MIMIR: A Customizable Agent Tuning Platform for Enhanced Scientific Applications

**Xiangru Tang[♠*], Chunyuan Deng[♣*], Hanmin Wang[♠*], Haoran Wang[♠*],
Yilun Zhao[♠], Wenqi Shi[♣], May Fung[♡], Wangchunshu Zhou, Jiannan Cao[♠],
Heng Ji[♡], Arman Cohan[♠], Mark Gerstein[♠]**
[♠] Yale University [♣] Georgia Tech [♡] UIUC
{xiangru.tang,arman.cohan,mark.gerstein}@yale.edu
https://github.com/gersteinlab/MIMIR

Abstract

Recently, large language models (LLMs) have evolved into interactive agents, proficient in planning, tool use, and task execution across various tasks. However, without agent-tuning, open-source models like LLaMA2 currently struggle to match the efficiency of larger models such as GPT-4 in scientific applications due to a lack of *agent tuning* datasets. In response, we introduce MIMIR, a streamlined platform that leverages large LLMs to generate agent-tuning data for fine-tuning smaller, specialized models. By employing a role-playing methodology, MIMIR enables larger models to simulate various roles and create interaction data, which can then be used to fine-tune open-source models like LLaMA2. This approach ensures that even smaller models can effectively serve as agents in scientific tasks. Integrating these features into an end-to-end platform, MIMIR facilitates everything from the uploading of scientific data to one-click agent fine-tuning. MIMIR is publicly released and actively maintained at https://github.com/gersteinlab/MIMIR, along with a demo video[1] for quick-start, calling for broader development.

1 Introduction

Recently, large language models (LLMs) have undergone a significant evolution, transitioning into interactive agents that have demonstrated considerable progress in many scientific scenarios (OpenAI, 2022, 2023; Anthropic, 2023; Google, 2023). The commendable performance of these models across various downstream tasks has incited researchers to propose methods for utilizing LLMs to generate instruction datasets (Peng et al., 2023; Wang et al., 2023c; Sun et al., 2023a). The quality and diversity of such data are instrumental in aligning, pre-training, and fine-tuning LLMs (Sun et al., 2023b; Chiang et al., 2023; Taori et al., 2023; Xu et al., 2023; Li et al., 2023; Shao et al., 2023; Ding et al., 2023). Besides promoting methods for general instruction tuning to enhance the capability of LLMs, there is increasing research emphasis on fine-tuning LLMs to acquire tool usage (Schick et al., 2023; Zhang et al., 2024; Zhou et al., 2023) and establish stronger agent abilities (Chen et al., 2023; Qin et al., 2023; Zeng et al., 2023) in scientific applications.

While there is an abundance of datasets available for instruction tuning (Wang et al., 2023c; Chiang et al., 2023; Xu et al., 2023; Li et al., 2023; Ivison et al., 2023), datasets specifically focused on agent tuning (Zeng et al., 2023) are in short supply. This imbalance has inculcated reliance on proprietary LLMs such as ChatGPT or GPT-4 as mainstay tools for reasoning and planning in scientific applications. This dependency raises significant privacy concerns, especially when integrating sensitive domain knowledge, such as EHR data, into model training, as highlighted by Kim et al. (2023) and Tian et al. (2023). Furthermore, concentrating solely on fine-tuning LLMs with tool-learning datasets might inadvertently compromise their ability to master rare domain-specific knowledge and perform complex reasoning, a concern raised by Zeng et al. (2023). For example, an LLM trained on a dataset focused on diagnosing cardiovascular diseases might overlook a critical anomaly that falls outside its training scope, such as an unusual symptom of a rare cancer. Additionally, if we finetune GPT-3.5, incorporating confidential patient data into a public model can heighten the risk of privacy breaches.

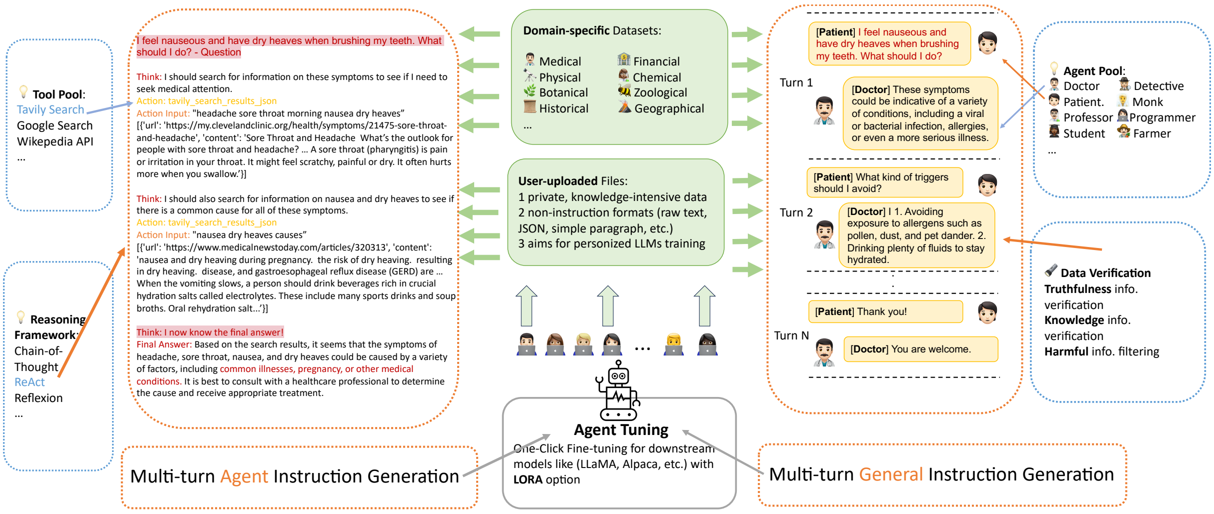
In this paper, we introduce MIMIR[2], a novel streamlined platform, as illustrated in Figure

---

[1] https://www.youtube.com/watch?v=7fVgv_T_xjc. * means contributed equally.

[2] https://github.com/gersteinlab/MIMIR. Detailed instructions and demonstrations for the platform's use and dataset integration can be found after deployment.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 486–496
November 12-16, 2024 ©2024 Association for Computational Linguistics


Tool Pool:
Tavily Search
Google Search
Wikepedia API
...
Reasoning Framework:
Chain-of-Thought
ReAct
Reflexion
...
Multi-turn Agent Instruction Generation
Domain-specific Datasets:
Medical
Physical
Botanical
Historical
Financial
Chemical
Zoological
Geographical
...
User-uploaded Files:
1 private, knowledge-intensive data
2 non-instruction formats (raw text, JSON, simple paragraph, etc.)
3 aims for personized LLMs training
...
Agent Tuning
One-Click Fine-tuning for downstream models like (LLaMA, Alpaca, etc.) with LORA option
Turn 1
Turn 2
Turn N
Agent Pool:
Doctor
Detective
Patient.
Monk
Professor
Programmer
Student
Farmer
...
Data Verification
Truthfulness info. verification
Knowledge info. verification
Harmful info. filtering
Multi-turn General Instruction Generation

I feel nauseous and have dry heaves when brushing my teeth. What should I do? - Question

Think: I should search for information on these symptoms to see if I need to seek medical attention.
Action: tavily_search_results_json
Action Input: "headache sore throat morning nausea dry heaves"
[{'url': 'https://my.clevelandclinic.org/health/symptoms/21475-sore-throat-and-headache', 'content': 'Sore Throat and Headache What's the outlook for people with sore throat and headache? ... A sore throat (pharyngitis) is pain or irritation in your throat. It might feel scratchy, painful or dry. It often hurts more when you swallow.'}]

Think: I should also search for information on nausea and dry heaves to see if there is a common cause for all of these symptoms.
Action: tavily_search_results_json
Action Input: "nausea dry heaves causes"
[{'url': 'https://www.medicalnewstoday.com/articles/320313', 'content': 'nausea and dry heaving during pregnancy. the risk of dry heaving. resulting in dry heaving. disease, and gastroesophageal reflux disease (GERD) are ... When the vomiting slows, a person should drink beverages rich in crucial hydration salts called electrolytes. These include many sports drinks and soup broths. Oral rehydration salt...'}]

Think: I now know the final answer!
Final Answer: Based on the search results, it seems that the symptoms of headache, sore throat, nausea, and dry heaves could be caused by a variety of factors, including common illnesses, pregnancy, or other medical conditions. It is best to consult with a healthcare professional to determine the cause and receive appropriate treatment.

[Patient] I feel nauseous and have dry heaves when brushing my teeth. What should I do?

[Doctor] These symptoms could be indicative of a variety of conditions, including a viral or bacterial infection, allergies, or even a more serious illness.

[Patient] What kind of triggers should I avoid?

[Doctor] I 1. Avoiding exposure to allergens such as pollen, dust, and pet dander. 2. Drinking plenty of fluids to stay hydrated.

[Patient] Thank you!

[Doctor] You are welcome.

Figure 1: *MIMIR* provides an integrated pipeline from generating **multi-role** and **multi-turn** instructions to a one-click fine-tuning process for downstream models. Users also could upload their files combined with existing domain-specific datasets to customize their instruction data.

1. MIMIR is adept at tackling the challenges in specialized scientific fields such as medicine, biology, physics, and chemistry. A primary obstacle in these domains is the significant variation in domain-specific knowledge and the need to perform complex reasoning on non-standard data formats. Additionally, some fields also involve handling sensitive data. MIMIR facilitates the integration of proprietary knowledge data with an established, external domain-specific knowledge base. Through this integration, MIMIR simultaneously generates a multi-turn agent tuning dataset, which includes multiple rounds of interactions between the user and the agent to enhance the agent's performance in complex, domain-specific scenarios. To construct the general multi-turn agent instruction tuning dataset, we adopt the method proposed in Park et al. (2023), employing LLMs as interactive agents in multi-round conversations. Specifically, MIMIR integrates reasoning frameworks and search tools to generate interaction trajectories. This allows users to tailor templates within these frameworks, offering demonstrations that align with their specific objectives.

Our pipeline seamlessly incorporates private and public knowledge bases, agent-tuning data generation protocols, multi-role agent configurations, and one-click fine-tuning into a unified flow. The tuning data generated through this pipeline is more accurate and credible. Notably, in comparison to original data and other agent tuning systems, like self-instruct (Wang et al., 2023c), and Baize (Xu et al., 2023), we achieved win or equal rates of **87%**, **75%**, and **77%**, respectively. We summarize the key features of MIMIR as follows:

- **Simple and User-friendly.** For users unfamiliar with agent tuning, activating agent capabilities using open-source models such as LLaMA2 is feasible, facilitating the creation of agents in scientific applications.
- **Private and Dataset Integration.** Users can seamlessly integrate public datasets with their proprietary knowledge bases using MIMIR offline, ensuring data privacy and avoiding leakage issues.
- **Domain-Specific Role-playing.** Our system supports domain-specific role-playing during the generation of domain-specific data. For example, it facilitates multi-turn interactions among various medical roles, including doctors, patients, and medical students, for the creation of medical domain data.
- **One-Click Fine-Tuning.** Using parameter visualization and LoRA technology (Hu et al., 2021), users can formulate and implement customized fine-tuning scripts for LLMs, thereby optimizing performance and efficiency.

2 Background and Related Work

Domain-Specific Instruction Data Generation Following the success of ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), open-source LLMs like LLaMA2, Alpaca (Taori et al., 2023) and

Mistral (Jiang et al., 2023) have arisen, all requiring instruction data for training. Although these LLMs exhibit remarkable performance in general domains, their lack of domain-specific knowledge results in inadequate performance in scientific areas that require specialized expertise. Several efforts have been made to adapt LLMs to these domain-specific scientific areas. This typically involves generating domain-specific data to fine-tune such LLMs, like medical HuaTuo (Wang et al., 2023a).

Agent Tuning Recently, LLMs excel in text understanding and following instructions (Qian et al., 2023; Chiang and Lee, 2023; Shen et al., 2023; Gao et al., 2023; Wang et al., 2023b). Beyond its single-agent capabilities, agents further allow for the customization of multi-agent systems. Such systems are valuable in specific domains like the medical domain (Tang et al., 2023). Research suggests that through mechanisms such as debate and cooperation, the collective capabilities of agents can not only be enhanced but also lead to the improvement in the quality of generated responses (Li et al., 2023; Liang et al., 2023). As a result, there is increasingly more work utilizing multi-agent systems for data generation (Du et al., 2023; Li et al., 2023; Qian et al., 2023; Wu et al., 2024). In a multi-agent-based data generation system, individual LLM agents can assume different roles and generate instruction data through role-playing prompting. Specifically, some frameworks employ multiple agents that engage in conversations with each other, producing instruction data in a chat-like format. This approach allows for the creation of more diverse and interactive instruction datasets. Besides utilizing the agent's ability to generate instruction tuning data, there are also some methods to generate data for agent tuning (Zeng et al., 2023; Chen et al., 2023), which focus more on tasks like web navigation.

3 System Design and Workflow

3.1 System Input

Self-defined topics Recognizing that users sometimes hold private data, MIMIR features an offline pipeline that allows users to import their sensitive knowledge. This approach is designed not only to safeguard privacy but also to meet distinct user requirements. Considering the prevalence of domain-specific data among users, MIMIR is adeptly designed to accommodate custom inputs from the user's side. As shown in Table 2, MIMIR accepts two types of inputs for file uploading. In the offline mode, the generated output consists of agent-tuning datasets constructed through multi-turn dialogues, the same as the standard mode.

Domain-specific Dataset Incorporation In addition to leveraging parametric knowledge in Large Language Models, MIMIR enhances its capabilities by incorporating **520** scientific-related domain-specific datasets available on the Hugging Face. This integration serves as a robust supplementary knowledge base for instructional data. For instance, in the medical domain, MIMIR includes several public medical datasets similar to the setting in Flan-PaLM (Singhal et al., 2023) into our pipeline: MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU Clinical Topics (Hendrycks et al., 2021). As Shown in Table 1, MIMIR integrates **520** datasets, utilizing a reasoning framework and retrieval tools to generate user-specific trajectory interactions for enhanced scientific applications.

3.2 Agent Tuning Data Generation

3.2.1 Multi-turn General Instruction Data

Multi-turn Dialogue After users select their self-uploaded topics and an existing domain-specific dataset for generating the instruction dataset, MIMIR seamlessly integrates these datasets in the backend to create an intermediate data pool. Each data point in this pool is utilized as a keyword or key sentence in the subsequent step. Building on previous work (Xu et al., 2023), we generate a multi-turn dataset based on multiple rounds of interaction between a human and an agent. Additionally, we provide the functionality for users to predefine the number of interaction rounds they wish to include in their instruction data, ensuring tailored dataset generation. Compared to Camel (Li et al., 2023), which employs role-playing and inception prompting for agent communication, our method focuses more on generating diverse, domain-specific instruction datasets rather than solving reasoning tasks. In our agent setting, we do not use complex communication to interact with environments. Instead, we use a role-playing approach to prompt LLMs to assume different roles, enabling them to generate representative data (e.g., role-playing as doctors, medical professors, and students) to create medical instruction data in the medical domain.

Resource	**MIMIR** (ours)	**Self-Instruct** (Wang et al., 2023c)	**Baize** (Xu et al., 2023)	**AgentInstruct** (Zeng et al., 2023)	**FireAct** (Chen et al., 2023)
Real API Call?	✓	✓	✓	✓	✓
Multi-step Reasoning?	✓	✗	✓	✓	✓
API Retrieval?	✓	✗	✗	✓	✓
Instruction Tuning for Tool Learning?	✓	✗	✗	✓	✓
Instruction Tuning for Alignment?	✓	✓	✓	✓	✗
Role-playing for Generation?	✓	✗	✗	✗	✗
Expertise Focus	Scientific Domains	General	General	Task-specific	Task-specific
Domain and Tasks	Medical, Physical, Chemical, ...	General	Chat	Web, KG, OS, Database	Question Answering
Number of Datasets	**520**	-	-	6	4
Avg. Reasoning Traces	Customized	1.0	Customized	5.24	Customized

Table 1: A system-wise comparison of our MIMIR to other instruction tuning datasets for tool use and general ability. KG and OS stand for knowledge graph and operation systems.

Topic Type	Examples of User-defined Input
Keyword-based	"Anatomy", "Biochemistry", "Biostatistics", "Cardiology", "Dermatology", "Emergency Medicine", "Endocrinology", "Epidemiology", "Gastroenterology".
Sentence-based	"In ophthalmology, cataracts, characterized by the clouding of the eye's natural lens, are a leading cause of visual impairment worldwide and can be effectively treated through a surgical procedure that replaces the clouded lens with an artificial one."

Table 2: User-Defined topic examples: Users can upload their private domain-specific knowledge by two types of input: keyword or sentence.

Domain-specific Role-playing We leverage LLMs to replicate specific domain roles via advanced inception prompting. Within the medical sphere, our system comprises **14** unique roles: *Doctor, Nurse, Pharmacist, Medical Laboratory Technician, Physical Therapist, Nutritionist, Psychologist, Radiology Technician, Medical Researcher, Medical Educator, Medical Administrator, Medical Interpreter, Medical Equipment Engineer, and Medical Librarian*. To ensure the comprehensive representation of scientific domain roles, taxonomy in Tang et al. (2024) guided our selection. Our methodology for role-specific prompting is simple yet efficient, particularly adept at producing multi-turn instructional data. We have crafted a bespoke prompt setting for each role in our MIMIR agent ensemble. This allows users to select the most relevant in-domain roles for generating multi-turn instruction tuning datasets. Our approach significantly surpasses the efficacy of previous configurations, as shown in Section 5.

3.2.2 Multi-turn Agent Instruction Data

Initial Trajectory For datasets specifically tailored for MIMIR, we primarily utilize their training split as our input source. In cases where datasets are not partitioned, we employ the entire dataset for training. The training set examples are used directly as the initial trajectory. For a limited number of datasets that do not follow an instruction-based format, we leverage GPT-4 for synthesizing the initial trajectory. For instance, in the medicine domain, a phrase like "headaches, sore throat, dry heaves" is transformed into a more contextualized statement: "Recently, I've been experiencing headaches and a sore throat. In the mornings, I feel nauseous, especially when brushing my teeth, accompanied by dry heaves. What should I do?".

Tool We augmented MIMIR with a suite of search tools. Following the approaches in Press et al. (2023) and Chen et al. (2023), we integrated SerpAPI[13] to develop a Google search tool. SerpApi is a real-time API to access Google search results. These tools aim to retrieve the relevant knowledge, prioritizing data from the "highlight words". Additionally, MIMIR is equipped with Tavily[24] as an alternative search API. Tavily's Search API is a search engine built specifically for AI agents (LLMs), delivering real-time, accurate, and factual results. These search tools empower models with the latest knowledge and information pertinent to their reasoning trajectory, facilitating robust agent tuning in the Scientific domains. This integration is crucial for ensuring that LMs remain up-to-date and effective in their responses.

Reasoning Framework Within our MIMIR framework, we incorporate ReAct (Yao et al., 2023) as our primary reasoning framework to generate rationales. For each interaction cycle, this

---

[31] `https://serpapi.com/`

[42] `https://tavily.com/`

framework outputs two components: the thought process, which reflects on previous results, and the action, which involves selecting and utilizing tools. For example, it might use the Google Search tool to acquire necessary information. Following this, the action yields a result, such as search outcomes, within our framework. If the thought process aligns with the correct direction, it concludes in the thinking phase, leading to the final answer. The system determines this alignment based on predefined criteria or heuristics within the ReAct framework, such as reaching a confidence threshold or exhausting all relevant actions. When these conditions are met, the 'think' step will output a conclusion indicating that the result is ready, and the 'act' step will directly output the answer. In this way, the 'think' step in the agent framework serves a function similar to the EOS (End of Sequence) token in traditional language models, signaling the completion of the reasoning process. Besides the default reasoning framework, MIMIR also supports user-customized Chain-of-Thought (CoT) (Wei et al., 2023) Templates and Reflexion (Shinn et al., 2023) mechanisms. These additional mechanisms cater to varying user preferences and contribute to the system's versatility. Importantly, the decision-making steps within MIMIR are directed by the internal reasoning processes provided by these frameworks. This design ensures a coherent and efficient reasoning pathway tailored to each interaction.

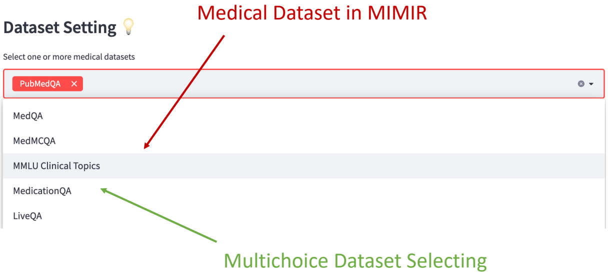
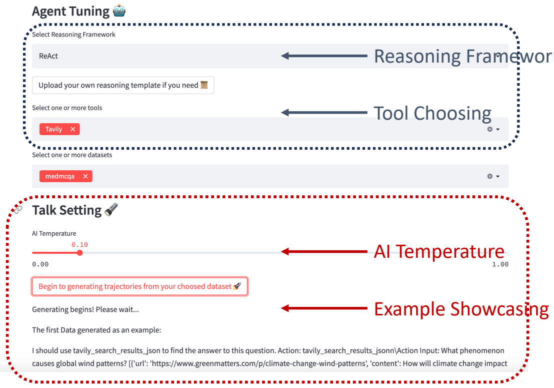


Figure 2: User can upload the custom data and select multiple domain-specific datasets in MIMIR. In this figure, we provide an example for selecting medical domain datasets.

4 The *MIMIR* UI

Our framework's system design focuses on enabling users to create instructional data to enhance the capabilities of LLMs. In this section, we present three interface screenshots (S1, S2, and S3), accompanied by detailed instructions, to demonstrate the design of the MIMIR UI.


Agent Tuning
Reasoning Framework
Upload your own reasoning template if you need
Tool Choosing
Talk Setting
AI Temperature
0.00
1.00
Begin to generating trajectories from your choosed dataset
Example Showcasing

Select Reasoning Framework: ReAct

Select one or more tools: Tavily

Select one or more datasets: medmcqa

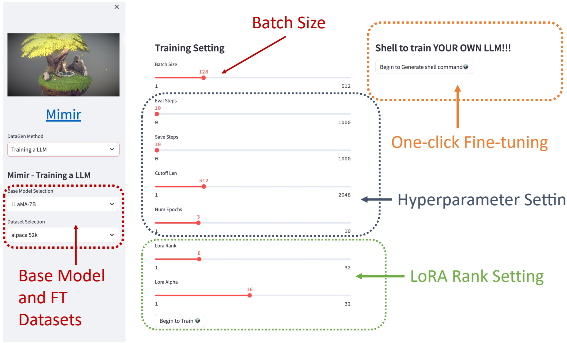
AI Temperature: 0.10

Generating begins! Please wait...

The first Data generated as an example:

I should use tavily_search_results_json to find the answer to this question. Action: tavily_search_results_jsonn\Action Input: What phenomenon causes global wind patterns? [{'url': 'https://www.greenmatters.com/p/climate-change-wind-patterns', 'content': How will climate change impact

Figure 3: Agent tuning interface: we commence by allowing the user to select a reasoning framework and designate their preferred tools. Subsequently, we empower the user to configure the hyperparameters for the models. Furthermore, to facilitate a comprehensive understanding, we provide an illustrative example as the user proceeds with dataset generation.



DataGen Method: Training a LLM

Base Model Selection: LLaMA-7B

Dataset Selection: alpaca 52k

Batch Size: 128

Eval Steps: 10

Save Steps: 10

Cutoff Len: 512

Num Epochs: 3

Lora Rank: 8

Lora Alpha: 16

Figure 4: User can fine-tune personalized models with a single click, selecting from a pre-defined list of models such as LLaMA2.

S1: Dataset Selection View As depicted in Figure 2, MIMIR facilitates the selection of multiple domain-specific datasets through an efficient and user-friendly multi-selection checkbox interface. Given the extensive collection of 520 datasets, users can conveniently search by entering dataset initials in the provided search box. This feature allows users to efficiently narrow their options and locate the most relevant datasets.

S2: Agent Tuning View In Figure 3, we present how we amalgamate various reasoning frameworks and tools to facilitate the generation of rationales for trajectory interactions. Initially, users select a reasoning framework from options including CoT, ReAct (Yao et al., 2023) (the default choice), and Reflexion (Shinn et al., 2023). Subsequently, they can choose from a suite of tools available in our tool pool. Additionally, users have the flexibility to upload custom templates to create CoT rationales tailored to their specific requirements.

The next step involves selecting and uploading multiple datasets as source input. Finally, users can create agent-tuning datasets by clicking the designated button at the bottom of the interface. We also provide an example showcasing the dataset generation process.

S3: One-Click Finetuning The training script interface, as depicted in Figure 4, enables users to fine-tune foundation models such as LLaMA2 with a single click, using datasets they have created. This can be done in our default or LoRA (Hu et al., 2021) settings. Furthermore, the interface provides the functionality to create data scripts for model fine-tuning, leveraging visualized parameters. This innovative feature empowers users to efficiently train large-scale models tailored to their specific domains, utilizing the dialogue data they have generated. The data format of our system output for fine-tuning follows the instruction tuning format, ensuring consistency and ease of integration across computing environments. After our system outputs the agent tuning datasets, we proceed with standard full-parameter or LoRA instruction fine-tuning, and we attach the fine-tuning scripts for ease of use.

5 Human Evaluation of the Generated Data

Experiments in the Biomedical Domain In our study, we selected a diverse set of data samples to form our investigation set for source input in MIMIR. Specifically, we chose 25 random samples from each of the following biomedical domain datasets we described before: MedQA, MedMCQA, and MMLU Medical Topics. Although these datasets traditionally consist of medical multiple-choice questions, for our evaluation, we removed the multiple-choice options and tasked the model with generating long-form answers. These answers required a detailed reasoning path to address the complex medical questions. This approach ensures a comprehensive dataset, facilitating an in-depth biomedical data analysis. In addition, to demonstrate the effectiveness of our agent tuning, we also conducted standard benchmark tests (multiple-choice) on these datasets.

Experiment Setting We utilized the default configuration in MIMIR to process the input source datasets for generating instruction data. MIMIR's token limit is set at 1000, with a temperature setting 0.1. Based on LLaMA2, we conducted a comparative analysis of MIMIR with Baize, Self-Instruct, AgentInstruct, and FireAct using identical settings. We engaged 13 medical students (in the MD program) to select the most appropriate output from these four methodologies. These experts were instructed to complete the evaluation sets independently, relying solely on their professional judgment, without any intercommunication, as presented in Figure 5. For generating dialogue data, we utilized Azure's GPT API calling, and for fine-tuning LLaMA2, we employed 4x 80GB A100 GPUs.

Method	MedQA	MedMCQA	MMLU-Med	Average
**LLaMA2*				
Zero-shot	35.2	36.3	46.3	39.3
Zero-shot + RAG	36.2	38.3	47.7	40.7
`MIMIR` **(Ours)**	**55.9**	**54.1**	**68.5**	**59.5**
**GPT-3.5*				
Zero-shot	53.6	51.0	67.3	57.2

Table 3: Model performance on the standard benchmarks. RAG means retrieval-augmented generation, using Jin et al. as the baseline.

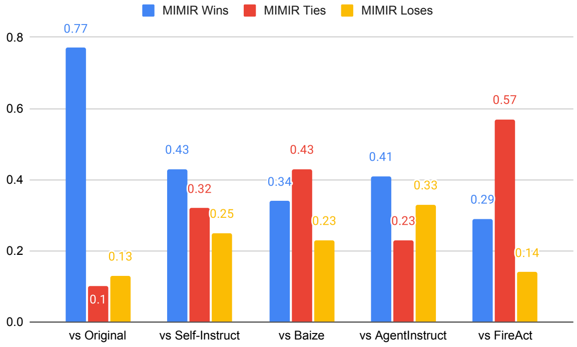


Figure 5: General Preference of data generated by MIMIR in 3-turn setting with *original topics*, *self-instruct*, *Baize*, *AgentInstruct*, and *FireAct*.

Result and Observation Our findings reveal that, compared to original datasets, MIMIR shows a marked preference. The reasoning behind this trend is that outputs from MIMIR were chosen more frequently by domain experts than those yielded from simpler topics. When set against methods that primarily follow instructions, like self-instruct and Baize, MIMIR displays significant strides by demonstrating enhanced capabilities when enabling agent learning. Moreover, MIMIR showcases considerable potential in assimilating external, domain-specific knowledge, particularly compared to other agent tuning frameworks such as FireAct and AgentInstruct. In addition, as shown in Table 3, our method significantly improves LLaMA2 performance across benchmark tests.

Though other scientific domains might present their own unique challenges, we manually experimented with several examples from chemistry, physics, and geographical science and found the results to be quite good. However, due to the cost associated with manual annotation and the lack of domain-specific benchmarks, we only present detailed results for the biomedical domain.

6 Conclusion and Future Work

MIMIR is a streamlined platform for agent tuning, focusing on scientific expertise and advanced applications. It integrates domain-specific datasets and user-uploaded topics, utilizing various contemporary reasoning frameworks and tools. Our platform is particularly useful in accelerating scientific discovery in biology and medicine by incorporating diverse tools and automating tool selection.

7 Ethics Statement

This paper introduces a streamlined platform for personalized agent tuning. It aims to empower users to refine their agents while ensuring the privacy of their data.

Privacy Excluding personal data, all datasets integrated into MIMIR are accompanied by licenses that authorize us to compile, adapt, and redistribute the original datasets. Additionally, we introduce a knowledge filtering method to eliminate potentially harmful and inaccurate information. The model and reasoning framework employed do not reveal sensitive information.

Data During interactions with human participants, we strictly adhered to ethical standards and prioritized their well-being. The datasets and output examples provided for selection are exclusively sourced from publicly available and legally compliant materials.

Recruitment of Domain Experts The 13 domain experts engaged in our study were recruited from a variety of sources. An open call for participation was announced in various professional medical forums, online groups, and mailing lists, directly targeting professionals in the medical community. The recruitment process ensured that all potential evaluators possessed relevant qualifications and expertise in the subject matter. Prior to participation, all participants provided informed consent and received a comprehensive briefing about the study's purpose, their expected role, and the data handling procedures to ensure anonymity and data privacy.

Processing of Evaluation Results To maintain the ethical standards of word anonymity and confidentiality, all obtained evaluations were anonymized and de-identified before analysis. Evaluators' identities were replaced with arbitrary numerical identifiers to protect their identities during the analysis and subsequent publication processes. Moreover, all data underwent privacy-preserving protocols, and secure, encrypted databases were used for storage to prevent unauthorized access and ensure data integrity.

8 Limitation

There are areas in which the system can still improve.

Agent-Tuning Data Generation Protocols: While MIMIR incorporates private and public knowledge bases and employs multi-role agent configurations, the underlying assumptions could be limiting. The generation protocols might not account for dynamic changes in real-world data or rapid advancements in the knowledge base of specific domains.

Dependence on External Tools: The efficiency of MIMIR heavily depends on the performance of external tools such as SerpAPI and Tavily. Any limitation inherent in these tools will directly impact the accuracy and results procured by MIMIR.

Domain-Specific Data: As the effectiveness of MIMIR is closely tied to the quality of the incorporated datasets, any errors, biases, or inconsistencies in these datasets may negatively impact the results generated by the system.

Limited Domain Experiments: We have not conducted extensive experiments across a wide range of scientific domains. However, the results in the biomedical domain are quite promising. While this provides a strong indication of the potential of our system, it is important to validate MIMIR's performance across other domains to fully understand its generalizability and robustness.

References

Anthropic. 2023. Claude.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. Fireact: Toward language agent fine-tuning. *ArXiv*, abs/2310.05915.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Google. 2023. Bard.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sung-Hoon Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *ArXiv*, abs/2307.01881.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

OpenAI. 2022. ChatGPT.

OpenAI. 2023. GPT-4 Technical Report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023a. Principle-driven self-alignment of language models from scratch with minimal human supervision.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023b. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. 2024. Prioritizing safeguarding over autonomy: Risks of llm agents for science. *arXiv preprint arXiv:2402.04247*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. 2023. Evil geniuses: Delving into the safety of llm-based agents.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge.

Qingyun Wang, Manling Li, Hou Pong Chan, Lifu Huang, Julia Hockenmaier, Girish Chowdhary, and Heng Ji. 2023b. Multimedia generative script learning for task planning.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2024. Aligning llms with individual preferences via interaction.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Xiangru Tang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. 2023. Agents: An open-source framework for autonomous language agents.

A BioMedical Dataset Format

As shown in Table 4, we use a variety of datasets to evaluate our model's performance across multiple domains and formats. Specifically, MedQA provides questions and answers from the US Medical Licensing Examination, while MedMCQA includes questions, answers, and explanations from AIIMS and NEET PG entrance exams. PubMedQA offers a different format with questions, context, and answers from PubMed paper abstracts. Lastly, the MMLU dataset includes questions and answers from the Graduate Record Examination and the US Medical Licensing Examination.

B Community Guidelines

These guidelines aim to establish a uniform framework for the development, validation, and application of agent-tuning instructions within the MIMIR system. They specifically focus on addressing the unique demands and challenges associated with domain expertise.

We recognize that the dynamic integration of public datasets into the system presents unique challenges, particularly regarding copyright and appropriate use. This implies that if a dataset encounters copyright issues or is unexpectedly removed, we must withdraw it through an automated process. In practical terms, this means we often face various issues daily due to these constraints. Therefore, as a community, we must adhere to these guidelines within the legal boundaries to ensure compliance and maintain the integrity of our system.

Continuing from the established guidelines, it is crucial to emphasize the importance of ethical considerations and data privacy in the handling of datasets. As we navigate the complexities of incorporating publicly sourced data, we must remain vigilant in protecting the privacy and rights of individuals represented within these datasets. In this way, we can ensure that our pursuit of technological advancement and domain expertise does not come at the expense of ethical responsibility and user trust.

C Knowledge Verification

According to Table 5, our analysis reveals an increase in the overall hallucination rate when generating extended turn instruction data. To address this, MIMIR incorporates a fine-grained knowledge verification feature for the generated datasets. Users can select any round of instruction data and verify it with a single-click action. For this purpose, we extract key QA pairs and topics and integrate them into our verification module. This module operates on a domain-specific state-of-the-art model. Utilizing this approach, we aim to generate more accurate and reliable responses. Currently, MIMIR employs GPT-4 as its verification model, leveraging its exceptional performance across various medical tasks.

D Implements Details for Role-Playing

D.1 Memory Setting

```
for name in picked_roles:
    prompt = "You are {}. {} You come to a chat room
    because you want to discuss the topic
    about {}. " \
    "The following people are in
    this chat room: {}.
    What is your main point? Be brief, " \
    "and use at most 20 words
    and answer from your
    perspective.".format(
        name, role_prompt[name], query,
        ’, ’.join(picked_roles))
    ideas[name] =generate(prompt_meta.format(prompt)
    , asure, ai_temperature)
```

Code 1: Memory setting for the running loop in MIMIR agent system.

In MIMIR, we present a framework to simulate interactive role-based dialogues in a chat room environment. Our methodology encompasses four key components: the initialization of a memory data structure for each role, the preparation of a compressed memory counterpart, the establishment of a placeholder for role-specific ideas, and the generation of these ideas through a sophisticated prompt formulation. By iterating over a predefined set of roles, our system dynamically constructs context-specific prompts, incorporating role-specific cues and a central discussion topic. This is followed by generating concise, perspective-driven responses using an advanced language model.

D.2 Memory Rater

```
import re
def get_rating(x):
    nums = [int(i) for i in re.findall(r’\d+’, x)]
    if len(nums) > 0:
        return min(nums)
    else:
        return None

memory_ratings = {}
for name in picked_roles:
    memory_ratings[name] = []
    for i, memory in enumerate(memories[name]):
        prompt = "You are {}. Your ideas are: {}.
```

Dataset	Format	Choice	Testing Size	Domain
MedQA	Question + Answer	A/B/C/D	1273	US Medical Licensing Examination
MedMCQA	Question + Answer	A/B/C/D and Explanations	6.1K	AIIMS and NEET PG entrance exams
PubMedQA	Question + Context + Answer	Yes/No/Maybe	500	PubMed paper abstracts
MMLU	Question + Answer	A/B/C/D	1089	Graduate Record Examination & US Medical Licensing Examination

Table 4: Summary of the Datasets we use.

Turn	Overall Hallucination Ratio
1	4.27
2	7.37
3	14.27
4	21.27
5	24.27

Table 5: Halluciation Ratio across all instruction data generated from domain-specific datasets. Overall scores are reported by averaging all the results by domain-specific expert evaluation.

```
You are currently in a chat room and
you are talk about {}. " \
         "You observe the following: {}.
         Give a rating, between 1 and 5,
         to how much you care about this. "
.format(name, ideas[name], query, memory)
res = generate(prompt_meta.format(prompt),
asure, ai_temperature)
rating = get_rating(res)
max_attempts = 2
current_attempt = 0
while rating is None and current_attempt
< max_attempts:
    rating = get_rating(res)
    current_attempt += 1
if rating is None:
    rating = 0
memory_ratings[name].append((res, rating))
```

Code 2: Memory setting for the running loop in MIMIR agent system.

In the given code, a function named get_rating is implemented using regular expressions to extract numerical values from a string. The smallest number in the string is returned as the rating, or None if no numbers are found. The script iterates over predefined roles stored in picked_roles, managing each role's memories. For every memory associated with a role, a specific prompt is generated that includes the role's name, ideas, a query, and the memory itself. An AI generation function then processes this prompt.

Subsequently, the gettextunderscore rating function is used to analyze the AI's response, extracting a numerical rating that reflects the relevance or importance of the memory in question. In cases where no rating can be determined after a few attempts, a default rating of zero is assigned. The outcomes, comprising both the AI's response and the extracted rating, are stored in a dictionary, with each entry corresponding to a different role.

Turn	Human	AI
1	Recently, I've been having headaches and a sore throat. In the morning, I feel nauseous and have dry heaves when brushing my teeth. What should I do?	These symptoms could be indicative of a variety of conditions, including a viral or bacterial infection, allergies, or even a more serious illness. Your healthcare provider can help determine the cause of your symptoms and recommend appropriate treatment. ...
2	What kind of triggers should I avoid?	The triggers you should avoid depend on the underlying cause of your symptoms. However, some general tips to help alleviate symptoms include: ... It's important to consult with your healthcare provider to determine the specific triggers you should avoid based on your symptoms and medical history.

Table 6: Two agents are talking about the topic, "Recently, I've been having headaches and a sore throat. In the morning, I feel nauseous and have dry heaves when brushing my teeth. What should I do?"

E Diologue Example

As shown in Table 6, we present a detailed interaction between a human and an AI regarding a health concern. The dialogue addresses symptoms, potential triggers, and the role of stress in exacerbating these symptoms.