
Exploration with Principles for Diverse AI Supervision

Anonymous Author(s)

Affiliation

Address

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 [39] using next token prediction has led to substantial AI
17 advancements, as evidenced by the groundbreaking results they have produced [32, 29].
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 [32]
20 along with a range of other models [6, 13, 8, *inter alia*], rely on fine-tuning through
21 human demonstrations, demanding significant human involvement and domain expertise.
22 This reliance on extensive human supervision presents a substantial chal-
23 lenge since human supervision requires domain expertise, is time con-
24 suming, and is tedious. Moreover, hu-
25 mans can struggle to provide reliable
26 supervision in highly specialized do-
27 mains. For instance, ChatGPT pos-
28 sesses a greater depth of knowledge
29 than the average human, which makes
30 it difficult to rely on humans to pro-
31 vide supervision for ChatGPT. More-
32 over, while our most advanced AI sys-
33 tems have made significant strides,
34 they still necessitate thorough, human-
35 guided processes to enhance their abil-
36 ity to answer factual or mathematical queries [19]. Yet, when it comes to more intricate and mission-
37
38

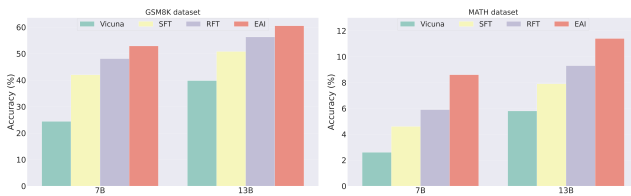


Figure 1: Test accuracy on mathematical reasoning benchmark GSM8K. Baselines include Vicuna, supervised finetuning Vicuna on training set (denoted as SFT), and supervised finetuning Vicuna on rejection sampled model generated diverse solutions on training set (denoted as RFT). Our Exploratory AI (EAI)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

critical tasks, such as navigating complex tax or law regulations, these challenges will demand even more specialized expertise and effort.

Prior works attempt to explore alternatives to human supervision, by using AI supervision instead. For example in mathematical reasoning, these studies propose sampling self generated solutions for human curated questions from large language models and employ techniques like rejection sampling, along with other techniques, to curate training data for the model [7, 28, 2, 15, 45, 42, *inter alia*]. While learning from such sampled content proves effective, a significant challenge persists: the sampled contents often lack the necessary diversity, resulting in a rapid saturation of the learning process [see e.g. 42, 45]. Moreover, the sampling approach has been confined to solutions exclusively, relying on human-curated questions, thus imposing constraints on the diversity of generated data.

To tackle these limitations, we propose a novel approach for using AI models to autonomously generate *diverse* data for learning purposes. This concept draws inspiration the APT algorithm [20] designed for unsupervised RL pretraining [35, 16, 31]. RL pretraining studies exploring in a reward-free environment to develop skills for quickly maximize various downstream rewards. APT allows training RL agent to learn skills by autonomously explore reward free environment based on evaluating novelty of encountered states using particle based entropy estimation [3, 34]. Adapting APT to large language models presents several challenges, including computational complexity and the difficulty of learning reward functions and exploration policies [12, 7]. Rather than relying on traditional RL techniques, we harness the unique capabilities of large language models, such as their ability to learn from context and follow instructions. In essence, we use them to perform the roles of both a reward function and an exploration policy. Our approach, which we term Exploratory AI (EAI), involves two key components: an actor and a critic. The actor is responsible for generating novel content in natural language, while the critic evaluates this generated content and provides critiques to guide the actor’s exploration. By evaluating the novelty of the generated contents, our method allows for effective exploration in the rich space of natural language. EAI can generate diverse data independently of human intervention. This makes it more scalable and automated, positioning it as a preferable alternative to methods like supervised finetuning or rejection sampling that depend on data curated by humans. Furthermore, EAI provides an interpretable window into the behavior and knowledge of the model. It sheds light on how well the model possesses knowledge and its reasoning behind generating novel questions. One can look at generations and their corresponding evaluations which provide valuable insights about how model generates and evaluates.

We evaluate our approach on mathematical reasoning benchmarks GSM8K [7] and MATH [14], EAI substantially improves performance on challenging reasoning tasks, outperforming both human supervision and AI supervision baselines. In contrast to human supervision, our approach is autonomous and more scalable. When compared to prior state-of-the-art AI supervision baselines including RFT [42] and WizardMath [23], our method provides a straightforward yet highly effective solution for the generation of high-quality and diverse data.

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

2 Exploratory AI for Diverse AI Supervision

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

Our method employs a dynamic interplay between generation and evaluation. This concept draws inspiration from unsupervised RL pretraining (URL) [16] and particularly the APT algorithm [20]. RL pretraining studies exploring in a reward-free environment to develop skills for quickly maximizing various downstream rewards. APT allows training RL agent to learn skills by autonomously exploring a reward free environment based on evaluating novelty of encountered states using particle based entropy estimation [3, 34].

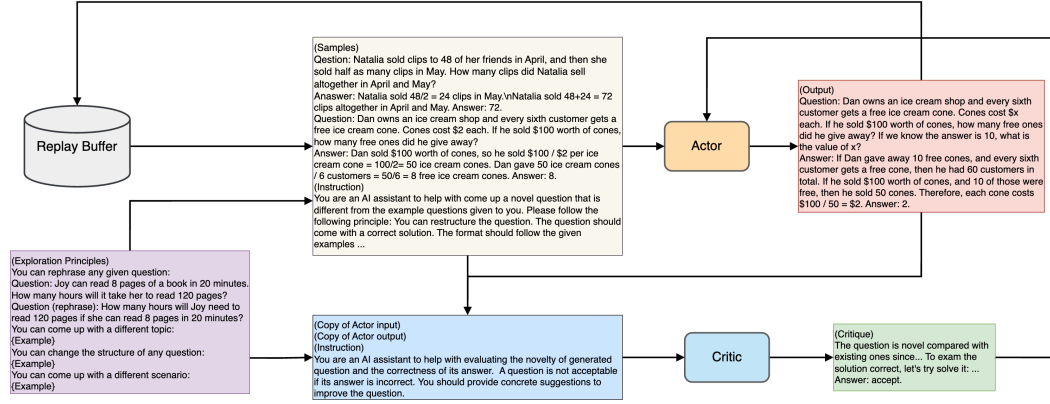


Figure 2: Generating diverse data in the *Exploratory AI Framework*. In the diagram, we demonstrate how the actor generates diverse content by conditioning on samples from the replay buffer. The replay buffer can be initialized with a pre-existing human-created dataset (e.g., GSM8K training set) or can remain empty for starting from scratch with zero-shot exploration. 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.

Adapting APT directly to large language models presents several challenges, including grappling with computational complexity and the difficulty of learning reward functions and exploration policies [12, 7]. In response, we propose Exploratory AI (EAI), a novel approach that circumvents the need for direct reinforcement learning (RL) by harnessing the power of large language models for exploration. Our method explore the natural language space by employing these language models to assess the novelty of generated content and guide the exploration process. Our approach consists of two key components: an “actor” responsible for generating novel content and a “critic” that evaluates the actor’s outputs and provides feedback to guide further content generation.

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

We equip both the actor and critic with a curated set of guiding principles to facilitate the generation and evaluation of diverse questions. 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. While it’s theoretically possible to provide all these principles to the model, in this study, we opt to a more controlled approach. To balance the quantity of generated data for each principle, we uniformly sample one principle at a time and input it to both the actor and critic. The actor is instructed to follow the principle (e.g., restructuring the question) during question generation. Similarly, the critic’s role is to evaluate the diversity, considering the selected principle. It’s worth noting that the critic’s principle is worded slightly differently from the exploration principle; for a detailed list, please refer to Appendix E. Our method is shown in Figure 2 and the algorithm is 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}

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}**

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?

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

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

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

References

- [1] Anthropic. Introducing claude, 2023. URL <https://www.anthropic.com/index/introducing-claude>.
- [2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [3] J Beirlant. Nonparametric entropy estimation: An overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39, 1997.
- [4] Jeff Bilmes. Submodularity in machine learning and artificial intelligence. *arXiv preprint arXiv:2202.00132*, 2022.
- [5] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023.
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org>, 2023.
- [7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [8] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm, 2023. URL <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- [9] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.
- [10] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. *arXiv preprint arXiv:2302.06692*, 2023.
- [11] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- [12] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.
- [13] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. *Blog post, April, 1, 2023*.
- [14] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [15] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.
- [16] Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*, 2021.
- [17] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

- [18] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.
- [19] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- [20] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems*, 2021.
- [21] Hao Liu, Tom Zahavy, Volodymyr Mnih, and Satinder Singh. Palm up: Playing in the latent manifold for unsupervised pretraining. *Advances in Neural Information Processing Systems*, 35: 35880–35893, 2022.
- [22] Hao Liu, Carlo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 2023.
- [23] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.
- [24] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.
- [25] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.
- [26] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL <http://probml.github.io/book2>.
- [27] Mirco Mutti, Lorenzo Pratissoli, and Marcello Restelli. A policy gradient method for task-agnostic exploration. *arXiv preprint arXiv:2007.04640*, 2020.
- [28] Ansong Ni, Jeevana Priya Inala, Chenglong Wang, Alex Polozov, Christopher Meek, Dragomir Radev, and Jianfeng Gao. Learning math reasoning from self-sampled correct and partially-correct solutions. In *The Eleventh International Conference on Learning Representations*, 2022.
- [29] OpenAI. Gpt-4 technical report, 2023.
- [30] Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. *arXiv preprint arXiv:2302.05103*, 2023.
- [31] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.
- [32] J. Schulman, B. Zoph, C. Kim, J. Hilton, J. Menick, J. Weng, J. F. C. Uribe, L. Fedus, L. Metz, M. Pokorný, R. G. Lopes, S. Zhao, A. Vijayvergiya, E. Sigler, A. Perelman, C. Voss, M. Heaton, J. Parish, D. Cummings, R. Nayak, V. Balcom, D. Schnurr, T. Kaftan, C. Hallacy, N. Turley, N. Deutsch, and V. Goel. Chatgpt: Optimizing language models for dialogue. *OpenAI Blog*, 2022. URL <https://openai.com/blog/chatgpt>.
- [33] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.
- [34] Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4):301–321, 2003.

- 232 [35] Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated
233 reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental*
234 *Development*, 2(2):70–82, 2010.
- 235 [36] Zhiqing Sun, Yikang Shen, Qinzhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming
236 Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with
237 minimal human supervision. *arXiv preprint arXiv:2305.03047*, 2023.
- 238 [37] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy
239 Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model,
240 2023.
- 241 [38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei,
242 Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open
243 foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- 244 [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
245 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information*
246 *processing systems*, 30, 2017.
- 247 [40] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan,
248 and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models.
249 *arXiv preprint arXiv:2305.16291*, 2023.
- 250 [41] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi,
251 and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instruc-
252 tions. *arXiv preprint arXiv:2212.10560*, 2022.
- 253 [42] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou.
254 Scaling relationship on learning mathematical reasoning with large language models. *arXiv*
255 *preprint arXiv:2308.01825*, 2023.
- 256 [43] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf:
257 Rank responses to align language models with human feedback without tears. *arXiv preprint*
258 *arXiv:2304.05302*, 2023.
- 259 [44] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao
260 Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv*
261 *preprint arXiv:2309.05653*, 2023.
- 262 [45] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with
263 reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

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 $j = 1$ **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

264 A Setting

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

267 **Benchmarks.** We evaluate our method on the mathematical reasoning tasks GSM8K. This bench-
268 marks exams model’s mathematical reasoning capabilities, we finetune model on the training split,
269 and evaluate model on the test split. The GSM8k dataset includes around 7,500 training and 1,319
270 test math problems for high school-level students, involving basic arithmetic operations. Problems
271 typically require 2 to 8 steps for a solution. The MATH dataset comprises 7,500 training and 5,000
272 challenging test problems from prestigious math competitions (AMC 10, AMC 12, AIME) covering
273 various academic domains, including prealgebra, algebra, number theory, counting and probability,
274 geometry, intermediate algebra, and precalculus.

275 **Baselines.** We compare our approach with (a) Base model including Vicuna 7B, 13B, and 30B [6].
276 Vicuna is LLaMA2 finetuned on user conversations shared online (ShareGPT). We use Vicuna as
277 base model for all baselines and our method; (b) Supervised finetuning (SFT) on training set of the
278 original GSM8K or MATH, in which a language model is finetuned on human written exemplars of
279 questions–answers pairs. SFT has been widely used in prior works for improving language models
280 mathematical reasoning [18, 38, 29, *inter alia*] and following user intention [13, 8, *inter alia*]. We
281 also compare with WizardMath [23] which does SFT on ChatGPT annotated questions and solutions,
282 as well as MAMmoTH [44] which uses GPT4 annotated solutions; (c) Rejection sampling finetuning
283 (RFT) [42] which applies supervised finetuning on rejection sampled model generated data. We
284 provide baseline scores for SFT and RFT from both their original papers and our implementations
285 using Vicuna, ensuring a fair and comprehensive comparison; (d) Proprietary models including
286 GPT-4 [29], ChatGPT [32], and Claude2 [1].

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

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)

B Results

Benchmark. In Table 1, we present the results of pass@1 (%) on GSM8K and MATH, showcasing the performance of various models across different supervision types and model sizes. Our approach, denoted as Vicuna + EAI, is compared with Vicuna and Vicuna + RFT, which serve as the closest baselines. Notably, our method outperforms Vicuna in all scenarios. For instance, in the 7B parameter setting, our model achieves 52.9% and 9.5% on GSM8K and MATH, respectively, surpassing Vicuna’s scores of 24.4% and 2.6%. This trend continues across larger model sizes and different datasets. Furthermore, compared to Vicuna + RFT, our approach maintains a consistent advantage, demonstrating the effectiveness of our self AI supervision method in improving model performance on these tasks. These results highlight the significant improvements our method brings to the table and its competitive edge in natural language understanding and generation tasks.

Comparison of diversity. We evaluate EAI in terms of the diversity of generated data. We compare RFT and EAI in terms of the submodularity diversity gain [4, 26]. This metric serves as an indicator of the extent to which the generated data contribute to the overall diver-

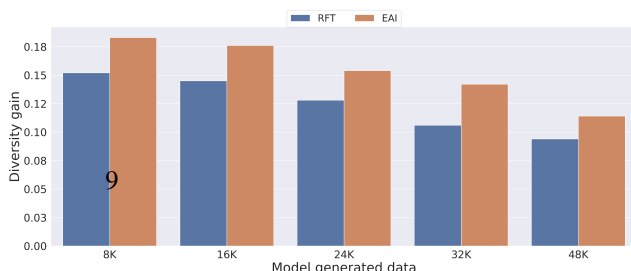


Figure 2: Comparison of diversity gain between RFT and EAI.

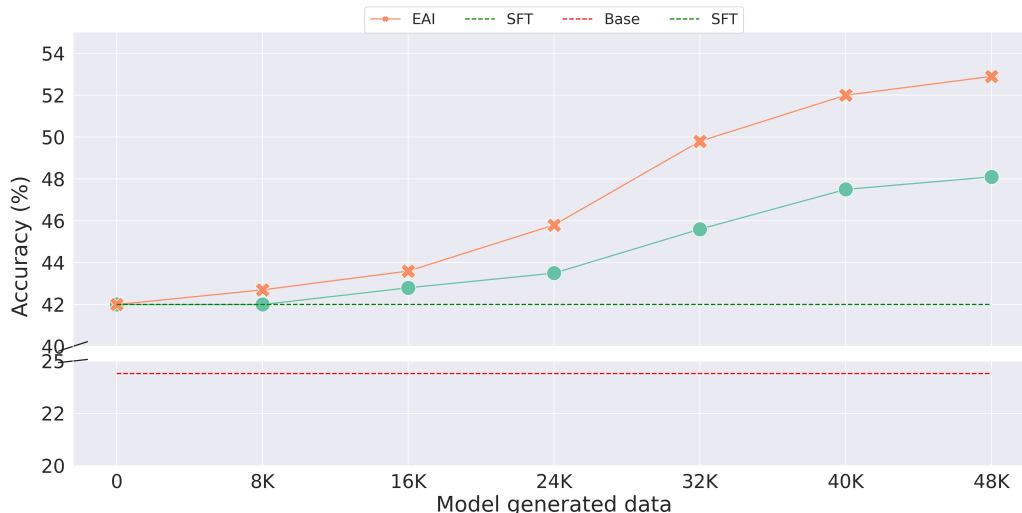


Figure 4: Data scaling on GSM8K. Shown are GSM8K accuracy with different amount of generated data. EAI generates high quality data for learning and scales well with data.

sity of the dataset. A higher diversity gain suggests that the newly generated questions exhibit greater dissimilarity from the existing dataset. We measure the gain over GSM8K training set by varying the amount of generated content. We use OpenAI GPT embedding `text-embedding-ada-002` to encode the data. The results depicted in Figure 3 clearly demonstrate that EAI consistently outperforms RFT in terms of diversity, thereby providing a more diverse set of generated data

Effect of sampled inputs. The Table 2 presents the results of an experiment examining the impact of varying the number of samples on GSM8K and MATH. As the number of samples increases from 0 to 8, we observe a steady incremental improvement on both GSM8K and MATH. On GSM8K, the performance rises from 50.1 to 52.9. On MATH, the effect is more pronounced. These results suggest that increasing the number of samples has a positive effect on both GSM8K and MATH, highlighting the significance of 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

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

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

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

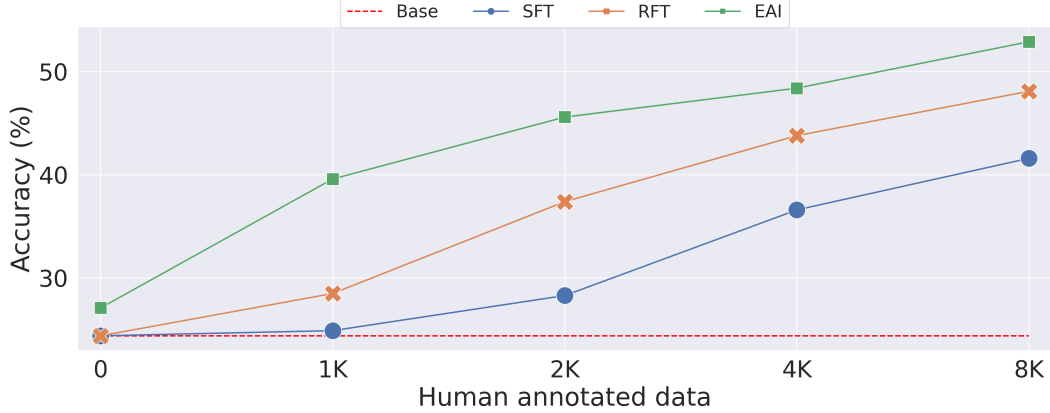


Figure 5: Performance on GSM8K with different amount of human annotated data. EAI performs well even without human annotation and scales well with more human provided annotations.

the ability to rephrase and generate diverse content is crucial. Similarly, the omission of "restructure" leads to a significant drop in MATH scores, highlighting the significance of introducing novel question-structuring approaches for solving more challenging problems.

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

C Related Work

Transformers [39] trained using next token prediction have gave rise to many state-of-the-art AI systems [32, 29]. The remarkable AI results achieved with this generative AI approach heavily hinge upon the availability of diverse and high-quality data. For instance, state-of-the-art AI models including ChatGPT [32] and GPT4 [29] along with a range of other open source models such as Vicuna, Koala, and Dolly [8, 13, 6, *inter alia*], require extensive finetuning through human demonstrations. This process involves human conversations with ChatGPT or written demonstrations, demanding significant human involvement and domain expertise. Previous research has explored various avenues to enhance performance and sample efficiency, as well as alternative sources of supervision. To align with human preferences, there has been active research into developing simple algorithms for learning from human preferences [22, 43, 9, 38, *inter alia*]. In contrast to human demonstrations or feedback, another line of work explores the utilization of environmental feedback, such as unit test errors [17, 5, 33], which has demonstrated improved results in coding tasks. Some works explore using LLMs to provide AI supervision based exploration techniques for applications in solving games [10, 21, 40] and demonstrate improved results. Furthermore, some prior research leveraged proprietary APIs to indirectly obtain high-quality human data, enhancing

model capabilities in areas like instruction following [41, 37, 13, 6, *inter alia*] and mathematical reasoning [23, 25, 44, *inter alia*]. Another line of research explores the use of models to supervise themselves [36, 24, 15, 2, 42], yielding improved results in reasoning tasks and alignment with human preferences. Our work focuses on generating diverse and high-quality data using AI models and we demonstrate applying our proposed approach to enhance open-source models by having them self-generate learning data. Our approach’s exploration technique is related to unsupervised RL based exploration [35, 16, 20, 31, 27, 11, 30, *inter alia*], however, our method does not require training RL agent. Additionally, some works have delved into more detailed forms of human supervision [19], demonstrating that LLMs benefit more from step-by-step process-based supervision than sparse outcome-based supervision. Our research uniquely centers on the data dimension, with a specific emphasis on harnessing AI models to generate diverse high -quality AI supervision. To this end, we introduce an actor-critic based approach for automating the exploration process.

D Conclusion

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

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

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

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

E Prompt

The critic's principle is based on the exploration principle but is worded slightly differently from the actor's exploration principle, in order to tell critic what kind of new questions are diverse and acceptable. The critic will use the principle to evaluate diversity of questions and evaluate the 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?

F Case Study of EAI

We conduct a qualitative study how actor and critic accomplish exploration. On GSM8K, we manually inspect 100 examples randomly sampled from EAI explored dataset to better understand how does the exploration work. We found that critic is good at evaluating the novelty of generated questions, and can give concrete suggestions to help revising. The actor can follow critic's feedback to improve the generated question. A representative example is shown in the quoted text below, where the critic gives revise suggestions to make the generated question substantially different, by suggesting to introduce mathematical concept of volumes and use a different scenario.

[Critic guides exploration]

(First round)

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

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

✗

(Second round)

Actor's generation: An artist is creating a sculpture using various geometric shapes. The sculpture consists of a cube, a sphere, a cylinder, and a pyramid. The cube has a side length of 6 units, the sphere has 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 base edge length of 5 units and a height of 7 units. Calculate the total volume of the artist's sculpture. A: omitted

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

445 We also conducted a study on how effective is critic at evaluating answer’s correctness. Among
446 the 100 critiques assessed, the critic successfully identified the correctness of 94 responses. Upon
447 inspection, we determined that out of the 94 instances in which the critic’s judgment was accurate, a
448 significant majority, specifically 76 cases, were of this nature. Our analysis revealed that the critic
449 excels in evaluation rather than generation. This observation is exemplified in the quoted example, it
450 can be seen that even when the model cannot directly answer the question, it can effectively assess
451 the correctness of the actor’s response. This discovery is intriguing, as the actor’s response effectively
452 serves as a ‘hint’. This observation aligns with prior works that found models are better at evaluating
453 answer (or improving an answer) than generating it from scratch in various settings [5, 24, 7, *inter*
454 *alia*].

455 [Critic cannot solve question directly]

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

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

463 [Critic can evaluate actor’s answer]

464 Input: Please review whether the answer for the following question is correct. Q: A magician has 5 hats,
465 each with a different number of balls inside. The first hat has 3 balls, the second hat has twice as many
466 balls as the first, the third hat has half as many balls as the second, the fourth hat has 5 balls, and the fifth
467 hat has 4 times as many balls as the fourth. How many balls are in the magician’s hats altogether? A: There
468 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
469 hat. There are $4 * 5 = 20$ balls in the fourth hat. In total, there are $6 + 3.5 + 20 = 29.5$ balls. Since you
470 can’t have a fraction of a ball, you would need to round up to the nearest whole number of balls. Therefore,
471 there are 29 balls in the magician’s hats altogether.

472 Output: The magician has 5 hats, each with a different number of balls inside. The first hat has 3 balls,
473 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
474 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
475 balls as the fourth, so it has $5 * 4 = 20$ balls. In total, there are $3 + 6 + 3 + 5 + 20 = 37$ balls. Since you
476 can’t have a fraction of a ball, you would need to round up to the nearest whole number of balls. Therefore,
477 there are 37 balls in the magician’s hats altogether. The answer provided is incorrect. ✔

478 G Experiment Details

479 We use a temperature of 0.7 for the actor during exploration, following Cobbe et al. [7], and we sample
480 10 actor generations for every batch of samples from the replay buffer. We use a temperature of 0.0
481 for the critic since we found that it performs best. The evaluations for all baselines and our approach
482 are conducted with deterministic sampling following prior work and report `maj1@1` (accuracy) across
483 all experiments. We follow prior work by conducting evaluations using deterministic sampling for
484 both our approach and the baseline methods. We report `maj1@1` accuracy across all experimental
485 setups. All models are trained with the same hyperparameters: global batch size = 128, learning rate
486 = $2e-5$, epochs = 3, sequence length = 2048. The training is done with 8x A100 80GB GPUs.