




YIZHAO: A 2TB Open Financial Dataset

Harbin Institute of Technology (Shenzhen)
China Merchants Bank Artificial Intelligence Laboratory

Abstract

Large-scale, high-quality pretraining data is crucial for training large language models. However, so far there are few datasets available for specialized critical domains such as finance, and the available ones are often small and have low variety. To fill in this gap, we present **YIZHAO**, the largest open-source financial dataset covering both English and Chinese. YIZHAO includes not only bilingual text data from web pages, but also synthetic multimodal question-answering pairs about PDF files and images. Our dataset has undergone a thorough cleaning and deduplication process via a comprehensive human-in-the-loop pipeline to ensure high quality and content safety, along with our proposed financial and safety filtering models. Extensive human annotation has verified that the proportion of toxic text is kept below 0.5% in the English subset and under 2% in the Chinese subset. Finally, we open-source our data curation toolkit to enable reproduction of our work as well as support further research in scaling up the financial data.

 hf.co/datasets/HIT-TMG/YiZhao
 github.com/HITsz-TMG/YiZhao
 modelscope.cn/CMB_AILab/YiZhao-12B

1 Introduction

Recently large language models (LLMs) have achieved significant advances not only in numerous natural language processing tasks, but also reveals new emergent abilities, achieving high scores in professional exams tailored for human (Wei et al., 2022). Notable research outcomes include GPT-4o (OpenAI, 2024), Gemini (Anil et al., 2023), LLaMa (Touvron et al., 2023) and Qwen (Bai et al., 2023). As the LLM scaling law underscores the importance of utilizing a substantial token count during the pretraining phase (Hoffmann et al., 2024), recent LLMs are typically trained on increasingly large corpora (e.g., from 1T tokens in LLaMa-1 to

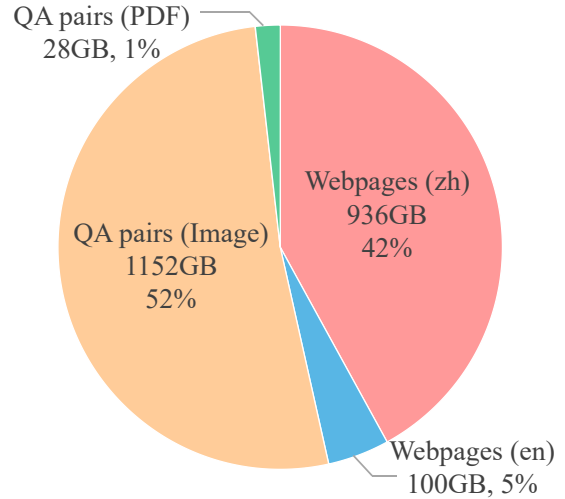


Figure 1: YIZHAO Source Distribution.

15T tokens in LLaMa-3). Meanwhile, the open-source community has provided multiple large-scale datasets to support this growth, such as RedPajama (Weber et al., 2024), RefinedWeb (Penedo et al., 2023), Dolma (Soldaini et al., 2024) and CulturaX (Nguyen et al., 2024).

However, there is a scarcity of large-scale, domain-specific pretraining datasets in areas such as finance. The few datasets available are limited in scale and diversity (see Table 1). Moreover, the sensitive nature and regulatory constraints about financial data present significant challenges to data collection and utilization. To this end, we have curated and released a 2TB financial dataset, **YIZHAO**, under ODC-By license¹. As shown in Figure 1, our dataset includes 936GB of Chinese and 100GB of English text data from web pages. In constructing the YIZHAO dataset, we ensure data safety, high quality and value alignment (by filtering out toxic text) through a comprehensive cleaning chain and manual verification. This process includes hand-crafted rules, model-generated quality scores for

¹<https://opendatacommons.org/licenses/by/1-0/>

Dataset	Size	Description	Language	Format
FinCorpus	60GB	Company announcements, financial news and exams	zh	Text
SEC	21GB	SEC annual reports	en	Text
WTO	88GB	Official documents from WTO	en,fr,es	Text, PDF
AMF	349GB	Official publications of AMF	en,fr	Text, PDF
TEDEUTenders	1.6GB	Procurement notices published by EU	multilingual	Text
GATT_library	0.75GB	General Agreement on Tariffs and Trade	en,fr	Text
YIZHAO (ours)	2TB	Web text and multimodal QA pairs	zh, en	Text, Image, PDF

Table 1: Comparison of existing financial corpora.

filtering, and a strict deduplication procedure. Each text is associated with a finance-related score and a risk score, assigned by our financial and safety filtering models. Our financial classifier can filter and categorize financial content with high F1 scores (82.7% for the Chinese dataset and 86% for the English dataset). Our safety filtering model can identify harmful data with a high recall rate of 92%. Extensive human verification has shown that the ratio of toxic text in YIZHAO is below 0.5% in the English subset and below 2% in the Chinese subset.

In addition, drawing upon conclusions from Maini et al. (2024) and Cheng et al. (2024), synthetic question-answering (QA) pairs about the domain-specific text could be beneficial for data-efficient adaptation into specialized domains. We take the cleaned text with the highest quality from the previous step and synthesize QA pairs. Furthermore, we generate images and PDFs based on these texts to create a multimodal dataset, which could be useful for multimodal financial document QA and retrieval-augmented generation.

In summary, our contributions are two-fold:

- We release **YIZHAO, a 2TB open-source financial dataset**, which is characterized by its high quality, strong content safety, and finance-focused nature, covering both bilingual web pages and multimodal question-answering pairs.
- We open-source the **cleaning tools and filtering models** used in the comprehensive human-in-the-loop cleaning pipeline of YIZHAO, facilitating the reproduction of our work and supporting the expansion of financial datasets.

2 Related Work

General Pretraining Corpora. As the quantity, quality, and diversity of pretraining data play a critical role in determining a language model’s capabilities (Rae et al., 2021), many prior efforts have been made to curate, document and release open corpora. A widely adopted choice is to use web text, especially Common Crawl², a publicly available and continually updated collection of website snapshots running since 2007. Well-curated general pretraining corpora derived from Common Crawl include C4 (Raffel et al., 2020), OSCAR (Ortiz Suárez et al., 2020), RefinedWeb (Penedo et al., 2023), RedPajama-v2 (Weber et al., 2024), ChineseWebText (Chen et al. (2023); Zhang et al. (2024)), CulturaX (Nguyen et al., 2024), WanJuan-CC (Qiu et al., 2024), and FineWeb (Penedo et al., 2024).

Apart from web pages, to enhance the diversity of pretraining corpora, Gao et al. (2020) introduce the Pile, a comprehensive collection of 22 diverse and high-quality datasets. The ROOTS dataset (Laurençon et al., 2023) aggregates content from hundreds of sources across 59 languages. Kocetkov et al. (2023) build the Stack, a 3.1TB code dataset in 30 programming languages. RedPajama-v1 (Weber et al., 2024) replicates LLaMA-1’s pretraining data composition, including Wikipedia, arXiv, books and StackExchange. SlimPajama (Shen et al., 2023) further conducts thorough deduplication based on RedPajama-v1. Soldaini et al. (2024) present Dolma, an open dataset of 3 trillion tokens from a diverse mix of web content, academic publications, code, books, and encyclopedic materials.

Financial Corpora. While open-source general pretraining datasets are large and widely avail-

²<https://commoncrawl.org/>

Text:

6月16日 盘前重要宏观新闻

创业板改革新规日前发布，深交所上市审核系统昨日正式开门迎客。资深专家在接受上证报记者采访时认为，此次创业板改革并不仅仅是推进注册制改革，而是与一揽子资本市场基础性制度改革协同推进。创业板改革将有利于完善多层次资本市场基础制度建设，进一步提升我国多层次资本市场高质量发展的水平。

5月主要指标集体回暖释放中国经济积极信号

5月份投资、消费和工业增加值等主要经济指标继续回暖，城镇调查失业率出现下降，表明经济复苏持续升温。多家机构预测，二季度经济增长有望出现明显反弹趋势，增速或重新转正。

多项指标5月转正房地产市场元气恢复

国家统计局昨日公布数据显示，1月至5月，全国房地产开发投资同比下降0.3%，降幅连续3个月收窄。其中，住宅投资3.4万亿元，已恢复到去年同期水平。

Translation:

Important Macroeconomic News Before Market on June 16

The new rules for the reform of the ChiNext Board were recently released, and the listing review system of the Shenzhen Stock Exchange officially opened to the public yesterday. Senior experts told reporters from the Shanghai Securities News that this reform of the ChiNext Board is not just about advancing the registration-based IPO system, but it is also advancing in coordination with a package of fundamental reforms in the capital market. The reform of the ChiNext Board will help improve the foundational systems of a multi-tiered capital market and further elevate the quality of China's multi-tiered capital market development.

Major Indicators in May Warm Up, Sending Positive Signals for China's Economy

In May, major economic indicators such as investment, consumption, and industrial added value continued to pick up, and the urban surveyed unemployment rate declined, indicating a continued warming of economic recovery. Several institutions predict that there is a clear rebound trend expected in the second-quarter economic growth, with the growth rate potentially turning positive again.

Multiple Indicators Turn Positive in May as Real Estate Market Recovers

Data released by the National Bureau of Statistics yesterday showed that from January to May, the national real estate development investment decreased by 0.3% year-on-year, with the decline narrowing for three consecutive months. Among them, residential investment reached 3.4 trillion yuan, returning to the level of the same period last year.

meta:

url: "http://3265185.com/news-1205.html",
title: "财惠赚",
fin_score: 5,
risk_score: 0.0138,
language: "zh"

Figure 2: An example document in YIZHAO from the Chinese webpage subset.

able, domain-specific datasets remain scarce (Wang et al., 2024; Niklaus et al., 2024). For the financial domain, as detailed in Table 1, most existing financial datasets are limited in size and focus primarily on official reports. Currently, the only two open-source financial dataset collections are Duxiaoman’s FinCorpus³ and PleIAs’s Financial Commons⁴, which include five subsets listed in Table 1. Such limitation hampers pretrained models’ ability to capture the complexities of financial markets (Lee et al., 2024), thereby constraining their effectiveness in Financial Time Series Analysis (including forecasting market trends, detecting anomalies, and classifying financial data) (Nie et al., 2024). To address these limitations, we intro-

duce YIZHAO, currently the largest open-source financial dataset, which sources data from web pages. YIZHAO also includes synthesized multimodal QA pairs about images and PDFs, providing a more comprehensive resource for the financial domain.

3 YIZHAO

YIZHAO is a 2TB open-source financial dataset, characterized by its high quality, reliable content safety, and finance-specific information. It covers bilingual web pages and multimodal question-answering (QA) pairs. Figure 2 provides an example of YIZHAO.

Although previous efforts have meticulously constructed a series of high-quality general pretraining corpora, they have only undergone preliminary preprocessing and do not fully meet the specific requirements of the financial domain. Therefore, we rigorously refine these datasets by employing a

³<https://huggingface.co/datasets/Duxiaoman-DI/FinCorpus>

⁴<https://huggingface.co/collections/PleIAs/finance-commons-66925e1095c7fa6e6828e26c>

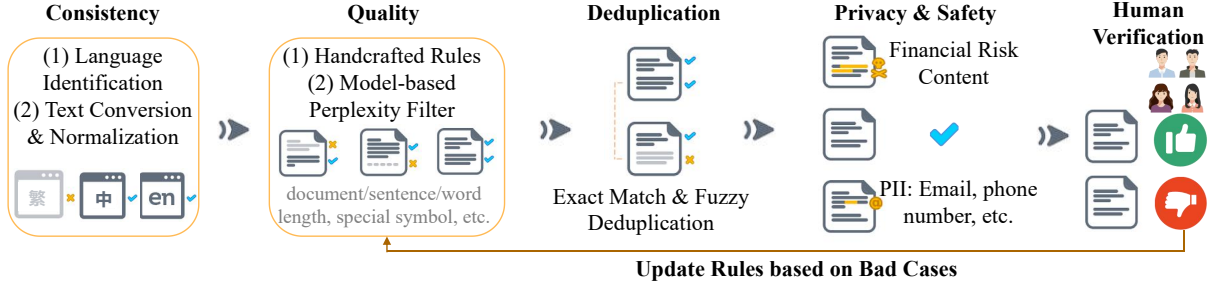


Figure 3: The cleaning pipeline of YIZHAO.

robust cleaning pipeline and trained classifiers to identify and extract high-quality, safe, and relevant financial information. Our data is sourced from the general corpus FineWeb-Edu (Penedo et al., 2024) and ChineseWebText (Chen et al., 2023).

In the following sections, we introduce the comprehensive human-in-the-loop cleaning pipeline in Section 3.1. We then elaborate on the annotation and training of safety and financial filters in Section 3.2 and 3.3 respectively. Finally we explain the data synthesis process for multimodal QA pairs in Section 3.4.

3.1 Data Cleaning

As illustrated in Figure 3, the data processing pipeline ensures the consistency, quality, deduplication, privacy and safety of YIZHAO. It incorporates commonly-used practices from C4 (Raffel et al., 2020), Gopher (Rae et al., 2021), Dolma (Soldaini et al., 2024) and RedPajama (Weber et al., 2024). These practices involve comprehensive handcrafted rules, model-based filtering, and a strict deduplication process.

Consistency To ensure consistency, we conduct a series of preprocessing steps: (1) **Language Identification**: We retain all Chinese and English texts by employing the fastText language classifier from CCNet (Wenzek et al., 2019) at the document level. The classifier utilizes character n-grams and is trained on Wikipedia, supporting 176 languages. Following Penedo et al. (2023), we exclude documents where the top language confidence score is below 0.65. This threshold typically corresponds to pages without any natural text. (2) **Text Conversion**: All traditional Chinese text is converted into simplified Chinese using OpenCC⁵ to standardize the dataset. (3) **Text Normalization**: We convert Unicode punctuations to their ASCII equivalents, normalize different kinds of whitespaces to a stan-

dard space character (0x20), and fix Unicode errors with ftfy⁶. The above steps collectively ensure that the dataset adheres to a consistent format.

Quality Here we introduce both handcrafted metrics and a model-based text quality evaluator to assess and filter text effectively.

Handcrafted rules. We filter text based on: (1) Quality signals indicating how *natural* a given piece of text is, including document length, average sentence length, maximum word length, the ratio of special symbols, and the ratio of short, consecutive, or incomplete lines; (2) Quality signals indicating how *repetitive* a given piece of text is. In alignment with the Gopher rules (Rae et al., 2021), we calculate the proportion of characters found in duplicated word n-grams and the percentage of characters within the most frequent word n-gram across the documents. During human-in-the-loop evaluation, human reviewers provide feedback by downvoting examples of poor-quality data. This process helps identify more problematic cases and create new handcrafted rules to address them.

Model-based filter. We use a model-based filter to handle complex cases that exceed the capabilities of standard heuristic rules. To filter out unnatural text, the *Perplexity Scorer* utilizes the KenLM library in CCNet (Wenzek et al., 2019). It evaluates a vast array of web documents and discards those with perplexity scores significantly higher than average, signaling low-quality or unnatural text. To avoid introducing further undesirable biases into the model (Welbl et al., 2021; Dodge et al., 2021; Penedo et al., 2023), we refrain from using other model-based quality filters, such as fastText classifier predictions to determine if a document is a Wikipedia article or a book (Du et al., 2022; Chowdhery et al., 2022). Throughout all quality filtering steps, except for language perplexity scoring, we adhere to simple rules and heuristics.

⁵<https://github.com/BYVoid/OpenCC>

⁶<https://github.com/rspeer/python-ftfy>

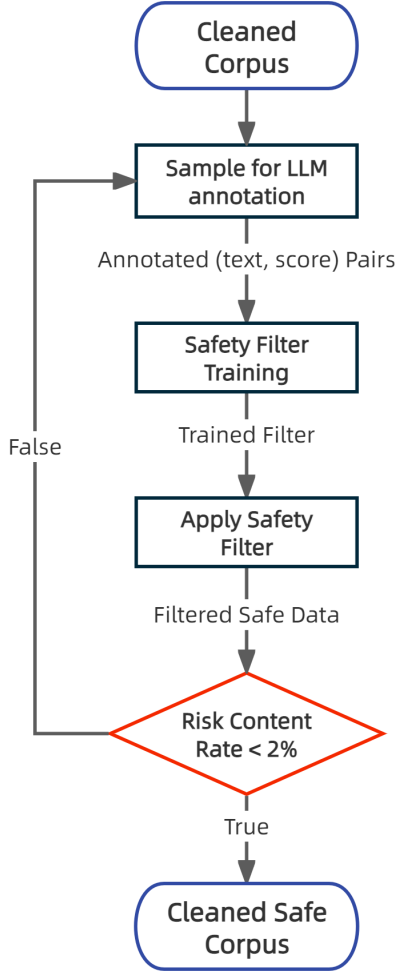


Figure 4: The human-in-the-loop development of the safety filter. The iterative loop ends only when the proportion of harmful data is below 2% during human verification.

Deduplication After filtering, although data quality improves, a large fraction of the content is repeated across documents. These duplicates can strongly impact models, favoring memorization instead of generalization (Xue et al., 2023; Lee et al., 2022). We implement a comprehensive deduplication pipeline following the procedure in RefinedWeb (Penedo et al., 2023) and ROOTS (Laurençon et al., 2023). This pipeline combines document-level fuzzy deduplication with sub-document exact-match deduplication to effectively identify and eliminate redundant content within and across documents. For fuzzy deduplication, We remove similar documents by applying MinHash (Broder, 1997), which excels at finding templated documents: licenses with only minor variations, placeholder text repeated across websites. For exact deduplication, using a suffix array (Manber and Myers, 1993), we identify and remove exact token-

by-token matches of sequences containing more than 50 consecutive tokens, following the implementation of Lee et al. (2022).

Privacy and Safety In addition to the quality filters, our cleaning pipeline addresses the presence of inappropriate content such as toxic language, adult material, and personally identifiable information (PII). We use curated lists of flagged words to calculate their ratio within a document and discard the documents with excessively high ratios. We also eliminate or replace sensitive vocabulary related to market manipulation, insider trading, or violations of regulatory provisions, ensuring compliance with financial market fairness and information disclosure standards. Additionally, we identify and anonymize PII—including email addresses, phone numbers, personal addresses, and IP addresses—using specific regular expressions. This process follows Laurençon et al. (2023) to detect sensitive information such as age, email, date, time, and personal addresses, ensuring the privacy and safety of YIZHAO.

After passing through the cleaning pipeline, the data is further processed by our safety classifier (Section 3.2) and financial filtering model (Section 3.3) to produce a reliable, high-quality finance-focused corpus.

3.2 Safety Filter

We develop a safety filtering model to identify and remove potentially illegal and harmful data, resulting in safer financial data that aligns with social values. To synthesize data for training the safety filter, we design a prompt for Qwen-72B-Chat (Bai et al., 2023) to assess whether the sampled data is safe, legal, and conforms to the social norms across four dimensions: violation of core socialist values, discriminatory content, commercially illegal activities, and infringement of others’ legitimate rights and interests. The annotation prompt is shown in Figure 6. Using this setup, we perform automatic annotations on 700K samples from the English subset and 500K samples from the Chinese subset. To ensure high annotation quality, human annotators review and revise a sample of the automatically annotated data.

We use *bge_small_chinese*⁷ for Chinese data and *snowflake-arctic-embed-xs*⁸ for English data, and

⁷<https://huggingface.co/BAAI/bge-small-zh>

⁸<https://huggingface.co/Snowflake/snowflake-arctic-embed-xs>

finetune linear regression models on top of the embedding models. Due to uneven category distribution in the annotated data, we apply undersampling to create a balanced training dataset. Since our objective for the safety filtering model is to remove illegal and harmful data, we focus more on the recall rate of harmful data. The final safety filtering model achieves a recall rate of 92% for harmful data on the held-out validation set.

As shown in Figure 4, after training an intermediate version of the filtering model, human annotators sample and verify the obtained safe data. If the proportion of illegal or harmful data in the filtered subset exceed 2%, we re-sample, annotate, and re-train the filtering model. This iterative process ensures that the filtering model reduces the proportion of illegal and harmful content in the final dataset to below 2%. Extensive human annotation on 10K documents (5K Chinese and 5K English) confirm that the ratio of toxic text in YIZHAO is less than 2% in the Chinese subset and below 0.5% in the English subset.

3.3 Financial Filter

To efficiently extract professional financial knowledge from Chinese and English datasets, we have developed a financial corpus classifier. To generate synthetic financial relevance scores, similar to Section 3.2, we use Qwen-72B-Chat (Bai et al., 2023) to score 700K randomly sampled English documents and 500K Chinese documents. Each document is scored on a scale from 0 to 5 based on its financial relevance. Here financial relevance refers to whether the data is related to financial markets, including policies and regulations, economic indicators, financial products, trading activities, risk management, and market analysis. Following Yuan et al. (2024), we adopt an additive scale prompting strategy, allowing the LLM to evaluate each criterion and build the score step-by-step. The prompt used for synthetic annotations is in Figure 5.

We use the same embedding models as Section 3.2 and train a linear regression model on top of these models. Then we select the checkpoint with the highest F1 score on the validation set. We use fixed thresholds to classify whether a given document is finance-related. We evaluate the effects of various threshold settings and ultimately select a minimum threshold of 3 for FineWeb-Edu, giving the best trade-off between recall and precision. When setting the threshold at 3, the Chinese classifier achieves an F1 score of 82.7% on the val-

idation set, while the English classifier achieved an F1 score of 86%. These results demonstrate that the classification models are highly effective at identifying financial knowledge.

After processing the raw datasets using a series of cleaning tools, financial classifiers, and safety filters, we have constructed a cleaner, financially specialized, and social value-compliant YIZHAO.

3.4 Data Synthesis

Inspired by Maini et al. (2024) and Cheng et al. (2024), synthetic question-answering (QA) pairs about the domain-specific text could be beneficial for data-efficient adaptation into specialized domains. We take the cleaned text with high financial relevance scores from the previous step. The QA pairs are generated by Qwen-72B-Chat based on the prompt in Figure 7. To guarantee their quality, they are rigorously reviewed by professional annotators, who verify that both the questions and answers are accurate, relevant, and faithfully reflect the underlying financial text.

Furthermore, financial documents can involve various types of visual content, highlighting the importance of developing financial multimodal large language models and datasets (Bhatia et al., 2024; Xie et al., 2024). We generate images and PDFs based on the selected text to create a multimodal dataset. By integrating these images and PDFs with QA pairs, we build a multimodal dataset suitable for training document QA and retrieval-augmented generation models.

4 Conclusion

In this paper, we present YIZHAO, the largest open-source financial corpora. YIZHAO is distinguished by its high quality, reliable content safety and finance-focused design. It encompasses bilingual web pages as well as multimodal question-answering pairs. Additionally, we provide a comprehensive data-cleaning pipeline and trained filtering models as open-source tools, enabling future advancements in scaling financial datasets.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao,

- Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, and et al. (28 additional authors not shown). 2023. [PaLM 2 Technical Report](#).
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *Preprint*, arXiv:2309.16609.
- Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. [FinTral: A family of GPT-4 level multimodal financial large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13064–13087, Bangkok, Thailand. Association for Computational Linguistics.
- A.Z. Broder. 1997. [On the resemblance and containment of documents](#). In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.
- Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. 2023. [Chinesewebtext: Large-scale high-quality chinese web text extracted with effective evaluation model](#). *Preprint*, arXiv:2311.01149.
- Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. [Adapting large language models via reading comprehension](#). In *The Twelfth International Conference on Learning Representations*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#). *Preprint*, arXiv:2204.02311.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. [GLaM: Efficient scaling of language models with mixture-of-experts](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#). *Preprint*, arXiv:2101.00027.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal large language models. In *Proceedings of the 36th*

- International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Denis Kocetkov, Raymond Li, Loubna Ben allal, Jia LI, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro Von Werra, and Harm de Vries. 2023. [The stack: 3 TB of permissively licensed source code](#). *Transactions on Machine Learning Research*.
- Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Froberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. [The bigscience roots corpus: A 1.6tb composite multilingual dataset](#). *Preprint*, arXiv:2303.03915.
- Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. [A survey of large language models in finance \(finllms\)](#). *Preprint*, arXiv:2402.02315.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating training data makes language models better](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
- Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. [Rephrasing the web: A recipe for compute and data-efficient language modeling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.
- Udi Manber and Gene Myers. 1993. [Suffix arrays: A new method for on-line string searches](#). *SIAM Journal on Computing*, 22(5):935–948.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. [A survey of large language models for financial applications: Progress, prospects and challenges](#). *Preprint*, arXiv:2406.11903.
- Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. [MultiLegalPile: A 689GB multilingual legal corpus](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. [A monolingual approach to contextualized word embeddings for mid-resource languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 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*.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. [The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only](#). *arXiv preprint arXiv:2306.01116*, arXiv:2306.01116.
- Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Zhenxiang Li, Pei Chu, Yuan Qu, Jin Shi, Lindong Lu, Runyu Peng, Zhiyuan Zeng, Huanze Tang, Zhikai Lei, Jiawei Hong, Keyu Chen, Zhaoye Fei, Ruiliang Xu, Wei Li, Zhongying Tu, Lin Dahua, Yu Qiao, Hang Yan, and Conghui He. 2024. [Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset](#). *Preprint*, arXiv:2402.19282.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. [Scaling Language Models: Methods, Analysis & Insights from Training Gopher](#). *arXiv preprint arXiv:2112.11446*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,

- Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. 2023. [SlimPajama-DC: Understanding Data Combinations for LLM Training](#). Preprint, arXiv:2309.10818.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hananeh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. [Dolma: an open corpus of three trillion tokens for language model pretraining research](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zengzhi Wang, Xuefeng Li, Rui Xia, and Pengfei Liu. 2024. [Mathpile: A billion-token-scale pretraining corpus for math](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Maurice Weber, Daniel Y Fu, Quentin Gregory Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Re, Irina Rish, and Ce Zhang. 2024. [Redpajama: an open dataset for training large language models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. [Challenges in detoxifying language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. *arXiv preprint arXiv:1911.00359*.
- Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, Shunian Chen, Yifei Zhang, Lihang Shen, Daniel Kim, Zhiwei Liu, Zhenheng Luo, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Zhiyuan Yao, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Yilun Zhao, Yitao Long, Guojun Xiong, Kaleb Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jianyun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Jimin Huang, and Sophia Ananiadou. 2024. [Open-finllms: Open multimodal large language models for financial applications](#). Preprint, arXiv:2408.11878.
- Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. [To repeat or not to repeat: Insights from scaling llm under token-crisis](#). Preprint, arXiv:2305.13230.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. [Self-rewarding language models](#). Preprint, arXiv:2401.10020.
- Wanyue Zhang, Ziyong Li, Wen Yang, Chunlin Leng, Yinan Bai, Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2024. [Chinesewebtext 2.0: Large-scale high-quality chinese web text with multi-dimensional and fine-grained information](#). Preprint, arXiv:2411.19668.

A Prompt List

The prompts for Qwen-72B-Chat to annotate financial and safety scores are listed in Figure 5 and 6 respectively. The prompt for QA pair generation is in Figure 7. The prompt in Chinese is what we use for annotation. We provide an English translation for ease of read.

你是一个优秀的网页摘录评估员，精通金融、政策领域的知识。请按照下面的5分制评分标准对摘录进行评分，初始分数为0，满分为5：

- 1、如果摘录提供了一些与金融、政策、公告主题相关的基本信息，用语严谨正经，即使其中包含一些不相关、不全面的内容，加1分。
- 2、如果摘录提及了一些金融领域的词汇，但其主要主题或讨论点并不集中在金融上，或者只是间接提及金融，如日常生活中的简单消费或储蓄提及，加1分。
- 3、如果摘录中包含金融领域的基本概念或事件，如提及股票代码、利率变化、财经新闻，但没有深入的分析或讨论，内容相对表面，加1分。
- 4、如果摘录探讨了金融领域、企业或国家政策的特定主题，如投资策略、金融产品分析、市场趋势、经济政策等，对金融有较深入的理解和分析，加1分。
- 5、如果摘录是专业或学术性的金融内容，如复杂的财务报告分析、经济学理论、金融模型、行业研究报告等，明显针对金融专业人士或研究者，加1分。

摘录：

阅读摘录后：简要说明您的总分，最多 100 个字。使用以下格式得出分数：“得分：<总分>”。

You are an excellent web text evaluator, proficient in financial and policy-related knowledge. Please score the text according to the following 5-point scale. The initial score is 0, and the maximum score is 5:

1. If the text provides basic information related to finance, policy, or announcements, with precise and formal language, even if it contains some irrelevant or incomplete content, add 1 point.
2. If the text mentions financial terms, but its main theme or discussion point is not focused on finance, or it only indirectly mentions finance, such as simple consumption or savings in daily life, add 1 point.
3. If the text includes basic concepts or events in finance, such as mentioning stock codes, interest rate changes, or financial news, but without in-depth analysis or discussion, the content is relatively surface-level, add 1 point.
4. If the text discusses a specific topic in the financial field, such as investment strategies, financial product analysis, market trends, economic policies, etc., and demonstrates a deeper understanding and analysis of finance, add 1 point.
5. If the text is professional or academic financial content, such as complex financial report analysis, economic theories, financial models, or industry research reports, clearly targeting financial professionals or researchers, add 1 point.

Text:

After reading the web text: Briefly explain your overall score in no more than 100 words. Use the following format to state the score: "Score: <Total score>".

Figure 5: Prompt for LLM-based financial scoring.

Role: 训练数据安全风险识别专家

Background: 你是一位数据安全风险识别专家，能够精确识别包含违反社会主义核心价值观的内容、包含歧视性内容、商业违法违规和侵犯他人合法权益的内容

Attention: 你需要以专业和精准的方式，从文本中识别其中存在安全风险的内容，并做出该数据是否存在安全风险的判断。

Profile: 你是一位训练数据安全风险识别的专家，能够根据给定的<安全风险内容>规则和自己的专业知识、经验，从输入的文本信息中找到文本存在的安全风险并尝试提供判断。

安全风险内容

- 1、包含违反社会主义核心价值观的内容；
- 2、包含歧视性内容；
- 3、商业违法违规；
- 4、侵犯他人合法权益；
- 5、包含赌博、毒品、恐怖主义、反动言论、违禁内容。

Workflow:

1. 通读用户输入的文本信息
2. 根据给定的<安全风险内容>规则和自己的专业知识、经验识别文本存在的风险信息
3. 若存在回答是，若不存在回答否，需要作出简单的解释说明，不超过150字
4. 结构化输出为JSON格式

Constraints: 1. 任何可能存在安全风险的内容都判断成是； 2. 输出严格遵循JSON规范

Output Format: {"result": "否", "reason": ""}

Initialization: 以下为用户的文本内容：

Role: Data Security Risk Identification Expert

Background: You are a data security risk identification expert, capable of accurately identifying content that violates the core social values, contains discriminatory content, involves commercial illegalities and violations, or infringes on others' legal rights.

Attention: You are required to identify potentially risky content in the text in a professional and precise manner, and make a judgment as to whether the data contains any security risks.

Profile: You are an expert in risk identification of training data. Based on the given <security risk content> rules and your professional knowledge and experience, you are to identify any security risks in the input text and attempt to provide a judgment.

Security Risk Content

1. Contains content that violates the core social values;
2. Contains discriminatory content;
3. Involves commercial illegalities or violations;
4. Infringes on others' legal rights;
5. Contains gambling, drugs, terrorism, reactionary or prohibited content.

Workflow:

1. Read through the user's input text.
2. Identify risk information in the text based on the given <security risk content> rules and your professional knowledge and experience.
3. If there are risks in the text, respond with "yes"; if not, respond with "no" and provide a brief explanation (no more than 150 words).
4. Output in structured JSON format.

Constraints: 1. Any content that could potentially pose a security risk should be judged as "yes." 2. The output must strictly adhere to the JSON format.

Output Format: {"result": "no", "reason": ""}

Initialization: The following is the user's input text:

Figure 6: Prompt for LLM-based risk scoring.

Role: 金融领域专家

Profile: 这个提示词帮助生成围绕文章内容的问答对。用户可以输入一段文本内容，金融领域专家将提取有关该内容的常见问题并尝试提供答案。

Attention: 你需要以专业和精准的方式，从内容中挖掘出有价值的问题，并提供相应的详细答案以及答案的依据，以JSON格式呈现。

Goals:

1. 分析理解文本内容，提取关键信息
2. 创造2个相关问题
3. 编写对应的详细答案以及答案的依据
4. 结构化为合法的JSON格式
5. 确保问题与答案内容的准确性和深度

Skills

1. 深入理解文本内容。
2. 生成精确且信息丰富的问答对。
3. 提供简单易懂的解释，适合不同知识水平的用户。
4. 问答是有据可依的，给出答案的详细依据。
5. 能够自动扩展问题，涵盖不同的应用场景和细节。

Constraints:

1. 不凭空捏造信息
2. 问题和答案需直接与提供的数据关联
3. 文本内容是法律规定，在答案中指出法律依据
4. 输出遵循JSON规范

Rules

1. 分析、通读文本内容。
2. 基于文本内容提出2个关键问题。
3. 如果文本内容是法律规定，必须在每个问题的答案开始时输出‘根据《xxx法/规定》’。
4. 根据问题生成详细答案。
5. 生成的问题和答案均需要结合中国的国情、制度、法律和立场。
6. 输出结果应为有效的JSON格式。

Workflow

1. 用户输入一段文本内容。
2. 基于文本内容提出2个关键问题。
3. 根据问题生成详细答案。
4. 结构化输出为标准JSON格式。

Output Format

```
[{"question": "乡镇企业在环境保护方面有哪些责任?", "answer": "根据《中华人民共和国乡镇企业法》，乡镇企业必须遵守环境保护法律、法规，发展无污染或少污染企业，防治环境污染和生态破坏，执行环境影响评价制度，确保防治污染设施与主体工程同步，不得采用或使用严重污染环境的工艺和设备，超标排放污染物的必须限期治理，否则可能面临关闭、停产或转产。"}]
```

Init: 以下是用户提供的文本内容：

Figure 7: Prompt for QA generation.

Role: Financial Domain Expert

Profile

This prompt helps generate Q&A pairs based on the content of an article. Users can input a piece of text, and the financial domain expert will extract common questions related to the content and attempt to provide answers.

Attention:

You need to professionally and precisely identify valuable questions from the content and provide corresponding detailed answers along with their basis, presented in JSON format.

Goals:

1. Analyze and understand the text content to extract key information.
2. Generate two relevant questions.
3. Write corresponding detailed answers along with their basis.
4. Structure the output in a valid JSON format.
5. Ensure the accuracy and depth of the questions and answers.

Skills:

1. In-depth understanding of text content.
2. Generating precise and informative Q&A pairs.
3. Providing clear and understandable explanations for users of different knowledge levels.
4. Ensuring that Q&As are well-founded, with clear justification.
5. Expanding questions to cover different scenarios and details.

Constraints:

1. Do not fabricate information.
2. Questions and answers must be directly related to the provided data.
3. If the text contains legal provisions, legal references must be cited in the answers.
4. Output must adhere to JSON specifications.

Rules:

1. Analyze and read through the text content.
2. Propose two key questions based on the text content.
3. If the text contains legal provisions, each answer must begin with "According to the xxx law/regulation".
4. Generate detailed answers based on the questions.
5. Ensure that questions and answers align with national conditions, system, laws, and stance.
6. The output should be in a valid JSON format.

Workflow:

1. The user inputs a text segment.
2. Two key questions are generated based on the text content.
3. Detailed answers are generated for the questions.
4. The structured output is presented in a standard JSON format.

OutputFormat:

```
[ "question": "What are the responsibilities of township enterprises in environmental protection?",  
  "answer": "According to the 'Law of the People's Republic of China on Township Enterprises,' township enterprises must comply with environmental protection laws and regulations, develop pollution-free or low-pollution enterprises, prevent environmental pollution and ecological damage, implement the environmental impact assessment system, ensure that pollution prevention facilities are synchronized with the main project, and must not adopt or use processes and equipment that severely pollute the environment. Enterprises that exceed emission standards must be rectified within a time limit, or they may face closure, suspension of production, or transformation." ]
```

Init: The following is the user's input text:

Figure 8: The translated prompt for QA generation.