considered, only **Top-1 ECE** and **Top-1 Brier** permit direct comparison between Single- and Multi-Answer RL. These metrics show that Multi-Answer training matches Single-Answer RL in calibration, indicating that explicitly optimizing for set-valued outputs does not degrade calibration of the top answer.

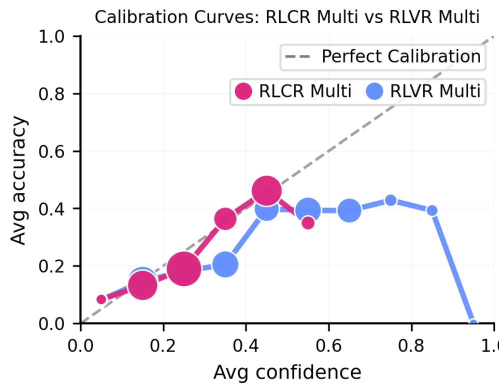


Figure 4: Calibration curves on DDXPlus. RLCR-Multi closely tracks perfect calibration across confidence levels, while RLVR-Multi exhibits systematic overconfidence.

On HotPotQA-Modified, trends from DDXPlus largely persist, with RLCR improving calibration over RLVR. However, despite outperforming Multi-Answer RLVR, Multi-Answer RLCR underperforms Single-Answer RLCR, indicating degraded calibration. This arises from a strong generative prior that enforces confidences across the answer set to sum to 1. In a single–gold-answer setting like HotPotQA-Modified (pass@K $<$ 30%), proper calibration instead requires total confidence < 1 for most questions. Although RL exploration could in principle overcome this prior, we do not observe this in practice and leave improved exploration strategies to future work.

Does multi-answer RL improve output diversity? Table 1 showed that multi-answer training improves both pass@K and coverage, indicating that models trained to generate answer sets output a broader space of plausible answers. To analyze diversity more directly, we compare RLVR-Single and RLVR-Multi under a matched sampling budget. We sample 30 generations from RLVR-Single, yielding 30 answers, and 10 generations from RLVR-Multi, yielding $10 \times K = 30$ answers total. We then compute the number of unique answers per set and plot the distribution in Figure 5, and plot average recall—the mean number of correct answers per set—as a function of K in Figure 3.

RLVR-Multi exhibits greater output diversity, producing nearly twice as many unique answers on average (8 vs. 4). Moreover, its higher average recall indicates that this increased diversity corresponds to a larger number of correct answers, rather than spurious variation. Taken together. these results demonstrate that multi-answer training increases diversity without compromising correctness, addressing a key failure mode of single-answer RL.

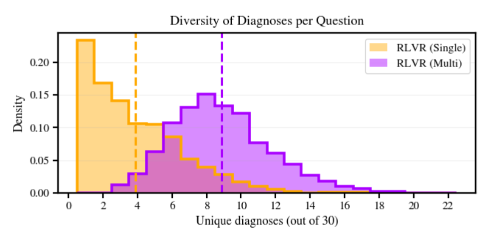


Figure 5: RLVR-Multi generates more unique diagnoses per question than RLVR-Single across 5,000 tests, improving recall.

Is multi-answer RL more compute efficient than single-answer RL? The standard approach to obtaining answer sets from single-answer models is inference-time scaling, most commonly by repeated sampling. Figure 2 illustrates a representative instance, showing three sampled generations from an RLVR-Single model. Although each generation produces a different final answer, the corresponding chains exhibit substantial overlap, repeatedly reproducing the same reasoning scaffolding and intermediate phrases. Quantitatively, this manifests as high n-gram overlap across samples (see Figure 2), indicating that inference-time sampling often incurs significant computational redundancy.

To compare this redundancy against multi-answer training, we measure the number of tokens for a set of K answers under both approaches. For RLVR-Single, we sample the model K times independently and sum the resulting token lengths. For RLVR-Multi, we generate K answers within a single response and measure its total token length. In the medical domain, the average token length of an RLVR-Multi response is only 56% of the total token length required by an RLVR-Single model to produce the same number of answers (Figure 7). These results show that training models to directly generate answer sets can eliminate a significant amount of redundant computation incurred by inference-time sampling.

Summary Together, these results show that training models to explicitly optimize multi-answer structure results in outputs that are more accurate, calibrated, and diverse, while simultaneously being more efficient to generate.

4.3 Does multi-answer training remain stable with increasing K?

To study how performance scales with K, we train Multi-Answer RLVR with $K = \{2, 3, 4, 5\}$. Because each medical example admits multiple correct diagnoses, increasing K expands the model's capacity to recover valid alternatives. As shown in Figure 6, the average number of correct answers per set increases monotonically with K. Since outputs are constrained to be unique, these gains reflect the discovery of additional correct hypotheses rather than repetition. Training remains stable across all values of K.

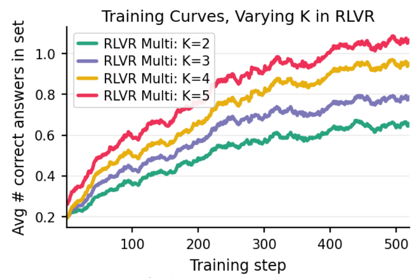


Figure 6: As $k \uparrow$, Multi-Answer RLVR stably recovers more unique correct answers per set.

5 Related Work

RLVR and Diversity. Recent work documents trade-offs between RL training and output diversity Kirk et al. (2024); Shypula et al. (2025); Yang & Holtzman (2025). In particular, Wu & Choi (2025) and Yue et al. (2025) show that RLVR gains in pass@1 often coincide with support contraction, reducing pass@k, while West & Potts (2025) find base models can outperform aligned models on creative tasks. To mitigate these effects, prior methods aim to preserve diversity during RL training: Chen et al. (2025) and Walder & Karkhanis (2025) optimize pass@k-style objectives, Li et al. (2025) add learned diversity rewards, and Song et al. (2025) introduce outcome-conditioned exploration bonuses. Most operate in the single-answer setting, encouraging diversity implicitly rather than explicitly training models to generate answer sets in one pass.

Calibration. Reliable deployment of language models requires accurate uncertainty estimation Kalai et al. (2025). Prior work has explored post-hoc confidence verbalization Xiong et al. (2024); Lin et al. (2022), but multiple studies show that such estimates are systematically overconfident Xiong et al. (2024); Mei et al. (2025); Kirichenko et al. (2025). As a result, recent work has focused on training models to produce calibrated confidence estimates. Lin et al. (2022) fine-tune models to predict confidence conditioned on a question and its answer, while others apply RL with proper scoring rules Brier (1950); Gneiting & Raftery (2007). Within this paradigm, Stangel et al. (2025) and Xu et al. (2024) optimize purely for calibration, whereas Damani et al. (2025) jointly optimize correctness and calibration.

Generating Answer Sets. Several recent efforts aim to elicit answer sets from language models. Wang et al. (2024) prompt models to output probability distributions over a fixed label set. Zhang et al. (2025) introduce verbalized sampling, prompting models to explicitly express a distribution over multiple responses. Similarly, Troshin et al. (2025) show that asking models to enumerate or iteratively sample answers increases output diversity. While related in spirit, these approaches are training-free and operate purely at inference time.

6 Conclusion

We propose Multi-Answer Reinforcement Learning, which trains language models to generate sets of plausible answers rather than a single mode. Across medical and QA benchmarks, this objective improves coverage and diversity, recovering correct alternatives missed by standard RL while using fewer total tokens. More broadly, our work extends a line of approaches that use language to serialize structured reasoning, moving from single answers to explicit representations of output distributions. Limitations remain: single-answer objectives still achieve higher top-1 accuracy, our experiments are confined to QA, and serial generation limits parallelism despite improved token efficiency. Nevertheless, these results suggest a path toward models that can explore their full internal distribution without sacrificing performance or calibration.

REFERENCES

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Beirami, A., Agarwal, A., Berant, J., D'Amour, A. N., Eisenstein, J., Nagpal, C., and Suresh, A. T. Theoretical guarantees on the best-of-n alignment policy. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=u3U8qzFV7w.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950.

Chen, Z., Qin, X., Wu, Y., Ling, Y., Ye, Q., Zhao, W. X., and Shi, G. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

Damani, M., Puri, I., Slocum, S., Shenfeld, I., Choshen, L., Kim, Y., and Andreas, J. Beyond binary rewards: Training lms to reason about their uncertainty, 2025. URL https://arxiv.org/abs/2507.16806.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jin, R., Gao, P., Ren, Y., Han, Z., Zhang, T., Huang, W., Liu, W., Luan, J., and Xiong, D. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*, 2025.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., and Wilson, A. G. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972, 2024.

Kirichenko, P., Ibrahim, M., Chaudhuri, K., and Bell, S. J. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=kYbojsAOBj.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Li, T., Zhang, Y., Yu, P., Saha, S., Khashabi, D., Weston, J., Lanchantin, J., and Wang, T. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025.

Lin, J., Wu, Z., and Sun, J. Training llms for ehr-based reasoning tasks via reinforcement learning. *arXiv preprint arXiv:2505.24105*, 2025.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Mei, Z., Zhang, C., Yin, T., Lidard, J., Shorinwa, O., and Majumdar, A. Reasoning about uncertainty: Do reasoning models know when they don't know? *arXiv preprint arXiv:2506.18183*, 2025.

Puri, I., Sudalairaj, S., Xu, G., Xu, K., and Srivastava, A. A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods. *arXiv preprint arXiv:2502.01618*, 2025.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Shypula, A., Li, S., Zhang, B., Padmakumar, V., Yin, K., and Bastani, O. Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*, 2025.

Song, Y., Kempe, J., and Munos, R. Outcome-based exploration for LLM reasoning. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025. URL https://openreview.net/forum?id=VORSpYLBJ6.

Stangel, P., Bani-Harouni, D., Pellegrini, C., Özsoy, E., Zaripova, K., Keicher, M., and Navab, N. Rewarding doubt: A reinforcement learning approach to confidence calibration of large language models. *CoRR*, abs/2503.02623, March 2025. URL https://doi.org/10.48550/arXiv.2503.02623.

Tchango, A. F., Goel, R., Wen, Z., Martel, J., and Ghosn, J. DDXPlus: A new dataset for automatic medical diagnosis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=heBKnuV42O.

Troshin, S., Saparina, I., Fokkens, A., and Niculae, V. Asking a language model for diverse responses. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 66–72, 2025.

Turtel, B., Franklin, D., Skotheim, K., Hewitt, L., and Schoenegger, P. Outcome-based reinforcement learning to predict the future. *arXiv preprint arXiv:2505.17989*, 2025.

Walder, C. and Karkhanis, D. Pass@ k policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*, 2025.

Wang, C., Szarvas, G., Balazs, G., Danchenko, P., and Ernst, P. Calibrating verbalized probabilities for large language models. *arXiv preprint arXiv:2410.06707*, 2024.

West, P. and Potts, C. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*, 2025.

Wu, F. and Choi, Y. The invisible leash: Why rlvr may not escape its origin. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

Xiao, C., Zhang, M., and Cao, Y. Bnpo: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*, 2025.

Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, J. X., Kan, M.-Y., He, J., and Xie, M. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36:41618–41650, 2023.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Xu, T., Wu, S., Diao, S., Liu, X., Wang, X., Chen, Y., and Gao, J. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, 2024.

Yang, C. and Holtzman, A. How alignment shrinks the generative horizon. *arXiv preprint arXiv:2506.17871*, 2025.

Yang, D., Tsai, Y.-H. H., and Yamada, M. On verbalized confidence scores for LLMs. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025. URL https://openreview.net/forum?id=CVRdNQvFPE.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., and Shi, W. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*, 2025.

A METRICS DETAILS

Correctness Metrics We evaluate correctness, diversity, and efficiency using the following metrics:

1. **Pass@1 (↑):** Measures the accuracy of a single selected answer from the generated set. For multi-answer methods, we use the first answer in the set, which we empirically find the model treats as most likely. For single-answer methods, we select an answer uniformly at random from the set.

2. **Pass@K (↑):** A binary metric, equals 1 if *any* answer in the set is correct, 0 otherwise.

3. **Recall (Avg. # Correct per Set) (↑):** Measures the average number of correct answers produced per example. Formally, for a set of K answers,

$$\text{Recall} \;=\; \frac{1}{K}\sum_{i=1}^{K} \mathbb{1}\{a_i \text{ is correct}\}. \tag{1}$$

4. **Avg. Token Count (↓):** average total token count per question across K generations.

5. **Uniqueness (↑):** The number of distinct answers within the generated set (i.e. output diversity).

Calibration Metrics All methods are prompted to output a confidence score $q_i \in [0, 1]$ for each generated answer a_i. We evaluate calibration using the following metrics:

1. **Brier Score (↓).** Measures the squared error between predicted confidence and binary correctness:

$$\text{Brier} = \frac{1}{K}\sum_{i=1}^{K}\big(q_i - \mathbb{1}\{a_i \in \mathcal{Y}^*\}\big)^2. \tag{2}$$

   We report Brier scores for the top-ranked answer (used for Pass@1) and for all answers pooled across the set.

2. **Expected Calibration Error (ECE) (↓).** Measures the discrepancy between predicted confidence and empirical accuracy by binning confidence scores:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{L} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{3}$$

   where M is the number of bins, B_m denotes the samples in bin m, and L is the total number of samples. We use $M = 10$, and report ECE both for the top-ranked answer and the entire set of answers.

3. **Set ECE (↓, diagnostic).** Measures calibration at the level of answer sets (see Appendix D)

B SYSTEM PROMPTS

.

RLCR Single Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind, provides the user with the final answer, then analyzes its confidence about the solution and then provides the user with its confidence level. The confidence level is a number between 0 and 1 (inclusive) enclosed within <confidence> </confidence> tags. The final answer is enclosed between <answer> </answer> tags. The analysis about confidence and uncertainty is enclosed within <analysis> </analysis> tags. Here are some guidelines for the analysis: 1. Your task is to point out things where the model could be wrong in its thinking, or things where there might be ambiguity in the solution steps, or in the reasoning process itself. 2. You should not suggest ways of fixing the response, your job is only to reason about uncertainties. 3. For some questions, the response might be correct. In these cases, it is also okay to have only a small number of uncertainties and then explicitly say that I am unable to spot more uncertainties. 4. Uncertainties might be different from errors. 5. If there are alternate potential approaches that may lead to different answers, you should mention them. 6. List out plausible uncertainties, do not make generic statements, be as specific as possible. 7. Enclose this uncertainty analysis within <analysis> </analysis> tags. The final format that must be followed is: <think> reasoning process here </think> <answer> final answer here </answer> <analysis> analysis about confidence and uncertainty here </analysis> <confidence> confidence level here (number between 0 and 1) </confidence>

RLVR Single Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>. Do NOT put any sentences or reasoning process within the <answer> </answer> tags - only put the final answer that will be verified with exact match score within the <answer> </answer> tags.

RLCR Multi Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The question may be ambiguous or difficult to answer, and you must propose multiple possible answers. You must assign a confidence score to each candidate. Make sure the confidences sum to less than or equal to 1. The confidences are allowed to sum to less than 1 if you are unsure about all candidates. You will be graded on how closely your confidences match the actual correctness of the candidates. Output EXACTLY {K} DISTINCT candidates with confidences that sum to less than or equal to 1. FORMAT ONLY (no extra text): <think> reasoning process about different candidate answers here </think> <answer1> candidate_answer_1 </answer1> <confidence1> confidence level for candidate 1 here (number between 0 and 1) </confidence1> ... exactly {K} pairs ...

RLVR Multi Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. You must propose multiple possible answers, not just one. For each candidate, think separately about why it could be correct or incorrect. Output EXACTLY {K} DISTINCT candidate answers. FORMAT ONLY (no extra text): <think> reasoning process about different candidate answers here </think> <answer1> candidate_answer_1 </answer1> <answer2> candidate_answer_2 </answer2> ... exactly {K} answers ...

C ADDITIONAL TRAINING DETAILS

We sample 32 responses per prompt using a temperature of 0.7 and train with an effective batch size of 1536. Optimization uses a constant learning rate with linear warmup, a base learning rate of 1×10^{-6}, and a warmup ratio of 0.05. All experiments are conducted on NVIDIA A100 and H100 GPUs, and we observe consistent performance trends across hardware types.

Following prior work Turtel et al. (2025), we remove the standard deviation normalization from the advantage computation, which can improve learning stability in the presence of extreme miscalibration. Training is performed with the BNPO objective, which aggregates token-level losses normalized by the number of active tokens in each local training batch (Xiao et al., 2025). For both datasets, we set a maximum completion length of 1536. We do 1 epoch of training. System prompts for RLCR Single, RLCR Multi, RLVR Single, and RLVR Multi are in Appendix B.

Format Reward: We use a format reward to encourage adherence to the structured format required by the system prompts. In RLVR Single, models must format their output in `<think>` and `<answer>` tags. In RLCR Single, in addition to `<think>` and `<answer>` tags, we require a `<confidence>` tag for verbalized confidence. RLVR Multi requires a `<think>` tag and then k `<answeri>` tags. RLCR Multi requires a `<think>` tag and then k sets of `<answeri>` and `<confidencei>` tags. A valid response must contain all these tags in the correct order. Both format and calibration rewards are weighted equally.

In the Multi-RLCR and Multi-RLVR setting, we enforce the uniqueness required by the prompt by zeroing out all rewards if the normalized answers extracted from the `<answeri>` tags are not unique. Additionally, in the N = 1 setting where our dataset assigns only one correct answer to each question, the format reward zeroes out if the sum of the confidences extracted from the `<confidencei>` tags is more than 1.

D EVALUATION METRICS

In addition to the metrics defined in Section 4.1, we define a set ECE metric to measure calibration at the set level.

Set-ECE: Measures calibration at the *answer set* level. For each example, we define set-level correctness $y_{\text{set}} = \mathbb{1}\{\exists i : a_i \in \mathcal{Y}^*\}$. For datasets where questions only admit one correct answer (and thus set level probabilities must sum to 1, we define set-level confidence $q_{\text{set}} = \sum_{i=1}^{K} q_i$. When there are multiple correct answers and probabilities can sum to > 1, set-level confidence is defined as $q_{\text{set}} = 1 - \prod_{i=1}^{K}(1 - q_i)$. Set ECE is computed by binning q_{set} and comparing empirical set accuracy to predicted set confidence, much like standard ECE.

E Token Efficiency Comparison: Single Answer vs. Multi Answer RL

Here, we provide a visual of the average token compute usage of a set of answers generated by RLVR Single, RLVR Multi, RLCR Single, and RLCR Multi. As one can see, Multi-Answer training *significantly* increases compute efficiency.

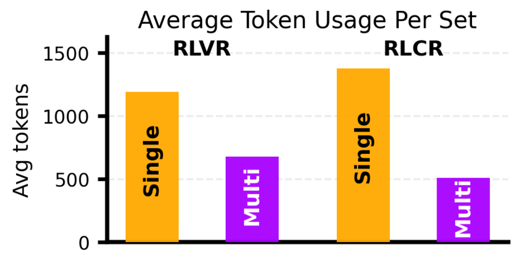


Figure 7: Average token usage for RLVR and RLCR in single and multi settings on DDXPlus.

F Unique Answer Analysis: RLVR Single vs. RLVR Multi

We create a Word Cloud comparing the answer diversity of RLVR Single and RLVR Multi on an example medical question. We run RLVR Single 30 times and RLVR Multi 10 times, with $k = 3$ answers per set. As one can see, RLVR single, even though it also collects 30 answers, only collects 3 unique answers. RLVR Multi, on the other hand, admits significantly more.

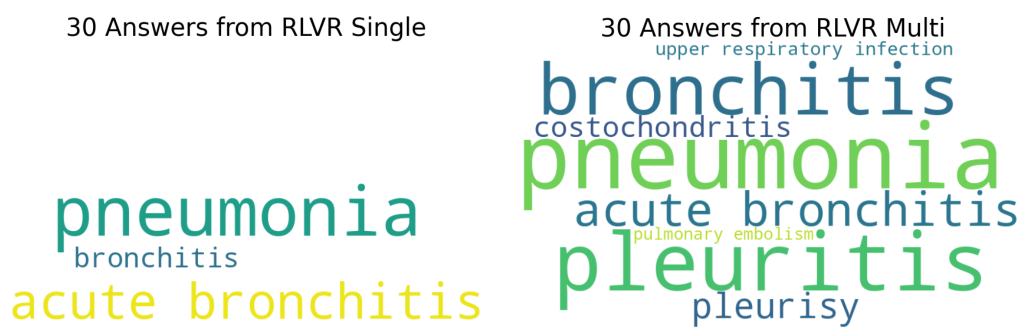


Figure 8: Word Cloud comparing Unique Answers gathered by RLVR Single and RLVR Multi

G Increasing k in RLVR-Multi

k	Evaluation Recall, RLVR Multi
2	0.78
3	0.68
4	0.65
5	0.62

Table 3: We report recall, or the average number of correct answers generated in a set, on an evaluation set of $5,000$ questions as k increases.

H FULL EXAMPLE - MEDICAL RLVR SINGLE, RLVR MULTI, RLCR SINGLE, RLCR MULTI

Question

```
Demographics:
  Age: 61
  Sex: M

================================================================================
SYMPTOMS AND ANTECEDENTS:
================================================================================

Symptoms:
  1. Symptom: Have you been coughing up blood?  Yes
  2. Symptom: Do you have pain somewhere, related to your reason for consulting?
     Yes
  3. Symptom: Characterize your pain:: a knife stroke
  4. Symptom: Do you feel pain somewhere?: lower chest
  5. Symptom: Do you feel pain somewhere?: posterior chest wall(R)
  6. Symptom: Do you feel pain somewhere?: posterior chest wall(L)
  7. Symptom: How intense is the pain?: 5
  8. Symptom: Does the pain radiate to another location?: nowhere
  9. Symptom: How precisely is the pain located?: 6
  10. Symptom: How fast did the pain appear?: 5
  11. Symptom: Are you experiencing shortness of breath or difficulty breathing in
     a significant way?  Yes
  12. Symptom: Do you have a fever (either felt or measured with a thermometer)?
     Yes
  13. Symptom: Have you had chills or shivers?  Yes
  14. Symptom: Do you have any lesions, redness or problems on your skin that you
     believe are related to the condition you are consulting for?  Yes
  15. Symptom: What color is the rash?: pink
  16. Symptom: Do your lesions peel off?: Y
  17. Symptom: Is the rash swollen?: 1
  18. Symptom: Where is the affected region located?: commissure(L)
  19. Symptom: How intense is the pain caused by the rash?: 0
  20. Symptom: Is the lesion (or are the lesions) larger than 1cm?: N
  21. Symptom: How severe is the itching?: 0
  22. Symptom: Do you have diffuse (widespread) muscle pain?  Yes
  23. Symptom: Have you noticed any new fatigue, generalized and vague discomfort,
     diffuse (widespread) muscle aches or a change in your general well-being
     related to your consultation today?  Yes
  24. Symptom: Do you have a cough?  Yes


Antecedents:
  1. Antecedent: Do you drink alcohol excessively or do you have an addiction to
     alcohol?  Yes
  2. Antecedent: Do you have heart failure?  Yes
  3. Antecedent: Do you have a chronic obstructive pulmonary disease (COPD)?  Yes
  4. Antecedent: Have you traveled out of the country in the last 4 weeks?: N
  5. Antecedent: Is your BMI less than 18.5, or are you underweight?  Yes

--------------------------------------------------------------------
CORRECT / GOLD ANSWERS:
  1. Pneumonia
  2. Pulmonary neoplasm
  3. Bronchitis
  4. Tuberculosis
  5. Possible NSTEMI / STEMI
  6. GERD
  7. Unstable angina
  8. Pericarditis
  9. Stable angina
```

RLVR MULTI

Okay, let's tackle this case. The patient is a 61-year-old male with several symptoms. He's coughing up blood, which is a red flag. He has chest pain described as a knife stroke, localized to the lower chest and right posterior chest wall. The pain is severe and sudden. He also has shortness of breath, fever, chills, and a rash on the left commissure that's pink and peeling. Additionally, he has diffuse muscle pain, fatigue, and a history of alcohol abuse, heart failure, COPD, and being underweight.

First, the hemoptysis (coughing up blood) and chest pain could point to a pulmonary issue. Given his COPD, maybe something like pulmonary embolism? But he also has heart failure, which could contribute to pulmonary issues. However, the sudden onset of severe pain and the presence of fever and chills might suggest an infection, like pneumonia. But the rash is another clue. The rash is on the commissure, pink, peeling, and swollen. That sounds like a fungal infection, maybe oral thrush, but the location is a bit unclear. Alternatively, could it be a drug reaction? He's on alcohol and has heart failure, so maybe medications are involved. But the main issues are the chest pain and hemoptysis.

Another possibility is tuberculosis, especially with the fever, night sweats (though not mentioned here), and hemoptysis. But the rash doesn't fit TB. Alternatively, a fungal infection like histoplasmosis, but that's more common in certain regions. The rash might be a secondary infection. Wait, the rash is on the commissure, which could be a sign of a fungal infection like candidiasis, but that's more oral. Maybe it's a skin manifestation of something else.

Considering the heart failure and COPD, could there be a pulmonary infection like pneumonia? The fever, chills, and shortness of breath support that. But the hemoptysis is more concerning for something like a pulmonary embolism or a tumor. However, the sudden onset and severe pain might point to a pulmonary embolism. But the rash is confusing. Alternatively, maybe the rash is a sign of a systemic infection, like sepsis, but that's a broader picture.

Another angle: the patient's alcohol use and underweight status could predispose him to infections or nutritional deficiencies. The diffuse muscle pain and fatigue could be due to systemic illness. The combination of hemoptysis, chest pain, fever, and rash makes me think of a fungal infection like coccidioidomycosis or histoplasmosis, but those usually have other symptoms. Alternatively, maybe a drug-induced reaction, but without knowing medications, it's hard. So the top three would be pulmonary embolism, pneumonia, and tuberculosis. Wait, the rash might be a clue for something else. Alternatively, the rash could be a sign of a systemic fungal infection. So maybe the three are pulmonary embolism, pneumonia, and fungal infection (like histoplasmosis). But the user wants three distinct diagnoses. Let me check again. The key symptoms are hemoptysis, chest pain, fever, chills, and rash. So possible diagnoses: 1. Pulmonary embolism, 2. Pneumonia, 3. Tuberculosis. Alternatively, 3. Fungal infection. But the rash might be more specific. Maybe the rash is a sign of a drug reaction, but without knowing medications, it's hard. So I'll go with pulmonary embolism, pneumonia, and tuberculosis.
\texttt{</think>}

\texttt{<answer1>} Pulmonary Embolism \texttt{</answer1>}
\texttt{<answer2>} Pneumonia \texttt{</answer2>}
\texttt{<answer3>} Tuberculosis \texttt{</answer3>}

RLCR MULTI

Okay, let's tackle this case. The patient is a 61-year-old male with several symptoms. He's coughing up blood, which is a red flag. He has chest pain described as a knife stroke, localized to the lower chest and right posterior chest wall. The pain is severe (intensity 5) and sudden (appeared quickly). He also has shortness of breath, fever, chills, and a rash on the left commissure that's pink, peeling, and swollen. Additionally, he has diffuse muscle pain, fatigue, and a history of alcohol addiction, heart failure, COPD, and being underweight.

First, the presence of hemoptysis (coughing up blood) and chest pain raises concerns for pulmonary issues. Given his COPD, maybe a pulmonary embolism? But he also has heart failure, which could contribute to pulmonary congestion. However, the sudden onset of pain and fever might point towards an infection or embolism. The rash is on the commissure, which is the corner of the mouth, and it's pink, peeling, and swollen. That could be a fungal infection like oral thrush, but given his alcohol use and underweight status, maybe something else. Wait, the rash is on the commissure, and he has a fever, which could indicate an infectious process. The diffuse muscle pain and fatigue could be part of a systemic infection or a viral illness. However, the combination of hemoptysis, chest pain, fever, and chills suggests a more serious condition.

Pulmonary embolism (PE) is a possibility, especially with the sudden onset of chest pain and shortness of breath. But PE usually doesn't present with hemoptysis unless it's large. Alternatively, pneumonia could explain the fever, chills, and cough, but the hemoptysis is more concerning. Another angle: tuberculosis? But TB is more chronic. The rash might be a sign of a fungal infection, but I'm not sure. His history of heart failure could lead to pulmonary edema, which might cause hemoptysis, but the pain and fever are less typical.

Another thought: maybe a lung abscess or infection. The sudden onset and systemic symptoms (fever, chills) could point towards an infection. However, the rash is localized to the mouth, which might be a separate issue.

So, the top differentials would be pulmonary embolism, pneumonia, and maybe tuberculosis. But given the sudden onset, PE is higher. However, the presence of hemoptysis and chest pain could also suggest a pulmonary issue like a ruptured lung cyst or a malignancy. But with the fever and chills, infection is more likely. Alternatively, maybe a viral infection with secondary bacterial pneumonia.

Another possibility is acute bronchitis, but that's less likely with hemoptysis. The diffuse muscle pain and fatigue could be part of a systemic infection. So, the most likely diagnoses are pulmonary embolism, pneumonia, and tuberculosis. But I need to rank them. Given the sudden onset of chest pain, shortness of breath, fever, and hemoptysis, pulmonary embolism is a strong candidate. Pneumonia could also fit, especially with the fever and chills. Tuberculosis is less likely without a chronic cough or weight loss, but he is underweight. However, the rash might be a red herring.

So the three main possibilities are pulmonary embolism, pneumonia, and tuberculosis. But I need to check if there's a better fit. Alternatively, acute myocardial infarction? But the pain is described as a knife stroke, which could be pleuritic, pointing more to a pulmonary issue. So I'll go with pulmonary embolism, pneumonia, and tuberculosis as the top three.
\texttt{</think>}

\texttt{<answer1>} Pulmonary Embolism \texttt{</answer1>}
\texttt{<confidence1>} 0.45 \texttt{</confidence1>}
\texttt{<answer2>} Pneumonia \texttt{</answer2>}
\texttt{<confidence2>} 0.35 \texttt{</confidence2>}
\texttt{<answer3>} Tuberculosis \texttt{</answer3>}
\texttt{<confidence3>} 0.20 \texttt{</confidence3>}

RLVR SINGLE

RLVR SINGLE:
\texttt{<think>}
Okay, let’s tackle this medical case. The patient is a 61-year-old male with several symptoms. He’s coughing up blood, which is a red flag. He has chest pain described as a knife stroke, localized to the lower chest and posterior walls on both sides. The pain is severe (intensity 5) and doesn’t radiate. He also reports shortness of breath, fever, chills, and a pink rash that peels on the left commissure. Additionally, he has widespread muscle pain, fatigue, and a cough worse at night. His medical history includes alcohol addiction, heart failure, COPD, and being underweight.

First, coughing up blood (hemoptysis) along with chest pain and respiratory symptoms makes me think of pulmonary issues. The presence of fever, chills, and night sweats could indicate an infection. Given his COPD, he’s already at risk for respiratory infections. But the combination of hemoptysis, fever, and chest pain might point towards something more serious like tuberculosis, especially since he’s underweight and has a history of alcohol use, which can weaken the immune system. However, tuberculosis typically presents with chronic symptoms, though acute cases can occur.

Another possibility is pneumonia, but hemoptysis is less common in typical pneumonia. Then there’s the rashpink, peeling on the commissure. That could be a sign of a systemic infection or an autoimmune condition. However, the systemic symptoms like fever, muscle pain, and fatigue align more with an infectious process.

His heart failure could contribute to fluid retention and make him more susceptible to infections. The combination of hemoptysis, fever, and chest pain also makes me consider lung cancer, but that’s more common in smokers and might present with different symptoms like weight loss and chronic cough. However, the acute onset of symptoms and the presence of fever might lean more towards an infection.

Tuberculosis should be considered, especially given the chronic nature of some symptoms, systemic involvement, and the patient’s risk factors. The rash might be a secondary manifestation. Alternatively, fungal infections could be a possibility, especially in someone with COPD and immunocompromised state. But tuberculosis is more likely given the symptoms.

I should also check for other causes like bronchiectasis, which can cause hemoptysis and recurrent infections. But again, the fever and systemic symptoms point towards an active infection. Considering all factors, the most probable diagnosis is tuberculosis, given the combination of hemoptysis, fever, night sweats, weight loss, and the presence of a rash as a possible extrapulmonary manifestation. \texttt{</think>} \texttt{<answer>}Tuberculosis
\texttt{</answer>}
\texttt{<confidence>}95.\texttt{</confidence>}

RLCR SINGLE

```
RLCR SINGLE (prefer correct; otherwise idx=0):
\texttt{<think>}
Okay, let's tackle this medical case. The patient is a 61-year-old male with several symptoms and antecedents. First, the key symptoms include coughing up blood, chest pain described as a knife stroke, shortness of breath, fever, chills, a pink rash that peels, muscle pain, fatigue, cough, night symptoms, and pain when inhaling. The antecedents are alcohol addiction, heart failure, COPD, and being underweight.

Coughing up blood (hemoptysis) is a red flag. Combined with chest pain and shortness of breath, this could point to a respiratory issue. The pain being a knife stroke suggests sharp, localized pain, which might indicate something like a pulmonary embolism (PE) or pleuritis. However, the presence of fever, chills, and a rash complicates things. The rash is pink, peels, and is on the commissure, which might be related to a systemic condition.

The patient has a history of COPD and heart failure, which are risk factors for PE. Alcohol use and underweight status could contribute to immune suppression or clotting issues. The diffuse muscle pain and fatigue could be due to a systemic infection or a condition like sepsis. However, the combination of hemoptysis, chest pain, and respiratory symptoms might lean towards tuberculosis or lung cancer, especially in someone with a history of smoking (though not mentioned here). Wait, the antecedents don't mention smoking, but COPD could be related.

I need to consider the most likely diagnosis. Given the symptoms, pulmonary embolism seems possible, especially with the history of COPD and heart failure. However, the rash and systemic symptoms might point towards something else like a systemic infection or even a drug reaction. Alternatively, tuberculosis could present with hemoptysis, fever, and night sweats, but the rash is less typical. Another angle: the patient's underweight status and alcohol use could lead to immunocompromise, making infections more likely. But the acute onset of symptoms (pain appearing quickly) might favor PE over a chronic condition like TB.

I should also check if the symptoms align with other conditions. The diffuse muscle pain and fatigue could be part of a systemic illness. The combination of hemoptysis, fever, chills, and chest pain is very concerning for PE. However, without imaging or more specific tests, it's challenging. The answer might be pulmonary embolism, but I'm not entirely sure due to the presence of the rash and other symptoms. Could it be something else like pneumonia with a rash? Or maybe a severe infection like sepsis with multi-system involvement. But the key symptom of hemoptysis points towards a pulmonary cause. I think PE is the most likely, but there's uncertainty due to the rash and other symptoms.
\texttt{</think>}

\texttt{<answer>} pulmonary embolism \texttt{</answer>}
\texttt{<analysis>} The diagnosis of pulmonary embolism (PE) is strongly suggested by hemoptysis, chest pain with a knife-like character, shortness of breath, fever, chills, and a history of COPD and heart failure, all of which increase PE risk. However, the presence of a pink, peeling rash on the commissure and diffuse muscle pain introduces ambiguity. These could indicate an alternative diagnosis such as a systemic infection, drug reaction, or tuberculosis. The patient's underweight status and alcohol use may contribute to immune compromise, complicating differential diagnosis. While PE remains the most likely explanation given the acute onset and respiratory symptoms, the rash and systemic manifestations introduce uncertainty. \texttt{</analysis>}
\texttt{<confidence>} 0.75 \texttt{</confidence>}
```