# Distilling LLMs' Decomposition Abilities into Compact Language Models

*Denis Tarasov, Kumar Shridhar*

ETH zürich

ICML International Conference On Machine Learning

**ABSTRACT** Large Language Models (LLMs) have demonstrated proficiency in their reasoning abilities, yet their large size presents scalability challenges and limits any further customization. In contrast, compact models offer customized training but often fall short in solving complex reasoning tasks. This study focuses on distilling the LLMs' decomposition skills into compact models using offline reinforcement learning. We leverage the advancements in the LLM`s capabilities to provide feedback and generate a specialized task-specific dataset for training compact models. The development of an AI generated dataset and the establishment of baselines constitute the primary contributions of our work, underscoring the potential of compact models in replicating complex problem-solving skills.

## MOTIVATION

distilling distinct components of the reasoning process into smaller models emerges as a promising avenue for research [1]. Decomposition, particularly in the context of teaching smaller models, proves advantageous due to their cost-effectiveness, reduced computational requirements, and accessibility. Reinforcement Learning with Human Feedback (RLHF) [2] is one of the most popular methods for solving NLP tasks, however it requires extensive interactions with the environment and does not use data directly. Offline RL serves as a promising direction for NLP problems while being heavily under-explored.

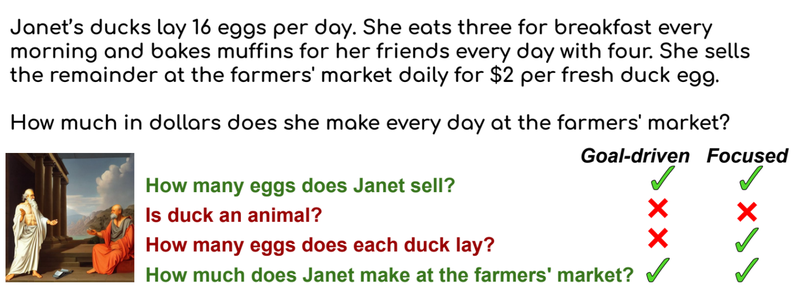


Fig 1. Demonstration of sub-questioning problem from [1].

## CONTRIBUTIONS

1) An AI-generated benchmark where math questions are broken down into simpler sub-questions based on the GSM8K [3].
2) Training smaller LMs for the same task using fine-tuning and offline RL techniques and providing baselines.
3) Exploring the potential benefits of using AI-generated feedback on its own responses in enhancing model performance for the reasoning problems. Analogous approach was successfull in improving LLMs [4].

## RESULTS

This work introduces a novel AI-generated benchmark tailored for evaluating sub-questioning in reasoning tasks. The outcomes reveal a significant performance gap between the best-performing approach and ChatGPT. The underwhelming performance of the offline RL approach underscores the need for further advancements in this domain, see **Table 2**.

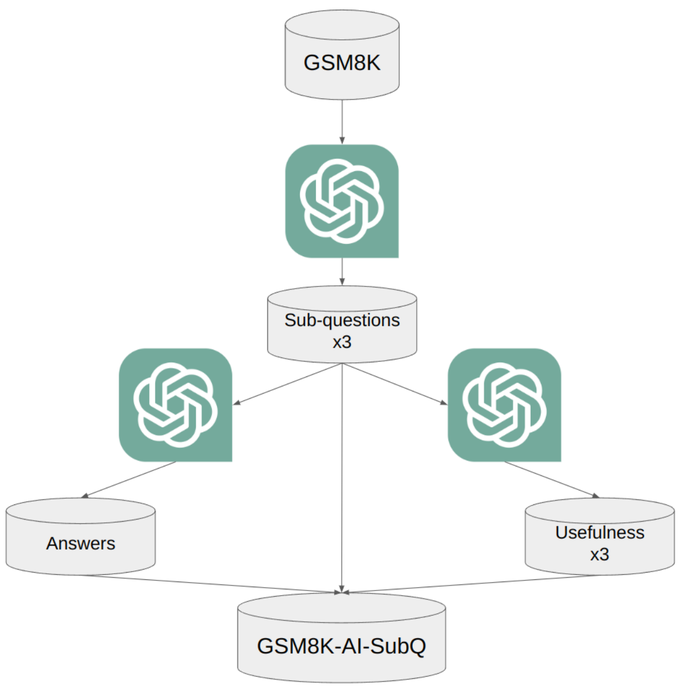


Fig 2. Dataset collection process.

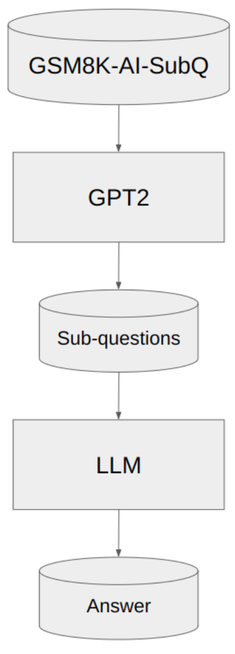


Fig 3. Experimental pipeline.

## METHODOLOGY

Our study leverages the widely adopted GPT-2 architecture, specifically selecting models of various sizes to tailor our experiments. For the overview of dataset collection process see **Fig. 2**. Behavioral Cloning (BC) and Filtered BC are employed as supervised baselines.

We use ILQL [5] as a powerful offline RL baseline and cast the text generation problem as a token-level POMDP. The agent's observations correspond to prefixes of tokens, and the agent's action pertains to the selection of the next token to be generated. ILQL objective is shown in **Eq. 1**

$$L_{Q,V}(\theta) = \mathbf{E}_{\tau\sim D}\left[\sum_{i=0}^{T}(R(h_i,a_i)+\gamma V_\theta(h_{i+1})-Q_\theta(h_i,a_i))^2 + L_2^\tau(Q_{\hat{\theta}}(h_i,a_i)-V_\theta(h_i))\right] \quad \textbf{(Eq. 1)}$$

**Sparse reward function:** agent gets reward of +1 if the answer to the problem was correct.

**Full reward function:** agent gets reward of +1 if the answer to the problem was correct and usefulness score for each sub-question.

For the evaluation of the quality of the generated sub-questions we use different LLMs by providing them with a prbolem and sub-questions and verifying the correctness of the final answer. ChatGPT, LLaMA (7B/13B) and Mistral 7B were used for this purpose. See **Fig. 3** for the pipeline overview.

| Metric | 0 correct | 1 correct | 2 correct | 3 correct |
|---|---|---|---|---|
| Number of problems | 1343 (18%) | 866 (11%) | 1139 (15%) | 4052 (54%) |
| Mean problem length | 269.4 ± 106.7 | 252.1 ± 100.3 | 240.2 ± 94.7 | 217.5 ± 82.9 |
| Median problem length | 250 | 235.5 | 225 | 201.5 |

Table 1. Statistics on ChatGPT abilities to solve problems with it's own sub-questions. Longer problems are harder to solve.

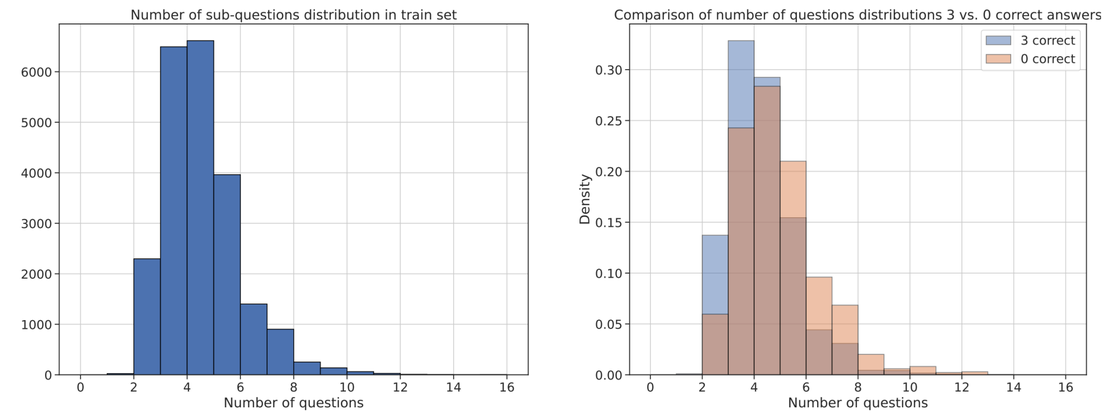


Fig 4. Number of sub-questions distribution and comparison between hard and easy problems. More questions are associated with lower performance.

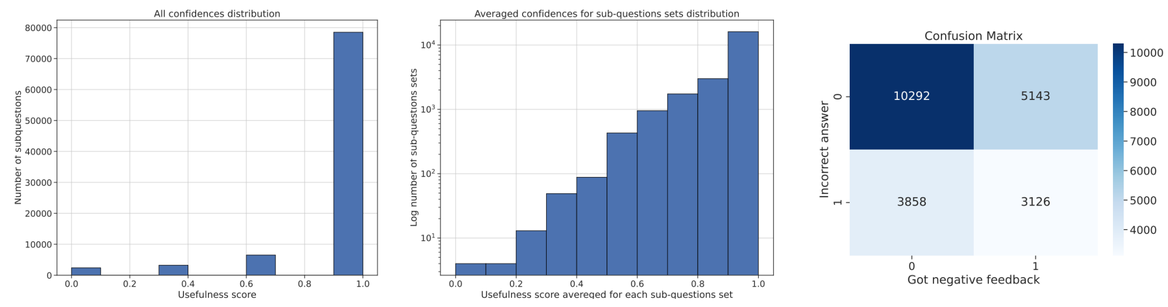


Fig 5. Self-feedback statistics. ChatGPT usually very confident that it's sub-questions are useful for solving the problem. Confusion matrix demonstrates that usfulness score is very noisy, however it contatins useful signal.

| Algorithm | DistillGPT | GPT-2 small | GPT-2 medium | Average |
|---|---|---|---|---|
| BC | 0.255 | 0.284 | 0.310 | 0.283 |
| Filtered BC | **0.260** | **0.293** | **0.319** | **0.291** |
| ILQL-sparse | 0.249 | 0.281 | 0.310 | 0.280 |
| ILQL-full | 0.253 | 0.277 | 0.309 | 0.280 |
| ChatGPT | N/A | N/A | N/A | 0.429 |

Table 2. Experimental results for different GPT sizes. Averaged over all LLM answerers.

arXiv GitHub

## REFERENCES

[1] Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. Automatic generation of socratic subquestions for teaching math word problems. arXiv preprint arXiv:2211.12835, 2022.
[2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35: 27730–27744, 2022.
[3] 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.
[4] 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, 2022b.
[5] Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning. arXiv preprint arXiv:2206.11871, 2022.