
Exploration with Principles for Diverse AI Supervision

Anonymous Author(s)
Affiliation
email

Abstract

1 Training large transformers using next-token prediction has given rise to groundbreaking ad-
2 vancements in AI. While this generative AI approach has produced impressive results, it heavily
3 leans on human supervision. Even state-of-the-art AI models like ChatGPT depend on fine-
4 tuning through human demonstrations, demanding extensive human input and domain expertise.
5 This strong reliance on human oversight poses a significant hurdle to the advancement of AI
6 innovation. To address this limitation, we propose a novel paradigm termed Exploratory AI
7 (EAI) aimed at autonomously generating high-quality training data. Drawing inspiration from
8 unsupervised reinforcement learning (RL) pretraining, EAI achieves exploration within the
9 natural language space. We accomplish this by harnessing large language models to assess
10 the novelty of generated content. Our approach employs two key components: an actor that
11 generates novel content following exploration principles and a critic that evaluates the generated
12 content, offering critiques to guide the actor. Empirical evaluations demonstrate that EAI sig-
13 nificantly boosts model performance on complex reasoning tasks, addressing the limitations of
14 human-intensive supervision.

15 1 Introduction

16 Training large transformers [41] using next token prediction has led to substantial AI
17 advancements, as evidenced by the groundbreaking results they have produced [33, 30].
18 While this generative AI approach has yielded remarkable AI results, it heavily relies
19 on human supervision. For instance, state-of-the-art AI models including ChatGPT [33]
20 along with a range of other models [7, 14, 9, *inter alia*], rely on fine-tuning through
21 human demonstrations, demanding significant human involvement and domain expertise.

22 This reliance on extensive human su-
23 pervision presents a substantial chal-
24 lenge since human supervision re-
25 quires domain expertise, is time con-
26 suming, and is tedious. Moreover, hu-
27 mans can struggle to provide reliable
28 supervision in highly specialized do-
29 mains. For instance, ChatGPT pos-
30 sesses a greater depth of knowledge
31 than the average human, which makes
32 it difficult to rely on humans to pro-
33 vide supervision for ChatGPT. More-
34 over, while our most advanced AI sys-
35 tems have made significant strides,
36 they still necessitate thorough, human-
37 guided processes to enhance their abil-
38 ity to answer factual or mathematical queries [20]. Yet, when it comes to more intricate and mission-
39 critical tasks, such as navigating complex tax or law regulations, these challenges will demand even
40 more specialized expertise and effort.

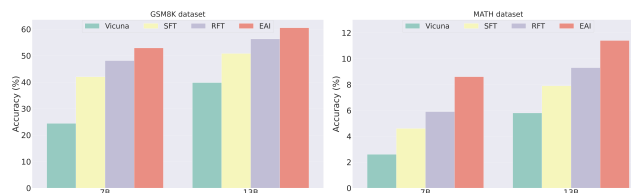


Figure 1: Test accuracy on mathematical reasoning benchmark GSM8K. Baselines include Vicuna, supervised finetuning Vicuna on training set (denoted as SFT), and supervised finetuning Vicuna on rejection sampled model generated diverse solutions on training set (denoted as RFT). Our Exploratory AI (EAI) substantially outperform all baselines.

41 Prior works attempt to explore alternatives to human supervision, by using AI supervision instead.
42 For example in mathematical reasoning, these studies propose sampling self generated solutions for
43 human curated questions from large language models and employ techniques like rejection sampling,
44 along with other techniques, to curate training data for the model [8, 29, 2, 16, 48, 45, *inter alia*].
45 While learning from such sampled content proves effective, a significant challenge persists: the
46 sampled contents often lack the necessary diversity, resulting in a rapid saturation of the learning
47 process [45, 48]. Moreover, the sampling approach has been confined to solutions exclusively, relying
48 on human-curated questions, thus imposing constraints on the diversity of generated data.

49 To tackle these limitations, we propose a novel approach for using AI models to autonomously
50 generate *diverse* data for learning purposes. This concept draws inspiration the APT algorithm [21]
51 designed for unsupervised RL pretraining [36, 17, 32]. RL pretraining studies exploring in a reward-
52 free environment to develop skills for quickly maximize various downstream rewards. APT allows
53 training RL agent to learn skills by autonomously explore reward free environment based on evaluating
54 novelty of encountered states using particle based entropy estimation [3, 35]. Adapting APT to large
55 language models presents several challenges, including computational complexity and the difficulty
56 of learning reward functions and exploration policies [13, 8]. Rather than relying on traditional RL
57 techniques, we harness the unique capabilities of large language models, such as their ability to
58 learn from context and follow instructions. In essence, we use them to perform the roles of both
59 a reward function and an exploration policy. Our approach, which we term Exploratory AI (EAI),
60 involves two key components: an actor and a critic. The actor is responsible for generating novel
61 content in natural language, while the critic evaluates this generated content and provides critiques
62 to guide the actor’s exploration. By evaluating the novelty of the generated contents, our method
63 allows for effective exploration in the rich space of natural language. EAI can generate diverse data
64 independently of human intervention. This makes it more scalable and automated, positioning it as
65 a preferable alternative to methods like supervised finetuning or rejection sampling that depend on
66 data curated by humans. Furthermore, EAI provides an interpretable window into the behavior and
67 knowledge of the model. It sheds light on how well the model possesses knowledge and its reasoning
68 behind generating novel questions. One can look at generations and their corresponding evaluations
69 which provide valuable insights about how model generates and evaluates.

70 We evaluate our approach on mathematical reasoning benchmarks GSM8K [8] and MATH [15],
71 EAI substantially improves performance on challenging reasoning tasks, outperforming both hu-
72 man supervision and AI supervision baselines. In contrast to human supervision, our approach is
73 autonomous and more scalable. When compared to prior state-of-the-art AI supervision baselines
74 including RFT [45] and WizardMath [24], our method provides a straightforward yet highly effective
75 solution for the generation of high-quality and diverse data.

76 Our contributions are two-fold: (a) In contrast to the predominant reliance on human supervision,
77 our novel approach, EAI, leverages the capabilities of large language models to autonomously
78 generate diverse high-quality training data. It achieves this by harnessing these models for self-
79 guided exploration, inspired by unsupervised reinforcement learning pretraining. (b) We conduct an
80 extensive series of experiments to systematically assess the effectiveness of EAI. Our results show
81 that EAI substantially outperform prior human supervision and AI supervision state-of-the-arts, and
82 significantly improve model performance.

83 **2 Exploratory AI for Diverse AI Supervision**

84 We present our approach for harnessing AI models to create AI supervision, in order to address the
85 reliance on extensive human supervision.

86 Our method employs a dynamic interplay between generation and evaluation. This concept draws
87 inspiration from unsupervised RL pretraining (URL) [17] and particularly the APT algorithm [21]. RL
88 pretraining studies exploring in a reward-free environment to develop skills for quickly maximizing
89 various downstream rewards. APT allows training RL agent to learn skills by autonomously exploring
90 a reward free environment based on evaluating novelty of encountered states using particle based
91 entropy estimation [3, 35].

92 Adapting APT directly to large language models presents several challenges, including grappling
93 with computational complexity and the difficulty of learning reward functions and exploration
94 policies [13, 8]. In response, we propose Exploratory AI (EAI), a novel approach that circumvents

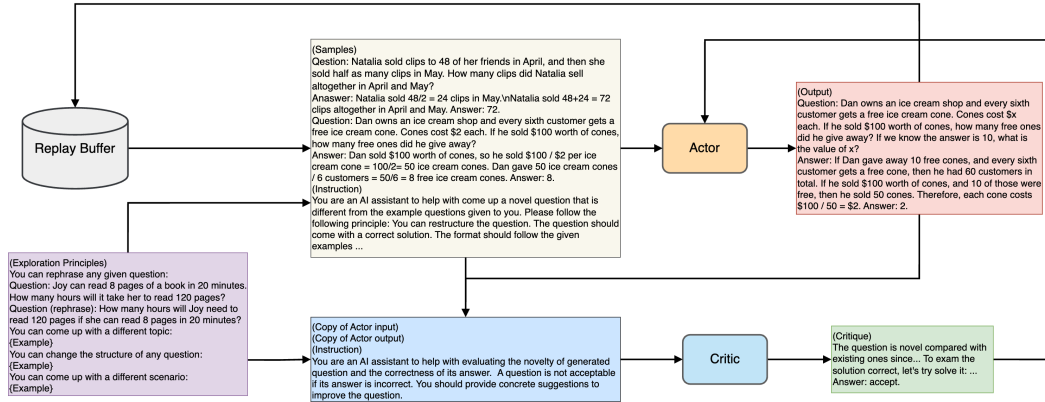


Figure 2: Generating diverse data in the Exploratory AI Framework. In the diagram, we demonstrate how the actor generates diverse content by conditioning on samples from the replay buffer and exploration principles. These principles include rephrasing question, coming up a novel topic, restructuring question, and coming up a new scenario, we provide examples associated with the principles to guide exploration. The actor’s input and its generated output undergo evaluation by the critic. The critic assesses the novelty of the generated data; when the evaluation is favorable, the data is stored in the replay buffer. In cases where the evaluation does not meet the criteria, the critic provides critiques to guide the actor. The replay buffer can be initialized with a pre-existing human-created dataset (e.g., GSM8K training set) or can remain empty for starting from scratch with zero-shot exploration.

95 the need for direct reinforcement learning (RL) by harnessing the power of large language models for
 96 exploration. Our method explore the natural language space by employing these language models to
 97 assess the novelty of generated content and guide the exploration process. Our approach consists of
 98 two key components: an “actor” responsible for generating novel content and a “critic” that evaluates
 99 the actor’s outputs and provides feedback to guide further content generation.

100 Concretely, we instruct the actor to generate content that diverges from samples from the replay buffer.
 101 The replay buffer can be initialized with a pre-existing human-created dataset (e.g., GSM8K training
 102 set) or can remain empty for zero-shot exploration. Similar to APT, we found having pre-existing
 103 samples accelerates learning and encourages the actor to have more long term exploratory behaviors.
 104 We then instruct the critic to assess the actor’s outputs and provides critiques. This feedback loop
 105 guides the actor in refining and enhancing its content. This iterative process continues until it reaches
 106 a predefined maximum number of iterations, and the resulting outputs are stored in a dataset. The
 107 data can then be used for finetuning AI models.

108 We equip both the actor and critic with a curated set of guiding principles to facilitate the generation
 109 and evaluation of diverse questions. These principles include rephrasing question, coming up a novel
 110 topic, restructuring question, and coming up a new scenario, we provide examples associated with
 111 the principles to guide exploration. While it’s theoretically possible to provide all these principles to
 112 the model, in this study, we opt to a more controlled approach. To balance the quantity of generated
 113 data for each principle, we uniformly sample one principle at a time and input it to both the actor and
 114 critic. The actor is instructed to follow the principle (e.g., restructuring the question) during question
 115 generation. Similarly, the critic’s role is to evaluate the diversity, considering the selected principle.
 116 It’s worth noting that the critic’s principle is worded slightly differently from the exploration principle;
 117 for a detailed list, please refer to Appendix A. Our method is shown in Figure 2 and the algorithm is
 118 shown in Algorithm 1.

Actor prompt

You are an AI assistant to help with come up a novel question that is different from the example questions given to you. The question should come with a correct solution. Please follow the given principle in generating the question. **{principle}**

119

Critic prompt

You are an AI assistant to help with evaluating the novelty of generated question and the correctness of its answer. A question is not acceptable if its answer is incorrect. You should provide concrete suggestions to improve the question. Explain your reasoning step by step and output final evaluation on novelty and correctness at the end. Follow the given principle on evaluating the novelty. **{principle}**

120

Principles for exploration

You can rephrase any given question:

Question: Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?

Question (rephrase): How many hours will Joy need to read 120 pages if she can read 8 pages in 20 minutes?

You can come up with a different topic:

Question: Jack is stranded on a desert island. He wants some salt to season his fish. He collects 2 liters of seawater in an old bucket. If the water is 20% salt, how many ml of salt will Jack get when all the water evaporates?

Question (topic): Samantha is designing a circular garden in her backyard. The garden has a diameter of 8 meters. She wants to build a path around the garden that is 1 meter wide. What is the area of the path, in square meters, that Samantha will need to pave with stones or concrete?

You can change the structure of any question:

Question: Dan owns an ice cream shop and every sixth customer gets a free ice cream cone. Cones cost \$2 each. If he sold \$100 worth of cones, how many free ones did he give away?

Question (restructured): Dan owns an ice cream shop and every sixth customer gets a free ice cream cone. Cones cost \$x each. If he sold \$100 worth of cones, how many free ones did he give away? If we know the answer is 10, what is the value of x?

You can come up with a different scenario:

Question: Ed has 2 dogs, 3 cats and twice as many fish as cats and dogs combined. How many pets does Ed have in total?

Question (scenario): Sarah owns 4 bicycles, 2 skateboards, and three times as many pairs of rollerblades as bicycles and skateboards combined. How many wheeled sports equipment items does Sarah have in total?

121

122 Exploratory AI has several attractive properties as an approach for facilitating AI supervision in
123 language models:

- 124 1. EAI can generate diverse AI supervision for learning, independently of human input, making
125 it more scalable compared with supervised finetuning or rejection sampling based on human
126 curated data.
- 127 2. EAI provides an interpretable window into the behavior and knowledge of the model. It sheds light
128 on how well the model possesses knowledge and its reasoning behind generating novel questions.
129 One can look at generations and their corresponding evaluations which provide valuable insights
130 about how model generates and evaluates.
- 131 3. EAI's versatility allows for a fusion of the best aspects of supervised finetuning and prompting.
132 Users can prompt the model to focus on certain topics or aspects by directing actor and critic with
133 different prompting strategies.
- 134 4. EAI demonstrates its effectiveness by excelling in mathematical reasoning tasks, as we will
135 demonstrate in our experiments. Moreover, its capabilities are not limited to mathematics; it
136 holds promise for a broad spectrum of language-related tasks in principle.

137 In empirical experiments, we will evaluate the utility of EAI for mathematical reasoning and analysis
138 its effectiveness.

139 3 Setting

140 We evaluate our method on the mathematical reasoning tasks, and achieve better results that EAI
141 largely improve results and significantly outperforms prior state-of-the-arts.

142 **Benchmarks.** We evaluate our method on the mathematical reasoning tasks GSM8K. This bench-
143 marks exams model's mathematical reasoning capabilities, we finetune model on the training split,
144 and evaluate model on the test split. The GSM8k dataset includes around 7,500 training and 1,319
145 test math problems for high school-level students, involving basic arithmetic operations. Problems
146 typically require 2 to 8 steps for a solution. The MATH dataset comprises 7,500 training and 5,000
147 challenging test problems from prestigious math competitions (AMC 10, AMC 12, AIME) covering

Algorithm 1 Exploratory AI for diverse AI supervision.

Required: One (or two) large language models M for actor and critic.
Replay Buffer B , empty or optionally initialized with pre-existing data.
Initialize
for $i = 1$ **to** max iterations **do**
 Randomly sample data points from B
 Use preassigned principle or sample one principle.
 for $i = j$ **to** max rounds **do**
 Prompt the actor with the principle to generate content (a question and its answer) that in the same domain but diverge from the sampled inputs (questions and answers) sampled from B
 Prompt the critic with the principle to evaluate the diversity of generated question and correctness of answer, and decide whether to accept
 if Accepted **then**
 Save generated question and answer to B
 break
 else
 Continue to prompt actor with the critique as additional input
 end if
 end for
end for

148 various academic domains, including prealgebra, algebra, number theory, counting and probability,
149 geometry, intermediate algebra, and precalculus.

150 **Baselines.** We compare our approach with (a) Base model including Vicuna 7B, 13B, and 30B [7].
151 Vicuna is LLaMA2 finetuned on user conversations shared online (ShareGPT). We use Vicuna as
152 base model for all baselines and our method; (b) Supervised finetuning (SFT) on training set of the
153 original GSM8K or MATH, in which a language model is finetuned on human written exemplars of
154 questions–answers pairs. SFT has been widely used in prior works for improving language models
155 mathematical reasoning [19, 40, 30, *inter alia*] and following user intention [14, 9, *inter alia*]. We
156 also compare with WizardMath [24] which does SFT on ChatGPT annotated questions and solutions,
157 as well as MAMMO TH [47] which uses GPT4 annotated solutions; (c) Rejection sampling finetuning
158 (RFT) [45] which applies supervised finetuning on rejection sampled model generated data. We
159 provide baseline scores for SFT and RFT from both their original papers and our implementations
160 using Vicuna, ensuring a fair and comprehensive comparison; (d) Proprietary models including
161 GPT-4 [30], ChatGPT [33], and Claude2 [1]. All baselines are evaluated by prompting them to output
162 step by step reasoning followed by final answers [44].

163 **Generated data size.** We sample roughly the same amount of data for each principle outlined
164 in Section 2. To optimize computational cost, we have set the number of interaction rounds in
165 Algorithm 1 to a maximum of two. Our preliminary experiments revealed that this two-round
166 interaction is typically sufficient for the actor to produce high-quality and diverse data. For each of the
167 four principles – ‘rephrase question’, ‘introduce a new topic’, ‘restructure the question’, and ‘introduce
168 a new scenario’ – we generate approximately 25,000 samples for GSM8K and approximately 15,000
169 samples for MATH. The generation on 8 A100 80GB GPUs take from 40 to 200 hours depending on
170 the model size and the specific principles applied.

171 4 Results

172 **Benchmark.** In Table 1, we present the results of pass@1 (%) on GSM8K and MATH, showcasing
173 the performance of various models across different supervision types and model sizes. Our approach,
174 denoted as Vicuna + EAI, is compared with Vicuna and Vicuna + RFT, which serve as the closest
175 baselines. Notably, our method outperforms Vicuna in all scenarios. For instance, in the 7B parameter
176 setting, our model achieves 52.9% and 9.5% on GSM8K and MATH, respectively, surpassing
177 Vicuna’s scores of 24.4% and 2.6%. This trend continues across larger model sizes and different
178 datasets. Furthermore, compared to Vicuna + RFT, our approach maintains a consistent advantage,
179 demonstrating the effectiveness of our self AI supervision method in improving model performance

Table 1: Results of pass@1 (%) on GSM8k and MATH. In this study, to ensure equitable and cohesive evaluations, we report the scores of all models under the same settings of greedy decoding. * denotes original LLaMA2 based SFT and RFT. Bold numbers are the absolute improvement of EAI over RFT which serves as prior state-of-the-art and the closet baseline.

Model	Supervision	Data	Params	GSM8K	MATH
GPT-4	-	-	-	92.0	42.5
ChatGPT	-	-	-	80.8	34.1
Claude 2	-	-	-	88.0	32.5
Minerva	-	-	8B	16.2	14.1
			62B	52.4	27.6
			540B	58.8	33.6
LLaMA2	-	-	7B	14.6	2.5
			13B	28.7	3.9
			35B	42.2	6.4
SFT*	Human+LLaMA	7.5K	7B	41.6	-
			13B	50.0	-
RFT*	Human+LLaMA	47K	7B	47.5	5.6
			13B	54.8	9.6
WizardMath	Human+ChatGPT	96K	7B	54.9	10.7
			13B	63.9	14.0
			70B	81.6	22.7
MAmmoTH	Human+GPT4	260K	7B	51.7	31.2
			13B	61.7	36.0
			70B	76.7	44.2
Vicuna	-	-	7B	24.4	2.6
			13B	39.8	5.8
			30B	47.5	7.7
SFT	Human	7.5K	7B	42.0	4.6
			13B	50.8	7.9
RFT	Human+Vicuna	48K	7B	48.1	5.9
			13B	56.3	9.3
EAI	Human+Vicuna	48K	7B	52.9 (+4.8)	8.6 (+2.7)
			13B	60.5(+4.2)	11.4(+2.1)

180 on these tasks. These results highlight the significant improvements our method brings to the table
 181 and its competitive edge in natural language understanding and generation tasks.

182 **Comparison of diversity.** We evaluate EAI in terms of the diversity of generated data.
 183 We compare RFT and EAI in terms
 184 of the submodularity diversity gain [4,
 185 27]. This metric serves as an indicator
 186 of the extent to which the generated
 187 data contribute to the overall diversity
 188 of the dataset. A higher diversity
 189 gain suggests that the newly generated
 190 questions exhibit greater dissimilarity
 191 from the existing dataset. We measure
 192 the gain over GSM8K training set by
 193 varying the amount of generated content.
 194 We use OpenAI GPT embedding text-embedding-ada-002
 195 to encode the data. The results depicted
 196 in Figure 3 clearly demonstrate that
 197

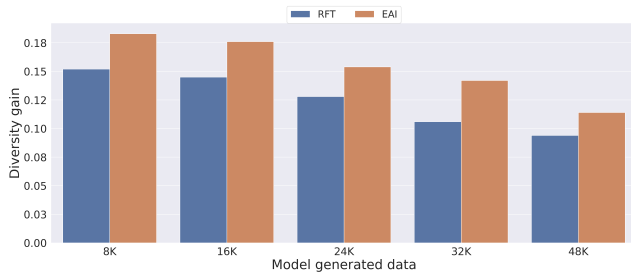


Figure 3: Comparison of diversity gain by adding generated data to GSM8K training set. EAI achieves higher diversity than other baselines.

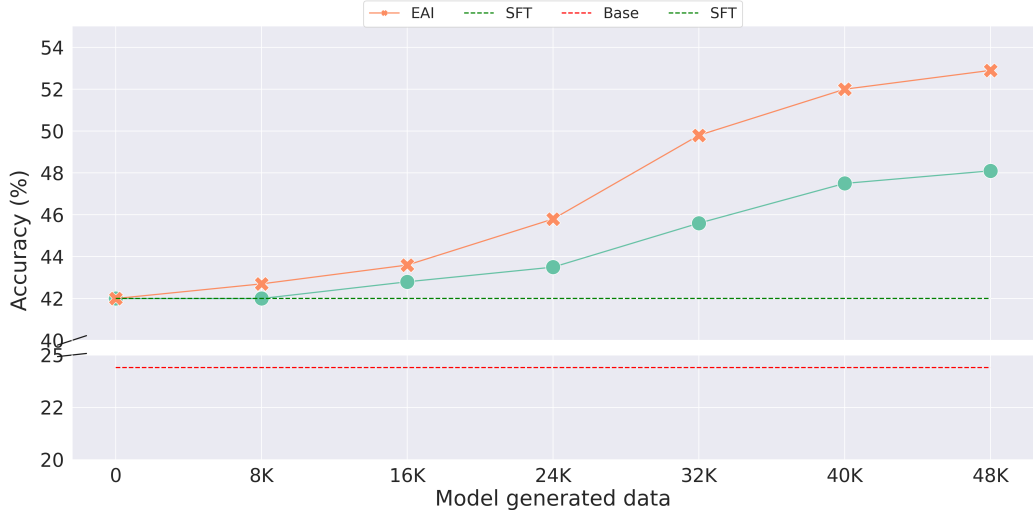


Figure 4: Data scaling on GSM8K. Shown are GSM8K accuracy with different amount of generated data. EAI generates high quality data for learning and scales well with data.

198 EAI consistently outperforms RFT in terms of diversity, thereby providing a more diverse set of
 199 generated data

200 **Effect of sampled inputs.** The Table 2 presents the results of an experiment
 201 examining the impact of varying the number of samples on GSM8K and MATH.
 202 As the number of samples increases from 0 to 8, we
 203 observe a steady incremental improvement on both
 204 GSM8K and MATH. On GSM8K, the performance
 205 rises from 50.1 to 52.9. On MATH, the effect is more
 206 pronounced. These results suggest that increasing
 207 the number of samples has a positive effect on both
 208 GSM8K and MATH, highlighting the significance of
 209 conditional input size in this experimental context.

Table 2: Effect of different number of samples from replay buffer.

Number	0	1	4	8
GSM8K	50.1	50.8	51.9	52.9
MATH	6.6	7.1	7.5	8.6

210 **Scaling with generated data.** We assess the performance of EAI in terms of sample efficiency on
 211 the GSM8K dataset. Our primary focus lies in understanding how the results evolve in response
 212 to varying amounts of generated data. Sample efficiency holds paramount importance, given that
 213 autoregressive data generation is inefficient. Enhanced sample efficiency broadens the practical
 214 utility of our approach in real-world applications. The results depicted in Figure 4 clearly illustrate
 215 a significant advantage for EAI over the previous state-of-the-art RFT. Notably, as more data is
 216 employed, RFT exhibits improved performance, but its sample efficiency lags behind EAI by a
 217 substantial margin. At just 16K data points, EAI outperforms RFT’s performance at 48K data points,
 218 achieving more than a 3x higher level of sample efficiency.

Table 3: Effect of different exploration principles on GSM8K and MATH.

rephrase	new topic	restructure	new scenario	GSM8K	MATH
✓	✓	✓	✓	52.9	8.6
✗	✓	✓	✓	48.8	7.1
✓	✗	✓	✓	49.7	7.8
✓	✓	✗	✓	48.9	6.9
✓	✓	✓	✗	49.5	7.5

219 **Evaluating the effect of exploration principles.** The results of varying exploration principles, as
 220 shown in Table 3, reveal some interesting insights. When all principles are in place (✓ for rephrase,
 221 new topic, restructure, and new scenario), the model performs at its best on GSM8K and MATH.
 222 This suggests that using all principles simultaneously leads to the most favorable outcomes. Among
 223 the principles, the most critical ones appear to be "rephrase" and "restructure", as seen when one of
 224 them is removed (✗). Without "rephrase" the performance drops on both datasets, emphasizing that
 225 the ability to rephrase and generate diverse content is crucial. Similarly, the omission of "restructure"

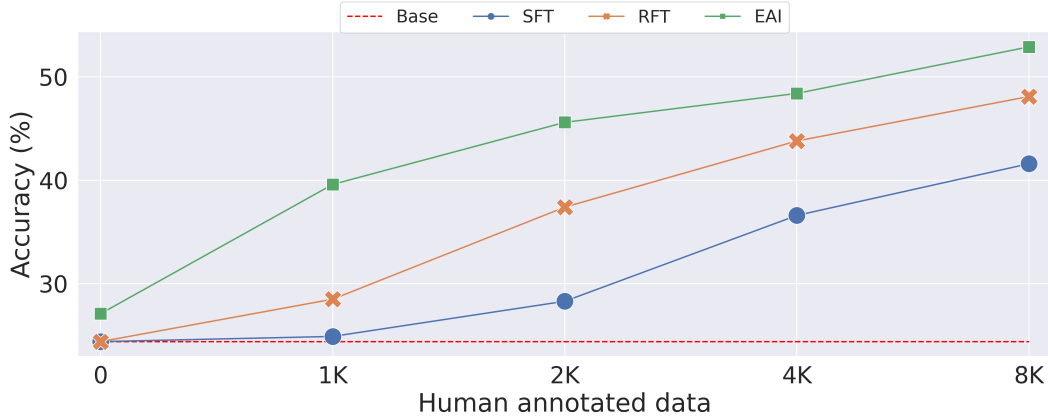


Figure 5: Performance on GSM8K with different amount of human annotated data. EAI performs well even without human annotation and scales well with more human provided annotations.

226 leads to a significant drop in MATH scores, highlighting the significance of introducing novel
 227 question-structuring approaches for solving more challenging problems.

228 **Scaling with human annotation size.** Figure 5 illustrates the results obtained when utilizing
 229 varying amounts of human annotation data from the GSM8K training set. We employ three different
 230 approaches in our experiments: SFT which directly finetunes the base model, Vicuna-7B, on the
 231 provided data. RFT which leverages the provided data to perform rejection sampling from the model.
 232 EAI which utilizes the provided data to initialize a replay buffer and explore new content for training.
 233 The results consistently demonstrate that EAI significantly outperforms all the baseline methods
 234 across various levels of human annotation data, underscoring its efficacy in generating high-quality
 235 training data. Remarkably, our experiments reveal that EAI performs admirably even in the absence of
 236 any human annotations, hinting at the potential to entirely eliminate the need for human intervention
 237 in the process.

238 5 Related Work

239 Transformers [41] trained using next token prediction have gave rise to many state-of-the-art AI
 240 systems [33, 30]. The remarkable AI results achieved with this generative AI approach heavily
 241 hinge upon the availability of diverse and high-quality data. For instance, state-of-the-art AI models
 242 including ChatGPT [33] and GPT4 [30] along with a range of other open source models such
 243 as Vicuna, Koala, and Dolly [9, 14, 7, *inter alia*], require extensive finetuning through human
 244 demonstrations. This process involves human conversations with ChatGPT or written demonstrations,
 245 demanding significant human involvement and domain expertise. Previous research has explored
 246 various avenues to enhance performance and sample efficiency, as well as alternative sources of
 247 supervision. To align with human preferences, there has been active research into developing simple
 248 algorithms for learning from human preferences [23, 46, 10, 40, *inter alia*]. In contrast to human
 249 demonstrations or feedback, another line of work explores the utilization of environmental feedback,
 250 such as unit test errors [18, 6, 34], which has demonstrated improved results in coding tasks. Some
 251 works explore using LLMs to provide AI supervision based exploration techniques for applications
 252 in solving games [11, 22, 42] and demonstrate improved results. Furthermore, some prior research
 253 leveraged proprietary APIs to indirectly obtain high-quality human data, enhancing model capabilities
 254 in areas like instruction following [43, 39, 14, 7] and mathematical reasoning [24, 26, 47, *inter alia*].
 255 Another line of research explores the use of models to supervise themselves [38, 25, 16, 2, 45],
 256 yielding improved results in reasoning tasks and alignment with human preferences. Our work
 257 focuses on generating diverse and high-quality data using AI models and we demonstrate applying
 258 our proposed approach to enhance open-source models by having them self-generate learning data.
 259 Our approach’s exploration technique is related to unsupervised RL based exploration [36, 37, 21, 5,
 260 32, 28, 12, 31, *inter alia*], however, our method does not require training RL agent. Additionally, some
 261 works have delved into more detailed forms of human supervision [20], demonstrating that LLMs
 262 benefit more from step-by-step process-based supervision than sparse outcome-based supervision.
 263 Our research uniquely centers on the data dimension, with a specific emphasis on harnessing AI

264 models to generate diverse high-quality AI supervision. To this end, we introduce an actor-critic
265 based approach for automating the exploration process.

266 **6 Conclusion**

267 In this work we propose an approach to automatically generate diverse, high-quality data from AI
268 models. Our approach Exploratory AI consists of prompting an actor model to generate diverse
269 contents that are different from existing contents, and using a critic model for evaluating the novelty
270 of generated data and providing critiques to guide the exploration process. Experimental evalua-
271 tions confirms the effectiveness of EAI, demonstrating its capacity to generate diverse content and
272 substantially enhance model performance on GSM8K and MATH datasets.

273 **Limitations and Future Work.** Although our method achieves state-of-the-art results using AI
274 supervision, it does have some limitations that need to be addressed:

- 275 • Our approach, while achieves substantial improvement with open source models, lags behind
276 results achieved using supervision by more powerful models such as ChatGPT.
- 277 • This work focuses on evaluation on mathematical reasoning domains. It worth applying EAI to
278 more diverse applications.

279 In terms of future prospects, our approach of generating diverse content with AI models opens up
280 interesting possibilities, such as extending EAI to evaluate novelty across the entire data buffer,
281 employing either a brute force approach (evaluating all data through a critic) or employing embedding
282 similarity search techniques. Moreover, there’s potential in extending EAI to incorporate multiple
283 actor and critic models, or in experimenting with various strategies to enhance exploration. It
284 would also be interesting to apply our method to proprietary APIs to source even more diverse and
285 higher-quality data.

286 References

- 287 [1] Anthropic. Introducing claude, 2023. URL [https://www.anthropic.com/index/](https://www.anthropic.com/index/introducing-claude)
288 [introducing-claude](https://www.anthropic.com/index/introducing-claude).
- 289 [2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones,
290 Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai:
291 Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- 292 [3] J Beirlant. Nonparametric entropy estimation: An overview. *International Journal of the*
293 *Mathematical Statistics Sciences*, 6:17–39, 1997.
- 294 [4] Jeff Bilmes. Submodularity in machine learning and artificial intelligence. *arXiv preprint*
295 *arXiv:2202.00132*, 2022.
- 296 [5] Víctor Campos, Pablo Sprechmann, Steven Hansen, Andre Barreto, Steven Kapturowski, Alex
297 Vitvitskyi, Adria Puigdomenech Badia, and Charles Blundell. Beyond fine-tuning: Transferring
298 behavior in reinforcement learning. *arXiv preprint arXiv:2102.13515*, 2021.
- 299 [6] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language
300 models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023.
- 301 [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,
302 Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot
303 impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org>, 2023.
- 304 [8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
305 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to
306 solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- 307 [9] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi,
308 Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first
309 truly open instruction-tuned llm, 2023. URL [https://www.databricks.com/blog/2023/](https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm)
310 [04/12/dolly-first-open-commercially-viable-instruction-tuned-llm](https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm).
- 311 [10] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun
312 Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model
313 alignment. *arXiv preprint arXiv:2304.06767*, 2023.
- 314 [11] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek
315 Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language
316 models. *arXiv preprint arXiv:2302.06692*, 2023.
- 317 [12] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you
318 need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- 319 [13] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization.
320 In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.
- 321 [14] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and
322 Dawn Song. Koala: A dialogue model for academic research. *Blog post, April*, 1, 2023.
- 323 [15] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn
324 Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset.
325 *arXiv preprint arXiv:2103.03874*, 2021.
- 326 [16] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei
327 Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.
- 328 [17] Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang,
329 Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. *arXiv*
330 *preprint arXiv:2110.15191*, 2021.

- 331 [18] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi.
332 Coderl: Mastering code generation through pretrained models and deep reinforcement learning.
333 *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.
- 334 [19] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay
335 Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving
336 quantitative reasoning problems with language models. *Advances in Neural Information
337 Processing Systems*, 35:3843–3857, 2022.
- 338 [20] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan
339 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint
340 arXiv:2305.20050*, 2023.
- 341 [21] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In
342 *Advances in Neural Information Processing Systems*, 2021.
- 343 [22] Hao Liu, Tom Zahavy, Volodymyr Mnih, and Satinder Singh. Palm up: Playing in the latent
344 manifold for unsupervised pretraining. *Advances in Neural Information Processing Systems*, 35:
345 35880–35893, 2022.
- 346 [23] Hao Liu, Carlo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with
347 feedback. *arXiv preprint arXiv:2302.02676*, 2023.
- 348 [24] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng,
349 Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reason-
350 ing for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*,
351 2023.
- 352 [25] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri
353 Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, et al. Self-refine: Iterative refinement
354 with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.
- 355 [26] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and
356 Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv
357 preprint arXiv:2306.02707*, 2023.
- 358 [27] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL
359 <http://probml.github.io/book2>.
- 360 [28] Mirco Mutti, Lorenzo Pratissoli, and Marcello Restelli. A policy gradient method for task-
361 agnostic exploration. *arXiv preprint arXiv:2007.04640*, 2020.
- 362 [29] Ansong Ni, Jeevana Priya Inala, Chenglong Wang, Alex Polozov, Christopher Meek, Dragomir
363 Radev, and Jianfeng Gao. Learning math reasoning from self-sampled correct and partially-
364 correct solutions. In *The Eleventh International Conference on Learning Representations*,
365 2022.
- 366 [30] OpenAI. Gpt-4 technical report, 2023.
- 367 [31] Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsuper-
368 vised skill discovery. *arXiv preprint arXiv:2302.05103*, 2023.
- 369 [32] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration
370 by self-supervised prediction. In *International conference on machine learning*, pages 2778–
371 2787. PMLR, 2017.
- 372 [33] J. Schulman, B. Zoph, C. Kim, J. Hilton, J. Menick, J. Weng, J. F. C. Uribe, L. Fedus, L. Metz,
373 M. Pokorny, R. G. Lopes, S. Zhao, A. Vijayvergiya, E. Sigler, A. Perelman, C. Voss, M. Heaton,
374 J. Parish, D. Cummings, R. Nayak, V. Balcom, D. Schnurr, T. Kaftan, C. Hallacy, N. Turley,
375 N. Deutsch, and V. Goel. Chatgpt: Optimizing language models for dialogue. *OpenAI Blog*,
376 2022. URL <https://openai.com/blog/chatgpt>.
- 377 [34] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with
378 dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

- 379 [35] Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk.
380 Nearest neighbor estimates of entropy. *American journal of mathematical and management*
381 *sciences*, 23(3-4):301–321, 2003.
- 382 [36] Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated
383 reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental*
384 *Development*, 2(2):70–82, 2010.
- 385 [37] Aravind Srinivas and Pieter Abbeel. Unsupervised learning for reinforcement learning, 2021.
386 URL https://icml.cc/media/icml-2021/Slides/10843_QHaHBNU.pdf.
- 387 [38] Zhiqing Sun, Yikang Shen, Qinzhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming
388 Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with
389 minimal human supervision. *arXiv preprint arXiv:2305.03047*, 2023.
- 390 [39] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy
391 Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model,
392 2023.
- 393 [40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei,
394 Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open
395 foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- 396 [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
397 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information*
398 *processing systems*, 30, 2017.
- 399 [42] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan,
400 and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models.
401 *arXiv preprint arXiv:2305.16291*, 2023.
- 402 [43] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi,
403 and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instruc-
404 tions. *arXiv preprint arXiv:2212.10560*, 2022.
- 405 [44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le,
406 Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models.
407 *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- 408 [45] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou.
409 Scaling relationship on learning mathematical reasoning with large language models. *arXiv*
410 *preprint arXiv:2308.01825*, 2023.
- 411 [46] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf:
412 Rank responses to align language models with human feedback without tears. *arXiv preprint*
413 *arXiv:2304.05302*, 2023.
- 414 [47] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao
415 Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv*
416 *preprint arXiv:2309.05653*, 2023.
- 417 [48] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with
418 reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

419 **A Prompt**

420 The critic’s principle is based on the exploration principle but is worded slightly differently from
421 the actor’s exploration principle, in order to tell critic what kind of new questions are diverse and
422 acceptable. The critic will use the principle to evaluate diversity of questions and evaluate the
423 correctness of answer.

Principles for critique

A different topic is acceptable:

Question: Jack is stranded on a desert island. He wants some salt to season his fish. He collects 2 liters of seawater in an old bucket. If the water is 20% salt, how many ml of salt will Jack get when all the water evaporates?

Question (topic): Samantha is designing a circular garden in her backyard. The garden has a diameter of 8 meters. She wants to build a path around the garden that is 1 meter wide. What is the area of the path, in square meters, that Samantha will need to pave with stones or concrete?

A question with different structure is acceptable:

Question: Dan owns an ice cream shop and every sixth customer gets a free ice cream cone. Cones cost \$2 each. If he sold \$100 worth of cones, how many free ones did he give away?

Question (restructured): Dan owns an ice cream shop and every sixth customer gets a free ice cream cone. Cones cost \$x each. If he sold \$100 worth of cones, how many free ones did he give away? If we know the answer is 10, what is the value of x?

Rephrased question is acceptable:

Question: Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?

Question (rephrase): How many hours will Joy need to read 120 pages if she can read 8 pages in 20 minutes?

A different scenario is acceptable:

Question: Ed has 2 dogs, 3 cats and twice as many fish as cats and dogs combined. How many pets does Ed have in total?

Question (scenario): Sarah owns 4 bicycles, 2 skateboards, and three times as many pairs of rollerblades as bicycles and skateboards combined. How many wheeled sports equipment items does Sarah have in total?

424

425 **B Case Study of EAI**

426 We conduct a qualitative study how actor and critic accomplish exploration. On GSM8K, we manually
427 inspect 100 examples randomly sampled from EAI explored dataset to better understand how does the
428 exploration work. We observed that the critic excels at assessing the novelty of generated questions,
429 providing precise feedback to enhance them. The actor effectively utilizes the critic’s feedback to
430 refine the generated questions. Of the 100 samples reviewed, in 24 instances, the critic initially
431 found the questions lacking in novelty and provided suggestions for improvement. Out of these,
432 19 were subsequently refined and accepted after revisions. Below is a representative example of
433 this interaction where the critic offers advice to significantly differentiate the generated question.
434 Here, the critic recommends incorporating the mathematical concept of volumes and presenting an
435 alternative scenario

436 [Critic guides exploration]

437 (First round)

438 Actor’s generation: Question: Jake is planning to buy a new skateboard which costs \$200. Jake only
439 has a quarter of the money he requires. His brother decided to give him \$20 for it, and his uncle three
440 times as much as his brother. How much more does Jake need to raise to purchase the skateboard? (answer
441 omitted)

442 Critic’s critique: The new question you’ve provided is somewhat similar in structure to the existing
443 questions, as it involves basic arithmetic operations (multiplication, division, addition, and subtraction).
444 While it introduces a different context with hats and balls, the underlying math principles are similar to
445 those used in the existing questions. To make the new question substantially different, you could consider
446 introducing more complex mathematical concepts such as volumes or using a different scenario altogether.

447 X

448 (Second round)

449 Actor’s generation: An artist is creating a sculpture using various geometric shapes. The sculpture
450 consists of a cube, a sphere, a cylinder, and a pyramid. The cube has a side length of 6 units, the sphere has
451 a radius of 4 units, the cylinder has a height of 8 units and a base radius of 3 units, and the pyramid has a
452 base edge length of 5 units and a height of 7 units. Calculate the total volume of the artist’s sculpture. A:
453 omitted

454 Critic’s critique: The new question is substantially different from the existing questions. It involves
455 a slightly more complex problem-solving process. It is good to be added to math questions. The decision
456 is accept. ✓

457 We also conducted a study to determine how effective the critic is at evaluating the correctness of
458 answers. Of the 100 critiques assessed, the critic accurately identified incorrectness in 13 actor
459 generations and offered critiques to correct 10 of them. However, there were 7 instances where the
460 critic failed to recognize the errors. The remaining actor generations were accurate. It’s worth noting
461 that, due to the uniform usage of the four exploration principles, out of the 100 actor generations, 50
462 were either paraphrased questions or questions with new scenarios, these 50 were less prone to errors
463 as they didn’t necessitate significant changes to the mathematical structure. In fact, only 2 solutions
464 among them were incorrect. On the other hand, those that involved restructuring or tackled new topics
465 were more error-prone: 7 from restructuring and 4 from new topics were incorrect. Intriguingly, the
466 critic corrected 6 out of the 7 restructured ones, but only 1 out of the 4 new topics. This might suggest
467 that questions on new topics are more open-ended and might need multiple rounds of interactions for
468 accurate evaluation. Further refinement in this verification process will be the focus of our future
469 work.

470 Our analysis revealed that the critic excels in evaluation rather than generation. This observation is
471 exemplified in the quoted example, it can be seen that even when the model cannot directly answer the
472 question, it can effectively assess the correctness of the actor’s response. This discovery is intriguing,
473 as the actor’s response effectively serves as a ‘hint’. This observation aligns with prior works that
474 found models are better at evaluating answer (or improving an answer) than generating it from scratch
475 in various settings [6, 25, 23, 8, *inter alia*].

476 [Critic cannot solve question directly]

477 Input: A magician has 5 hats, each with a different number of balls inside. The first hat has 3 balls,
478 the second hat has twice as many balls as the first, the third hat has half as many balls as the second, the
479 fourth hat has 5 balls, and the fifth hat has 4 times as many balls as the fourth. How many balls are in the
480 magician’s hats altogether?

481 Output: There are $3 + 2 * 3 = 7$ balls in the first two hats. There are $5 - 2 = 3$ balls in the third hat. There
482 are $4 * 3 = 12$ balls in the fourth and fifth hats. In total, there are $7 + 3 + 12 = 22$ balls in the magician’s
483 hats. Answer: 22. ✗

484 [Critic can evaluate actor’s answer]

485 Input: Please review whether the answer for the following question is correct. Q: A magician has 5 hats,
486 each with a different number of balls inside. The first hat has 3 balls, the second hat has twice as many
487 balls as the first, the third hat has half as many balls as the second, the fourth hat has 5 balls, and the fifth
488 hat has 4 times as many balls as the fourth. How many balls are in the magician’s hats altogether? A: There
489 are $3 + 2 * 3/2 = 3 + 3 = 6$ balls in the first two hats. There are $5 - 3/2 = 5 - 1.5 = 3.5$ balls in the third
490 hat. There are $4 * 5 = 20$ balls in the fourth hat. In total, there are $6 + 3.5 + 20 = 29.5$ balls. Since you
491 can’t have a fraction of a ball, you would need to round up to the nearest whole number of balls. Therefore,
492 there are 29 balls in the magician’s hats altogether.

493 Output: The magician has 5 hats, each with a different number of balls inside. The first hat has 3 balls,
494 the second hat has twice as many balls as the first, so it has $3 * 2 = 6$ balls. The third hat has half as many
495 balls as the second, so it has $6 / 2 = 3$ balls. The fourth hat has 5 balls. The fifth hat has 4 times as many
496 balls as the fourth, so it has $5 * 4 = 20$ balls. In total, there are $3 + 6 + 3 + 5 + 20 = 37$ balls. Since you
497 can’t have a fraction of a ball, you would need to round up to the nearest whole number of balls. Therefore,
498 there are 37 balls in the magician’s hats altogether. The answer provided is incorrect. ✓

499 C Experiment Details

500 We use a temperature of 0.7 for the actor during exploration as in prior work [8], and we sample
501 10 actor generations for every batch of samples from the replay buffer. We use a temperature
502 of 0.0 for the critic since we found that it performs best. Following prior work [45], we filter
503 out reasoning paths with incorrect answers or calculations—based on Python evaluation—for the
504 ‘paraphrasing’ and ‘new scenarios’ exploration categories. However, we do not apply this filter to
505 the ‘restructuring’ or ‘new topics’ exploration categories, as we do not have ground truth answers
506 for these two categories. The evaluations for all baselines and our approach are conducted with
507 deterministic sampling following prior work and report maj1@1 (accuracy) across all experiments.
508 We follow prior work by conducting evaluations using deterministic sampling for both our approach

509 and the baseline methods. We report maj1@1 accuracy across all experimental setups. All models
510 are trained with the same hyperparameters: global batch size = 128, learning rate = 2e-5, epochs = 3,
511 sequence length = 2048. The training is done with 8x A100 80GB GPUs.