MANTA: A Scalable Pipeline for Transmuting Massive Web Corpora into Instruction Datasets

Heuiyeen Yeen **Seokee Hong** **Hyeongu Yun** **Jinsik Lee**
LG AI Research
{heuiyeen214, seokhee.hong, hyeongu.yun, jinsik.lee}@lgresearch.ai

Abstract

We introduce MANTA, an automated pipeline that generates high-quality large-scale instruction fine-tuning datasets from massive web corpora while preserving their diversity and scalability. By extracting structured syllabi from web documents and leveraging high-performance LLMs, our approach enables highly effective query-response generation with minimal human intervention. Extensive experiments on 8B-scale LLMs demonstrate that fine-tuning on the MANTA-1M dataset significantly outperforms other massive dataset generation methodologies, particularly in knowledge-intensive tasks such as MMLU and MMLU-Pro. Our approach also delivers superior performance across a broad spectrum of other tasks, such as Math and Coding. Moreover, MANTA supports seamless scalability by allowing the continuous integration of web corpus data, enabling expansion into domains requiring intensive knowledge. [1]

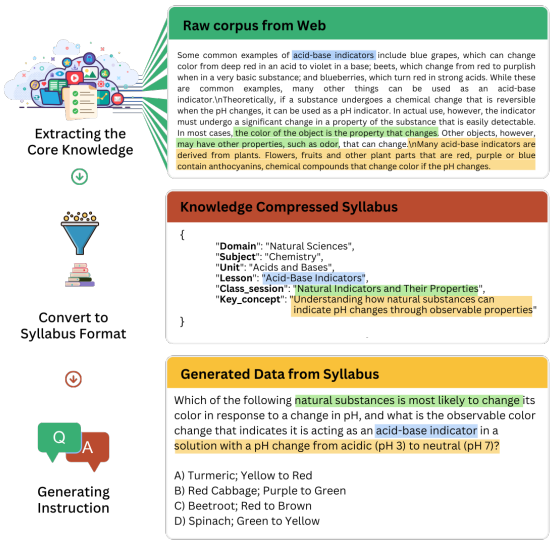


Figure 1: A step-by-step example of the MANTA pipeline that extracts core knowledge from a large web corpus and refines it into a syllabus to generate instructions. It shows that it is possible to generate instructions that reflect the world knowledge contained in vast Web Corpus.

1 Introduction

The scalability of the dataset size is a critical factor not only in the pre-training stage but also in the instruction fine-tuning stage (Yue et al., 2024a; Honovich et al., 2023; Lambert et al., 2024; Yue et al., 2024b; Mitra et al., 2024b). A major challenge in large-scale instruction fine-tuning dataset generation is that directly constructing a supervised dataset, consisting of user queries and assistant responses, is highly human-labor intensive (Zheng et al., 2024a; Zhao et al., 2024; Köpf et al., 2024). To more effectively construct high-quality large-scale fine-tuning datasets, recent researches have been conducted on developing automated pipelines with less human intervention (Wang et al., 2023; Mitra et al., 2024a; Xu et al., 2024; Li et al., 2024; Yue et al., 2024b). In this context, we aim to explore how the vast and diverse knowledge within massive web corpora can be effectively synthesized into an instruction dataset.

In this paper, we present a dataset generation pipeline for instruction fine-tuning from massive web corpora. Our proposed pipeline, MANTA, preserves the diversity of web corpora and scales up to the size of the source web corpora, enabling the construction of large-scale high quality instruction fine-tuning datasets with minimal human effort. This is achieved through the systematic employment of publicly accessible high-performance LLMs and web corpora filtered for educational value.

MANTA pipeline starts by encapsulating the core knowledge of each large-scale web document into a JSON-formatted structured syllabus, as shown in Figure 1. This approach enables us to achieve both the high quality derived from edu-

[1]The dataset is available at https://huggingface.co/datasets/LGAI-EXAONE/MANTA-1M.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18755–18770

cational value and the robustness in generating instructions facilitated by the knowledge-friendly syllabus format. The large-scale syllabi, constructed with near-zero human effort, offer a significant advantage over previous massive instruction dataset generation methodologies in that their topic distribution is not confined to predefined taxonomies (Li et al., 2024), curated seed examples (Wang et al., 2023; Mitra et al., 2023), or the LLM's learned knowledge base (Xu et al., 2024).

By utilizing a publicly accessible fine-tuned LLM, MANTA pipeline generates a large-scale supervised dataset composed of query-response pairs generated from the syllabi. Additionally, we propose a method for generating more complex instruction pairs. *Syllabus Fusion* allows us to increase instruction difficulty by combining multiple syllabi, resulting in enhanced performance of models trained on this dataset. *Multi-turn Expansion* enables the generation of realistic follow-up conversations, improving the usability of the trained model.

We rigorously validate the MANTA pipeline through extensive experiments by fine-tuning various pre-trained LLMs with 8B parameters on 1 million MANTA-generated data points, while also conducting comparative experiments with 1 million data points from other massive dataset generation methodologies. In our experiments, three pre-trained models—Mistral-3-7B-v0.3 (Jiang et al., 2023), Llama-3.1-8B (Dubey et al., 2024), and EXAONE-3.5-7.8B (LG AI Research, 2024)-were used for baseline.

This results in a significant performance gap compared to fine-tuning on other massively generated datasets, demonstrating a clear performance advantage. A notable finding is that the performance gap is particularly pronounced in MMLU (Hendrycks et al., 2020) and MMLU-Pro (Wang et al., 2024), which require diverse knowledge. The models fine-tuned on the MANTA-1M not only excel in the academic knowledge domain but also achieve high performance across various domains, spanning from a mathematics such as MATH-500 (Lightman et al., 2024) to instruction-following ability like MT-Bench (Zheng et al., 2024b). In addition, the diversity and difficulty analyses conducted on the MANTA-1M dataset itself suggest that the MANTA pipeline effectively mimics the topic distribution of massive web corpora while generating more challenging queries.

We highlight our main contributions as follows: (1) we present MANTA pipeline, an automated pipeline for instruction dataset generation from massive web corpora with minimal human effort; (2) with extensive experiments and various analyses, we demonstrate that the MANTA pipeline outperforms other massive dataset generation methodologies; and (3) for reproducibility, we make a subset of the dataset generated by the MANTA pipeline publicly accessible.

2 Related work

As the importance of diverse and large-scale instruction tuning datasets becomes more pronounced, recent studies predominantly focus on synthesizing mass datasets handling diverse domains by prompting LLMs. Starting from a few data points used as the seed dataset, many studies utilize language models' in-context learning ability to generate new datasets (Wang et al., 2023; Yue et al., 2023; Toshniwal et al., 2024). While these methods are scalable, they need well-crafted human-generated seed demonstrations.

Another line of research synthesizes instruction dataset based on the pre-defined topics or taxonomies of the instruction. UltraChat (Ding et al., 2023) first builds the list of topics or materials used for starting points for building instruction dataset by interacting with LLMs. GLAN (Li et al., 2024) also generates datasets by building hierarchical taxonomy of human knowledge. ORCA3 (Mitra et al., 2024a) converts raw text according to 17 defined skills to used as seed skills to generate dataset. While these methodolgies have advantage to focus on pre-defined abilities that the dataset builders expect their LLMs to have, they are also required to enumerate the list of topics. Furthermore, the list of the capabilities or topics is limited to the knowledge of a few human experts, even they utilize powerful LLMs when brainstorming.

To bypass construction of seed datasets or knowledge taxonomy, recent studies leverage enormous amount of web corpus to extract and synthesize instruction dataset. Cheng et al. (2024) collects raw document from web corpus and directly extracts question-answer pairs from them. While they rely on web corpus to represent world knowledge distribution, the web corpus that is able to be used for source of the instruct is limited as the format or quality. On the other hand, since our method compresses the raw document and extracts syllabus from it, MANTA pipeline is not constrained to the

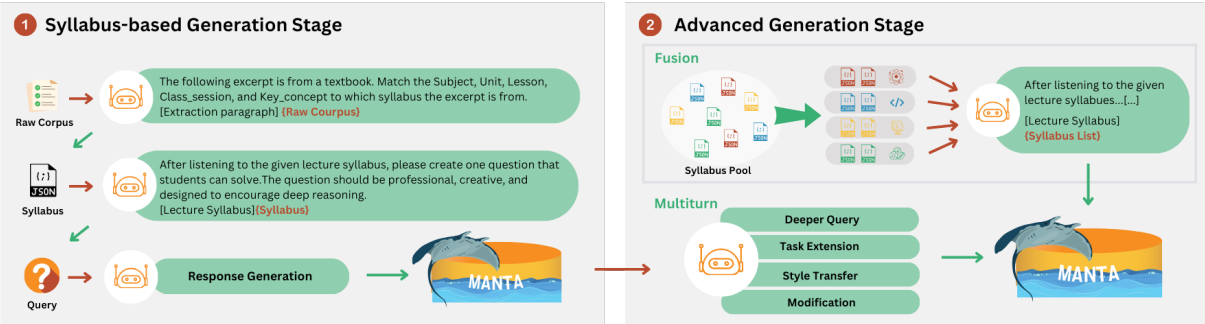


Figure 2: The MANTA pipeline consists of two major stage: the extraction of syllabi from the web to generate instructions, and the process of enhancing high-quality instructions through a fusion method and multi-turn extension. In each stage, brown arrows represent inputs to the LLMs, while green arrows indicate outputs. The datasets generated at both stages are included in the MANTA.

specific format or characteristics of the document.

3 MANTA Pipeline

We develop a massive instruction data generation pipeline, MANTA, that consists of three major subprocesses. First, we reconstruct well-formatted texts playing role as syllabus that encapsulate the core knowledge from crawled web documents. Then we generate query-response pairs from the syllabus. Lastly, we enrich the generalization performance through difficulty enhancement and multi-turn expansion. MANTA prioritizes generating a high-quality instruction fine-tuning dataset from large-scale web corpora with minimal human effort. By utilizing this pipeline, we have generated 200k of syllabus and 3M query-response pairs from them. The overall pipeline for these processes is shown in Figure 2.

3.1 Web Documents to Structured Syllabi

MANTA pipeline takes massive amounts of web-crawled corpora as its primary input. To construct high-quality dataset while preserving the diversity of web-crawled corpora, we leverage LLMs. Specifically, LLMs encapsulate information included in each web document to a formatted text, which we refer to as a *Syllabus*. A syllabus is in JSON-format and represents the core knowledge of the original web document. Inspired by GLAN (Li et al., 2024), we prompt the LLMs and generate a syllabus containing the following attributes: *Subject, Unit, Lesson, Class session, and Key concept*.

To ensure the quality of the syllabi, we use 20 million web documents from high-quality web-crawled corpora, including Wikipedia, FineWeb (Penedo et al., 2024), and Knowledge Pile (Fei et al., 2024). Since each web document is mapped to a single syllabus, we build a syllabi pool of 20 million entries. Note that no human intervention is involved up to this stage of the process. Unlike existing studies where the topic distribution of the generated dataset was restricted by predefined taxonomies (Li et al., 2024) or curated seed examples (Wang et al., 2023; Mukherjee et al., 2023), MANTA can fully reflect the diverse topic distribution of the massive web corpus.

We further enhance the syllabi pool by adopting *Domain* attribute through a bottom-up categorization approach. Specifically, we perform k-means clustering on the text-embedding vectors of each syllabus, partitioning them into 2,000 clusters. Next, we sample 100 syllabi from each cluster in order of proximity to the centroid, which serve as the foundation for domain categorization. We utilize nine domains from Dewey decimal classification (Dewey, 1876) and additionally introduce 'Computer Science and Coding', as well as 'Mathematics'. After classifying the clusters, we update the domain attribute of each syllabus, resulting in 200k syllabus from diverse web corpus.

3.2 Syllabus-based Instruction Generation

To generate user queries relevant to a given syllabus and corresponding responses, MANTA employs a LLM again. Given a syllabus, we first prompt the LLM to generate a query. And then, LLM receives generated query to create response. To be specific, the EXAONE-3.5-32B-Instruct (LG AI Research, 2024) is used for MANTA Pipeline. Also, various models can be used to query generator and response generator. We utilize different kinds of LLMs as generators in each step shown in Section 4.2.1 and Appendix F.

Generating data through a syllabus instead of

directly from raw web documents offers two key advantages. First, since raw web documents often focus on granular topics, we observed that directly generating queries are prone to yielding overly narrow and specialized questions. This tendency also makes the query generator more prompt-sensitive, which is why our pipeline, incorporating the syllabus, is more robust and requires less human effort. Another advantage is that MANTA can generate high-quality queries even from web documents with low educational value when processed through the syllabus-based generation. This is possible because the LLM draws upon its own knowledge base to construct a well-structured syllabus, allowing it to extract meaningful content even from documents with limited knowledge.

By following the processes described above, MANTA pipeline generates large-scale supervised fine-tuning data from massive web corpora with minimal human effort.

3.3 Advanced Instruction Generation

Difficulty Control Recent studies observe that augmenting datasets with complex and difficult queries improves the overall performances of the fine-tuned model (Xu et al., 2023; Luo et al., 2024, 2023; Guo et al., 2023). In contrast to existing methods which aim to enhance the difficulty of the instruction itself, we design a *Syllabus Fusion* method, which utilizes the multiple syllabus to create a new, complex instruction. We first randomly sample syllabus within the same domain. Then we prompt a LLM to generate a fused syllabus for complex query generation. Using the prompts introduced in the Appendix 6, we explore the Syllabus Fusion with different numbers of syllabi (2, 3, 4, 5, and 10). This step assumes that a higher number will naturally reflect complex knowledge by integrating each piece of simple knowledge.

Consequently, we believe that the Syllabus Fusion method allows MANTA to automatically create diverse complexity without additional prompting effort for difficulty enhancement, unlike previous methods (Xu et al., 2023; Luo et al., 2024, 2023).

Multi-turn Expansion To address real-world scenarios where multi-turn conversations occur, we expand our dataset to multi-turn scenario. We categorized various multi-turn scenarios into four categories and expanded them accordingly in the pipeline. We define four types of follow-up queries:

- Deepening Question: Requests the same task but with additional specific conditions.
- Task Extension: Demands a different task compared to the preceding query.
- Style Transfer: Requests to change the tone and manner of expression in the preceding response.
- Modification: Requests to change the format of writing (*e.g.* business email, academic paper, programming code, *etc.*), paragraph structure (*e.g.* line breaks, delimiters, markdown format, *etc.*), or main content of the preceding response.

Give the specific query type and the previous conversation context, a LLM is prompted to generate follow-up query sound natural to the context.

4 Experiment

4.1 Setup

Baseline Dataset and Model To minimize human intervention, we compare our approach with competitive datasets generated through large-scale, multi-domain dataset generation using LLMs. MAmmoTH2 (Yue et al., 2024b) utilizes 10 million naturally occurring instructional data points from a pre-training web corpus and refines them by LLMs. MAGPIE-PRO (Xu et al., 2024) is a fully automated method for synthesizing alignment data from instruction-tuned LLMs. It is adaptable for creating multi-turn, preference-based, and domain-specific datasets without human involvement. MAGPIE-PRO is constructed using Llama-3-70B-instruct (AI@Meta, 2024) model, resulting in 1M publicly available data. ORCA 3 (Mitra et al., 2024a) introduces an automated data generation agent framework using LLMs, utilizing GPT-4 to create substantial high-quality synthetic data. Although the size of the full dataset is approximately 22M, only 1M is publicly available, thus we use the 1M dataset for comparison in this study.

For a fair comparison, Supervised Fine-tuning experiments utilize 1M data points for each dataset, training similarly sized pretrained models: EXAONE-3.5-8B, Llama-3.1-8B, and Mistral-7B-v0.3.

Supervised Fine-Tuning Setup Supervised Fine-Tuning is conducted on 16 NVIDIA A100-SXM4-40GB GPUs with a cosine learning rate schedule,

Fine-tuning Dataset	Total Average	Academic General			Domain Specific			Instruction Following	
		MMLU	MMLU-PRO	ARC-C	MATH 500	GPQA-DIA	HUMANEVAL	MT-BENCH	ALPHACAEVAL2
fine-tuned from Mistral-7B-v0.3									
MAmmoTH2-1M	28.74	43.20	29.60	37.03	16.60	24.74	28.00	4.60	4.75
MAGPIE-PRO-1M	21.38	10.20	29.05	57.76	10.00	3.54	38.41	5.38	26.04
ORCA 3-1M	43.01	58.10	34.16	74.15	29.20	32.32	40.85	6.74	7.92
MANTA-1M	**47.88**	**64.51**	**37.71**	**77.73**	**31.80**	**32.83**	**41.46**	**7.08**	**26.22**
fine-tuned from Llama-3.1-8B									
MAmmoTH2-1M	37.29	52.73	30.86	65.44	18.40	29.29	39.00	5.61	6.48
MAGPIE-PRO-1M	44.43	56.12	34.09	61.01	23.40	30.30	46.95	6.98	33.73
ORCA 3-1M	48.76	64.41	39.80	77.65	31.20	29.29	**48.17**	7.36	25.97
MANTA-1M	**53.20**	**66.61**	**45.81**	**79.69**	**39.60**	**34.34**	46.34	**7.83**	**34.93**
fine-tuned from EXAONE-3.5-7.8B									
MAmmoTH2-1M	32.79	55.71	33.44	54.10	31.00	29.80	55.49	4.91	7.77
MAGPIE-PRO-1M	47.12	56.73	36.29	69.97	38.00	32.83	56.71	5.52	31.21
ORCA 3-1M	49.06	61.94	39.08	80.12	42.00	20.20	53.66	7.16	23.90
MANTA-1M	**57.58**	**68.67**	**48.21**	**80.55**	**54.00**	**39.39**	**58.54**	**7.45**	**36.75**
MANTA-3M	58.42	68.73	47.85	84.04	57.40	34.85	60.98	7.77	35.79

Table 1: This result stems from the SFT on EXAONE-3.5-7.8B, Llama-3.1-8B, and Mistral-3-7B-v0.3, utilizing equal amounts of datasets that were created with LLMs. It outperformed all the benchmarks, particularly in the Academic General benchmark, indicating that MANTA exhibits robust performance across various disciplines and difficulty levels, while also being effective with all tested models. When calculating the average score, we multiply the MT-Bench score by 10 to scale it to a 100-point scale.

an initial learning rate of 2e-5, the AdamW optimizer, and a batch size of 256. All training processes are carried out up to a maximum of 4 epochs. All models are trained under the same setup.

Benchmark and Evaluation The evaluation encompasses a total of 10 benchmarks, which can be broadly categorized into Academic General, Domain Specific, and Instruction Following fields.

- Academic General: Assesses academic world knowledge. Includes MMLU (Hendrycks et al., 2020) for 57 subjects, MMLU-PRO (Wang et al., 2024) for an enhanced version, and ARC-C for grade 3-9 science exams.
- Domain Specific: Targets specialized domains.
  - **Math:** Uses GSM8K (Cobbe et al., 2021) for diverse math problems and MATH 500 (Lightman et al., 2023) for advanced problems.
  - **Science:** Includes THEOREMQA (Chen et al., 2023) for complex science problems and GPQA (Rein et al., 2023) for expert-level questions in biology, physics, and chemistry.
  - **Code:** Evaluated using HUMANEVAL (Chen et al., 2021).
- Instruction Following: Tests model responses based on real-world instructions.
  - MT-BENCH (Zheng et al., 2023) Assesses multi-turn dialogue coherence and engagement.
  - ALPACAEVAL 2 LC (Dubois et al., 2024) Measures win-rates for various NLP and instruction-following tasks.

For a detailed description of the evaluation benchmarks and assessment methods, please refer to Table 8 in Appendix.

4.2 Experimental Results

Main Result The Table 1 presents the results of trained models on four comparative datasets, each comprising 1M entries. The additional results of GSM8K and THEORMQA are shown in the Table 7. MANTA-1M outperforms in all benchmark areas. Specifically, in Academic General tasks, the significant performance gaps are noted in MMLU and MMLU-PRO. We believe that MMLU, a benchmark consisting of elementary to expert level knowledge across 57 diverse fields, reflects the impact of our dataset, which encompasses various fields, tasks, and difficulty levels.

In Domain Specific tasks, the models finetuned on MANTA-1M show robust performances across diverse domain-specific benchmarks. It is worth to note that we do not explicitly collect math or coding-related documents. However, thanks to the diverse distribution inherent in the web corpus, which is also reflected in the instruction distribution, diverse set of instructions across different lev-

Datasets	MMLU			
	STEM	Others	Social Sciences	Humanities
EXAONE 3.5 2B instruction model *(as a Response Generator)*				
MAGPIE-PRO	60.32	62.92	66.24	55.70
ORCA 3	54.45	56.95	60.91	58.90
MANTA	**63.40**	**65.15**	**68.97**	**60.56**
QWEN 2.5 32B instruction model *(as a Response Generator)*				
MAGPIE-PRO	64.08	65.71	69.98	62.09
ORCA 3	62.83	63.88	69.42	63.78
MANTA	**65.71**	**68.05**	**72.41**	**67.50**
gpt-4o-2024-08-06 *(as a Response Generator)*				
MAGPIE-PRO	62.10	66.07	70.33	64.87
ORCA 3	57.03	59.73	66.74	62.52
MANTA	**64.90**	**66.49**	**72.21**	**65.50**

Table 2: Performance changes in each MMLU subtask with different Response Generators. Even if the Answer quality improves, the MANTA query-based dataset shows the best results.

els of difficulty appears to have been created.

Finally, the model also exhibited superior performance in the Instruction Following benchmark. This indicates that MANTA-1M effectively incorporates real-world scenario.

To scale the training, we fine-tune EXAONE-3.5B-7.8B on MANTA-3M data points. Apart from GPQA-DIAMOND and ALPACAEVAL2 LC, performance improves with increased training data volume. This indicates that a merely extracting core knowledge from the web corpus and constructing instructions from it can lead to automatic generation of diverse and high-quality datasets.

4.2.1 The Importance of Query Organization

In Table 2, by varying the response generating models, we demonstrate that our approach to generate queries is the most dominant factor for robust model performance. We re-generate the responses using different parameter size models for 300k samples randomply drawn from ORCA3 and MAPIE-PRO, and MANTA. Subsequently, each dataset was trained on EXAONE-3.5-7.8B. Regardless of changes in the answer generating models, MANTA consistently outperformed in every MMLU subtask.

This suggests that even if the quality of answers improve, performance can vary depending on how the dataset queries are composed. Thus, we believe that the composition of queries significantly contributes to the finetuned model's performance. Moreover, as described in Section 4.1, although both ORCA3 and MAPIE-PRO are built using models with larger size compared to those used in the MANTA Pipeline (i.e., EXAONE-3.5-32B-Instruct), they demonstrate lower performance than MANTA. This result proves that MANTA's pipeline helps queries to generalize well across various domains while maintaining robust performance.

Dataset	TTR	MLTD	INSTAG	Query Length
MAGPIE-PRO	0.94	19.61	27.62	72.20
ORCA 3	0.62	60.45	50.02	1935.83
MANTA	0.77	88.56	44.87	509.43

Table 3: Diversity evaluation results for 500k samples in each of the comparison datasets.

5 Dataset Analysis

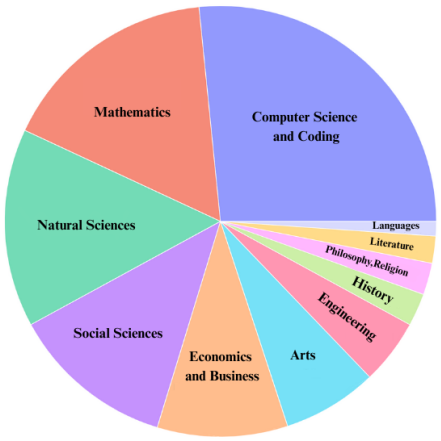
5.1 Diversity Analysis

Diversity of Instructions We conduct a diversity evaluation using various metrics to assess the diversity of queries. First, the TTR (Type-Token Ratio) was used as a measure of linguistic diversity. TTR (Li et al., 2016) can vary significantly with the length of the text, tending to be higher for shorter texts, and lower for longer ones as the proportion of unique words decreases. MAGPIE-PRO showed the highest TTR, attributed to its shorter query length compared to other datasets. When compared to ORCA 3, which has a longer average query length than MANTA but a lower TTR, MAGPIE-PRO is assessed to have a higher incidence of unique words. The MTLD (McCarthy and Jarvis, 2010) metric, Measure of lexical textual diversity, designed to address TTR's limitations. By this measure, it demonstrated significantly higher values compared to other datasets, indicating a broader use of diverse linguistic expressions.

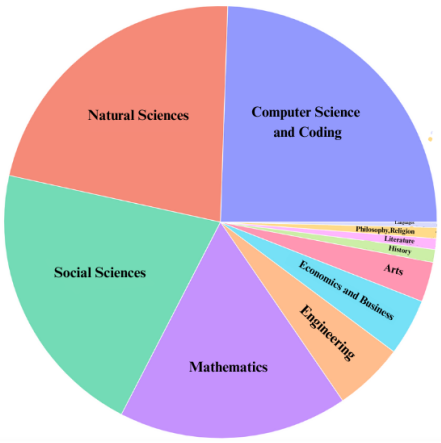
In terms of task and domain diversity, an evaluation was conducted using Lu et al. (2023) approach. Each data entry was tagged, and the evaluation was based on the ratio of the number of unique tags in each dataset to the number of unique tags appearing across all three datasets. Even in this case, MANTA has the second highest unique tag ratio.

Domain Distribution We demonstrate how the knowledge domains of a Web corpus can be reflected in the instruction distribution by comparing the distribution of extracted syllabus from the Web with the generated instruction domains in Figure 3.

The top side of the figure shows the domain distribution of the syllabus extracted from the corpus. Since the syllabus was extracted from approximately 20M corpora, it can be interpreted as the



(a) Domain Distribution of Extracted Syllabus



(b) Domain Distribution of Generated Instruction

Figure 3: Comparison of the distribution of Syllabus extracted from 20M Web Corpus and the distribution of MANTA-1M generated from it.

knowledge distribution covered by these corpora. It has been observed that a significant portion falls within the fields of CODE, MATH, Natural Science, and Social Science, reflecting the nature of these corpora. The bottom side of figure is that distribution of MANTA-1M. It follows the distribution within the fields of extracted syllabus. This demonstrates that the distribution of knowledge available on the Web can naturally be reflected in the distribution of instructions.

Furthermore, by incorporating a new Web corpus, MANTA that it is possible to expand the distribution of instructions into new domains or adjust it according to desired distributions.

5.2 Difficulty Analysis

Difficulty of Instructions The difficulty judgement is conducted using two methods. Similar to the approach used by Xu et al. (2023), we employed gpt-4o-2024-08-06 to assess the difficulty level of each of the 500k instructions. The detailed results are provided in Table 4, and the prompt used for the Difficulty judgement is available in Appendix 8. Instructions are rated on a scalar value from 1 to 10, with higher values indicating instructions of greater difficulty.

Dataset	LLM-as-a-Judge	INSTAG
MAGPIE-PRO	5.84	2.56 ± 0.84
ORCA 3	8.06	3.97 ± 1.52
MANTA	8.52	4.17 ± 1.09

Table 4: Difficulty judgements for 500K samples in each of the comparison datasets.

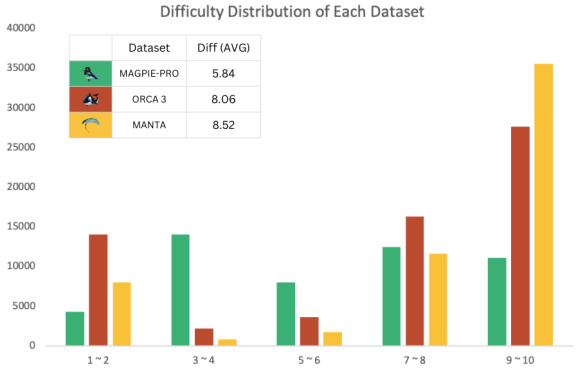


Figure 4: Histogram of the distribution of Difficulty judgements (LLM-as-a-Judge) for 500K samples in each of the comparison datasets.

Figure 4 illustrates that MANTA has an average difficulty level of 8.52, the highest among the datasets. This indicates that by simply combining syllabi from the same domain, it is possible to efficiently generate high-difficulty data without needing additional steps to increase the difficulty.

Leverage of Fused Syllabus We conduct an experiment on Fusion ratio to analyze the impact of the syllabi fusion method. We adjust the proportion of the fused dataset on a 300K dataset and trained it on EXAONE-3.5-7.8B, then report the averge performance on a specialized benchmark(e.g., MATH-500, GPQA-DIAMOND, MMLU-PRO, THEOREM-QA). As shown in Figure 5, the average difficulty of the dataset increases with a higher proportion of the fused data. However, the best performance on the benchmark is achieved with a uniform ratio distribution.

This result aligns with the findings of the study in Sun et al. (2024), suggesting that instead of solely training with challenging data, a dataset capable of generalizing across diverse difficulty contributes

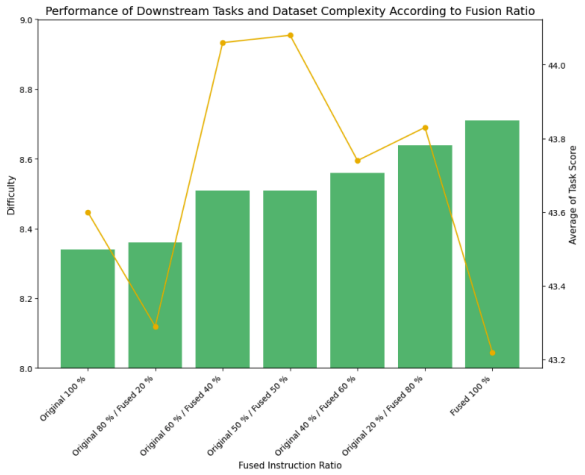


Figure 5: Average benchmark performance based on fusion ratio changes. The histogram shows dataset difficulty by fusion ratio, while the line graph displays model performance after training. Average of difficulty increases with a higher fusion ratio, but downstream tasks perform best with a balanced ratio.

Models	BASE	+ Science 200k
MATH-500	58.60	57.80 (-0.8)
GPQA-DIAMOND	32.83	34.34 (+1.51)
MMLU-STEM	65.40	65.70 (+0.3)
THEOREMQA	21.37	24.50 (+3.13)

Table 5: Baseline performance of the EXAONE-3.5-7.8B model trained on a random 300k dataset, with observed performance improvement when 200k additional data from a science-related syllabus is incorporated.

more positively to performance improvement in practical downstream tasks.

5.3 Generating Domain Specific Data by Targeted Syllabus

Sampling the syllabus of a desired domain to create instructions in a controllable environment is another significant advantage of a syllabus database reflecting diverse core knowledge from enormous Web corpus.

We conduct an experiment to determine whether creating instructions based on the Science domain syllabus would enhance performance in related downstream tasks. This evaluation is carried out using four downstream tasks from the Science-Leaderboard [2], which primarily consists of Science domain tasks. We compare the baseline model, EXAONE-3.5-7.8B, which is trained on a random 300k sample from MANTA as baseline and an additional 200k instructions made by the Natural Science syllabus.

Detailed performance results are shown in Table 5. Except for the Mathematics domain benchmark, MATH-500, improvements are observed across all science-related benchmarks. Notably, significant enhancements were found in GPQA-DIAMOND and THEOREMQA, which consist of problems at an expert level. This shows that utilizing instructions generated based on the relevant domain syllabus for the targeted domain can improve performance.

6 Conclusion

We present MANTA, an automated pipeline designed to generate high-quality large-scale instruction fine-tuning datasets from massive web corpora, while maintaining their diversity and scalability. With a syllabus-formatted structure distilled from massive web corpora, the MANTA pipeline robustly generates effective query-response pairs with minimal human effort. Extensive evaluations on 8B-scale LLMs show that fine-tuning on the MANTA-1M dataset consistently outperforms other large-scale dataset generation approaches, excelling in knowledge-intensive tasks like MMLU and MMLU-PRO, while also demonstrating strong results across a wide range of benchmarks.

7 Limitations & Ethics Statement

Limitations MANTA confirmed that extracting core knowledge from a vast corpus in the form of a syllabus and generating large volumes of instructions based on this improve the performance of downstream tasks. However, since the experiment was started with only 20M corpora, we specify that there may be performance changes when the MANTA pipeline is applied to larger corpora. We suggest a research direction on the feasibility of the MANTA pipeline across various forms and volumes of raw formats. Additionally, we experimented with the MANTA pipeline based on a few specific models. However, we hope that future research will explore this pipeline in depth using data derived from various models.

Ethics Statement MANTA aims to reduce human effort and create diverse, high-quality datasets by utilizing LLM throughout the entire process of generating instructions. However, it is noted that harmful datasets may be included in some stages of the process. Therefore, by labeling each data

[2] https://huggingface.co/spaces/wenhu/Science-Leaderboard

instance with the results of the safety evaluation annotations conducted in the Appendix D, users can be fully aware of the risks while using the data. Approximately 5% of the entire 1M dataset is identified as risky, and 3% of those require expert-related advice. We also specify that the datasets we release are for research use only to assess and develop the learning capabilities of LLMs for this reason. Therefore, we release this data in accordance with the terms of the CC-BY-NC-4.0 license.

References

AI@Meta. 2024. Llama 3 model card.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. *arXiv preprint arXiv:2406.14491*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Melvil Dewey. 1876. *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library*. Brick row book shop, Incorporated.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Zhaoye Fei, Yunfan Shao, Linyang Li, Zhiyuan Zeng, Conghui He, Hang Yan, Dahua Lin, and Xipeng Qiu. 2024. Query of cc: unearthing large scale domain-specific knowledge from public corpora. *arXiv preprint arXiv:2401.14624*.

Weidong Guo, Jiuding Yang, Kaitong Yang, Xiangyang Li, Zhuwei Rao, Yu Xu, and Di Niu. 2023. Instruction fusion: advancing prompt evolution through hybridization. *arXiv preprint arXiv:2312.15692*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

LG AI Research. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. 2024. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, and Chang Zhou. 2023. # instag: Instruction tagging for diversity and complexity analysis. *arXiv preprint arXiv:2308.07074*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2024. Wizardcoder: Empowering code large language models with evol-instruct. In *The Twelfth International Conference on Learning Representations*.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei-ge Chen, Olga Vrousgos, Corby Rosset, et al. 2024a. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024b. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*.

Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. 2024. Easy-to-hard generalization: Scalable alignment beyond human supervision. *arXiv preprint arXiv:2403.09472*.

Llama Team. 2024. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E Weston, et al. 2025. Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions. *arXiv preprint arXiv:2502.13124*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mammoth: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024b. Mammoth2: Scaling instructions from the web. *arXiv preprint arXiv:2405.03548*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Statistics of MANTA with Other Instruction Datasets.

Table 6 presents the statistical analysis of the MAmmoTH2 (Yue et al., 2024b) MAGPIE-PRO (Xu et al., 2024) and ORCA 3 (Mitra et al., 2024a) datasets, in addition to the MANTA dataset.

B Additional Evaluation Result

Apart from the TheoremQA results on Mistral 7B v0.3, it shows excellent results in additional benchmarks as well. This demonstrates that MANTA has satisfactory generalization capabilities. The results can be found in Table 7.

C Evaluation Details

For fair evaluation of a instruction tuned model, we evaluate all benchmarks in the Academic General and Domain Specific category using the *0-shot setting*. As an exception, TheoremQA followed the original measurement method. To accomplish this, we direct language models using prompts that demand answers in particular formats, and then extract the final answer from their responses. For a fair comparison, we use the same prompts across all models. We make public all the prompts we used in (LG AI Research, 2024) for transparent reproducibility.

In the Instruction Following benchmark that evaluate by LLM-as-a-judge, gpt-4o-2024-08-06 is used for MT-bench, and gpt-4-1106-preview is used for Alpaca Eval 2.

The detailed evaluation metrics and methodology are shown in Table 8. All results were reported by taking the maximum result from the models that 3 epochs or 4 epochs.

D Safety Analysis

We conduct additional experiments focusing on cost and safety concerns. We utilized the (Team, 2024) to evaluate the safety of datasets and the toxicity of responses from learning models to assess the safety of MANTA-1M. The overall dataset was found to be 95% safe, with approximately 5% deemed harmful. All types of toxicity were below 1%, except for the Specialized Advice category, which stood at 3%. This category includes responses with professional financial, medical, or legal advice, or those claiming that dangerous activities or items are safe. Given the dataset includes a substantial amount of data from medical, legal, and other specialized domains, it is analyzed that this category was detected in relevant query responses.

E Cost Analysis

E.1 API Call Perspective (LLM Inference Count)

The cost can be correlated with the number of API calls, equating to the number of LLM inferences. While datasets based on LLM generation do not explicitly state the cost details, with the exception of MAGPIE-PRO, we can calculate the API call count necessary to generate a single instruction as follows:

- **MANTA** : Total of 3 calls for syllabus extraction, instruction generation, and response generation.
- **ORCA 3** : Due to iterations aimed at enhancing instruction quality and difficulty, it is associated with high costs. The ORCA 3 (Mitra et al., 2024a) notes this limitation, "Generating synthetic data with multiple agents using LLMs and tools can be resource intensive."
- **MAmmoTH2** : Requires complex processes involving various LLMs, such as URL classification, question extraction, and refining.
- **MAGPIE-PRO** : Similar to MANTA, 3 calls are needed for instruction generation, response generation, and filtering, but it necessarily undergoes a deduplication step.

In conclusion, these calculations demonstrate the ability to rapidly scale to large volumes with fewer calls compared to other automatic generation methodologies.

E.2 Generation time for the MANTA

We perform experiments on a server with two NVIDIA H100 SXM5 80GB GPUs using the VLLM inference framework. The models are loaded in the float8 format.

When creating the 1M MANTA dataset, it spent 24 hours to generate the initial instructions (Step 1) and fusion - multi turn (Step 2). However, the extraction of the syllabus from the web corpus required approximately 100 hours, as it was conducted on 20 million documents.

Statistics of MANTA						
	Source	Answer	Turn (Avg)	Total Tokens	Query Length (AVG)	Response Length (AVG)
MAmmoTH2-1M	Raw Corpus	Mixtral-22B×8, Qwen-72B	1	343M	58.57±52.18	272.45 ±210.32
MAGPIE-PRO-1M	-	Llama-3-70B-instruct	1	476M	15.16 ±9.7	460.68 ±210.71
ORCA 3-1M	Raw Corpus	GPT-4	1.17 ±0.69	964M	383.18 ±861.64	439.92 ±277.44
MANTA-1M	Raw Corpus	EXAONE-3.5-32B-instruct	1.1 ±0.31	791M	86.44 ±52.16	625.17 ±211.02

Table 6: Statistics of MANTA with Other Instruction Datasets.

Models	Domain Specific	
	GSM8K	THEORMQA
Mistral-7B-v0.3		
MAmmoTH2-1M	31.31	20.88
MAGPIE-PRO 1M	51.93	18.00
ORCA 3-1M	52.53	20.50
MANTA-1M	74.68	19.63
Llama-3.1-8B		
MAmmoTH2-1M	23.35	21.63
MAGPIE-PRO 1M	61.25	22.50
ORCA 3-1M	61.56	22.63
MANTA-1M	79.38	25.25
EXAONE-3.5-7.8B		
MAmmoTH2-1M	53.90	24.50
MAGPIE-PRO 1M	73.92	28.88
ORCA 3-1M	72.78	26.50
MANTA-1M	82.64	29.75

Table 7: Additional Result in Academic General benchmark (e.g. GSM8K) and Domain Specific (e.g TheormQA)

Benchmark	Evaluation Settings	Metric
Academic General		
MMLU	0-shot / CoT	Accuracy
MMLU-PRO	0-shot / CoT	Accuracy
ARC-C	0-shot	Accuracy
Domain Specific		
GSM8K	0-shot / CoT	Accuracy
MATH 500	0-shot / CoT	Accuracy
GPQA-DIAMOND	0-shot / CoT	Accuracy
THEORMQA	5-shot	Accuracy
HUMANEVAL	0-shot	pass@1
Instruction Following		
MT-BENCH	LLM-as-a-judge (gpt-4o-2024-08-06)	LLM score
ALPACAEVAL 2.0	LLM-as-a-judge (gpt-4-1106-preview)	Win rate

Table 8: The benchmarks used to evaluate and their evaluation settings with metric.

MMLU			
STEM	Others	Social Sciences	Humanities
EXAONE-3.5-32B-instruct *(as a Query Generator)*			
63.87	62.61	69.97	60.26
Qwen2.5-32B-Instruct *(as a Query Generator)*			
65.62	65.89	69.95	60.8
Gemma-3-27b-it *(as a Query Generator)*			
66.06	68.47	72.28	**66.19**
Deepseek v3 *(as a Query Generator)*			
66.87	**67.99**	**72.33**	64.66

Table 9: The results of the MMLU for each sub-task, After processing through the MANTA pipeline with Query Generators of different sizes and model families, we generated responses using the same Response Generator (i.e., EXAONE-3.5-32B-instruct).

F Analysis of Changes in the Query Generator Model of the MANTA Pipeline

Based on EXAONE-3.5-32B-Instruct, Qwen2.5-32B-Instruct, Gemma-3-27b-it, and Deepseek v3, we analyze the results obtained by following the MANTA pipeline, as described in Section 3. We used the same syllabi, sampling 120k to generate queries and the Response Generator (i.e., EXAONE-3.5-32B-instruct). These were uniformly trained using EXAONE-3.5-7.8B. As a result in the Table 9, we can see that as the size and capability of the query-generating model improve, the performance improves proportionally. It becomes evident that the composition of queries by the trained model is even more crucial.

G Performance Variation Based on Model Size

The results aiming to observe the impact of changes in model size within the Academic General Benchmark are presented in Table 10. When comparing EXAONE-3.5-2.4B and EXAONE-3.5-7.8B, with each comparison set comprising 1 million data, all results proportionally increase with the size of the model parameters. However, the trends in the results for each dataset remain consistent, demon-

Academic General			
	MMLU	MMLU-PRO	ARC-C
fine-tuned from EXAONE-3.5-2.4B			
MAmmoTH2-1M	46.5	30.66	66.47
MAGPIE-PRO-1M	46.05	28.84	5.46
ORCA 3-1M	51.28	33.69	70.22
MANTA-1M	**56.74**	**38.33**	**74.15**
fine-tuned from EXAONE-3.5-7.8B			
MAmmoTH2-1M	55.71	33.44	54.10
MAGPIE-PRO-1M	56.73	36.29	69.97
ORCA 3-1M	61.94	39.08	80.12
MANTA-1M	**68.67**	**48.21**	**80.55**

Table 10: Performance Variation Across Different Datasets According to Model Size.

Academic General			
	MMLU	MMLU-PRO	ARC-C
fine-tuned from EXAONE-3.5-7.8B			
MANTA-1M	**68.67**	48.21	80.55
MANTA-1M$_{upgrade}$	68.12	**48.81**	**83.79**

Table 11: Performance Changes of MANTA After Logical Filtering

strating the robustness of MANTA.

H Performance Variation Based on Logical Filtering Results

Based on the automatically generated results, filtering was conducted to address more logical tasks, followed by an analysis of the filtering effect. The filtering prompt used was Yuan et al. (2025). In this filtering, data were assessed for logical complexity on a 1-10 scale. Data scoring 6 or higher were kept as is, while those below 6 were used to generate a new improved question to enhance complexity. The MANTA-1M$_{upgrade}$ is refined responses based on the improved questions reviewed by the EXAONE-3.5-32B-instruct model. Both original MANTA and MANTA-1M$_{upgrade}$ were then trained and implemented in the EXAONE-3.5-7.8B model. These results can be seen in Table 11. The MANTA-1M$_{upgrade}$ showed improved performance on the Academic General benchmark, except for the MMLU. This is likely because the MMLU covers a wide range of fields and includes questions at various difficulty levels, from elementary onwards. Training only on more challenging data seems to weaken generalization. These results align with those discussed in Section 5.2.

Generating Instruction from fused Syllabus

System:

After listening to the given all lecture syllabi, please create one question that students can solve. The question should be professional, creative, and designed to encourage deep reasoning.
Create it in JSON format as follows {"Generated_Question": str }

User:
`[Lecture Syllabi]`
{{*Given Syllabus list*}}

Figure 6: Prompt for Generating Instruction using fused syllabi

Generating Multi-turn Conversation

System:

Create a follow up question that follows a given Q-A. However, the follow up question must satisfy given condition.

Create it in JSON format as follows {"Generated_Question": str }

User:
`Given Q-A Question :` {{*Previous Turn Query*}}
`Answer :` {{*Previous Turn Response*}}

`Given Condition :`
{{*Given Condition for Multi-turn Generation*}}

Figure 7: Prompt for Generating Multi-turn Conversation

I Examples of Prompts for Genrating Step

The prompts for Advanced Instruction Generation step are shown in Figures 6-7

J Difficulty judge Prompt

The prompts for Difficulty judge Prompt is shown in Figures 8

K Example of instructions according to the number of fused syllabi and instruction generation from the Difference corpus

Examples of instructions based on a fustion of syllabi and examples of instructions extracted from different web corpora.

Difficulty Judge prompt

System:

Evaluate and rate the difficulty of the following question. You should give an overall score on a scale of 1 to 10, where a higher score indicates higher difficulty and complexity. Please rate questions at the level of a toddler as 1-2, questions at the level of an elementary school student as 3-4, questions at the level of a middle school student as 5-6, questions at the level of a high school student as 7-8, and questions at the level of a college student or professional as 9-10. You must just give a score without any other reasons.

User:
`Given Query`
{{*Generated Query*}}

Figure 8: Prompt for Difficulty judgement of Instruction by LLM-as-a-judge

Example of Generated Instructions Utilizing a Fused Syllabus in Mathematics

Syllabi in Mathematics Domain

[Lecture Syllabus]

PROVIDED SYLLABUS 1
Domain : Mathematics
Subject : Algebra
Unit : Linear Equations
Lesson : Point-Slope Form
Class_session : Finding the Equation of a Line Through Two Points
Key_concept : Calculating slope and using it to find the equation of a line in point-slope form

PROVIDED SYLLABUS 2
Domain : Mathematics
Subject : Algebra
Unit : Rationalization
Lesson : Rationalizing the Denominator
Class_session : Using Conjugates to Rationalize Denominators
Key_concept : Rationalization is the process of eliminating radicals from the denominator of a fraction by multiplying the numerator and denominator by the conjugate of the denominator.

Generated Query

In the context of algebraic linear equations, consider two points on a Cartesian plane: A(2, 3) and B(5, 11). Calculate the slope of the line passing through these points and use the point-slope form to construct the equation of the line. Provide your answer in the format y - y1 = m(x - x1), clearly stating all the steps and values used in your calculation.

Figure 9: Example of Generated Instructions utilizing a fused syllabi in Mathematics domain

Examples of Generated Instructions based on the Number of Fused Syllabus in Mathematics

 # 0
In a regular octagon, three vertices are chosen at random. Using concepts from geometric probability, explain how you would calculate the probability that the triangle formed by these vertices includes at least one side that coincides with a side of the octagon. Additionally, discuss any symmetries in the octagon that simplify this calculation.

 # 2
In a geometry problem involving a circle, if a tangent line intersects the circle at point P and a radius drawn to point P measures 4 cm, forming a right angle with the tangent, and you are given that another angle within the circle, adjacent to the radius but not directly involving the tangent, measures 60 degrees, use the properties of circles and angles to fill-in-the-blank: The measure of the angle formed between the tangent at point P and a line segment extending from the center of the circle through point P to an external point Q is ___ degrees.

 # 4
In a small town, two weather stations, A and B, independently record daily temperatures in degrees Celsius. Station A has a daily temperature variation with a mean of 20°C and a variance of 9, while Station B has a mean of 22°C and a variance of 16. On a particular day, the combined temperature measurement used by the town hall is calculated as the sum of the temperatures recorded by both stations. Given this setup:\n\n1. Calculate the expected combined temperature (mean) for that day.\n2. Determine the variance of the combined temperature measurement.\n3. If the town wants to ensure that the combined temperature rarely exceeds 45°C, calculate the probability that the combined temperature will be greater than 45°C, assuming temperatures follow a normal distribution.\n\nUse your understanding of combining random variables and their variances from AP®/College Statistics to address these points.

Figure 10: Example of instructions generated per fusion count in the field of mathematics. #n where n signifies the number of combined syllabi.

Example of MANTA from Fine-web Code

Raw Corpus

Message: session_start(): Server 18.104.22.168 (tcp 11211) failed with: Connection refused (111)\nLine Number: 18\nWe just sent you an email, containing instructions for how to reset your password.\nI am a Cowbird citizen. I've told 17 stories. I joined Cowbird on September 29, 2014.\nJust in the point of boiling

Extracted Syllabus

```
{
    "Domain": "Computer Science and Coding",
    "Subject": "Computer Science with PHP",
    "Unit": "Web Development",
    "Lesson": "Session Management and Error Handling",
    "Class Session": "Troubleshooting Server Connection Issues",
    "Key Concept": "Session Initialization and Server Communication"
}
```

Generated Query

Given a PHP web application that experiences intermittent session initialization failures due to server communication issues, complete the following code snippet to enhance error handling and ensure proper session management. Specifically, implement a mechanism to log errors related to session start failures and attempt reconnection a specified number of times before giving an error message to the user. Use session_start() and appropriate error logging functions.

Figure 11: Example of MANTA pipeline from Fine web Code Corpus.

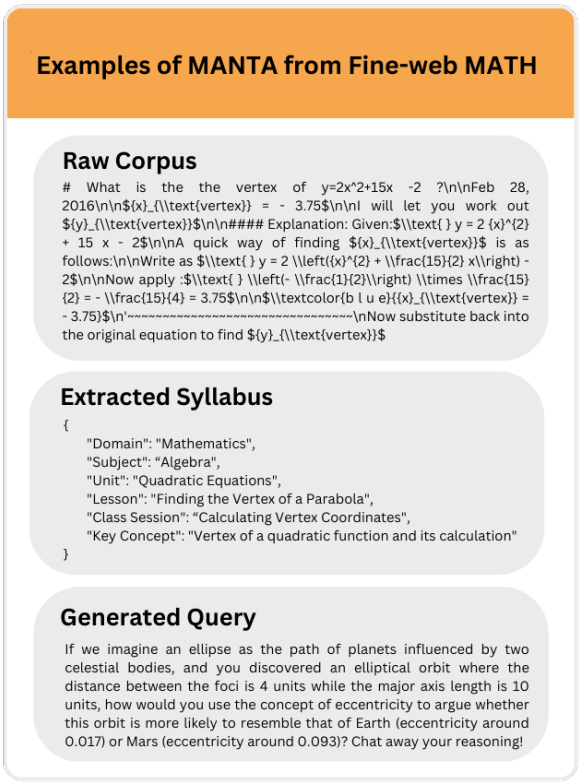


Figure 12: Example of MANTA pipeline from Fine Web MATH Corpus.

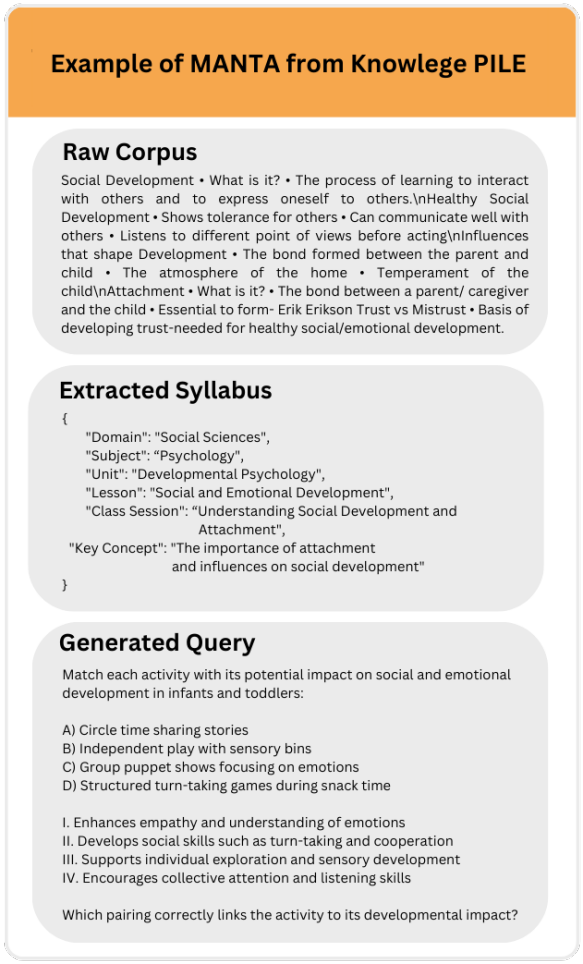


Figure 13: Example of MANTA pipeline from Knowledge PILE.