Automatically Generating Numerous Context-Driven SFT Data for LLMs Across Diverse Granularity

Shanghaoran Quan

School of Computer Science and Engineering, Beihang University
Beijing, Haidian, 100191, China
shrquan@buaa.edu.cn

Abstract

Constructing high-quality query-response pairs from custom corpora is crucial for supervised fine-tuning (SFT) large language models (LLMs) in many applications, like creating vertical-domain AI assistants or roleplaying agents. However, sourcing this data through human annotation is costly, and existing automated methods often fail to capture the diverse range of contextual granularity and tend to produce homogeneous data. To tackle these issues, we introduce a novel method named AUGCON, capable of **au**tomatically **g**enerating **con**text-driven SFT data across multiple levels of granularity with high diversity, quality and fidelity. AUGCON begins by generating queries using the Context-Split-Tree (CST), an innovative approach for recursively deriving queries and splitting context to cover full granularity. Then, we train a scorer through contrastive learning to collaborate with CST to rank and refine queries. Finally, a synergistic integration of self-alignment and self-improving is introduced to obtain high-fidelity responses.

Extensive experiments are conducted incorporating both automatic and human evaluations, encompassing four widely-used benchmarks and a test scenario in English and Chinese. The results highlight the significant advantages of AUGCON in producing high diversity, quality, and fidelity SFT data against several state-of-the-art methods.

Code — https://github.com/quanshr/AugCon

1 Introduction

With the rise of impressive capabilities of large language models (LLMs), a variety of vertical-domain LLM-based AI assistants have been introduced (Cheng, Huang, and Wei 2023; Chen et al. 2023; Luo et al. 2023). By incorporating specialized knowledge into LLMs, these custom models have been shown to outperform their general-purpose counterparts in their respective areas. These models can be developed through two strategies: building them from scratch (Yang et al. 2023; Liu et al. 2023d) or adapting existing general LLMs through supervised fine-tuning (SFT) (Shi et al. 2023; Zaiem et al. 2023), with the latter approach often favored for its efficiency and the foundational advantages offered by the general LLMs (Jiang et al. 2024; Dong et al. 2023; Cheng, Huang, and Wei 2023).

Directly supervised fine-tuning on the raw, custom corpora, also known as domain-adaptive pre-training (DAPT) (Gururangan et al. 2020), has proven beneficial (Bayer et al. 2024; Krieger et al. 2022) but revealed to be insufficient and may impair prompting ability on domain-specific tasks (Liu et al. 2023e; Pal et al. 2024). To better leverage the privatized knowledge and customize the outputs of LLMs, supervised fine-tuning using custom query-response pairs has become common practice (Shaikh et al. 2024; Chang, Peng, and Chen 2023). However, sourcing these pairs through human annotation is very costly and can't generate at scale. Recent studies have explored automated methods for creating these pairs from custom corpora. AdaptLLM (Cheng, Huang, and Wei 2023), for instance, has used regex-based patterns to generate query-response pairs, but this approach tends to produce a limited variety of SFT data, which may not significantly enhance prompting capabilities and risks overfitting due to the narrow range of query types predefined. ETRC (Jiang et al. 2024) and Context-Instruct (Wang et al. 2023b) improved this by employing delicately designed prompts to generate queries from context using an LLM. However, those existing methods using the same workflow repeatedly on the same context tend to produce redundant queries without adequately covering the entire context at various levels of granularity. To automatically construct synthetic custom SFT data incorporating a wide range of contextual **granularity** (queries range from detailed questions to macro topics) with high **diversity** (queries need to be diversified to cover as much as possible the provided corpus), **quality** (responses are correct and efficient in answering the queries), and **fidelity** (data needs to follow human values and conform to predetermined tone and formats) still remain challenges.

To address these challenges, we propose AUGCON, which automatically generates multi-granularity context-driven SFT data for LLMs at scale with high diversity, quality, and fidelity. AUGCON performs the following three essential steps:

1. **Recursively Deriving Queries via Context-Split-Tree:** Considering that it is challenging for predetermined prompts to generate non-repetitive queries with broad granularity from the same context, we propose a novel method called Context-Split-Tree (CST). Starting from a context (which is a continuous text chunk extracted from

the corpus), we use an LLM to derive a query from it. At the same time, we ask the LLM to split this context into two contexts that are as independent as possible. Each context will recursively continue to derive queries and splits until it cannot be further divided. At the end, we will obtain a binary tree rooted in the initial context, and each node represents a context and contains a query that matches the granularity of it.

2. **Contrastive-Learning-Based Scorer to Rank Queries and Filtering:** To further ensure the quality and diversity of the queries, we use contrastive learning to train a scorer to evaluate the query by taking the obtained queries as positive examples and manipulating the prompt in Step 1 (*e.g.*, by using suboptimal instruction or attaching fewer few-shot examples) to generate negative examples. Then, we sort the derived queries under the same context using the scorer and only retain the queries that get high scores and the diversity evaluated by ROUGE-L reaching a specific threshold. To ensure high quality and high diversity of queries while reaching the certain quantity requirements, the filtering stage will be iterated with CST until the requirements are met.
3. **Obtaining High-Fidelity Responses:** Inspired by the significant impact principles (Sun et al. 2024, 2023) have on LLMs, we employ a principle-driven self-alignment approach to guide the LLM in producing high-fidelity responses to filtered queries and their respective contexts. To enhance the quality of the generated answers further, we apply random search and conduct the LLM to self-evaluate its responses and discover the best in-context learning (ICL) examples from those annotated by humans. Ultimately, all context, ICL examples, and principles are discarded, leaving only the query-response pairs to supervised fine-tune the LLM.

The entire process only requires a handful of few-shot CST examples, alignment principles, and query response examples. We can also achieve impressive results by just utilizing the open-source model, which will later be fine-tuned with synthetic data, eliminating the necessity of distilling more powerful LLMs like ChatGPT.

To assess the efficacy of our approach, we meticulously construct a test scenario and carefully assemble a dataset consisting of high-quality Chinese magazine articles centered around daily topics, along with corresponding test queries. Human evaluation demonstrates that our method excels in generating queries of superior quality and in enhancing the performance of fine-tuned models. Additionally, automatic evaluations conducted on four popularly used English benchmarks with relevant metrics further highlight the significant advantages our method holds in capturing contextual knowledge when compared to other state-of-the-art context-driven SFT data generation approaches.

Specifically, the contributions of our work lie on:

- We propose AUGCON, which can automatically generate multi-granularity context-driven SFT data from the corpus for LLMs at scale with high diversity, quality, and fidelity, providing the solution to a realistic industrial and academic problem worth studying.
- Our ideas of deriving queries via CST, training the scorer using contrastive learning to collaborate with the generation process to refine data, and synergistic integrating self-alignment and self-improving to obtain high-fidelity responses, are very novel and may inspire further works.
- Extensive experiments incorporating both automatic and human evaluations, encompassing four widely-used benchmarks and a test scenario in English and Chinese compared with other state-of-the-art methods demonstrate the effectiveness and advantages of AUGCON.
- To boost the academy and for others to generate high-diversity SFT data on their own corpus without effort, we open-source all of our code, dataset, and fine-tuned model at: https://github.com/quanshr/AugCon.

2 Related Work

Synthetic Data for Language Models Due to the challenges of data scarcity (Babbar and Schölkopf 2019), privacy concerns (Abay et al. 2019), and the sheer cost of data collection and annotation (Gilardi, Alizadeh, and Kubli 2023), synthetic data has emerged as a promising solution to build large, diverse, and high-quality datasets at scale (Liu et al. 2024b). One benefit of synthetic data is it can be tailored to specific requirements (Cheng, Huang, and Wei 2023; Jiang et al. 2024), with practical applications having been employed in various domains. WizardMath (Luo et al. 2023) leverages a series of operations to increase the complexity of questions and answers using GPT-3.5, while Reflexion (Shinn et al. 2024) employs external or internally simulated linguistic feedback to improve the code reasoning capabilities of language models. Similarly, Toolformer (Schick et al. 2024) learns to decide which APIs to call and what arguments to pass by training on template-generated data. In addition, synthesized data has been proven effective in mitigating hallucination (Wei et al. 2023) and aligning with shared human preferences and values (Bai et al. 2022). While the generation of context-driven synthetic data has proven to be a powerful substitute for manual annotation, the challenge of ensuring high-quality synthetic data, which encompasses the complexity of queries (Liu et al. 2023a), the diversity of semantics (Ding et al. 2023), and the scale of the synthetic datasets (Yuan et al. 2023; Li et al. 2023), has been a consistent pursuit.

Context-Driven Synthetic Data Numerous studies have developed techniques for creating synthetic data informed by contextual cues (Shaikh et al. 2024; Chang, Peng, and Chen 2023; Liu et al. 2023e; Pal et al. 2024). UltraChat (Ding et al. 2023) leverages user-specified topics and supplements these with existing textual material to craft instructional conversations aimed at enhancing chatbot performance. SPIN (Chen et al. 2024), on the other hand, autonomously generates training data from its previous iterations, employing this approach to progressively refine its capabilities. RECOST (Zhang et al. 2024) selects top-tier instructional content by incorporating external knowledge to assess synthesized examples using an in-context relative predictive entropy measure. Additionally, various methods have been devised to extract character profiles and personas from

collected books or scripts for the purpose of producing role-playing dialogues (Shao et al. 2023), and several initiatives focus on mining domain-specific data from specialized corpora to construct domain-specific language models (Cheng, Huang, and Wei 2023). While alternative approaches employ retrieval augmented generation (RAG) (Ram et al. 2023; Borgeaud et al. 2022) or integrate auxiliary knowledge in vast context windows (Xiong et al. 2023; An et al. 2024), issues like entity susceptibility (Du et al. 2024), high inference computational demand (Liu et al. 2022; Hao et al. 2022), and alignment difficulties with formats and preferences (Qi et al. 2023; Mosbach et al. 2023) highlight the crucial role of context-driven SFT in effectively incorporating corpus knowledge internally.

3 Our Method: AUGCON

In this section, we delve into the details of our proposed AUGCON.

3.1 Preliminary

We have a raw custom corpus $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ with each context C_i represents a continuous text chunk extracted from corpus $\mathcal{C}$, the instruct prompt I_{CST} and few-shot examples E_{CST} for Context-Split-Tree and I_R and E_R for answering the queries, and several response principles $\mathcal{P}$ representing the human demands on responses when answering questions [1]. The E_R are context-query-response triplets and will follow the response principles, represented as $E_R \sim \mathcal{P}$.

Our task is to generate a specific number of SFT query-response pairs $\mathcal{D} = \{(q_{i,j}, r_{i,j})\}$ that each pair derives from either the whole or part of context C_i. The derived triplet (C, q, r) should also follow the response principles $\mathcal{P}$, and the generated $\mathcal{D}$ is expected to have high diversity, quality, and fidelity.

3.2 Recursively Deriving Queries via Context-Split-Tree

This step is to derive context-query pairs (C, q) from the given corpus $\mathcal{C}$. Previous approaches applied regex-based or predetermined prompts for query generation, which often led to queries that were relatively monotonous in structure and granularity. We believe that this type of approach did not fully exploit the context, leading to queries incapable of effectively provoking the model's capability to comprehend and differentiate between various levels of detail within the context, resulting in suboptimal outcomes.

To address this issue, we propose a very novel and effective method called Context-Split-Tree (CST), with the pseudocode shown in Algorithm 1. CST starts with an entire context C, with each attached with the instruct prompt I_{CST} and few-shot examples E_{CST} to call an LLM to generate a query q deriving from the entire context. At the same time, we ask the LLM to semantically divide the context into two child contexts C_1 and C_2, and the instruct prompt is designed with hints to let the LLM polish the two split contexts to make them as independent as possible and collectively encompass the entirety of the original context. Each child context will continue to recursively derive query and split until reaching a point where one of its split child context lengths is not less than itself or the length falls below a predetermined threshold λ. At this point, we consider it to have been split into the minimum granularity and cannot be further divided. Upon the completion of this recursive process, a binary tree structure is formed, with the initial context at the root, and each node representing a context along with its corresponding query tailored to its specific granularity. We collect data from all nodes as the outcome of this step.

Algorithm 1: Context Split Tree

Input: A corpus $\mathcal{C}$, CST prompt instruction I_{CST}, CST few-shot examples E_{CST}
Output: Query dataset $Data$ comprises of split context and derived query pairs

function CONTEXTSPLITTREE($C, Data$)
  if $len(C) < \lambda$ **then**
    return ▷ Below the minimum granularity
  end if
  Call LLM to get $C_1, C_2, q \leftarrow$ LLM($I_{\text{CST}}, E_{\text{CST}}, C$)
  Append (C, q) to $Data$
  if $len(C_1) \geq len(C)$ **or** $len(C_2) \geq len(C)$ **or** ROUGE-L[P] < 0.7 **then**
    return ▷ The signs of hallucinations
  end if
  CONTEXTSPLITTREE($C_1, Data$) ▷ Recursive calling
  CONTEXTSPLITTREE($C_2, Data$)
end function

Initialize $Data \leftarrow$ empty list
for each extracted context $C \in \mathcal{C}$ **do**
  CONTEXTSPLITTREE($C, Data$)
end for
return $Data$

The minimum length threshold λ and the initial context length l are like the lower bound and upper bound to control the granularity distribution of generated questions. One can easily adjust the overall average granularity of generated queries by adjusting the length threshold. Similarly, if we seek to address more global questions, we can do it by simply increasing the initial context length, as long as the model's context window permits. One beneficial property of CST is that the number of questions ultimately generated will maintain a linear relationship with the length of the initial text provided. This ensures that adjusting the length of the segmented contexts in the corpus does not lead to significant fluctuations in the total number of queries obtained, but rather merely shifts the distribution of query granularity. By employing CST, we can produce queries that span across different levels of details in the context, and these queries naturally have little redundancy or repetition, enabling more efficient use of the context information and stimulating the model's capability to comprehend and grasp the context in

[1] In this work, we use query-response and question-answer interchangeably.

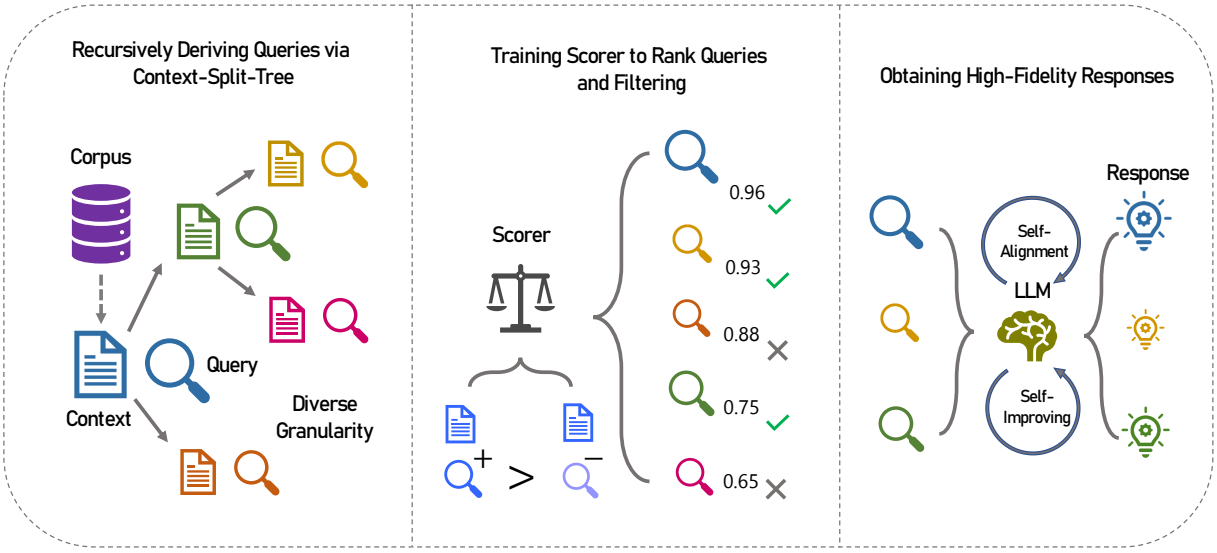

Figure 1: An overview of the proposed AUGCON.

different granularities. Moreover, another benefit of CST is that the derived queries just match the split context, making the later generated response to these queries more accurate and pertinent with less unrelated information.

3.3 Contrastive-Learning-Based Scorer to Rank Queries and Filtering

To further enhance the quality and diversity of the generated data, we introduce an effective ranking and filtering strategy collaborating with CST. Previous works have attempted to filter training data via heuristic algorithms, such as filtering out queries that are too long or too short (Wang et al. 2023a). Other works that are more relevant to us attempt to train scorers to judge the complexity and quality of question-response pairs (Liu et al. 2023a), but they need to have a step of distillation on stronger LLM APIs like ChatGPT, and their training methods are less effective. For example, they put a series of responses and ask for direct ranking, suffering from the positional bias (Liu et al. 2024a) in LLMs, or ask LLMs to directly assign a scalar score to a response, which is unstable. In this work, we apply contrastive learning to train a scorer to judge the degree of adherence to instruct prompt and few-shot examples, which is data-efficient and can achieve effective performance without the need for stronger LLMs.

The structure of our scorer is obtained by adding a linear head after the base model to map the last hidden state to a one-dimensional space. We take context-query pairs as inputs, applying scorer Sc to yield a scalar score $s = Sc(C, q)$. We use the context query pairs obtained from Step 1 as positive samples: $q^+ = \text{LLM}(I_{\text{CST}}, E_{\text{CST}}, C)$, and obtain negative samples by manipulating the instruct prompt (use suboptimal instructions): $q^- = \text{LLM}(I_{\text{CST}}{}^-, E_{\text{CST}}, C)$, few-shot examples (reduce ICL examples count): $q^- = \text{LLM}(I_{\text{CST}}, E_{\text{CST}}{}^-, C)$ or both of them: $q^- = \text{LLM}(I_{\text{CST}}{}^-, E_{\text{CST}}{}^-, C)$. Note that we do not generate all corresponding negative examples for positive data for training scorer, but rather randomly select a very small number of samples (*e.g.* only 500 pairs for each negative types in our implementation) to form the training set D_{Sc}. Then, the loss function of scorer can be represented as:

$$\mathcal{L} = -\mathbb{E}_{(C,q^+,q^-)\sim D_{Sc}}[\log(\sigma(Sc(C,q^+) - Sc(C,q^-)))] \quad (1)$$

We use the trained scorer applied on all the context query pairs obtained in Step 1 to get their scores. For each root context, we rank all queries from its CST in descending order of scores. Then, we start with an empty set and add one training query each time, only if the current query has a ROUGE-L precision score of less than 0.7 compared to any previously added queries. We will stop adding as the count reaches the limit. Each context will form such a set, and ultimately, we consolidate and retain the training data from all the sets. Through this approach, we can obtain diverse data and easily control the quantity, for it makes it possible to apply multi-times CST in the same context and filter the repeated one.

3.4 Obtaining High-Fidelity Responses

Inspired by the significant impact principles (Sun et al. 2024, 2023) have on LLMs, this principle-driven self-alignment step begins by appending the context and a set of helpful, ethical, and reliable principles to the LLM. These principles are meticulously crafted to ensure the LLM's outputs are closely aligned with human preferences or mimic certain response tones. Before initiating the response generation, we deploy a self-improving pipeline that makes the LLM self-evaluate its response and sift through the entire set of human-annotated Q&A pairs E_R, where random search is applied to find the most fitting few-shot examples to help

LLM generate high-fidelity responses under the predetermined principles, denoted as $E_R{}'$.

Our innovative synergistic integration of the principle-driven self-alignment with self-improving methodology effectively improves the fidelity of generated responses. Following this, we execute LLM($I_R, E_R{}', C, q$) to elicit each response r, ensuring that each response is not only in high quality but also in good alignment with the established principles. Notably, due to the precise matching of each query with its context's granularity within the CST framework, the LLM can effortlessly provide accurate and pertinent responses to the queries.

After obtaining all generated data, we prune all context, instruction, and response principles and only retain synthetic query response pairs as SFT data. This approach allows the fine-tuned LLM to potentially learn the methods and nuances of responding to queries in a manner that naturally aligns with human expectations, enabling the LLM to directly generate responses that are well-aligned with reliable principles and optimal ICL exemplars across a wide range of queries. It's important to note that the fine-tuned LLM can generate high-quality responses without the need to explicitly reference the principles set and ICL exemplars.

4 Evaluations

4.1 Baselines

To demonstrate the advantages of our method, we meticulously collect the following relevant baselines from a wide range of research. The set of contexts, base language models, and quantity of retained query-response pairs are maintained the same (if applicable) on both the baselines and our method to ensure a fair comparison.

(1) **Chat Model** (Bai et al. 2023; Touvron et al. 2023) applies instruction tuning and alignment tuning after pre-training. We utilize it both as the basic baseline and as the base model for calling and fine-tuning across all other baselines and our methods for fair comparison.

(2) **DAPT** (Gururangan et al. 2020) continuously pre-trains directly on the raw custom corpus to adapt and grasp domain-specific knowledge.

(3) **AdaptLLM** (Cheng, Huang, and Wei 2023) builds SFT samples by converting the raw corpora into reading comprehension tasks via regex-based mining patterns. Tasks they design include summarization, word-to-text, natural language inference, commonsense reasoning, and paragraph detection.

(4) **ETRC** (Jiang et al. 2024) derives question-answer pairs from extracted contexts with an LLM and augments data by ensembling contexts and their corresponding question-answer pairs with a length-based clustering algorithm.

(5) **Context-Instruct** (Wang et al. 2023b) is a context-driven instruction generation method that contains three parts: 1) partition text into manageable segments, 2) use an LLM to generate question, response, and confidence score triplets based on the segments, and 3) apply confidence-score-based filtering and deduplication to ensure data quality and diversity.

We also notice that there are several alternative methodologies such as RAG and long context LLMs, but we don't compare them as we have a huge difference both in training and inference (Mosbach et al. 2023; Liu et al. 2022). We encourage interested readers to refer to Section 2 for more relevant information.

4.2 Automatic Evaluation

To objectively assess the performance of our approach, we conduct automatic evaluations on four widely used benchmarks: SQuAD1.1 (Rajpurkar et al. 2016), TriviaQA (Joshi et al. 2017), DROP (Dua et al. 2019), and WebGLM-QA (Liu et al. 2023b). All these benchmarks contain a variety of test QA pairs with specific contextual references. In the evaluation, we compile the context from each benchmark into a single corpus, and then apply all baselines and our AUGCON on it and test on the test QA pairs in benchmarks.

Metrics For datasets featuring short-form responses (applied to the SQuAD1.1, TriviaQA, and DROP datasets), we measure the model's performance using exact matching (EM) accuracy. A response is considered correct if and only if it matches any of the possible answers. For datasets with long-form responses (applied to the WebGLM-QA dataset), we employ BERTScore (BS) (Zhang et al. 2019) (we use Roberta-Large (Liu et al. 2019) for calculation) to evaluate the semantic similarity between the generated outputs and the reference responses.

Results We use Llama3-70B-Instruct (AI@Meta 2024) as the base model for calling and conducting fine-tuning for automatic evaluations for all our baselines and the proposed AUGCON. The detailed results are shown in Table 1. The results illustrate that our proposed method consistently outperforms the established baselines across all four datasets. Specifically, when analyzing short-form datasets, it becomes evident that the data generated by AUGCON surpasses the comparative methods in extracting pivotal information and knowledge from the corpus, thus improving the question-answering accuracy of fine-tuned models. Meanwhile, the exceptional performance of AUGCON on datasets emphasizing long-form responses showcases its proficiency in generating high-fidelity query-response pairs. This capability directly contributes to enhancing the effectiveness of chat models, enabling them to deliver more relevant, engaging, and contextually appropriate responses based on the given corpus. This, in turn, significantly improves the overall user experience by ensuring that interactions are not only informative but also closely aligned with the user's specific curiosities and requirements.

Furthermore, the consistency of AUGCON in achieving top results across all four datasets, each with unique query patterns and focuses, speaks volumes about its versatility and adaptability. Such consistent performance across varied datasets also underscores the robust generalization ability of our method, making it a highly effective tool for a broad spectrum of corpora types and catering to diverse user interests and inquiries.

Method	Short-form (EM) SQuAD1.1	TriviaQA	DROP	Long-form (BS) WebGLM-QA
Llama3-c$_{70B}$	0.212±0.004	0.723±0.003	0.220±0.004	0.837±0.002
DAPT	0.258±0.004	0.767±0.003	0.266±0.004	0.851±0.002
AdaptLLM	0.273±0.003	0.791±0.004	0.284±0.003	0.842±0.001
ETRC	0.301±0.004	0.812±0.003	0.326±0.004	0.903±0.001
Context-Instruct	0.314±0.003	0.825±0.003	0.334±0.003	0.885±0.001
AUGCON*(Ours)*	**0.336**±0.004	**0.849**±0.003	**0.350**±0.003	**0.924**±0.002

Table 1: The results of automatic evaluation on four benchmarks. We run 10 times for each test and report the mean value and standard deviation, with the best results shown in bold.

4.3 Human Evaluation

In human evaluation on the test scenario, we meticulously curate a corpus dataset, referred to as the *DailyM* dataset, which consists of 1,000 articles carefully selected from a variety of high-quality Chinese magazines closely related to daily life. These articles extensively cover issues of widespread public concern such as basic livelihood, politics, economics, and law, with each article containing approximately 4,000 Chinese characters. Then, we test how well our method and baselines build an AI chat assistant specialized in this daily concern corpus. We apply our method on *DailyM* to generated SFT data called *DailyM-SFT* and use these data to fine-tune Qwen1.5-32B-Chat (Bai et al. 2023) to get fine-tuned model Qwen-DailyM-32B. To further test our method, we conduct annotators to write a total of 1,000 queries they are interested in related to these articles, forming the *DailyM* test set.

Metrics In our comprehensive evaluation framework, we assess both the generated queries and the outputs under the *DailyM* test set of the fine-tuned models to ensure a holistic understanding of the method's performance. Specifically, we evaluate the realism and diversity of generated queries and the relevance, accuracy, and satisfaction of fine-tuned models' outputs.

For both generated queries and model outputs, evaluators are provided with detailed scoring rubrics and examples to promote consistency in evaluation. The queries and outputs will be reviewed by multiple independent evaluators to ensure a balanced and objective assessment, with average scores calculated for each metric to determine the overall performance.

Results For all our baselines and the proposed AUGCON, we employ Qwen1.5-32B-Chat (Bai et al. 2023) as the base model for calling and conducting fine-tuning for later evaluations. For methods such as AdaptLLM, ETRC, Context-Instruct, and our AUGCON which generate query-response pairs based on context, we adhere to a standard where every 35 Chinese characters derive one query-response pair to ensure a fair comparison. We limit the number of generated entries to the same in the comparison because we find that all methods spend much more time on final fine-tuning process compared to the previous generation.

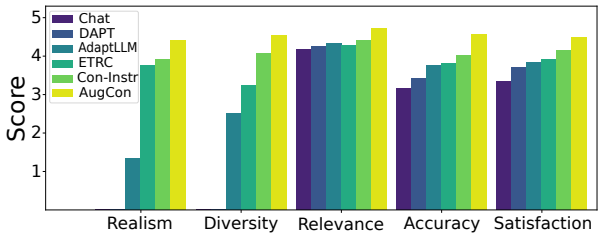

Figure 2: The results of human evaluation on *DailyM*. Query metrics are not applicable for the base chat model and DAPT so we don't show them.

Figure 2 presents the results of the human evaluation on the *DailyM* test set. The results demonstrate that AUGCON consistently surpasses the baseline methods across all evaluation metrics. Specifically, the superior performance in terms of query realism and diversity underscores our method's ability to produce human-like and high-diversity queries. Since our CST and filtering process effectively gain multi-granularity queries that are more effective in covering all granularity levels of context, the derived data will extract more useful knowledge from the corpus. Furthermore, the impressive performance in judging relevance, accuracy, and satisfaction in responses from fine-tuned models further validates that our method's high-quality and diverse queries, coupled with high-fidelity responses, can indeed enhance the performance of subsequently fine-tuned models and achieve higher satisfaction scores from humans. This suggests that AUGCON is particularly adept at constructing high-quality supervised fine-tuning data for LLMs from a given corpus.

4.4 Abalation Study

In this section, we conduct ablation experiments to assess the indispensability and impact of the three essential steps in our proposed method. Specifically, we develop four distinct variations of our method, with each one specifically tailored to concentrate on a fundamental step: (1) **AUGCON**$^{w/o}_{\textbf{CST1}}$

drops the CST part and replaces it by directly iteratively deriving queries from the extracted context until the desired number of queries is obtained. (2) **AUGCON**$^{w/o}_{\textbf{CST2}}$ removes the use of LLM to split in the CST process by directly splitting the contexts in the middle (we will set the whole sentence in the middle all belongs to the first sub-context to maintain semantic integrity). (3) **AUGCON**$^{w/o}_{\textbf{filter}}$ eliminates the contrastive-learning-based score training and filtering process and randomly samples a sufficient number of queries. (4) **AUGCON**$^{w/o}_{\textbf{fidelity}}$ obtains the answers to the queries without adhering to self-alignment and self-improving but utilizes fixed few-shot examples along with a straightforward prompt design devoid of guiding principles.

We implement the four variants on TriviaQA (short-form) and WebGLM-QA (long-form) datasets and conduct a comparison with our AUGCON. The results are shown in Table 2.

Variant	Short-form (EM) TriviaQA	Long-form (BS) WebGLM-QA
AUGCON$^{w/o}_{\text{CST1}}$	0.793±0.003	0.912±0.001
AUGCON$^{w/o}_{\text{CST2}}$	0.826±0.003	0.910±0.001
AUGCON$^{w/o}_{\text{filter}}$	0.828±0.003	0.915±0.001
AUGCON$^{w/o}_{\text{fidelity}}$	0.833±0.004	0.907±0.002
AUGCON	**0.849**±0.003	**0.924**±0.002

Table 2: The results of ablation study.

We find that all variants yield suboptimal outcomes, underscoring the fact that the three essential steps are all crucial and collectively contribute to achieving superior performance.

4.5 Training Phase

We present the loss curve training on the generated *DailyM-SFT* in Figure 3. An interesting observation is that the training loss appears to plateau within epochs from Epoch 2 onwards, yet we observe sudden drops in loss at the boundaries between two consecutive epochs. This pattern strongly signals that our training dataset is characterized by extremely low similarity and exceptionally high diversity, meaning that training on one segment of data does not have an impact on the loss associated with another segment.

We also conduct a human evaluation at each checkpoint during model training, on *DailyM* test set and a widely-used general alignment benchmark AlignBench (Liu et al. 2023c). The overall satisfaction scores increase steadily in both the *DailyM* test set and AlignBench, showcasing that our methods can increase the specific conversation ability without sacrificing general performance.

4.6 Granularity Comparison

A key innovative advantage of AUGCON is its ability to generate queries of varying granularity. To assess the performance of this feature in comparison to baseline methods, we categorize questions into three distinct types based on their scope and depth: detail, concept, and macro, and compare our method with Context-Instruct.

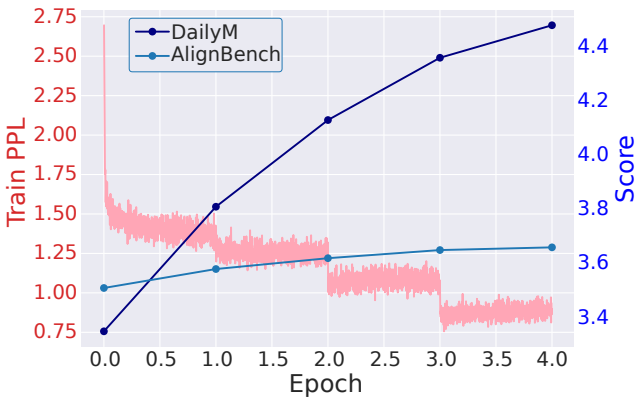

Figure 3: The training loss and human evaluation results during training phase.

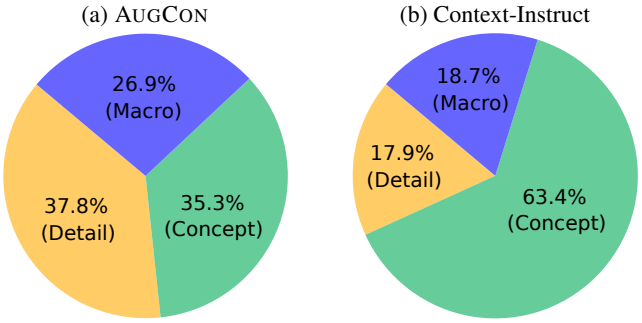

Figure 4: The proportion of three levels of granularity questions generated by AUGCON and Context-Instruct.

The proportions of the three types of questions are illustrated in Figure 4. Our approach achieves a more balanced distribution of question granularities, demonstrating its advantage in covering a diverse range of user inquiries and providing an intuitive explanation for our superior performance.

5 Conclusion

In this work, we propose AUGCON, a highly innovative and effective method to build vertical-domain AI assistants from custom corpora by deriving SFT query-response pairs with diverse granularity. AUGCON starts with query generation through the Context-Split-Tree (CST), an innovative approach for recursively deriving queries and splitting context to cover full granularity. We then employ contrastive learning to develop a scorer that works with CST to rank and refine queries. Finally, we introduce a synergistic integration of self-alignment and self-improving to obtain high-fidelity responses. We conduct extensive experiments on Qwen1.5-32B-Chat and Llama3-70B-Instruct models. The automatic evaluation on four benchmarks and human evaluation on a test scenario demonstrate the significant advantages of our method in producing high diversity, quality, and fidelity context-driven SFT data and improving the performance of custom fine-tuned models against existing methods.

References

Abay, N. C.; Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; and Sweeney, L. 2019. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, 510–526. Springer.

AI@Meta. 2024. Llama 3 Model Card.

An, C.; Huang, F.; Zhang, J.; Gong, S.; Qiu, X.; Zhou, C.; and Kong, L. 2024. Training-Free Long-Context Scaling of Large Language Models. *arXiv preprint arXiv:2402.17463*.

Babbar, R.; and Schölkopf, B. 2019. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, 108(8): 1329–1351.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bayer, M.; Kuehn, P.; Shanehsaz, R.; and Reuter, C. 2024. Cysecbert: A domain-adapted language model for the cybersecurity domain. *ACM Transactions on Privacy and Security*, 27(2): 1–20.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Chang, C.; Peng, W.-C.; and Chen, T.-F. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*.

Chen, W.; Wang, Q.; Long, Z.; Zhang, X.; Lu, Z.; Li, B.; Wang, S.; Xu, J.; Bai, X.; Huang, X.; et al. 2023. DISC-FinLLM: A Chinese Financial Large Language Model based on Multiple Experts Fine-tuning. *arXiv preprint arXiv:2310.15205*.

Chen, Z.; Deng, Y.; Yuan, H.; Ji, K.; and Gu, Q. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Cheng, D.; Huang, S.; and Wei, F. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Dong, G.; Yuan, H.; Lu, K.; Li, C.; Xue, M.; Liu, D.; Wang, W.; Yuan, Z.; Zhou, C.; and Zhou, J. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Du, K.; Snæbjarnarson, V.; Stoehr, N.; White, J. C.; Schein, A.; and Cotterell, R. 2024. Context versus Prior Knowledge in Language Models. *arXiv preprint arXiv:2404.04633*.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference on NAACL*.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Hao, Y.; Sun, Y.; Dong, L.; Han, Z.; Gu, Y.; and Wei, F. 2022. Structured prompting: Scaling in-context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*.

Jiang, T.; Huang, S.; Luo, S.; Zhang, Z.; Huang, H.; Wei, F.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; et al. 2024. Improving Domain Adaptation through Extended-Text Reading Comprehension. *arXiv preprint arXiv:2401.07284*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Krieger, J.-D.; Spinde, T.; Ruas, T.; Kulshrestha, J.; and Gipp, B. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE joint conference on digital libraries*, 1–7.

Li, Y.; Bubeck, S.; Eldan, R.; Del Giorno, A.; Gunasekar, S.; and Lee, Y. T. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; and Raffel, C. A. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35: 1950–1965.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Liu, R.; Wei, J.; Liu, F.; Si, C.; Zhang, Y.; Rao, J.; Zheng, S.; Peng, D.; Yang, D.; Zhou, D.; et al. 2024b. Best Practices and Lessons Learned on Synthetic Data for Language Models. *arXiv preprint arXiv:2404.07503*.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2023a. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Liu, X.; Lai, H.; Yu, H.; Xu, Y.; Zeng, A.; Du, Z.; Zhang, P.; Dong, Y.; and Tang, J. 2023b. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4549–4560.

Liu, X.; Lei, X.; Wang, S.; Huang, Y.; Feng, Z.; Wen, B.; Cheng, J.; Ke, P.; Xu, Y.; Tam, W. L.; et al. 2023c. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; He, X.; Liu, L.; Liu, T.; and Zhai, X. 2023d. Context matters: A strategy to pre-train language model for science education. In *International Conference on Artificial Intelligence in Education*, 666–674. Springer.

Liu, Z.; Zhong, A.; Li, Y.; Yang, L.; Ju, C.; Wu, Z.; Ma, C.; Shu, P.; Chen, C.; Kim, S.; et al. 2023e. Tailoring large language models to radiology: A preliminary approach to llm adaptation for a highly specialized domain. In *International Workshop on Machine Learning in Medical Imaging*, 464–473. Springer.

Luo, H.; Sun, Q.; Xu, C.; Zhao, P.; Lou, J.; Tao, C.; Geng, X.; Lin, Q.; Chen, S.; and Zhang, D. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Mosbach, M.; Pimentel, T.; Ravfogel, S.; Klakow, D.; and Elazar, Y. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Pal, S.; Bhattacharya, M.; Lee, S.-S.; and Chakraborty, C. 2024. A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research. *Annals of Biomedical Engineering*, 52(3): 451–454.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Shaikh, O.; Chai, V. E.; Gelfand, M.; Yang, D.; and Bernstein, M. S. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20.

Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Shi, W.; Min, S.; Lomeli, M.; Zhou, C.; Li, M.; Lin, V.; Smith, N. A.; Zettlemoyer, L.; Yih, S.; and Lewis, M. 2023. In-Context Pretraining: Language Modeling Beyond Document Boundaries. *arXiv preprint arXiv:2310.10638*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Sun, Z.; Shen, Y.; Zhang, H.; Zhou, Q.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2023. Salmon: Self-alignment with principle-following reward models. *arXiv preprint arXiv:2310.05910*.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023a. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting Of The Association For Computational Linguistics*.

Wang, Z. M.; Peng, Z.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Zhang, M.; et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Wei, J.; Hou, L.; Lampinen, A.; Chen, X.; Huang, D.; Tay, Y.; Chen, X.; Lu, Y.; Zhou, D.; Ma, T.; et al. 2023. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*.

Xiong, W.; Liu, J.; Molybog, I.; Zhang, H.; Bhargava, P.; Hou, R.; Martin, L.; Rungta, R.; Sankararaman, K. A.; Oguz, B.; et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Yang, Y.; Sun, H.; Li, J.; Liu, R.; Li, Y.; Liu, Y.; Huang, H.; and Gao, Y. 2023. MindLLM: Pre-training Lightweight Large Language Model from Scratch, Evaluations and Domain Applications. *arXiv preprint arXiv:2310.15777*.

Yuan, Z.; Yuan, H.; Li, C.; Dong, G.; Tan, C.; and Zhou, C. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Zaiem, S.; Algayres, R.; Parcollet, T.; Essid, S.; and Ravanelli, M. 2023. Fine-tuning strategies for faster inference using speech self-supervised models: a comparative study. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5. IEEE.

Zhang, Q.; Zhang, Y.; Wang, H.; and Zhao, J. 2024. RECOST: External Knowledge Guided Data-efficient Instruction Tuning. *arXiv preprint arXiv:2402.17355*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.