



# Xinyu: An Efficient LLM-based System for Commentary Generation

Yiquan Wu\*  
wuyiquan@zju.edu.cn  
Zhejiang University  
Hangzhou, China

Bo Tang\*  
tangb@iaar.ac.cn  
Institute for Advanced Algorithms  
Research  
Shanghai, China

Chenyang Xi†  
xicy@iaar.ac.cn  
Institute for Advanced Algorithms  
Research  
Shanghai, China

Yu Yu†  
yuy@iaar.ac.cn  
Institute for Advanced Algorithms  
Research  
Shanghai, China

Pengyu Wang  
2371407@stu.neu.edu.cn  
Northeastern University  
Shenyang, China

Yifei Liu  
liuyifei@zju.edu.cn  
Zhejiang University  
Hangzhou, China

Kun Kuang‡  
kunkuang@zju.edu.cn  
Zhejiang University  
Hangzhou, China

Haiying Deng  
denghaiying@xinhua.com  
State Key Laboratory of Media  
Convergence Production Technology  
and Systems  
Beijing, China

Zhiyu Li  
lizy@iaar.ac.cn  
Institute for Advanced Algorithms  
Research  
Shanghai, China

Feiyu Xiong  
xiongfy@iaar.ac.cn  
Institute for Advanced Algorithms  
Research  
Shanghai, China

Jie Hu  
hujie1@chinatelecom.cn  
Research Institute of China Telecom  
Beijing, China

Peng Cheng  
chengpeng@xinhua.org  
State Key Laboratory of Media  
Convergence Production Technology  
and Systems  
Beijing, China

Zhonghao Wang  
wangzhonghao@xinhua.org  
State Key Laboratory of Media  
Convergence Production Technology  
and Systems  
Beijing, China

Yi Wang  
wangyi08@xinhua.org  
State Key Laboratory of Media  
Convergence Production Technology  
and Systems  
Beijing, China

Yi Luo  
luoyi@xinhua.org  
State Key Laboratory of Media  
Convergence Production Technology  
and Systems  
Beijing, China

Mingchuan Yang  
yangmch@chinatelecom.cn  
Research Institute of China Telecom  
Beijing, China

## ABSTRACT

Commentary provides readers with a deep understanding of events by presenting diverse arguments and evidence. However, creating

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

KDD '24, August 25–29, 2024, Barcelona, Spain

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0490-1/24/08

<https://doi.org/10.1145/3637528.3671537>

commentary is a time-consuming task, even for skilled commentators. Large language models (LLMs) have simplified the process of natural language generation, but their direct application in commentary creation still faces challenges due to unique task requirements. These requirements can be categorized into two levels: 1) fundamental requirements, which include creating well-structured and logically consistent narratives, and 2) advanced requirements, which involve generating quality arguments and providing convincing evidence. In this paper, we introduce Xinyu, an efficient LLM-based system designed to assist commentators in generating Chinese commentaries. To meet the fundamental requirements, we deconstruct the generation process into sequential steps, proposing

targeted strategies and supervised fine-tuning (SFT) for each step. To address the advanced requirements, we present an argument ranking model for arguments and establish a comprehensive evidence database that includes up-to-date events and classic books, thereby strengthening the substantiation of the evidence with retrieval augmented generation (RAG) technology. To evaluate the generated commentaries more fairly, corresponding to the two-level requirements, we introduce a comprehensive evaluation metric that considers five distinct perspectives in commentary generation. Our experiments confirm the effectiveness of our proposed system. We also observe a significant increase in the efficiency of commentators in real-world scenarios, with the average time spent on creating a commentary dropping from 4 hours to 20 minutes. Importantly, such an increase in efficiency does not compromise the quality of the commentaries.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language generation.**

## KEYWORDS

LLM-based System, Commentary Generation, Supervised Fine-tuning, Retrieval Augmented Generation

### ACM Reference Format:

Yiquan Wu[1], Bo Tang[1], Chenyang Xi[2], Yu Yu[2], Pengyu Wang, Yifei Liu, Kun Kuang[3], Haiying Deng, Zhiyu Li, Feiyu Xiong, Jie Hu, Peng Cheng, Zhonghao Wang, Yi Wang, Yi Luo, and Mingchuan Yang. 2024. Xinyu: An Efficient LLM-based System for Commentary Generation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3637528.3671537>

## 1 INTRODUCTION

With the advancement of natural language processing (NLP), particularly large language models (LLMs), numerous text-generation systems have been proposed to enhance the effectiveness and efficiency of individuals across various fields, such as education, medicine and law [1, 4, 32]. Commentary is a type of article that contains diverse arguments and compelling evidence, which aims to provide readers with a deep understanding of certain events. As Figure. 1 shows, a commentator usually spends several hours writing a commentary, which includes mining arguments, searching for evidence, and embellishing the article. Given the continuous nature of news, their workload is substantial. Therefore, exploring the application of LLMs in commentary generation is worthwhile.

Although LLMs have benefited many generative tasks, they face challenges when directly applied to commentary generation due to unique task requirements. Broadly, the requirements for a commentary can be divided into two levels:

### 1) Fundamental requirements:

- The structure should be regular and complete. As Figure. 1 shows, the commentary should follow a total division structure.

- The content should be self-consistent. For example, the arguments in the commentary should not be contradictory, and the evidence must support the arguments.

### 2) Advanced requirements:

- Arguments should be specific and original. The argument is key to the commentary, representing the author's stance.
- Evidence should be convincing, which means the LLMs can't generate fake evidence, and the evidence is preferably new.

In this paper, we propose Xinyu, an efficient LLM-based system to assist commentators in Chinese commentary generation. Specifically, for the fundamental requirements, we decompose the generation into several sequential steps, ensuring the generated text is well-structured. We also design targeted supervised fine-tuning (SFT) and strategies for each step to maintain content consistency. For the advanced requirements, we propose an argument ranking model for ranking candidate arguments to ensure quality. Moreover, we construct a comprehensive evidence database, which maintains up-to-date events and books, and then use the technology of retrieval augmented generation (RAG) to generate convincing evidence.

Given the dynamic nature of commentary, traditional metrics for text generation tasks, such as ROUGE or BLEU, fall short in evaluating the overall quality of the commentary. Thus, we propose a comprehensive evaluation metric for commentary generation that considers five distinct perspectives. In our pilot study, GPT-4 demonstrated performance on par with human annotators, so we employ GPT-4 as the evaluator. The experimental results underscore the quality of the content generated by our system, Xinyu. In practical terms, we also examined how Xinyu could enhance the work efficiency of human commentators and the result shows that with Xinyu, the speed of commentary generation increased dramatically, reducing the average creation time from 4 hours to a mere 20 minutes. Importantly, this increase in efficiency does not sacrifice the quality of the commentaries.

To sum up, our main contributions are as follows:

- We leverage LLMs for the task of commentary generation and propose a system named Xinyu that can assist commentators in generating Chinese commentary 10 times faster with even quality.
- We decompose the commentary and generate it in steps, applying targeted supervised fine-tuning (SFT) for each. This approach ensures the commentary meets its fundamental requirements: it is well-structured and self-consistent.
- We propose an argument ranking module to improve the quality of the arguments and construct a comprehensive knowledge database (e.g., up-to-date events and books) for the generation of evidence with the help of retrieval augmented generation (RAG). This approach ensures the commentary meets the advanced requirements: it is specific and convincing.
- We design a comprehensive evaluation method for the commentary generation task with 5 distinct perspectives. The experimental results demonstrate the effectiveness of our proposed Xinyu.

\* These authors contributed equally to this work.

† These authors contributed equally to this work.

‡ Corresponding Author.

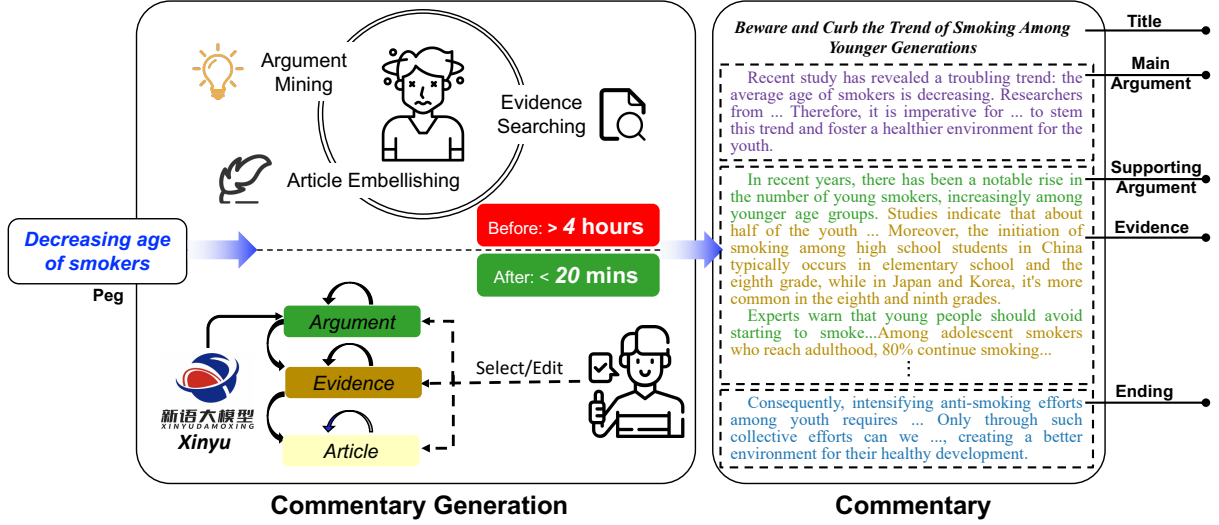


Figure 1: The illustration of the commentary generation task. To generate a commentary, it usually requires argument mining, evidence searching, and article embellishing. With the Xinyu, the intermediate steps can be sped up. The right part demonstrates the structure of a commentary, which consists of a title, a main argument, several supporting arguments and evidence, and an ending. This example is translated from Chinese.

## 2 RELATED WORK

### 2.1 Large Language Models

The domain of Natural Language Processing (NLP) has witnessed substantial progress [14, 21, 30, 31, 42], especially through the advent of Large Language Models (LLMs) [2, 17, 18, 27, 35]. These models show exceptional text generation proficiency, yielding high fluency and readability outputs [32, 39]. Their ability to adapt to downstream tasks with minimal in-context examples is particularly noteworthy. To further augment the efficacy of LLMs in downstream tasks, two main methods have been identified: supervised fine-tuning (SFT) and retrieval augmented generation (RAG).

**Supervised Fine-Tuning (SFT)** entails the adaptation of an LLM to a specific downstream task. This process refines the model’s parameters to align with the data distribution and task requirements, ensuring the model’s behavior mirrors human behavior within the given domain. The topic of SFT has been extensively explored in numerous research. Ouyang et al. [18] pioneered the introduction of supervised fine-tuning and reinforcement learning to align language models with human intent. Zhou et al. [41] compiled a dataset of merely 1K examples for SFT, demonstrating that the success of SFT depends on the quality and diversity of data.

**Retrieval Augmented Generation (RAG)** amalgamates LLMs with content retrieved from external databases. This approach offers a promising solution to the challenges encountered by LLMs, such as hallucination, outdated knowledge, and untraceable reasoning processes. The conventional RAG process encompasses indexing, retrieval, and generation [9, 15]. RAG has been further enhanced by a range of innovative techniques: fine-tuning retrieval models to obtain precise semantic representations [11, 28, 33], reformulating queries to align with the semantic space of queries and documents

[8, 20, 29], fine-tuning LLMs to harmonize the output of the retriever with the LLM’s preference [10, 22, 34].

In our work, we leverage the advances of both SFT and RAG to enhance the performance of the Xinyu.

### 2.2 Domain-specific LLMs

Large Language Models (LLMs) have advanced the field of natural language processing, providing a task-agnostic foundation for a wide range of applications. However, directly applying LLMs to solve sophisticated problems in specific domains meets many hurdles, caused by the heterogeneity of domain data, the sophistication of domain knowledge, and the diversity of the constraints [40].

Numerous researchers have devoted their efforts to domain-specific Language Models (LLMs) tailored for various fields. These specialized LLMs have been designed to cater to the unique requirements of domains such as medicine [12, 23, 24] for medical diagnosis, law [6, 36, 37] for handling legal documents, counselling [13] for mental health support, education [7, 16] for teaching assistance, science[3] for crafting scientific journalism, and so on.

The former domain-specific LLMs mainly focus on injecting domain knowledge (e.g., medical or legal knowledge) into LLMs. In this paper, our focus is on commentary generation, to support commentators in their writing process and produce well-structured, logically consistent commentaries that present novel arguments and convincing evidence.

## 3 PRELIMINARIES

This section is dedicated to defining key concepts that will be consistently referenced throughout this paper.

**Peg**, within the scope of commentary generation, denotes the specific aspect of the event that the commentary is responding to or

building upon. It acts as an anchor for the commentary. For instance, in Figure. 1, the peg is ‘Age of smokers decrease’.

**Main Argument**, in the context of commentary generation, signifies the central point that the generated commentary seeks to communicate. It forms the core message around which the commentary is structured.

**Supporting Argument**, is an additional point that helps to substantiate the main argument. Typically, a commentary will contain several supporting arguments that collectively contribute to the strength and depth of the main argument.

**Evidence** refers to the data, facts, or information employed to support the argument. In the process of commentary generation, evidence can be derived from the content itself or external sources.

Given a peg, the corresponding commentary will include one main argument and several supporting arguments, all of which are supported by evidence.

## 4 TECHNICAL ROUTE OF XINYU

In this section, we delve into Xinyu’s comprehensive technical route. Figure. 2 shows the overall framework of Xinyu. In Section 4.1, we introduce the five main generative components used in detail during the commentary generation process. In Section 4.2, we shift our focus to the two auxiliary components essential for meeting the advanced requirements of the commentary, including the argument ranking model and the construction of an evidence database. Note that without these two auxiliary components, the system can still fulfill the fundamental requirements.

### 4.1 Main Components

Based on the structure, we decompose the commentary generation into five steps: peg generation, main argument generation, supporting argument generation, evidence generation, and finally, article combination. This sequential approach is implemented with the help of SFT and RAG.

**4.1.1 Peg Generation.** The peg generation serves as a preliminary step in the commentary generation process, designed to swiftly summarize event details for the user. Utilizing a search engine, this component retrieves event details based on given keywords to generate a peg. Alternatively, users have the option to manually compose the peg, bypassing this automated step. Specifically, the content from the top three most relevant search results is processed as input, and the LLM condenses this information into a concise peg. The procedure is exemplified as follows:

*[You are a commentary writing expert, and here are the details of an event. Event detail: {event detail}. Please refine it into a concise and well-articulated peg:]*

To enhance the model’s proficiency in condensing event details during peg generation, we develop SFT data specifically for this step. By inputting event details and using the peg as a label, this method trains the model to more effectively summarize and pinpoint essential information, resulting in pegs that are informationally dense.

**4.1.2 Main Argument Generation.** This step aims to provide the main argument. Due to the variety of pegs, main arguments can

be driven in different directions. Here the strategy involves directing the LLM to generate across ten distinct directions: technology, finance, society, politics, literature and arts, lifestyle, environment, sports, education, and science. Each direction emphasizes its specific thematic elements, such as highlighting technological advancements or economic trends. To operationalize this strategy, we combine the peg, event details, and a chosen direction as input for the LLM, which then generates candidate main arguments one at a time. An example of the usage is below:

*[You are a commentary writing expert. Please complete the main argument in the direction of {direction} based on the peg: {peg} and event detail: {event detail}. The main argument should be profound, concise, and strongly related to the peg. Please provide the main argument:]*

To enhance the model’s ability to generate the main arguments, we design the corresponding SFT data. The input for this SFT data includes retrieved event details based on the peg and the direction of the article. The label for this SFT data is the main argument derived from the input. This data construction approach not only facilitates the generation of a helpful main argument for diverse article directions but also ensures its consistency with the initial peg, thus guaranteeing both relevance and alignment in the narrative.

To meet the advanced requirements of argument, these generated main arguments will then get a score from the argument ranking model, based on their novelty, and objectivity. Then these candidate main arguments will be ranked based on the scores.

**4.1.3 Supporting Argument Generation.** This step aims to generate supporting arguments that seamlessly align with both the main argument and the event’s details. To achieve this, the system synthesizes the main argument, event details, and a predefined argument structure. Available argument structures include parallel, progressive, and contrasting formats, each facilitating a unique commentary structure. This integration process enables the LLM to produce relevant supporting arguments. The LLM will decide the number of supporting arguments  $m$  itself. An example of usage could be:

*[You are a commentary writing expert. Based on the given main argument {main argument} of the commentary and event detail {event detail}, generate multiple supporting arguments for the commentary. The supporting arguments form {structure} structure, refining around the {main argument} with multi-level, multi-faceted, and multi-angle perspectives. Please provide the supporting arguments:]*

The SFT data for this step is constructed to facilitate this process. We utilize inputs comprising event details, main arguments, and argument structures. The labels are the marked corresponding supporting arguments. This data construction approach ensures that the model is fine-tuned to produce supporting arguments that enrich and substantiate the main argument effectively.

**4.1.4 Evidence Generation.** In the Evidence Generation step, the system aims to generate accurate and contextually relevant evidence. The process begins with accessing reference information of a supporting argument from the evidence database to ensure veracity, effectively mitigating the hallucinations. This reference information, along with the provided main and supporting arguments, serves as the input. Then, the LLM will generate evidence that is tailored to align precisely with the given supporting argument. An example of the usage is below:

Note the user can interact with Xinyu by providing additional input or editing the output at each step and here we mainly describe the automatic process.

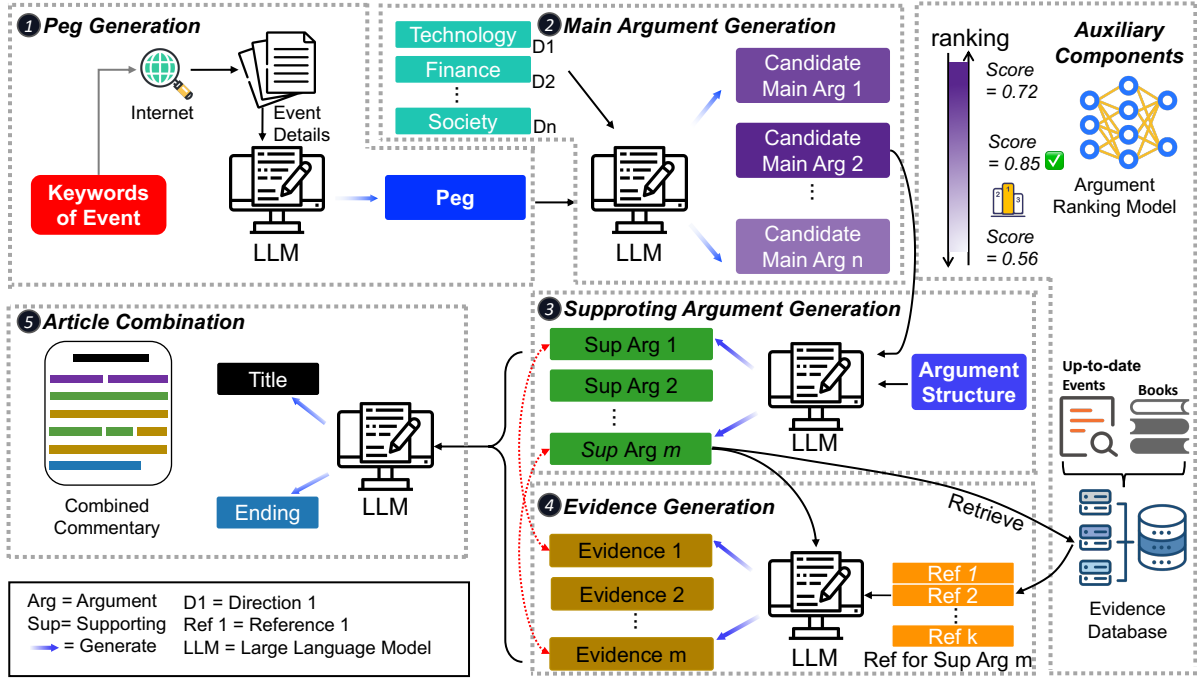


Figure 2: The overall framework of Xinyu. The generation process is divided into 5 steps.

[You are a commentary writing expert. Surrounding the main argument {main argument}, please use the evidence provided in the reference information, including dates, data, viewpoints, core content, etc., to continue writing evidence in the commentary to support the supporting argument {supporting argument}. Please annotate the corresponding reference information numbers in the continuation. Reference information: {reference}. Please provide the evidence: ]

The SFT data for evidence generation is structured with inputs including reference information, main and supporting arguments. The output label is the evidence associated with the corresponding supporting argument. These elements guide the model in generating evidence that is precise and contextually relevant to the provided supporting arguments.

**4.1.5 Article Combination.** This step is aimed at generating the title and ending, then forming the overall commentary. To achieve this, the system integrates the preceding event details, main arguments, supporting arguments, and evidence as inputs. Following this integration, the LLM then outputs the title and ending. An example of the usage for ending generation is below:

[You are a commentary writing expert. Please write a conclusion for the article, maintaining smooth language, consistent style, and logical coherence with the preceding text. The preceding text is as follows: {preceding text}. Please provide the ending:]

The usage for title generation is similar to the ending generation.

To facilitate the model in generating context-appropriate and coherent title and ending, the SFT data is constructed with inputs including the event details, main argument, combined supporting arguments and evidence. The label for this data is the corresponding title and ending. This structured approach ensures that the model

is adept at crafting titles and endings that effectively encapsulate the various dimensions of the commentary, providing a fitting start and end to the narrative.

After generating the title and ending, the system will combine all the output to form a complete commentary.

## 4.2 Auxiliary Components

In this section, we introduce two auxiliary components that assist Xinyu in meeting advanced requirements: the argument ranking model and the evidence database.

**4.2.1 Argument Ranking Model.** The argument ranking model plays a crucial role in the main argument generation process by assessing and ranking candidate main arguments. This aids users in selecting the most compelling argument.

Developing the ranking model presents a central challenge due to the subjective nature of assessing arguments, which lack universally accepted standards, unlike quantifiable metrics. For example, evaluating the generated arguments based on factors like novelty is cumbersome.

To address this challenge, we train a BERT-based scoring model with a pairwise loss function. This approach converts the ranking challenge into a series of binary comparisons, simplifying the task to discerning relative superiority between pairs of arguments. The loss function is defined as:

$$\mathcal{L}(x) = \sum \Phi(f(x_a) - f(x_b)) \quad (1)$$

where  $f(x)$  represents the scoring function for a given argument  $x$ , and  $\Phi$  is a non-linear transformation applied to the calculated difference between the two arguments  $x_a$  and  $x_b$ .

The quantity of *likes* on articles from opinion-sharing platforms, such as Zhihu, is often indicative of the novelty and objectivity of the arguments they present. Consequently, this metric is leveraged to assess the quality of the arguments within these articles. In the process of constructing the training dataset, we collect articles from such platforms, utilizing the number of likes as a criterion to establish a partial order among pairs of texts, and the dataset consists of 240,000 text pairs. This order serves to reflect their relative quality. In the inference phase, we translate the order to numerical score, and the scores of the candidate main arguments are utilized to rank these arguments.

**4.2.2 Evidence Database Construction.** In pursuit of generating convincing evidence, we construct an Evidence Database to store Chinese knowledge sourced from events and books for retrieval.

For the events knowledge, we first legally collect the daily updated article titles on the website’s hot list, and then prompt the LLM to complete the following four tasks given the article title: (1) summarize the event related to the article title; (2) determine which direction (e.g., technology, finance) the event belongs to; (3) extract the six elements of the event, including time, location, person, cause, process and result; (4) describe the event in a paragraph based on the six elements.

For knowledge from books, we gather classic works in law, finance, and various other subjects legally, segmenting the contents of these books into chunks and storing them within the database.

The evidence database is built upon 200,000 event knowledge data and 110,000 book knowledge data. Following the construction phase, we implement the ElasticSearch engine, anchored to the evidence database, to enhance retrieval capabilities. During retrieval, the supporting argument is inputted, prompting the fetching of the  $k$  most pertinent references from the database. These references are then fed into the Large Language Model (LLM) to generate evidence in support of the arguments. In practical application, the value of  $k$  is determined by the similarity score between the input argument and the existing knowledge, and we set a predefined threshold at 0.6 in the experiment. In addition, to maintain access to the most current event knowledge, we continuously collect data from online platforms and update our database daily.

## 5 EXPERIMENT

### 5.1 Evaluation Metrics of Commentary

**Automatic evaluation.** Existing automatic generation evaluation metrics, including but not limited to ROUGE and BLEU[19], mainly focus on the degree of similarity to a reference text. However, in the context of commentary generation, the inherent diversity of the commentary content poses a significant challenge to these similarity-based metrics, often leading to an incomplete evaluation. To address this limitation, we propose a novel evaluation metric that assesses commentaries across five distinct dimensions: • **Structure Soundness**: clarity of the hierarchy, compactness of the writing, and rationality of the layout; • **Logic Consistency**: consistency of the content, rationality of the argument, and thoughtfulness; • **Argument Quality**: freshness and directionality of the topic conception; • **Evidence Support**: specificity and appropriateness of the evidence used; and • **Language Finesse**: fluency, depth, and

Dimension	Structure	Logic	Argu.	Evidence	Language
Pearson’s r	0.66	0.69	0.73	0.66	0.64

**Table 1: Consistency analysis of Human and GPT-4 on five dimensions. Arug. refers to Argument, and r means the Pearson correlation coefficient.**

vividness of the expression style. Besides, we calculate the average of the five scores as **Overall**.

The prompt templates are as follows:

*[You are an expert in scoring generated commentaries. Please rate your answers from the {perspective} perspective based on the provided commentary. The scoring criteria are:*

*(1) 10 points represent... (2) 8 points represent... (3) 6 points ...*

*Please output a line that contains only one value representing the score. Please avoid any potential biases, and ensure that there are no factors other than the text that affect your judgment.]*

These dimensions are chosen to encompass both the fundamental and advanced requirements of commentary. Structural soundness and logical consistency constitute the fundamental requirements, ensuring a well-organized and logically coherent commentary. Conversely, the quality of argumentation and the adequacy of evidentiary support represent the advanced requirements, reflecting the depth and persuasiveness of the commentary. The dimensions are scored on a scale of 1-10, with 1 being the lowest and 10 the highest.

In our experiments, we utilize GPT-4 for automatic evaluation by crafting specific prompts for each dimension. To validate GPT-4’s accuracy, we compare its scores against those from human annotators for 30 randomly selected commentaries, calculating the Pearson correlation coefficient [5] for each dimension. As Tab. 1 illustrates, the Pearson correlation coefficient of each dimension surpasses 0.6, which proves GPT-4 is competent for this task.

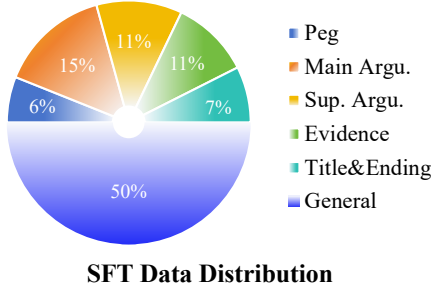
**Human evaluation.** In our ablation study, we assess the **Timeliness** of evidence. Due to the training limitations of GPT-4, which is based on data available only up to a specific date, it is not equipped to accurately ascertain the recency of evidence. Therefore, this aspect is evaluated through human judgment. The scoring for this metric ranges from 1 to 10, where 1 represents the lowest and 10 the highest possible score.

### 5.2 Experiment Settings

**Implementation.** The base model of Xinyu is LLaMA2-13B [27], and we specifically adapted it to better accommodate the nuances of the Chinese language. This adaptation involved expanding the LLaMA-13B tokenizer with an additional 28,000 Chinese tokens. To further optimize the model, we continued pre-training on LLaMA-13B using a corpus comprising 500B tokens, which contains both English and Chinese corpus. For supervised fine-tuning, we not only utilized the dataset introduced in Section 4 but also incorporated the general SFT dataset to maintain consistency with the data distribution of previous training phases. The amount of SFT data is 400,000 and the distribution of it is shown in Figure. 3. Our training process leveraged the Megatron-DeepSpeed framework.

[https://huggingface.co/datasets/BelleGroup/train\\_2M\\_CN](https://huggingface.co/datasets/BelleGroup/train_2M_CN)





**Figure 3: The distribution of SFT dataset. Argu. refers to Argument, Sup. means Supporting.**

The continued pre-training phase lasted 20 days on 128 Nvidia A800 80G GPUs, while the supervised fine-tuning (SFT) process took 2 days on 8 Nvidia A800 80G GPUs.

**Baselines.** We employ the following methods as our baselines:

Baichuan2 [35] represents a series of large-scale, multilingual language models trained from scratch on 2.6 trillion tokens. We select **Baichuan2-13B-Chat** and **Baichuan2-Turbo** as our baseline models. Qwen [2] is a comprehensive series of language models featuring a range of models with varying parameter counts. In this context, we choose **Qwen-72B** and **Qwen-72B-Chat** as our baseline models. InternLM [26] consists of a series of multilingual foundation models and chat models, with **InternLM-20B-Chat** selected as the baseline model. GLM [38] is a series of bilingual (English and Chinese) pre-trained language models, for which we use **GLM-4** as the baseline model. ERNIE [25] serves as a unified framework for pre-training large-scale knowledge-enhanced models, with **ERNIE-4** chosen as our baseline model. Finally, GPT is a series of large language models released by OpenAI, with **GPT-3.5-Turbo** [18] and **GPT-4** [17] used as baseline models.

**Test Cases.** For our test cases, we have carefully chosen 41 commentaries from the “Three Commentaries” section of People’s Daily Online. This selection encompasses a diverse range of topics including economics, livelihood, technology, culture, social issues, sports, and art, reflecting current news and societal trends. The prominence of the site and the authoritative nature of the series ensure that these articles represent high-quality journalistic commentary. In our baselines, we employ a one-step generation process using Event Detail, and extra Title, Argument, and Evidence, which are from the real commentary of the event. This setup is designed to emulate real-world scenarios where commentators use LLMs. The translated prompt in English is:

[I will provide you with a news background: {Event detail}]

Based on this news, with the title ‘Title’, please create a commentary article. The article should have clear and profound arguments, true and abundant evidence, smooth logical reasoning, reasonable structure, and appropriate commentary language. Your article should reference the following argument and evidence: Argument 1: {Argument 1} Evidence 1: {Evidence 1} Argument 2: {Argument 2} Evidence 2: {Evidence 2} Argument 3: {Argument 3} Evidence 3: {Evidence 3}]

<http://opinion.people.com.cn/GB/8213/420650/index.html>

The number of *Arguments* and *Evidence* will be 0 to 3.

However, for Xinyu, we only supply the event details to evaluate the effectiveness of our approach in generating commentary.

**Ablation.** In our ablation experiments, we adopt the following configurations to assess specific components:

1) Evaluating the impact of the framework: While maintaining the overall framework intact, we replace the Xinyu-13B model with alternative large language models (LLMs) to determine the effectiveness of the framework.

2) Assessing ranking efficiency: We eliminate the argument ranking component, allowing the LLM to directly generate a main argument without a predefined direction, to evaluate the ranking model’s contribution.

3) Investigating the role of RAG: By omitting the retrieval process from the evidence database, we let the LLM independently produce the evidence, aiming to understand the significance of the RAG component in enhancing model performance and the effectiveness of our construed evidence database. Moreover, we assess the timeliness of the generated evidence especially.

4) Measuring the practicality of Xinyu: We compared the quality of commentary articles generated by human experts with a journalism background with the aid of Xinyu, against those written purely manually, while also comparing the time spent. We randomly select 10 test cases.

### 5.3 Experimental Results and Analysis

Based on the model’s size, we split them into two types: LLMs larger than 20B and LLMs smaller than 20B. Tab. 2 shows the results of commentary generation with GPT-4’s evaluation. We report the results of the ablation study in Tab. 3, Tab. 4, Tab. 5, and Figure. 5.

**Results of commentary generation.** From Tab. 2, we can conclude that: (1) Generally, the bigger the language model’s size, the better it does. However, GPT-4’s leading advantage in this task is not as pronounced as in other generative tasks. (2) When looking at models within the 20 billion parameter size, our method achieved the best results in most of the metrics. (3) Compared to large-scale LLMs such as GPT-4, our method attained the best overall score, primarily due to our superior performance in the advanced requirements of argument and evidence. (4) There’s not a huge gap between the scores of the different methods. This is largely because GPT-4 is generally not harsh in its scoring, rarely giving out very low scores.

**Result of ablation study.** From Tab. 3, we have the following observations: (1) Our framework significantly enhances the performance of large-sized base models. For instance, the overall score of Qwen-72B increased from 7.37 to 7.82. (2) GPT-4 achieved the best performance with an overall score of 8.3, and our Xinyu ranks just behind GPT-4. Considering the size of the model, our method has greater potential in practice. (3) For the 20B scale base models such as Qwen-14B-Chat, using our framework actually decreased their performance. This might be due to these base models’ inherent limitations in generating text step-by-step. This also demonstrates the effectiveness of our SFT.

From Tab. 4, we have the following observations: (1) The implementation of the argument ranking model has significantly improved the effectiveness of argumentation, underscoring its impact.

Methods	Overall	Structure	Logic	Argument	Evidence	Language
Baichuan2-Turbo [35]	7.65	8.11*	8.28	8.06	5.63	8.17
Qwen-72B [2]	7.37	7.90	7.88	7.87	5.28	7.90
GLM-4 [38]	7.72	8.11*	8.35	8.08*	5.82	8.24
ERNIE-4 [25]	7.71	8.05	8.35	8.03	5.73	8.38*
GPT-3.5-Turbo [18]	7.70	8.00	8.08	8.00	6.39	8.05
GPT-4 [17]	7.78	8.05	8.40*	8.05	6.02	8.38*
Baichuan2-13B-Chat [35]	7.37	<b>8.11*</b>	8.00	7.92	4.78	<b>8.05</b>
Qwen-14B-Chat [2]	7.25	8.05	7.88	7.85	4.55	7.95
InternLM-20B-Chat [26]	7.26	7.80	7.83	7.88	4.83	8.00
Xinyu (based on fine-tuned LLaMA2-13B)	<b>7.93*</b>	8.00	<b>8.20</b>	<b>8.00</b>	<b>7.41*</b>	<b>8.05</b>

**Table 2: Evaluation of commentary generated by baseline LLMs using GPT-4. Reference is the published commentaries. ‘Bold’ indicates the highest score within the 20B scale baselines, and an asterisk (\*) denotes the highest score among all baselines.**

Base Model	Overall	Structure	Logic	Argument	Evidence	Language
Qwen-72B-Chat [2]	7.82	7.80	7.85	7.75	7.80	7.90
Baichuan2-Turbo [35]	7.91	8.00	7.93	7.70	8.11	7.80
GPT-4 [17]	8.30*	8.10*	8.58*	8.15*	8.17*	8.50*
Baichuan2-13B [35]	6.31	5.50	6.55	6.78	6.49	6.23
Qwen-14B-Chat [2]	6.22	5.60	6.25	6.25	6.80	6.20
Xinyu	<b>7.93</b>	<b>8.00</b>	<b>8.20</b>	<b>8.00</b>	<b>7.41</b>	<b>8.05</b>

**Table 3: GPT-4’s evaluation of commentary generated by our framework with different base models. ‘Bold’ indicates the highest score within the 20B scale baselines, and an asterisk (\*) denotes the highest score among all baselines.**

Methods	Overall	Structure	Logic	Argument	Evidence	Language
w/o Ranking	7.85	7.93	8.20	7.80	7.31	8.02
Xinyu	<b>7.93</b>	<b>8.00</b>	<b>8.20</b>	<b>8.00</b>	<b>7.41</b>	<b>8.05</b>

**Table 4: Results of ablation study on Argument Ranking Model.**

<b>Peg:</b> Dense nets in a national reserve are killing protected birds, some served as food in restaurants.		
<b>Xinyu:</b> In national nature reserves, the occurrence of widespread and persistent illegal hunting activities represents a profound desecration of the ecosystem, a critical waypoint for migratory birds... Currently, during the migratory season, the bird-catching nets pose a significant threat ... According to the national laws ... The situation reported reflects the evolution of illegal hunting into a complex chain of interests ... As media undercover investigations have found, bird carcasses are being supplied to local restaurants ... This phenomenon indicates a severe lack of management and supervision in the reserves ... The public has already shown extreme concern on ... To fundamentally reduce illegal hunting activities, it is necessary for all departments to join forces and form a tight regulatory network. The successful experience in Guangdong can serve as a reference... It is hoped that through such measures, we can effectively protect our natural heritage and ...	<b>GPT-4:</b> The migration of migratory birds is a natural wonder, yet the presence of bird-catching nets disrupts this natural order, causing severe damage to the ecological environment ... In the face of this situation, we cannot stand idly by. We must enhance the promotion and education of laws and regulations to make the public understand that bird catching is not only illegal but also harmful to the ecological environment ... Furthermore, we need to strengthen international cooperation to combat illegal bird catching together. The migration of birds knows no borders, and protecting migratory birds is equivalent to protecting our shared home, the Earth ... Additionally, we must utilize modern technological methods, such as drones and satellite remote sensing, to conduct comprehensive, no-blind-spot monitoring of nature reserves ... We are confident that with our collective efforts, we can build a beautiful home where humans and nature coexist harmoniously.	<b>Baichuan2-13B-Chat:</b> Recent news reports on serious bird catching issues on some islands of national nature reserves have sparked widespread concern and prompted deep reflection across society on the protection of wildlife ... Firstly, the migration of migratory birds is a natural marvel, essential not only for the reproduction of bird species but also as a crucial component of species migration and energy flow within ecosystems ... Moreover, there's a need to enhance public awareness of environmental protection. Protecting wildlife is not only the government's responsibility but also that of every citizen. ... ‘The problem of forests covered with bird-catching nets’ is not just an isolated case but reflects broader issues in wildlife protection efforts. We should learn from this issue, strengthen regulatory mechanisms, increase public environmental awareness, and collectively safeguard our ecological environment. Only through joint efforts can we achieve harmonious coexistence between humans and nature ...

**Figure 4: Case study. The content is translated from Chinese.**

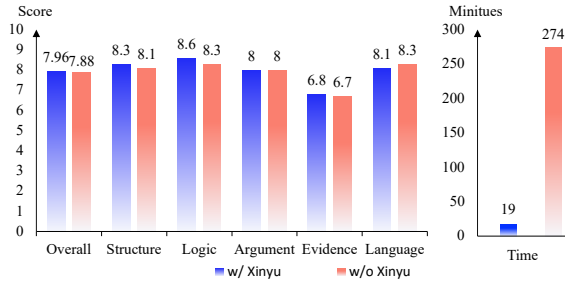
(2) Enhancements are observed across all metrics, illustrating the interrelationship among these aspects of commentary.

From Tab. 5, we can find that: (1) The implementation of retrieval augmented generation (RAG) significantly enhances the generation of evidence, with scores improving from 5.60 to 7.41. Additionally,



Methods	Evidence	Timeliness
w/o RAG	5.60	8.85
w/ Event	7.14	9.10
w/ Event + Book	<b>7.41</b>	<b>9.30</b>

**Table 5: Results of ablation study on Evidence Database. The Kappa value of Timeliness exceeds 0.78.**



**Figure 5: Human vs. Xinyu-Assisted.**

the timeliness of the generated evidence also saw an increase, rising from 8.85 to 9.30. (2) After incorporating the book dataset into RAG, its performance experienced further improvements. (3) The improvement proves the effectiveness of our evidence dataset.

From Figure. 5, we have the following observations: (1) Utilizing Xinyu’s assistance can significantly increase writing speed, and the average time for a commentary speeds up from more than 4 hours to 20 mins. (2) Moreover, commentaries generated with LLMs have achieved the same scores as manual writing, demonstrating the practicality of our system.

*Case study.* Figure 4 presents three commentaries on a certain peg generated by Xinyu, GPT-4, and Baichuan2-13B-Chat, respectively. All three commentaries exhibit good language fluency and structural coherence, highlighting the capabilities of these large language models (LLMs). However, the commentary from Baichuan2-13B-Chat focuses solely on facts without offering specific arguments. In contrast, both GPT-4 and Xinyu provide detailed arguments. Notably, Xinyu’s commentary stands out by presenting more convincing evidence and demonstrating a logical correlation in its supporting arguments.

## 6 CONCLUSION

In this paper, we introduce Xinyu, an innovative commentary generation system based on large language models (LLMs) designed to enhance the efficiency of commentators. Our approach involves breaking down the generation process into five steps, with supervised fine-tuning (SFT) applied to each step to ensure the output is well-structured and coherent, addressing the basic requirements of commentary. To fulfill the higher demands for novelty and persuasiveness, we develop an argument ranking model and employ retrieval-augmented generation (RAG) techniques for evidence generation. For RAG, we have compiled an evidence database comprising both current events and classical books. To better measure

the generated commentaries, we design a comprehensive evaluation method with 5 distinct perspectives. Our comprehensive experiments demonstrate the system’s effectiveness. Remarkably, in practical applications, Xinyu has reduced the average commentary creation time from 4 hours to just 20 minutes and maintained the quality.

In the future, we will consider the following directions to enhance our system: 1) Improve evidence recall accuracy, ensuring relevance to the arguments; 2) Utilize Reinforcement Learning with Human Feedback (RLHF) to better align commentaries with human preferences and specific writing styles.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (62441605, 62376243, 62037001, U20A20387), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010).

Finally, we would like to thank the anonymous reviewers for their helpful feedback and suggestions.

## REFERENCES

- [1] Ibrahim Adeshola and Adeola Praise Adepoju. 2023. The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments* (2023), 1–14.
- [2] 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. CoRR abs/2309.16609 (2023). <https://doi.org/10.48550/ARXIV.2309.16609> arXiv:2309.16609
- [3] Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. 'Don't Get Too Technical with Me': A Discourse Structure-Based Framework for Automatic Science Journalism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 1186–1202. <https://aclanthology.org/2023.emnlp-main.76>
- [4] Szu-Wei Cheng, Chung-Wen Chang, Wan-Jung Chang, Hao-Wei Wang, Chih-Sung Liang, Taishiro Kishimoto, Jane Pei-Chen Chang, John S Kuo, and Kuan-Pin Su. 2023. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and clinical neurosciences* 77, 11 (2023), 592–596.
- [5] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing* (2009), 1–4.
- [6] Jiayi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. CoRR abs/2306.16092 (2023). <https://doi.org/10.48550/ARXIV.2306.16092> arXiv:2306.16092
- [7] Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large Language Models in Education: Vision and Opportunities. CoRR abs/2311.13160 (2023). <https://doi.org/10.48550/ARXIV.2311.13160> arXiv:2311.13160
- [8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv:2212.10496 [cs.IR]
- [9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. CoRR abs/2312.10997 (2023). <https://doi.org/10.48550/ARXIV.2312.10997> arXiv:2312.10997
- [10] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot Learning with Retrieval Augmented Language Models. arXiv:2208.03299 [cs.CL]
- [11] Xianming Li and Jing Li. 2023. AngIE-optimized Text Embeddings. CoRR abs/2309.12871 (2023). <https://doi.org/10.48550/ARXIV.2309.12871> arXiv:2309.12871
- [12] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain

- Knowledge. *CoRR* abs/2303.14070 (2023). <https://doi.org/10.48550/ARXIV.2303.14070> arXiv:2303.14070
- [13] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. ChatCounselor: A Large Language Models for Mental Health Support. *CoRR* abs/2309.15461 (2023). <https://doi.org/10.48550/ARXIV.2309.15461> arXiv:2309.15461
  - [14] Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Ml-lip: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1023–1034.
  - [15] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting for Retrieval-Augmented Large Language Models. *CoRR* abs/2305.14283 (2023). <https://doi.org/10.48550/ARXIV.2305.14283> arXiv:2305.14283
  - [16] Kamil Malinka, Martin Peresini, Anton Firc, Ondrej Hujnak, and Filip Janus. 2023. On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree?. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2023, Turku, Finland, July 7-12, 2023*, Mikko-Jussi Laakso, Mattia Monga, Simon, and Judithe Sheard (Eds.). ACM, 47–53. <https://doi.org/10.1145/3587102.3588827>
  - [17] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). <https://doi.org/10.48550/ARXIV.2303.08774> arXiv:2303.08774
  - [18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Madie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). [http://papers.nips.cc/paper\\_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html)
  - [19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. <https://doi.org/10.3115/1073083.1073135>
  - [20] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. arXiv:2305.15294 [cs.CL]
  - [21] Kai Shen, Yichong Leng, Xu Tan, Siliang Tang, Yuan Zhang, Wenjie Liu, and Edward Lin. 2022. Mask the correct tokens: An embarrassingly simple approach for error correction. arXiv preprint arXiv:2211.13252 (2022).
  - [22] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv:2301.12652 [cs.CL]
  - [23] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathanael Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkumar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. *CoRR* abs/2212.13138 (2022). <https://doi.org/10.48550/ARXIV.2212.13138> arXiv:2212.13138
  - [24] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. *CoRR* abs/2305.09617 (2023). <https://doi.org/10.48550/ARXIV.2305.09617> arXiv:2305.09617
  - [25] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chaopang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *CoRR* abs/2107.02137 (2021). arXiv:2107.02137 <https://arxiv.org/abs/2107.02137>
  - [26] InternLM Team. 2023. InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities. <https://github.com/InternLM/InternLM>.
  - [27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). <https://doi.org/10.48550/ARXIV.2302.13971> arXiv:2302.13971
  - [28] VoyageAI. 2023. VoyageAI. Voyage’s embedding models. <https://docs.voyageai.com/embeddings/>.
  - [29] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. arXiv:2303.07678 [cs.IR]
  - [30] Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court’s view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 763–780.
  - [31] Yiquan Wu, Weiming Lu, Yating Zhang, Adam Jatowt, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Focus-aware response generation in inquiry conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 12585–12599.
  - [32] Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration. arXiv preprint arXiv:2310.09241 (2023).
  - [33] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. *arXiv:2309.07597* [cs.CL]
  - [34] Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. arXiv:2310.04408 [cs.CL]
  - [35] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyi Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open Large-scale Language Models. *CoRR* abs/2309.10305 (2023). <https://doi.org/10.48550/ARXIV.2309.10305> arXiv:2309.10305
  - [36] Linan Yue, Qi Liu, Yichao Du, Weibo Gao, Ye Liu, and Fangzhou Yao. 2023. FedJudge: Federated Legal Large Language Model. *CoRR* abs/2309.08173 (2023). <https://doi.org/10.48550/ARXIV.2309.08173> arXiv:2309.08173
  - [37] Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services. *CoRR* abs/2309.11325 (2023). <https://doi.org/10.48550/ARXIV.2309.11325> arXiv:2309.11325
  - [38] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=Aw0rrrPUF>
  - [39] Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Haorui Wang, Zhen Qin, Feng Han, Jialu Liu, Simon Baumgartner, Michael Bendersky, and Chao Zhang. 2024. PLAD: Preference-based Large Language Model Distillation with Pseudo-Preference Pairs. arXiv:2406.02886 [cs.CL]
  - [40] Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Li Yun, Hejie Cui, Zhang Xuchao, Tianjiao Zhao, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. arXiv preprint arXiv:2305.18703 (2023).
  - [41] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. *CoRR* abs/2305.11206 (2023). <https://doi.org/10.48550/ARXIV.2305.11206> arXiv:2305.11206
  - [42] Siying Zhou, Yifei Liu, Yiquan Wu, Kun Kuang, Chunyan Zheng, and Fei Wu. 2022. Similar case based prison term prediction. In *CAAI International Conference on Artificial Intelligence*. Springer, 284–297.

## A APPENDIX

### A.1 Overall Generation Process

This section presents a complete example corresponding to each step in section 4.1. The overall process consists of Peg Generation -> Main Argument Generation -> Supporting Argument Generation -> Evidence Generation -> Ending Generation & Title Generation. All the content is translated from Chinese.

<b>Prompt:</b> You are a commentary writing expert, and here are the details of an event. Event detail: { <i>event detail</i> }. Please refine it into a concise and well-articulated peg:
<b>Event Detail:</b> Recently, the National Medical Insurance Administration released the "Management Measures for the Evaluation of Disability Levels in Long term Care Insurance (Trial)" and the "Designated Management Measures for the Evaluation of Disability Levels in Long term Care Insurance (Trial)" to solicit public opinions and attract attention.
<b>Generated Peg:</b> The National Medical Insurance Administration has released the Management Measures for Disability Level Evaluation of Long term Care Insurance and the Management Measures for Designated Evaluation Institutions, publicly soliciting opinions.

Figure 6: Peg Generation.

<b>Prompt:</b> You are a commentary writing expert. Please complete the main argument in the direction of { <i>direction</i> } based on the peg: { <i>peg</i> } and event detail: { <i>event detail</i> }. The main argument should be profound, concise, and strongly related to the peg. Please provide the main argument:
<b>Direction:</b> Social
<b>Generated Main Argument:</b> This move aims to promote the improvement and development of the long-term care insurance system, in order to address the pain points of the people and promote industrial development. In the future, we will further improve and develop the long-term care insurance system, provide better medical security for the people, and create more development opportunities for society.

Figure 7: Main Argument Generation.

<b>Prompt:</b> You are a commentary writing expert. Based on the given main argument { <i>main argument</i> } of the commentary and event detail { <i>event detail</i> }, generate multiple supporting arguments for the commentary. The supporting arguments form { <i>structure</i> } structure, refining around the { <i>main argument</i> } with multi-level, multi-faceted, and multi-angle perspectives. Please provide the supporting arguments:
<b>Structure:</b> Parallel
<b>Generated Supporting Argument:</b> (1) Long-term care insurance is a system arrangement that provides nursing protection and economic compensation for long-term disabled individuals. It is a significant livelihood project aimed at alleviating the burden on families and institutions, enhancing the sense of gain and happiness among the people. (2) The development of long-term care insurance not only addresses the pain points of the people but also serves as an effective means to promote social development. (3) Currently, the pilot program for long-term care insurance is steadily advancing, with continuously improving service quality, meeting the diverse and differentiated long-term care needs of the masses. (4) To promote the high-quality development of long-term care insurance, there is still a long way to go.

Figure 8: Supporting Argument Generation.

<b>Prompt:</b> You are a commentary writing expert. Surrounding the main argument { <i>main argument</i> }, please use the evidence provided in the reference information, including dates, data, viewpoints, core content, etc., to continue writing evidence in the commentary to support the supporting argument { <i>supporting argument</i> }. Please annotate the corresponding reference information numbers in the continuation. Reference information: { <i>reference</i> }. Please provide the evidence:
<b>Supporting Argument:</b> The National Healthcare Security Administration has released the assessment and management measures for long-term care insurance, aiming to refine the system, address public concerns, and stimulate industry growth, thereby enhancing medical security. Long-term care insurance is designed to offer nursing care and financial compensation for individuals with long-term disabilities, serving as a significant welfare project to alleviate the burden on families and institutions, and to increase the sense of fulfillment and well-being among the populace.
<b>Reference:</b> [Xm5LIRp0]: The third involves insurance companies building comprehensive long-term care service platforms (referred to as the "P-side"), engaging in comprehensive commercial models related to long-term care insurance, such as disability status assessment, nursing service coordination, health management, and industry integration. [JSeZwQPE]: "Developing long-term care insurance cannot be achieved by individual companies alone. It requires the entire industry to unite and jointly promote its development," said Jiang Hao, from the Development Strategy Department of Taikang Anlian. The commercial health insurance market in China is a highly prospective future blue ocean. We believe that long-term care insurance tailored to market demands will soon emerge, filling market gaps. ... [VAqkcV45]: Original Title: China Insurance Industry Association and the Population and Labor Economics Research Institute of the Chinese Academy of Social Sciences jointly released the "2018-2019 China Long-Term Care Research Report." [CRTsiVcL]: In recent years, many pilot cities have actively explored the long-term care insurance (hereinafter referred to as "long-term care insurance") system, ensuring basic living care for many long-term disabled individuals and closely related medical care, leading to an increase in residents' happiness and sense of gain.
<b>Generated Evidence:</b> The release of long-term care insurance assessment and management methods by the National Healthcare Security Administration marks an important step in addressing the challenges of an aging population and actively constructing a senior care service system. Currently, the population aged 60 and above has approached 250 million, with more than 40 million elderly individuals living with disabilities. The demand for professional care services is immense, yet supply falls short... As policies continue to be refined and practices deepen, long-term care insurance will better meet societal needs, promote the healthy and sustainable development of the elder care industry, and bring tangible benefits to the elderly population.

Figure 9: Evidence Generation.

<p><b>Prompt:</b> You are a commentary writing expert. Please write an ending for the article, maintaining the smooth language, consistent style, and logical coherence with the preceding text. The preceding text is as follows: {<b>preceding text</b>}. Please provide the ending:</p>
<p><b>Preceding Text:</b> The National Medical Insurance Administration has issued the "Management Measures for the Evaluation of Disability Levels and Designated Management Measures for Evaluation Agencies of Long-term Care Insurance" for public comments... In the future, we will further improve and develop the long-term care insurance system to provide better medical security for the people and create more development opportunities for society.</p> <p>Long-term care insurance is a system arrangement that provides care protection and economic compensation for long-term disabled individuals... With continuous policy improvement and deepening of practices, long-term care insurance will better meet social needs, promote the healthy and sustainable development of the elderly care industry, and bring tangible benefits to the late-life of the general public.</p> <p>The development of long-term care insurance not only addresses the pain points of the people but also serves as an effective means to promote social development... We have reason to believe that with the continuous improvement and promotion of the long-term care insurance system, it will become an important force to promote social development and enhance the well-being of the people in China.</p> <p>Currently, the pilot implementation of the long-term care insurance system is steadily advancing, with continuously improving service quality, meeting the diversified and differentiated long-term care needs of the masses... With the continuous improvement of the long-term care insurance system, we have reason to believe that it will better adapt to the diversified and differentiated needs of the elderly, bringing more dignity and comfort to their later years.</p> <p>To promote the high-quality development of long-term care insurance, there is still a long way to go... Only in this way can we ensure that every elderly person in need can enjoy a dignified and decent late life.</p>
<p><b>Generated Ending:</b> Long-term care insurance is an important livelihood project and a complex systemic project. Its improvement and development require joint efforts from the government, society, businesses, and individuals. The government needs to strengthen policy guidance and regulation, enhance service quality, and guarantee levels of protection. Society needs to enhance publicity and education, increase public awareness, and participation. Businesses and individuals need to strengthen cooperation and communication to jointly promote the development and innovation of long-term care insurance. Only in this way can we truly make long-term care insurance a benefit for the people and create more opportunities and value for society.</p>

Figure 10: Ending Generation.

<p><b>Prompt:</b> You are a commentary writing expert. Please write a title for the article and provide insightful summaries. You should use your experience to thoughtfully summarize important matters, support claims with facts, and distill the essence of the text. The preceding text is as follows: {<b>preceding text</b>}. Please provide the title:</p>
<p><b>Generated Title:</b> Long-term care insurance is an important livelihood project and a complex systemic project. Its improvement and development require joint efforts from the government, society, businesses, and individuals. The government needs to strengthen policy guidance and regulation, enhance service quality, and guarantee levels of protection. Society needs to enhance publicity and education, increase public awareness, and participation. Businesses and individuals need to strengthen cooperation and communication to jointly promote the development and innovation of long-term care insurance. Only in this way can we truly make long-term care insurance a benefit for the people and create more opportunities and value for society.</p>

Figure 11: Title Generation.