

Easy Dataset: A Unified and Extensible Framework for Synthesizing LLM Fine-Tuning Data from Unstructured Documents

Ziyang Miao¹, Qiyu Sun¹, Jingyuan Wang¹, Yuchen Gong¹,
Yaowei Zheng¹, Shiqi Li^{2*}, Richong Zhang^{1†}

¹School of Computer Science and Engineering, Beihang University, China

²Independent Researcher

Open-source repository: <https://github.com/ConardLi/easy-dataset>

Demonstration video: <https://youtu.be/HlyvdE1ASrk>

Abstract

Large language models (LLMs) have shown impressive performance on general-purpose tasks, yet adapting them to specific domains remains challenging due to the scarcity of high-quality domain data. Existing data synthesis tools often struggle to extract reliable fine-tuning data from heterogeneous documents effectively. To address this limitation, we propose Easy Dataset, a unified framework for synthesizing fine-tuning data from unstructured documents via an intuitive graphical user interface (GUI). Specifically, Easy Dataset allows users to easily configure text extraction models and chunking strategies to transform raw documents into coherent text chunks. It then leverages a persona-driven prompting approach to generate diverse question-answer pairs using public-available LLMs. Throughout the pipeline, a human-in-the-loop visual interface facilitates the review and refinement of intermediate outputs to ensure data quality. Experiments on a financial question-answering task show that fine-tuning LLMs on the synthesized dataset significantly improves domain-specific performance while preserving general knowledge. The source code and installable package are available at <https://github.com/ConardLi/easy-dataset> and have garnered over 10,000 GitHub stars.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across general-purpose applications (Achiam et al., 2023; Yang et al., 2025), leading to their widespread deployment in both academia and industry. Nevertheless, as real-world applications increasingly demand domain-specific knowledge, effectively adapting LLMs to specialized scenarios remains a persistent challenge. Among existing approaches, supervised fine-tuning (Ouyang et al., 2022; Taori et al., 2023) has

emerged as a practical and effective method for domain adaptation, enabling models to internalize domain-specific knowledge through learning from curated labeled data. Yet the effectiveness of fine-tuning depends on the availability of large-scale, high-quality datasets tailored to the target domain. In practice, constructing such datasets manually is often costly and time-consuming, particularly in domains that require expert annotations. As a result, automated data synthesis (Wang et al., 2024b; Tan et al., 2024) has gained increasing attention as an attractive alternative to reduce human effort while preserving data diversity and domain relevance.

Despite the promise of automated data synthesis, implementing it in practice remains non-trivial due to two primary challenges: effectively parsing heterogeneous and noisy source documents and generating large-scale, high-quality supervised data. The first challenge stems from the structural complexity of real-world documents, which often contain a combination of unstructured and semi-structured elements, including free text, tables, and figures. As a result, standard parsing methods often struggle to handle this variability robustly and consistently (Zhang et al., 2024), resulting in incomplete or inaccurate information extraction. The second challenge involves generating diverse and faithful question-answer (QA) pairs that accurately capture the underlying domain knowledge. Simply reusing or repeating synthesized QA pairs can lead to overfitting and degrade downstream performance after fine-tuning. Therefore, an effective data augmentation strategy is crucial to generate varied examples while preserving semantic correctness and consistency with domain-specific content.

Recent studies have investigated automated data synthesis methods, such as constructing fine-tuning datasets based on predefined topics (Chesterfield Laboratories Inc., 2025) or problem sets (Canto et al., 2024; Marten et al., 2025). However, these tools generally lack reliable capabilities for parsing

*Corresponding author: 1009903985@qq.com

†Corresponding author: zhangrc@act.buaa.edu.cn

Table 1: Comparison of existing synthetic data generation tools.

Feature	Distilabel	Kiln	Curator	Data-Juicer	GraphGen	Easy Dataset
Adaptive Parsing	✓		✓	✓	✓	✓
Hybrid Chunking			✓	✓	✓	✓
Persona-Driven Data Synthesis						✓
Human-In-The-Loop		✓				✓
End-to-End GUI		✓		✓	✓	✓

unstructured documents to extract domain-specific data. [Chen et al. \(2024a,b\)](#) introduce various document processing techniques, but they fall short of yielding high-quality QA pairs that are directly suitable for fine-tuning. Consequently, there remains a clear need for an end-to-end solution for synthesizing fine-tuning data from raw documents to effectively improve the performance of LLMs in domain-specific applications.

We introduce Easy Dataset, a unified framework for synthesizing fine-tuning data from unstructured documents through an intuitive graphical user interface (GUI). It comprises two main components: **adaptive document processing** and **persona-driven data synthesis**. To transform raw documents into coherent text chunks, we integrate a suite of text extraction models, enabling robust parsing across various formats. The extracted textual content is subsequently segmented into semantically coherent chunks using a hybrid chunking strategy. To construct high-quality datasets from these chunks, we adopt a persona-driven prompting approach that generates diverse and large-scale question-answer (QA) pairs by leveraging publicly available LLMs. Importantly, the GUI facilitates human-in-the-loop refinement, allowing users to review and improve the quality of the synthesized data. It further allows users to execute the entire data synthesis workflow without writing any code, greatly enhancing usability for non-technical users.

Our contributions are summarized as follows:

- We introduce Easy Dataset, an extensible end-to-end framework that unifies adaptive document processing and persona-driven data synthesis to automate the curation of high-quality fine-tuning data from unstructured documents, while minimizing manual effort.
- We develop an intuitive visual interface that enhances accessibility for non-technical users and facilitates efficient human-in-the-loop refinement to ensure data quality.

- Empirical results on a financial QA task show that fine-tuning LLMs on the synthesized data significantly improves domain-specific performance while retaining general knowledge.

2 Related Work

In recent years, numerous systems have been proposed to facilitate automated synthetic data generation. Distilabel ([Canto et al., 2024](#)) provides a modular framework with practical components for generating various types of synthetic data, including instruction-tuning and preference datasets. Kiln ([Chesterfield Laboratories Inc., 2025](#)) features a GUI-based platform that supports zero-shot data generation, topic tree construction, and reasoning data creation, along with an interactive interface for human review and rating. Curator ([Marten et al., 2025](#)) facilitates the synthesis of diverse data types, including reasoning, code execution, chart generation, and function-calling data, to support fine-tuning across a broad spectrum of downstream tasks. Data-Juicer ([Chen et al., 2024a,b](#)) supports multiple user interfaces and provides a rich set of data processing operators, making it highly customizable. GraphGen ([Chen et al., 2025](#)) leverages knowledge graphs to generate question-answer pairs from source documents through a visual interface. Despite their contributions, these tools still exhibit several notable limitations (Table 1). Many lack support for human-in-the-loop interaction, hindering opportunities for iterative quality refinement and alignment with user preferences. Others do not support synthesizing fine-tuning data based on source documents, making it difficult to efficiently construct domain-specific datasets. Moreover, few provide a fully end-to-end graphical user interface (GUI) that integrates the entire data synthesis pipeline, resulting in steep learning curves and limited usability for users with diverse technical backgrounds. These gaps underscore the need for a unified, accessible solution.

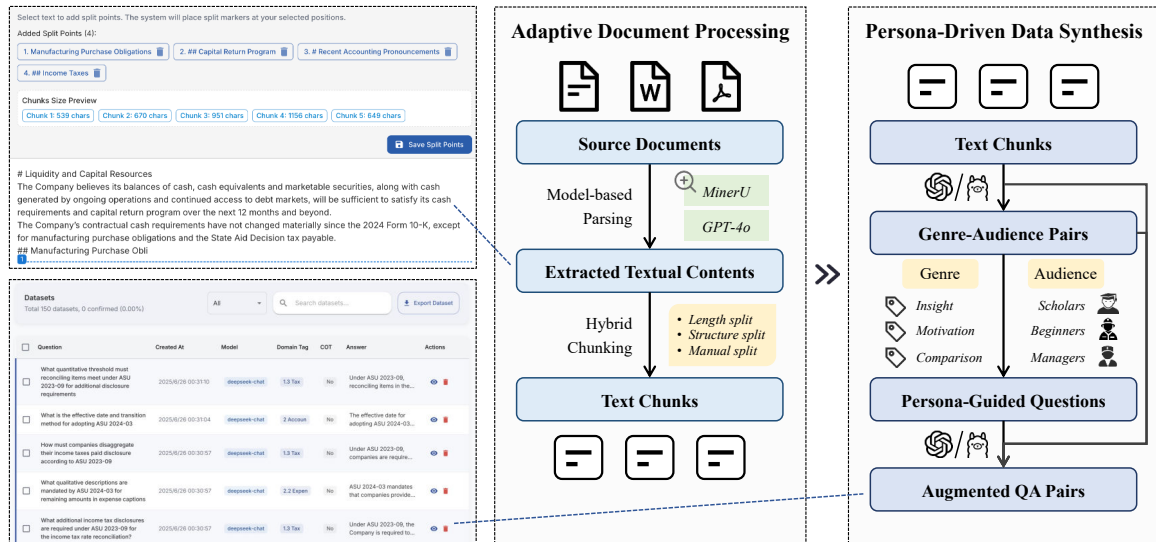


Figure 1: Overview of the Easy Dataset framework. The framework consists of two primary components: adaptive document processing and persona-driven data synthesis. In the first component, documents of different formats are processed via model-based parsing, followed by hybrid chunking to produce text chunks. In the second component, *Genre-Audience* pairs are created for each document to guide the construction of diverse question-answer pairs. Persona-guided questions are then generated to further increase diversity, and final augmented QA pairs are synthesized via knowledge-enhanced prompting to ensure factual consistency. The entire framework provides interactive human-in-the-loop refinement to maintain data quality.

3 Framework

We present Easy Dataset, a unified and extensible framework designed to synthesize high-quality fine-tuning datasets from unstructured documents. As illustrated in Figure 1, the framework implements a visual human-in-the-loop pipeline that enables interactive refinement at all stages of dataset construction. It begins with adaptive document processing, leveraging vision-language models (VLMs) to extract textual content from raw documents. The extracted text is then segmented using a hybrid chunking strategy that combines length-based, structure-based, and manual splitting. In the QA generation stage, Easy Dataset supports both naive and persona-driven data synthesis to generate stylistically diverse and semantically grounded QA pairs. A modular model configuration interface enables flexible access to both API-based and locally deployed LLMs and offers fine-grained control over generation parameters. Finally, the framework supports export of datasets in multiple formats and is natively compatible with fine-tuning frameworks such as LlamaFactory (Zheng et al., 2024).

3.1 Adaptive Document Processing

Document Parsing Our framework supports extracting textual content from unstructured documents of various formats, including plain text,

Markdown, DOCX, and PDF, which is essential for constructing high-quality datasets for downstream tasks. For plain text and Markdown documents, the original content is retained with minimal processing to preserve inherent semantics. In contrast, DOCX documents often lack explicit structure, necessitating preprocessing to extract meaningful text. To this end, we employ the lightweight Mammoth library (Williamson, 2015) to convert DOCX files into Markdown, which helps preserve the original semantics while avoiding unnecessary formatting noise. For PDF documents with relatively simple layouts, we use the pdf2md tool (Zillmann, 2017) to extract textual content efficiently. However, for PDFs with complex or mixed text-image compositions, rule-based methods often result in significant information loss. To address this limitation, we first apply layout analysis to automatically detect content regions within the document. Once the layout structure is established, textual regions are extracted directly, while visual regions are parsed using a vision-language model (VLM), which can accurately capture the content. This strategy ensures accurate text extraction and preserves key information, providing a reliable foundation for subsequent downstream data processing. In addition, our framework integrates with state-of-the-art PDF processing tools, including MinerU (Wang

et al., 2024a; He et al., 2024), providing users with greater flexibility to handle documents of varying complexity and scale.

Text Chunking To enable precise control over data sampling and ensure compatibility with existing LLM context windows, we propose *Hybrid-Chunking*, a structure-aware and adaptive preprocessing strategy that flexibly segments source documents into coherent text chunks. It allows users to configure chunk sizes and customize text delimiters, enabling adaptation to various content types, such as plain text, code snippets, or tabular data. The process begins with initial coarse-grained segmentation based on line breaks, followed by a hybrid splitting-and-merging routine: lengthy chunks are recursively split using user-defined delimiters, while neighboring short segments are merged if necessary to meet length constraints without breaking semantic units. Furthermore, for edge cases where automated rules may fail to work optimally, we provide a visual text-chunking interface that enables users to perform fine-grained manual adjustments, ensuring precise text segmentation. Users can flexibly tailor chunking policies and thresholds according to specific document content. This hybrid design achieves a balance between automation and user control, significantly improving the consistency and reliability of the resulting text chunks.

3.2 Persona-Driven Data Synthesis

Easy Dataset provides a flexible pipeline for generating question-answer (QA) pairs from text chunks. The process is highly configurable, supporting two generation modes: naive QA-pair generation and persona-driven QA-pair generation, the latter enabling the synthesis of diverse and stylistically controlled fine-tuning data.

Question Generation Based on the text chunks, questions are automatically constructed by prompting the LLMs. Specifically, each text chunk is combined with a customizable system prompt before being passed to the LLM for question generation. This system prompt enables fine-grained control over question style, target audience, and tone. For example, it can specify whether questions should be concise, elaborative, or directive in nature. To further improve robustness in real-world scenarios, we introduce a stochastic punctuation dropout mechanism. In particular, question marks are randomly removed to prevent the model from over-relying on punctuation cues, thereby enhancing the

generalization performance of models fine-tuned on the synthesized QA dataset.

Answer Generation Based on the generated questions, we derive high-quality answers that remain faithful to the source content. We employ a knowledge-enhanced prompting strategy, where the prompt contains the question and its corresponding source content. This ensures that the LLM-generated answers remain semantically aligned with the source content, factually consistent, and relevant to the overall domain context. The generation style is configurable to meet task-specific requirements, such as producing concise, detailed, or explanatory answers, allowing users to tailor the output style and format as needed. When using reasoning LLMs with chain-of-thought (CoT) capabilities, such as DeepSeek-R1 (Guo et al., 2025), the intermediate reasoning steps are also included in the QA pairs to provide transparency and support more interpretable fine-tuning. This substantially improves interpretability and facilitates downstream error analysis. To ensure the quality and reliability of the generated QA pairs, we provide a post-generation refinement interface that allows users to manually review, edit, and verify answers. In addition, LLMs are used to automatically refine both final answers and their corresponding CoT reasoning traces, serving as an additional step to improve overall robustness and factual accuracy.

Persona-Driven QA-Pair Generation Ge et al. (2024) first introduced the persona-driven data synthesis approach, demonstrating that prompting LLMs with different communicative roles (personas) enables the generation of stylistically diverse outputs, which can significantly increase dataset size. Building on this insight, we integrate persona-driven data synthesis into our Easy Dataset framework via a two-stage pipeline inspired by the Massive Genre-Audience (MGA) method (Hao et al., 2025), which provides a systematic way to diversify output styles. Specifically, the pipeline first automatically generates diverse Genre-Audience (GA) pairs to guide stylistic conditioning and then produces multiple QA pairs conditioned on these personas, enriching the dataset with varied perspectives and tones. While MGA was originally developed to expand pretraining corpora by reformulating documents and reducing data redundancy, we adapt its core principles for QA pair synthesis in Easy Dataset to boost both the diversity and scale of the generated data for domain-specific fine-tuning.

Table 2: Persona-driven QA pairs synthesized by Easy Dataset. By varying the GA pair, our framework generates a diverse set of data examples that maintain high fidelity to the source documents.

Example 1: Financial News Summary for Executives and Business Readers
<p>Genre: <i>Financial News Summary</i>. This genre condenses complex financial information into concise, accessible updates. It highlights key points like liquidity status, capital returns, and regulatory changes, making it suitable for readers who need quick, high-level insights without deep technical details.</p> <p>Audience: <i>Busy Executives and General Business Readers</i>. Professionals who need to stay informed about financial developments but lack the time for in-depth analysis. They prefer succinct summaries that provide the essential takeaways without overwhelming detail.</p>
<p>Generated Question: <i>When is the Company expected to adopt ASU 2023-09, and what transition method will be used?</i></p>
<p>Generated Answer: The Company will adopt ASU 2023-09 in its fourth quarter of 2026 using a prospective transition method. This update will enhance income tax disclosures by requiring additional details in the tax rate reconciliation and disaggregating taxes paid by federal, state, and foreign jurisdictions.</p>
Example 2: In-depth Policy Discussion for Professionals and Policy Makers
<p>Genre: <i>Tax and Accounting Policy Discussion</i>. This genre focuses on the implications of new tax and accounting policies, such as FASB updates. It explores how these changes affect corporate reporting, tax strategies, and financial transparency, often inviting debate and expert opinions.</p> <p>Audience: <i>Tax Professionals and Policy Makers</i>. Experts in tax law and accounting standards who need to understand the practical and strategic impacts of regulatory changes. They are interested in how policies like ASU 2023-09 influence corporate behavior and financial disclosures.</p>
<p>Generated Question: <i>How does ASU 2023-09 require companies to disaggregate income taxes paid by jurisdiction?</i></p>
<p>Generated Answer: ASU 2023-09 mandates that companies disaggregate their income taxes paid disclosure by federal, state, and foreign taxes. Additionally, for foreign taxes, further disaggregation is required for significant individual jurisdictions. This enhanced disclosure framework aims to provide greater transparency into the geographic distribution of tax payments, enabling stakeholders to better assess the company’s tax profile and jurisdictional exposures. (<i>truncated</i>)</p>

Our adapted persona-driven pipeline consists of two stages for synthesizing diverse, persona-guided QA pairs from unstructured documents. In the **Persona Synthesis Stage**, for each source document, we first leverage an LLM to automatically generate a set of unique (*Genre, Audience*) pairs based on the content. Each pair jointly defines a persona for the generation task. Specifically, *Genre* characterizes the inquisitive intent and dialogue style, while *Audience* profiles the cognitive state and knowledge background of the questioner. For example, a (Motivation, Beginners) persona guides the model to produce simple, encouragement-oriented questions that help novices build confidence. By introducing diverse Genre-Audience combinations, the synthesis process enables more comprehensive coverage of the original content from multiple perspectives. To enhance flexibility, users can also manually specify or refine GA pairs to better target specific domains or tasks. In the **Persona-Guided QA Generation Stage**, the synthesized personas guide the LLMs to generate diverse questions based on the text chunks from various perspectives. For each generated question, the model subsequently produces an answer conditioned on the question, the corresponding source text chunk, and the associated persona. This yields an augmented QA pair in which both the question and the answer are semantically grounded in the original content and stylistically consistent with the intended persona. This adaptation effectively transforms the original MGA

method from a pretraining augmentation strategy into a powerful framework for synthesizing diverse, persona-guided fine-tuning data. As illustrated in Table 2, this persona-driven approach enables the synthesis of diverse and faithful QA pairs from a single source document, thereby enhancing the effective utilization of the original source corpora.

3.3 Model Configuration

To seamlessly integrate LLMs into the data generation pipeline, we propose a flexible model configuration module. This module allows users to configure and manage multiple LLMs for data synthesis through an intuitive visual interface with minimal setup effort. By specifying the model provider, API endpoint, API key, and model name, users can easily utilize their target models. The module also supports locally deployed models via Ollama (Ollama, 2023), ensuring flexibility for different deployment scenarios. Additionally, it provides fine-grained control over generation parameters (e.g., temperature and top-p sampling), enabling users to adapt the generation process to diverse data requirements and domain-specific constraints.

3.4 Dataset Export

Easy Dataset enables export of the generated QA pairs as standard dataset files in formats such as JSON, JSONL, and CSV. It also supports widely adopted data schemas such as Alpaca (Taori et al., 2023) and ShareGPT (Zheng et al., 2023). Users

Table 3: Performance comparison of the Qwen2.5-7B-Instruct model before and after fine-tuning with naive and persona-driven data synthesis on general-purpose and domain-specific benchmarks. **Bold** indicates the best result, and underline indicates the second-best.

Method	MMLU	CMMLU	HellaSwag	MATH	HumanEval	Avg.	Domain Knowledge
Base Model	<u>62.0</u>	77.2	82.9	74.8	84.8	76.3	3.2
Naive Data Synthesis	60.2	78.4	<u>80.8</u>	72.9	80.5	74.6	<u>57.0</u>
Persona-Driven Data Synthesis	63.7	<u>77.7</u>	<u>80.8</u>	<u>74.1</u>	<u>81.1</u>	<u>75.5</u>	59.6

can define custom export templates by specifying key fields such as question, answer, reasoning steps, and domain labels, allowing them to flexibly adapt outputs to diverse task-specific fine-tuning pipelines and community standards. Additionally, Easy Dataset offers seamless integration with the LlamaFactory training framework (Zheng et al., 2024) by automatically generating a compatible configuration file. Users can simply specify the path to this configuration file to enable direct use in LlamaFactory, eliminating manual setup and significantly lowering the barriers to fine-tuning.

4 Evaluation

4.1 Experimental Setup

To evaluate the quality of data synthesized by Easy Dataset, we collected five up-to-date (later than the knowledge cutoff) financial reports from public online sources as a representative domain example. We also built a domain-specific evaluation dataset consisting of 100 questions derived from the source documents. Then we used Easy Dataset to generate a training dataset and fine-tuned the Qwen2.5-7B-Instruct (Yang et al., 2024) model using the LlamaFactory framework. We compared the model’s performance before and after fine-tuning using either the naive or persona-driven data synthesis pipeline. We evaluated the model on the domain-specific dataset and general-purpose benchmarks, including MMLU (Hendrycks et al., 2021), CMMLU (Li et al., 2024), HellaSwag (Zellers et al., 2019), MATH (Lightman et al., 2024), and HumanEval (Chen et al., 2021). For domain-specific evaluation, we employed the LLM-as-a-judge approach (Zheng et al., 2023) using the DeepSeek-V3 API (Liu et al., 2024) to provide reliable scores, with details provided in Appendix A. Other fine-tuning details are provided in Appendix B.

4.2 Results

Table 3 shows the evaluation results of Qwen2.5-7B-Instruct on several general-purpose benchmarks

and the domain-specific evaluation dataset, comparing the base model with its fine-tuned variants trained using either the naive or persona-driven data synthesis methods. On general-purpose benchmarks, we observe that fine-tuning on domain-specific data preserves the model’s general language capabilities, demonstrating that the synthetic dataset produced by Easy Dataset enables the model to acquire domain-specific knowledge while maintaining robust generalization. Notably, the persona-driven variant achieves the best results on the MMLU benchmark and competitive performance across most tasks, even outperforming the naive variant on several challenging benchmarks, indicating its potential for improving generalization through stylistic and semantic diversity. On the domain-specific evaluation dataset, both fine-tuned models demonstrate substantial gains over the base model, which performs poorly without exposure to up-to-date financial documents. Specifically, fine-tuning with the dataset generated by Easy Dataset using the naive synthesis pipeline achieves a score of 57.0, while incorporating persona-driven data synthesis further improves the score to 59.6. These results highlight the effectiveness of Easy Dataset in injecting timely, domain-specific knowledge while preserving the model’s general capabilities.

5 Conclusion and Future Work

We introduced Easy Dataset, an end-to-end framework that automates the synthesis of domain-specific fine-tuning datasets. With human-in-the-loop control and an intuitive interface, it enables efficient generation of datasets with high quality, diversity, and usability. Empirical results demonstrate that Easy Dataset significantly improves domain adaptation for LLMs while preserving general capabilities and robustness. We plan to extend Easy Dataset in several directions, including supporting broader modalities (e.g., SQL, tables, multi-modal), integrating automatic quality monitoring, and developing advanced enrichment strategies to further enhance data variety and fidelity.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Álvaro Bartolomé Del Canto, Gabriel Martín Blázquez, Agustín Piqueres Lajarín, and Daniel Vila Suero. 2024. Distilabel: An ai feedback (aif) framework for building datasets with and for llms. <https://github.com/argilla-io/distilabel>.
- Daoyuan Chen, Yilun Huang, Zhijian Ma, Heseng Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024a. Data-juicer: A one-stop data processing system for large language models. In *International Conference on Management of Data*.
- Daoyuan Chen, Yilun Huang, Xuchen Pan, Nana Jiang, Haibin Wang, Ce Ge, Yushuo Chen, Wenhao Zhang, Zhijian Ma, Yilei Zhang, Jun Huang, Wei Lin, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024b. Data-juicer 2.0: Cloud-scale adaptive data processing for foundation models. *arXiv preprint arXiv:2501.14755*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Zihong Chen, Wanli Jiang, Jinzhe Li, Zhonghang Yuan, Huanjun Kong, Wanli Ouyang, and Nanqing Dong. 2025. Graphgen: Enhancing supervised fine-tuning for llms with knowledge-driven synthetic data generation. *arXiv preprint arXiv:2505.20416*.
- Chesterfield Laboratories Inc. 2025. Kiln: Rapid ai prototyping and dataset collaboration tool. <https://github.com/Kiln-AI/Kiln>.
- Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Xintong Hao, Ruijie Zhu, Ge Zhang, Ke Shen, and Chenggang Li. 2025. Reformulation for pretraining data augmentation. *arXiv preprint arXiv:2502.04235*.
- Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets. *arXiv preprint arXiv:2407.13773*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.
- 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*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Ryan Marten, Trung Vu, Charlie Cheng-Jie Ji, Kartik Sharma, Shreyas Pimpalgaonkar, Alex Dimakis, and Maheswaran Sathiamoorthy. 2025. Curator: A tool for synthetic data creation. <https://github.com/bespokelabsai/curator>.
- Ollama. 2023. Ollama: Run large language models locally. <https://github.com/ollama/ollama>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.
- Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and 1 others. 2024b. A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

- Michael Williamson. 2015. Mammoth .docx to html converter. <https://github.com/mwilliamson/mammoth.js>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. **HellaSwag: Can a machine really finish your sentence?** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Conghui He, and Wentao Zhang. 2024. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. **LlamaFactory: Unified efficient fine-tuning of 100+ language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Johannes Zillmann. 2017. Pdf-to-markdown converter. <https://github.com/jzillmann/pdf-to-markdown>.

A LLM-as-a-Judge Prompt

Please act as an impartial evaluator and assess the quality of the AI assistant's response to the user's question. You will be provided with the following information:

1. The original user question (Question)
2. A standard answer containing information directly relevant to the user's question (Ground Truth)
3. The AI assistant's response (Prediction)

Your task is to conduct a thorough evaluation focusing on correctness, scoring from 0 to 5 points.

Evaluation Method:

1. Carefully read the question, the assistant's response, and the ground truth answer.
2. Identify and list all key factual statements from the ground truth.
3. For each fact, determine whether it is correctly reflected in the assistant's response.
4. Assign a final correctness score based on the degree of fact matching. If all facts from the ground truth are correctly reflected in the AI response, assign 5 points. If none are correct, assign 0 points.

Please carefully analyze the correctness of the answer. Finally, provide the score result in JSON format as follows:

```
[
  {
    "correctness": "3"
  }
]
```

Question

```
{ Question }
```

Prediction

```
{ Prediction }
```

Ground Truth

```
{ Ground Truth }
```

B Fine-Tuning Details

All the experiments were conducted on 2 NVIDIA A800 80GB GPUs. For all fine-tuning runs, we used a batch size of 64 and a total of 10,000 training samples. The model was trained for 2 epochs with a learning rate of 10^{-5} using a cosine learning rate schedule. A warmup ratio of 0.1 was applied to stabilize training and mitigate overfitting.