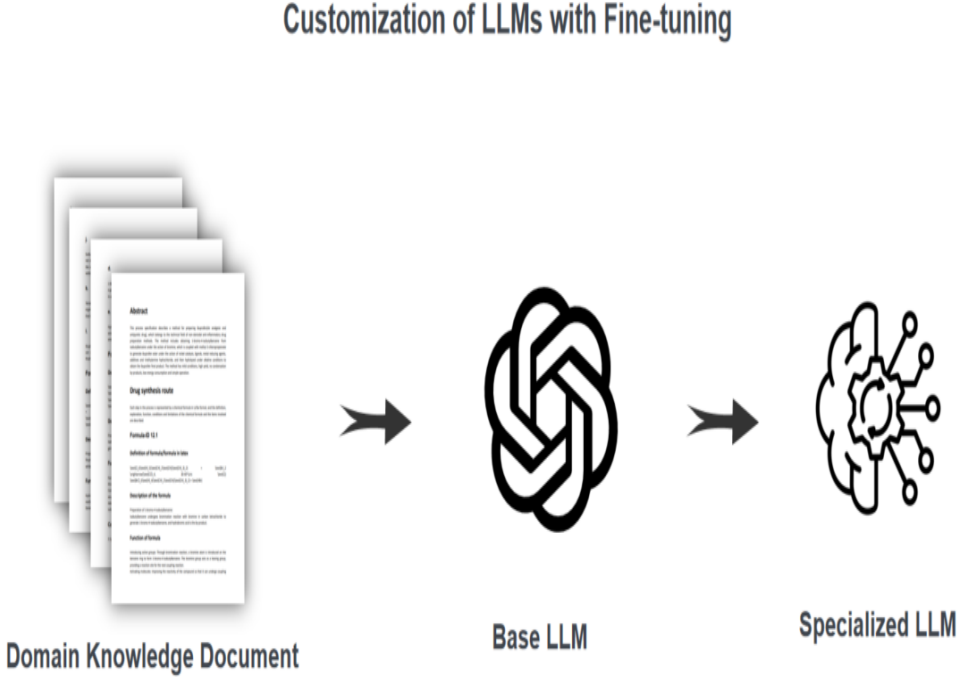
University of Stuttgart
Institute of Industrial Automation
and Software Engineering



synthetic data creation for training LLMs on domain-specific knowledge

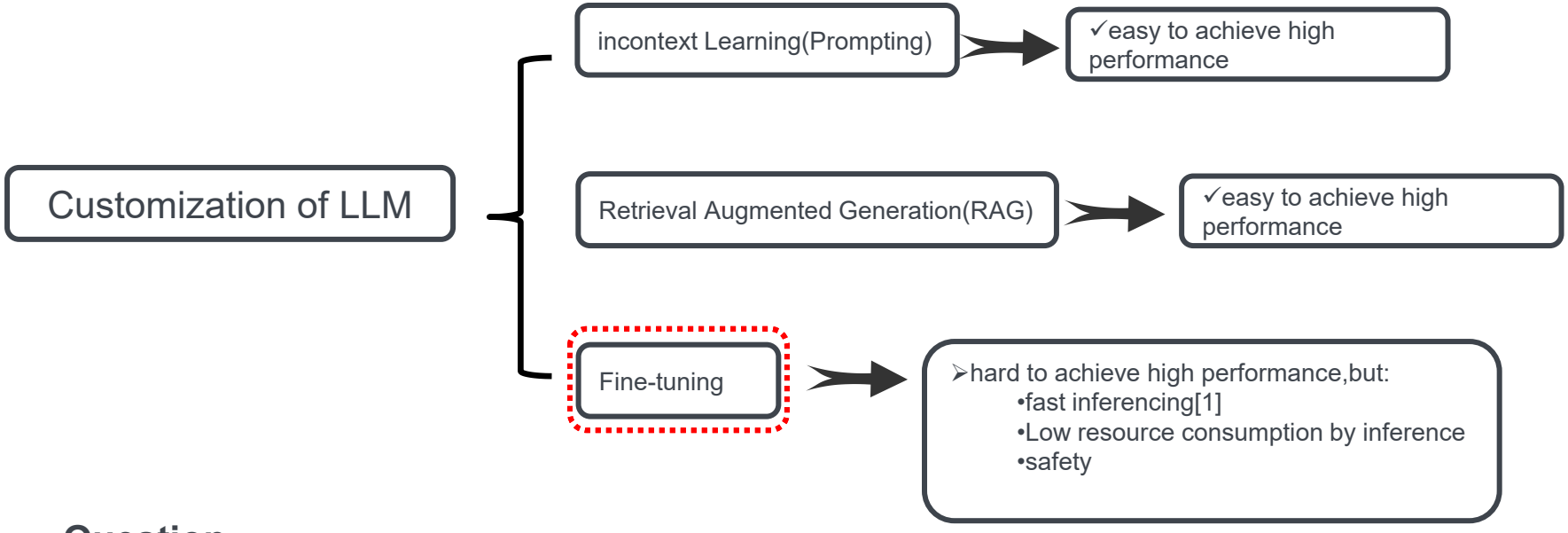
Presenter: Juntao Lin
Supervisor: Yuchen Xia
Examiner: Prof. Dr. Ing. Michael Weyrich

Contents

- Introduction
- Basics
- Training experiment design
- Testing and Evaluation
- Overall Results
- Findings and Outlook

Introduction

Current solution to achieve LLM customization



Question

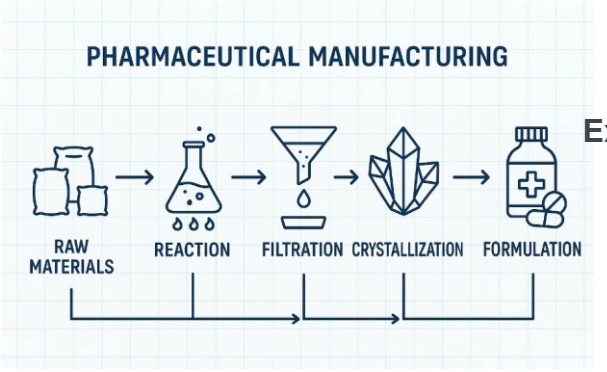
How to finetune an LLM?

Use Case

Scenario:Pharmaceutical Manufacturing

POC

Proof the feasibility of Fine-tuning



Extracted

Fine-tuning

Actual pharmaceutical industry processes

- Academic papers
- Patent Specification
- Authoritative databases
- etc.

A production process of a medicine(ibuprofen)

- Reaction Formulas
- Description
- Function
- Conditions of reaction
- Reactant properties
- Influential factors

Customization of LLM

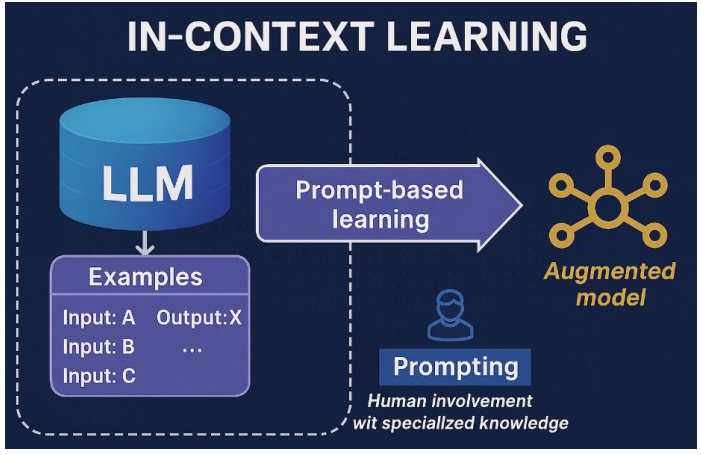
QA Chatbot

Basics

- Incontext Learning(Prompting)
- Retrieval-Augmented Generation (RAG)
- Fine Tuning

incontext Learning(Prompting)

In 2020

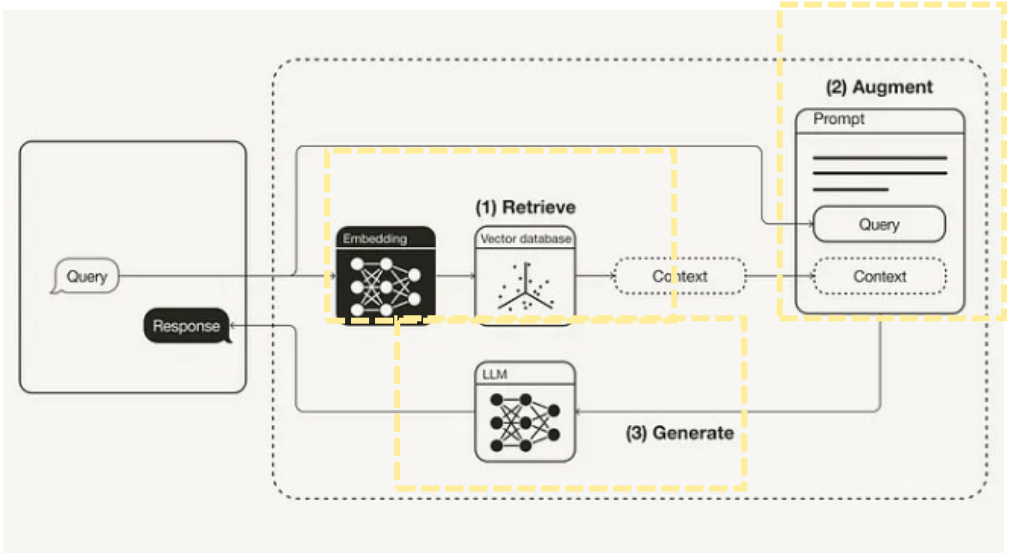


Characteristic of incontext learning in Customized LLMs

- No need to change model parameters
- Fast and easy to deploy
- High requirements on example design

Retrieval-Augmented Generation (RAG)

In 2021

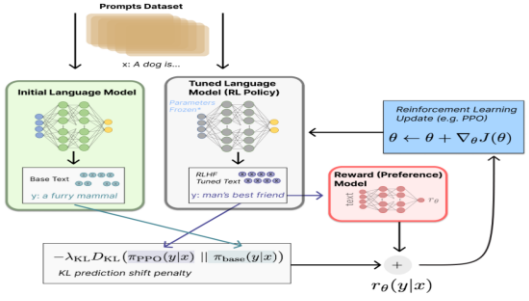


Characteristic of RAG in Customized LLMs

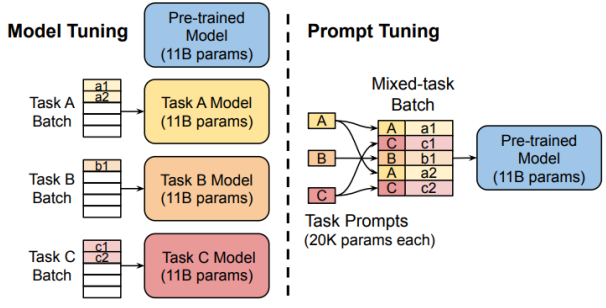
- No need to change model parameters
- Suitable for knowledge-intensive tasks
- High requirements for the retrieval system

Fine-tuning Approach

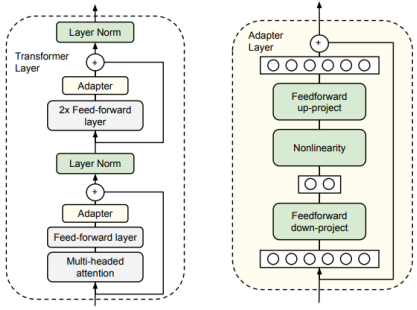
Literature Review



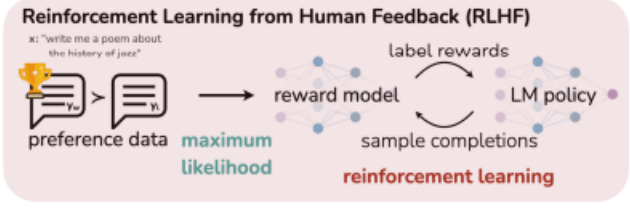
[2] Proximal Policy Optimization(PPO)

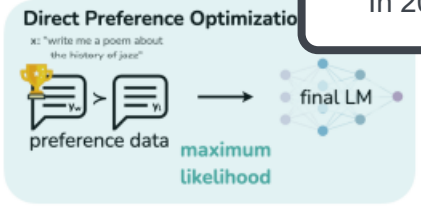


[3] Prompt Tuning



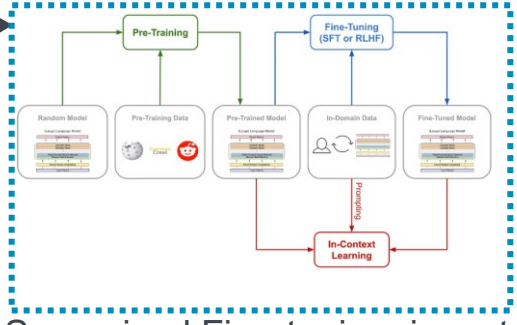
[4] Adapter Tuning





[5] Reinforcement Learning from Human Feedback

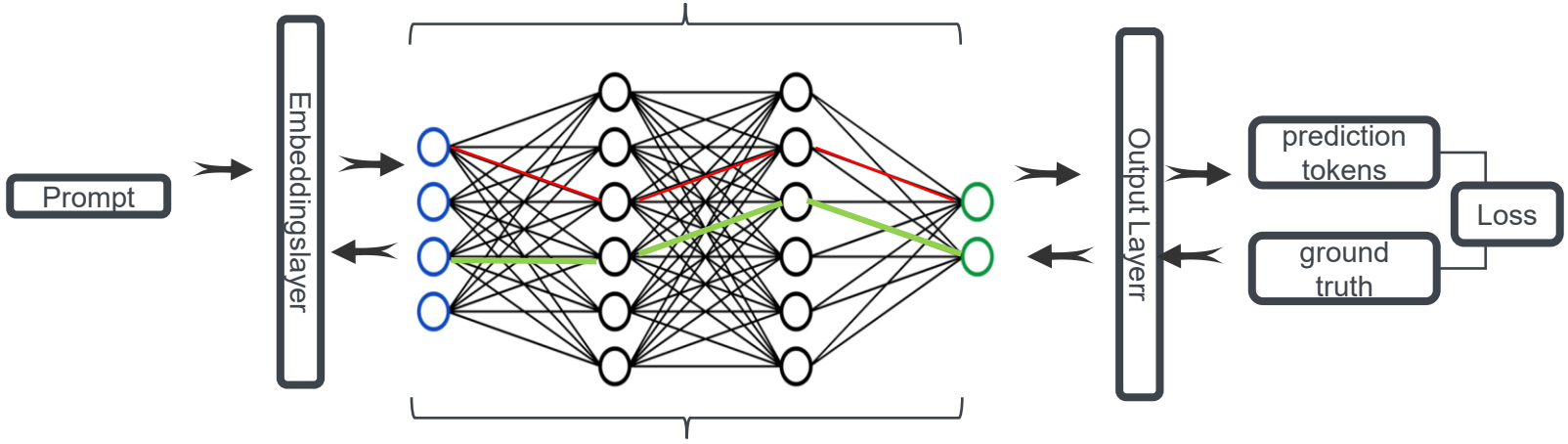
In 2022



[6] Supervised Fine-tuning+incontext learning(prompting)

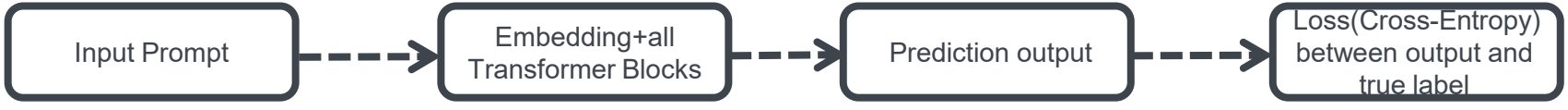
Supervised Fine-Tuning

$$\theta \leftarrow \theta - \text{learning rate} \times \frac{\partial \text{Loss}}{\partial \theta}$$

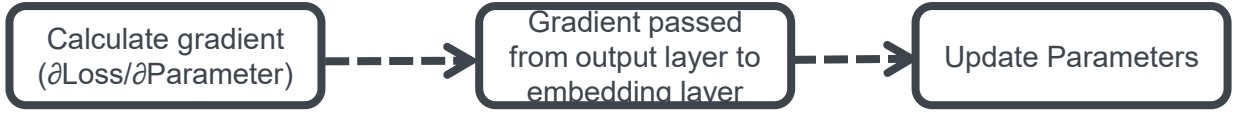


Training mechanism of SFT

- Forward Propagation



- Backward Propagation



Training experiment design

- Base model
- Training resources
- Training Data
- Experiments execution

Base Model

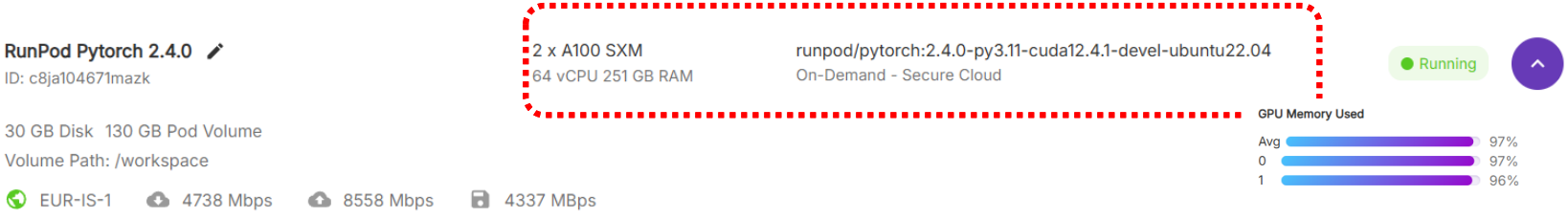
GPT-4o-mini vs. Qwen2.5-7B-Instruct

index	GPT4o-mini	Qwen2.5-7B-Instruct
Model parameter size	about 8B	7B
Developer	OpenAI(2024)	Alibaba(2024)
Call resource consumption	$0.26 /1M tokens	$0.03/1M tokens
Max. Context length	128k	32k
Specialize in task types	Question answering programming tasks reasoning tasks	Question answering dialogue generation
Pre-Trained data source	Wide range Covering many fields	Wide range Covering many fields Focus on Chinese corpus

Training resources

- For Qwen2.5-7B-Instruct
  - Cloud Computing Platform:**Runpod**
  - Fine-tuning Platform:**LLaMa-Factory**
  - Training environment configuration:

GPU	2 x A100 SXM
vRAM	160GB
pytorch version	2.4.0
python version	3.11
CUDA	12.4.1
Ubuntu	22.04



  - 70GB disk volume for each check-point step of 16-bit quantized model

- For GPT-4o-mini
  - Fine-tuning Platform:**OpenAI Platform**

Training Data(1/2)

Question

Where does the fine-tuning data come from?

- PDF-synthetic dataset
- Knowledge-distillation dataset

PDF-synthetic dataset

Knowledge documents of pharmaceutical production

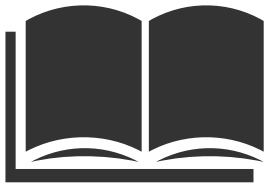
Extraction

87 Extracted question topics

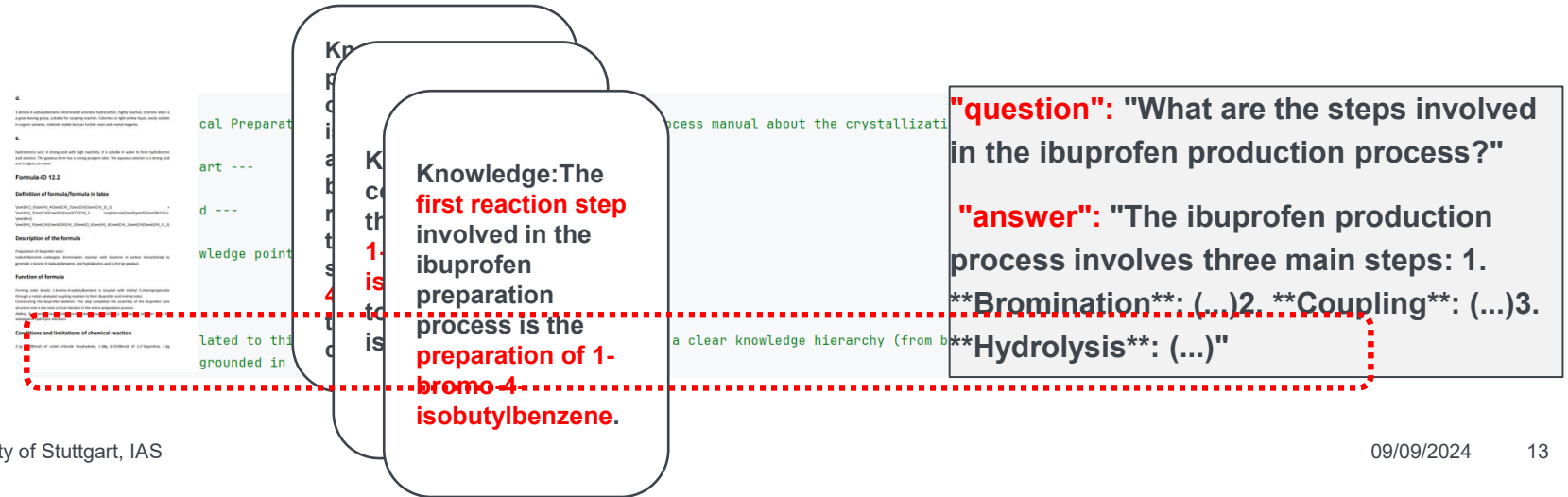
Input

LLM:GPT-4o

Generate



PDF-synthetic dataset:155 QA pairs



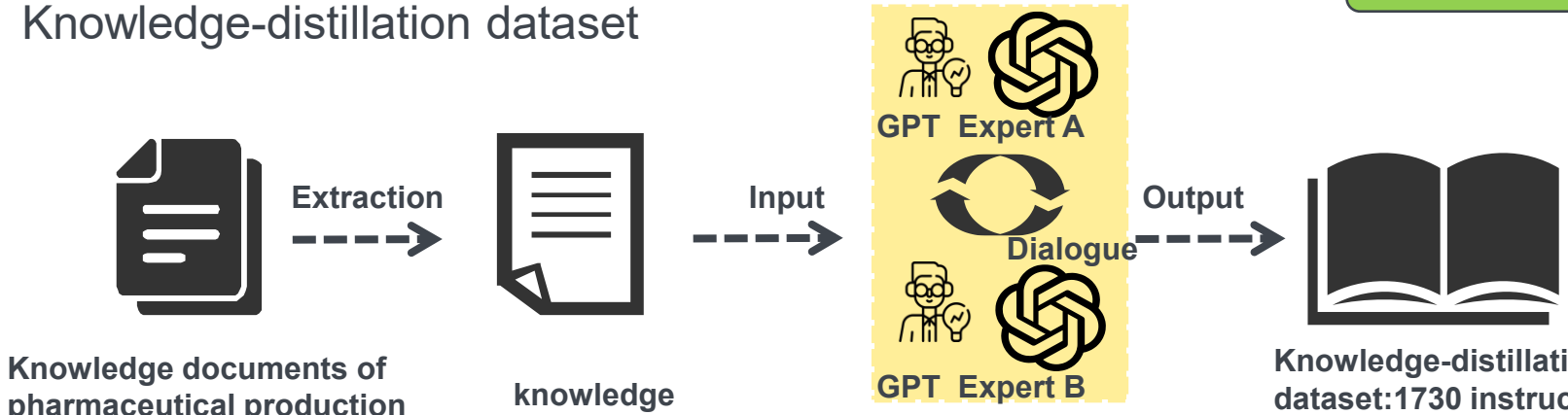
Training Data(2/2)

Question

Where does the fine-tuning data come from?

- PDF-synthetic dataset
- Knowledge-distillation dataset

Knowledge-distillation dataset




```python
def generate_reply(context, topic, last_reply, speaker):
    if speaker == "GPT-A":
        prompt = [
            {"role": "system", "content": "You are GPT-A, a rigorous and knowledgeable expert in pharmaceutical crystallization process."},
            {"role": "user", "content": f"following is context: \n{context}\n\nThe topic we want to explore is: {topic}.\nPlease contribute
             science and engineering..Do not discuss beyond the document content."}
        ]
    else:
        prompt = [
            {"role": "system", "content": "You are GPT-B, an expert with strong critical thinking ability in pharmaceutical crystallizatio
            {"role": "user", "content": f"following is context: \n{context}\n\nThe topic we want to explore is: {topic}。\nGPT-A just said the following: \n“{last_reply}”\nPlease respond, question, supplement, or extend
             the discussion on the high-quality discussions on technical issues that address pharmaceutical science and engineering..Do not discuss beyond the document content."}
        ]
```

"GPT A": "Explain the role of bromine in the bromination reaction during the preparation of 1-bromo-4-isobutylbenzene."

"GPT B": "This reaction is an electrophilic aromatic substitution where bromine acts as the electrophile. The light or UV conditions facilitate the formation of bromine radicals, which attack the electron-rich aromatic ring of isobutylbenzene."

Training Data

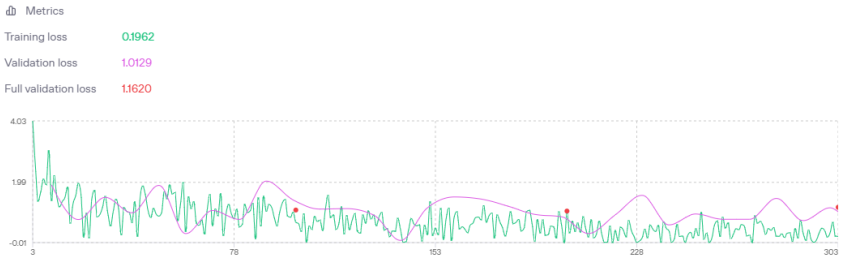
Dataset comparison

index	PDF-synthetic dataset	Knowledge-distillation dataset
Sample number	155 trainset(101) validation set(30) testset(25)	1,730 Trainset(1,440) validation set(180) testset(110)
Dataset size	12,764	172,113
Dataset Format	JSONL sharegpt 	JSONL sharegpt

Experiments execution

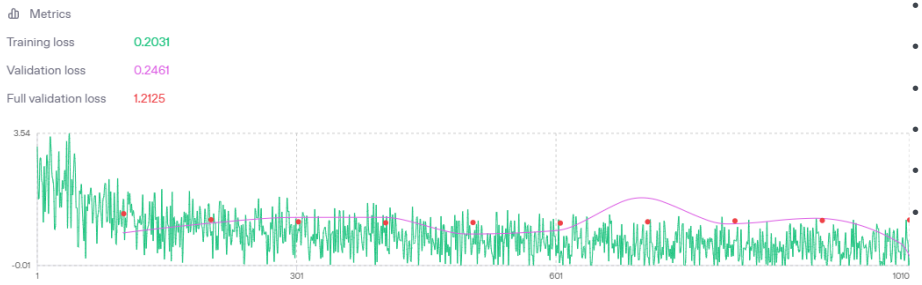
hyperparameters	GPT4o-mini +PDF synthetic datase	GPT4o-mini+Knowledge-distillation dataset	Qwen2.5-7B+Knowledge-distillation dataset
epoch	3	1	1
batch_size	1	1	1
learning_rate	1.8e-5	1.8e-5	1.0e-5

- GPT4o-mini with PDF-synthetic dataset



- Best performance of 6 expe
  - Exp1:Epoch=1,Batch_s
  - **Exp2:Epoch=3,Batch_**
  - Exp3:Epoch=5,Batch_size=1,Learning_rate=1.8e-5
  - Exp4:Epoch=3,Batch_size=2,Learning_rate=1.8e-5
  - Exp5:Epoch=3,Batch_size=4,Learning_rate=1.8e-5
  - Exp6:Epoch=3,Batch_size=1,Learning_rate=1.0e-5

- GPT4o-mini with Knowledge-distillation dataset



- Best performance of 6 experiments:
  - **Exp1:Epoch=1,Batch_size=1,Learning_rate=1.8e-5**
  - Exp2:Epoch=3,Batch_size=1,Learning_rate=1.8e-5
  - Exp3:Epoch=5,Batch_size=1,Learning_rate=1.8e-5
  - Exp4:Epoch=3,Batch_size=2,Learning_rate=1.8e-5
  - Exp5:Epoch=3,Batch_size=4,Learning_rate=1.8e-5
  - Exp6:Epoch=3,Batch_size=1,Learning_rate=1.0e-5

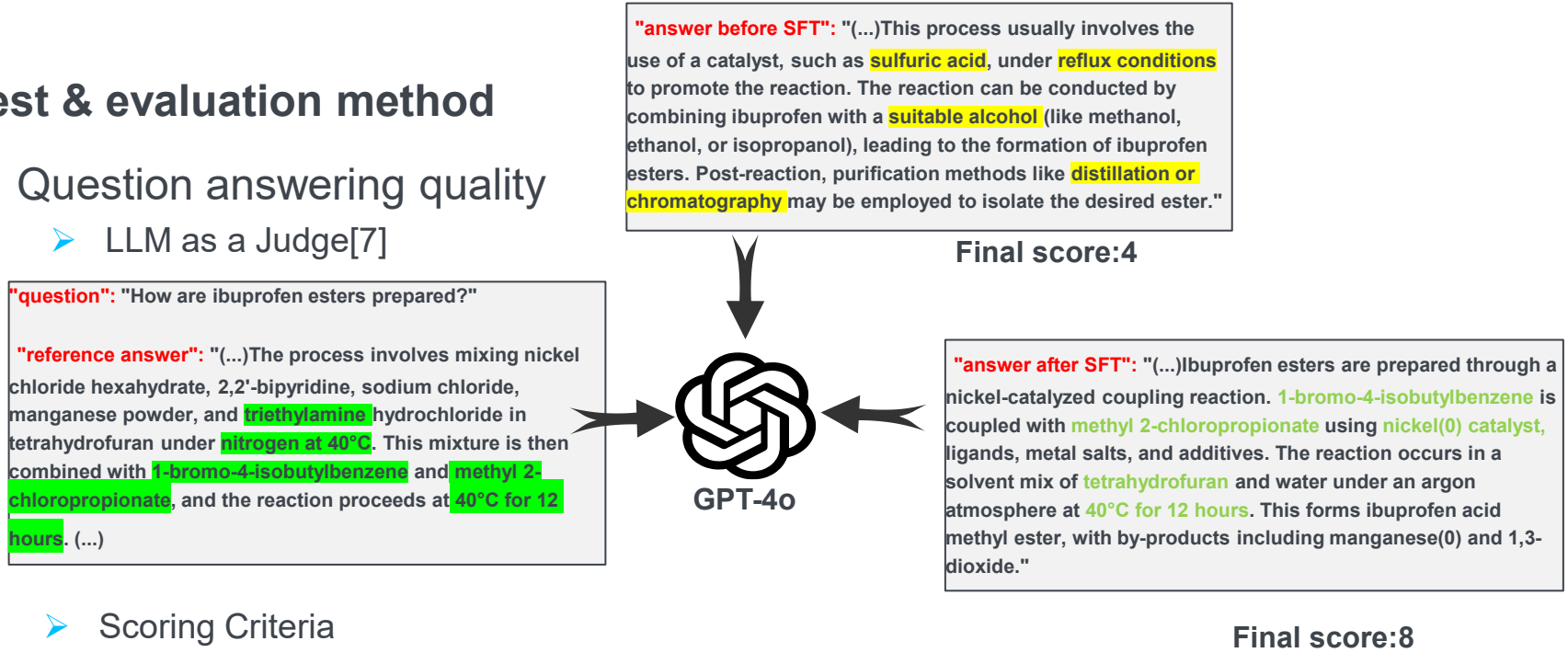
- Qwen2.5-7B with Knowledge-distillation dataset
  - Best performance of 6 experiments:
    - Exp1:Epoch=1,Batch_size=1,Learning_rate=1.8e-5
    - Exp2:Epoch=3,Batch_size=1,Learning_rate=1.8e-5
    - Exp3:Epoch=5,Batch_size=1,Learning_rate=1.8e-5
    - Exp4:Epoch=3,Batch_size=2,Learning_rate=1.8e-5
    - Exp5:Epoch=3,Batch_size=4,Learning_rate=1.8e-5
    - **Exp6:Epoch=1,Batch_size=1,Learning_rate=1.0e-5**

Testing and Evaluation

- Question answering quality (LLM-as-a-judge)
- Task solving (success rate)

Test & evaluation method

- Question answering quality
  - LLM as a Judge[7]



  - Scoring Criteria

Metrics	Description
Factual Accuracy(0-10)	Is the answer correct and consistent with the reference?
Completeness(0-10)	Does the answer cover the main points?
Clarity(0-10)	Is the answer clear and well-structured?

Test result by training set

LLMs	Dataset	Improved performance
		training set question
GPT4o-mini	PDF_synthetic	5.86-->7.29 +14.3%
	Knowledge_ distillation	5.86-->7.18 +13.2%
Qwen2.5-7B	Knowledge_ distillation	4.40-->6.18 +17.8%

Response before SFT

User

What is the solvent used in the preparation of 1-bromo-4-isobutylbenzene?

Assistant

In the preparation of 1-bromo-4-isobutylbenzene, a common solvent used is acetone or dichloromethane (DCM). These solvents facilitate the bromination reaction by dissolving the reactants and enabling better interaction between them. Depending on the specific reaction conditions, other solvents such as toluene may also be used. Always consider the reaction scale and conditions when choosing a solvent.

X

Response after SFT

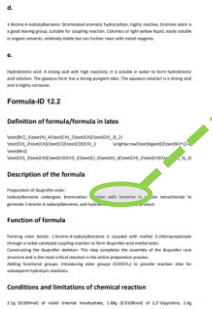
User

What is the solvent used in the preparation of 1-bromo-4-isobutylbenzene?

Assistant

The solvent used in the preparation of 1-bromo-4-isobutylbenzene is carbon tetrachloride.

Good Bad



Question

How to test the **generalizability** of models after SFT?

Generalization

Training generalization:Ask questions on the test set

LLMs	Dataset	Improved performance	
		training set question	test set question
GPT4o-mini	PDF_synthetic	5.86-->7.29 +14.3%	5.25-->5.96 +7.1%
	Knowledge_ distillation	5.86-->7.18 +13.2%	5.25-->7.29 +20.4%
Qwen2.5-7B	Knowledge_ distillation	4.40-->6.18 +17.8%	4.38-->5.58 +12%

➢ Response before SFT

User

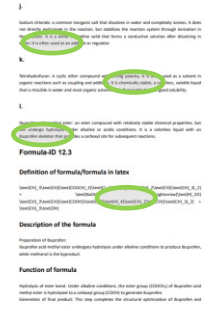
how to improve the final yield and purity of ibuprofen?

Assistant

To improve the final yield and purity of ibuprofen, consider the following strategies:

1. **Optimize Reaction Conditions:** Control temperature, pH, and reaction time carefully.
2. **Purification Techniques:** Utilize recrystallization or chromatography for better separatio
3. **Use High-Quality Reagents:** Ensure raw materials are pure and compatible.
4. **Workup Efficiency:** Improve extraction methods to reduce losses during purification.
5. **Monitor Reaction Progress:** Use analytical techniques (e.g., HPLC) to track purity and yi
6. **Prevent Side Reactions:** Avoid conditions that lead to byproducts.

Implementing these strategies can enhance both yield and purity.



➢ Response after SFT

User

how to improve the final yield and purity of ibuprofen?

Assistant

To enhance the yield and purity of ibuprofen, ensure precise control of reaction temperatures (30-40°C for bromination and 25-40°C for hydrolysis) and maintain uniform stirring. Adjust pH (1-2 for ibuprofen precipitation) to optimally form the product. Thoroughly wash with solvents (water, $NaHCO_3$, brine) to remove impurities and ensure complete methanol solvent removal to prevent side reactions or solvent interference. Employ careful distillation to obtain pure ibuprofen. These improvements can enhance product quality and yield by minimizing side reactions and impurities.

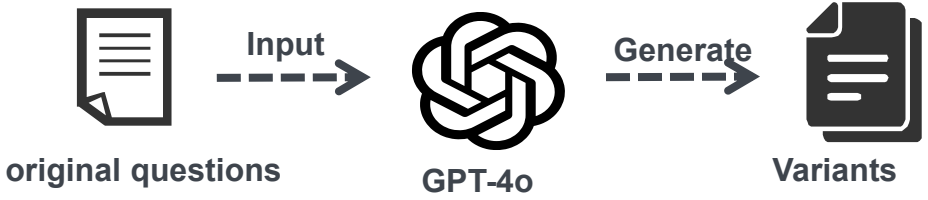
Verification: Improved **training generalization** of fine-tuned models

Test & evaluation method

Question

How to test the **generalizability** of models after SFT?

- **Language generalization**:ask different questions on same topic
  - Variants generation



"original question": **"What is the solvent used in the preparation of 1-bromo-4-isobutylbenzene?"**

Rephrasing

"variant 1": **" Which solvent is employed to prepare 1-bromo-4-isobutylbenzene?"**

"variant 2": **"In the preparation of 1-bromo-4-isobutylbenzene, what solvent is utilized?"**

Test result by rephrased question

LLMs	Dataset	Improved performa		
		training set question	test set question	rephrased question
GPT4o-mini	PDF_synthetic	5.86-->7.29 +14.3%	5.25-->5.96 +7.1%	5.88-->7.16 +12.8%
	Knowledge_ distillation	5.86-->7.18 +13.2%	5.25-->7.29 +20.4%	
Qwen2.5-7B	Knowledge_ distillation	4.40-->6.18 +17.8%	4.38-->5.58 +12%	

Verification: Improved **language generalization** of fine-tuned models

Test & evaluation method

Question

How to test the **generalizability** of models after SFT?

- **Knowledge generalization**:ask for task solving

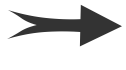
Quiz example

Question: In the preparation of 1-bromo-4-isobutylbenzene, if you start with 26.8g (0.2 mol) of isobutylbenzene, how much bromine (Br2) in moles is required to ensure the correct molar ratio is maintained?

A. 0.100 mol
B. 0.200 mol
C. 0.210 mol
D. 0.105 mol

Reasoning process

Step 1: From the provided knowledge, the molar ratio of isobutylbenzene to bromine is 1:1.05. This means that for every mole of isobutylbenzene, 1.05 moles of bromine are required.
Step 2: You start with 0.2 moles of isobutylbenzene. To find the required moles of bromine, multiply the moles of isobutylbenzene by the ratio:
\[\text{Required moles of Br}_2 = 0.2 \, \text{mol} \times 1.05 = 0.21 \, \text{mol} \]
Step 3: Therefore, 0.210 moles of bromine are required to maintain the correct molar ratio in the reaction. Thus, the correct choice is C.



success rate

the knowledge is required as an intermediate step

Test result by task solving

LLMs	Dataset	Improved performance	
		Task	Task+ prompting
GPT4o-mini	PDF_synthetic	Success Rate: 58/87-->50/87	Success Rate:65/87 ↑
	Knowledge_ distillation	Success Rate: 58/87-->57/87	Success Rate: 70/87 ↑
Qwen2.5-7B	Knowledge_ distillation	Success Rate: 54/87-->58/87	Success Rate: 70/87 ↑

Verification: The fine-tuned model + **incontextual learning(prompting)** enhances knowledge generalization

➢ Task solving without prompting

User

please choose the option you think is correct from the following multiple-choice questions and tell me the reasoning process.
Question: In the preparation of 1-bromo-4-isobutylbenzene, if you start with 26.8g (0.2 mol) of isobutylbenzene, how much bromine (Br2) in moles is required to ensure the correct molar ratio is maintained?

A. 0.100 mol
B. 0.200 mol
C. 0.210 mol
D. 0.105 mol

Assistant

To maintain the correct molar ratio in the preparation of 1-bromo-4-isobutylbenzene, isobutylbenzene (0.2 mol) reacts with bromine. The reaction requires a 1:1 molar ratio. Therefore, to match the 0.2 mol of isobutylbenzene, you need 0.2 mol of bromine.

Option B (0.200 mol) is correct.

➢ task solving with prompting

User

please choose the option you think is correct from the following multiple-choice questions and tell me the reasoning process.Please use the knowledge to solve the problem.The knowledge applied in this question is:In the preparation reaction of 1-bromo-4-isobutylbenzene, the molar ratio of isobutylbenzene: bromine = 1:1.05.
Question: In the preparation of 1-bromo-4-isobutylbenzene, if you start with 26.8g (0.2 mol) of isobutylbenzene, how much bromine (Br2) in moles is required to ensure the correct molar ratio is maintained?
A. 0.100 mol B. 0.200 mol C. 0.210 mol D. 0.105 mol

Assistant

To ensure the molar ratio of isobutylbenzene to bromine is 1:1.05, we start with 0.2 mol of isobutylbenzene. Calculate the required moles of bromine:

[
\text{Moles of Br}_2 = 0.2 \text{ mol} \times 1.05 = 0.21 \text{ mol}
]

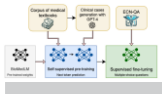
Thus, the correct choice is C. 0.210 mol, ensuring the appropriate molar ratio is maintained for the reaction.

Overall Results

Orthogonal experiment results

Catastrophic forgetting?

Models may forget old knowledge or tasks while learning a new task.

LLMs	Dataset	Improved performance					
		train set QA	test set QA	test with rephrased QA	Task	Task+ prompting	Task(MMLU)
GPT4o-mini	PDF_synthetic	5.86-->7.29 +14.3%	5.25-->5.96 +7.1%	5.88-->7.16 +12.8%	Success Rate: 58/87-->50/87	Success Rate:65/87	56.0-->51.8 -4.2%
	Knowledge_ distillation	5.86-->7.18 +13.2%	5.25-->7.29 +20.4%	5.88-->7.31 +14.3%	Success Rate: 58/87-->57/87	Success Rate: 70/87	56.0-->55.5 -0.5%
Qwen2.5-7B	Knowledge_ distillation	4.40-->6.18 +17.8%	4.38-->5.58 +12%	3.62-->5.89 +22.7%	Success Rate: 54/87-->58/87	Success Rate: 70/87	20.3-->23.7 +3.7%
BioMedLM[8] 	Clinical cases generation with GPT-4	3.13	2.71	3.01	Success Rate: 44/87	N.A.	N.A.

- Conclusion: The phenomenon of catastrophic forgetting is **NOT** significant

Findings and Outlook

Findings and Outlook

- Achievement:
  - ➢ SFT successfully experimented:
    - ➢ 3 fine-tuned Models:
      - language generalization↑,
      - training generalization↑,
      - knowledge generalization-,
    - ➢ fine-tuning+incotext learning(Prompting):
      - knowledge generalization↑
    - ➢ Cost:182€

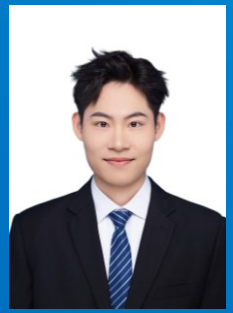
Findings and Outlook

- Findings:
  - Comparative experiment:knowledge_distillation dataset performs **better than** PDF synthesis dataset
  - Vertical domain-specific LLM (BioMedLM) does not have sufficient performance in our use case.
  - In task solving: incontext learning(prompting) **outperforms** fine-tuning approach

- Outlooks:
  - Better finetune LLM with **all data**
  - Powerful base LLM+incontext learning(prompting)

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!



Juntao Lin

e-mail st176526@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart
Institute of Automation and Software Systems
Pfaffenwaldring 47, 70550 Stuttgart

References

[1] O. Ovadia, M. Brief, M. Mishaeli, O. Elisha, et al., “Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs,” arXiv e-prints, 2023, arXiv:2312.05934.

[2] Z. Wang, B. Bi, S. Kumar Pentyala, K. Ramnath, S. Chaudhuri, S. Mehrotra, Z. (James)Zhu, X. Mao, S. Asur, N. (Claire)Cheng,et al., “A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More,” arXiv e-prints, 2024, arXiv:2407.16216.

[3] C. Peng, X. Yang, K. E Smith, Z. Yu, A. Chen, J. Bian, Y. Wu,et al., “Model Tuning or Prompt Tuning? A Study of Large Language Models for Clinical Concept and Relation Extraction,” arXiv e-prints, 2023, arXiv:2310.06239.

[4] D. Yin, X. Han, B. Li, H. Feng, J. Bai,et al., “Parameter-efficient is not sufficient: Exploring Parameter, Memory, and Time Efficient Adapter Tuning for Dense Predictions,” arXiv e-prints, 2023, arXiv:2306.09729.

[5] H. Zhong, Z. Shan, G. Feng, W. Xiong, X. Cheng, L.Zhao, D. He, J. Bian, L. Wang, “DPO Meets PPO: Reinforced Token Optimization for RLHF,” arXiv e-prints, 2024, arXiv:2404.18922.

[6] Q. Yin, X. He, L. Deng, C. Tou Leong, F. Wang, Y. Yan, X. Shen, Q. Zhang,et al., “Deeper Insights Without Updates: The Power of In-Context Learning Over Fine-Tuning,” arXiv e-prints, 2024, arXiv:2410.04691.

[7] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo,et al., “A Survey on LLM-as-a-Judge,” arXiv e-prints, 2025, arXiv:2411.15594.

[8] J. Khlaut, C. Dancette, E. Ferreres, A. Bennani, P. Hérent, P. Manceron, “Efficient Medical Question Answering with Knowledge-Augmented Question Generation,” arXiv e-prints, 2024, arXiv:2405.14654.