CSC6203/CIE6021: Large Language Model

Lecture 10: Specialized LLMs

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

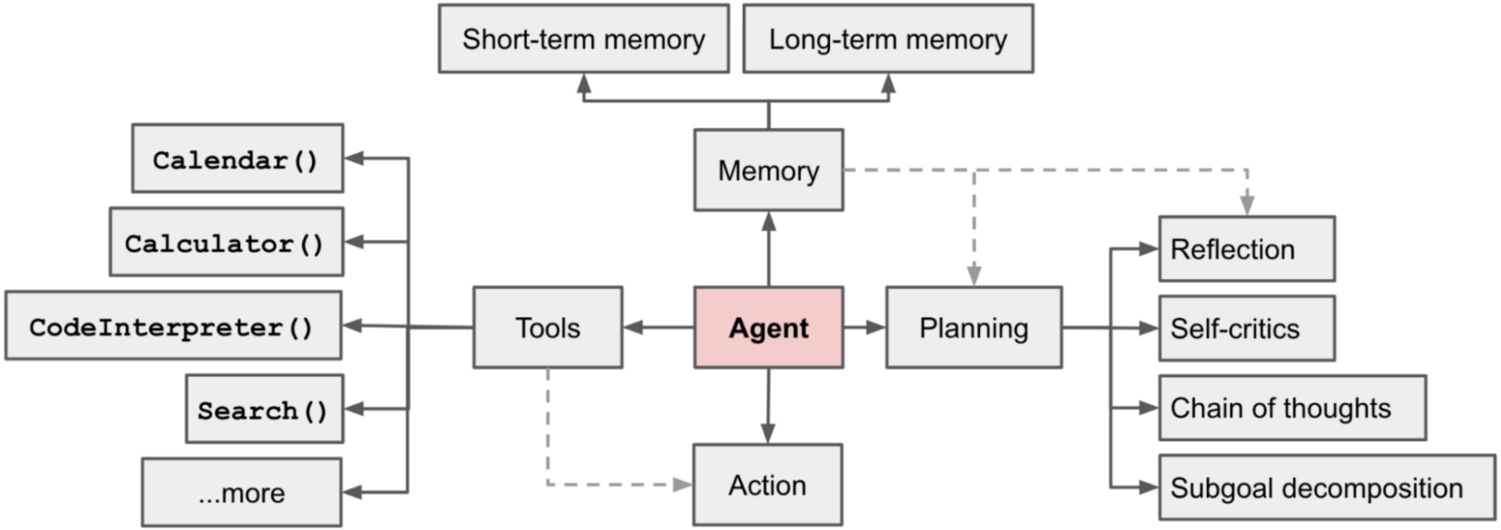
Winter 2023

Benyou Wang
School of Data Science

Recap…

What is LLM Agents

https://gptpluginz.com/llm-agents/

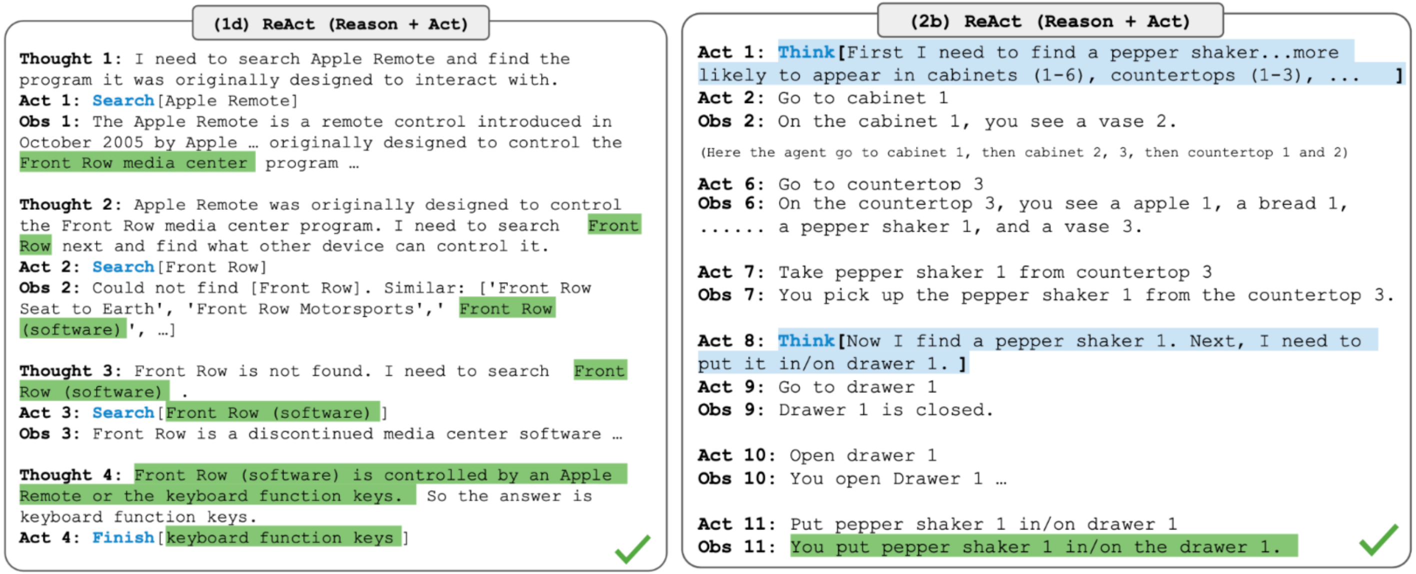


https://lilianweng.github.io/posts/2023-06-23-agent/

Examples of Planning

HotpotQA

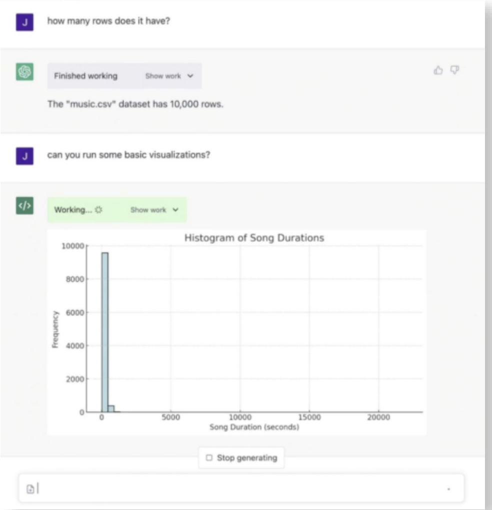
AlfWorld Env



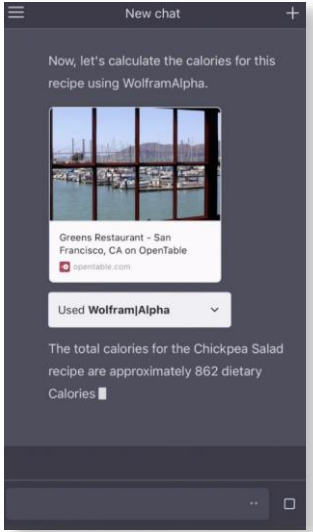
In both experiments on knowledge-intensive tasks and decision-making tasks, *ReAct* works better than the *Act*-only baseline where *Thought: …* step is removed.

[ReAct: Synergizing Reasoning and Acting in Language Models]

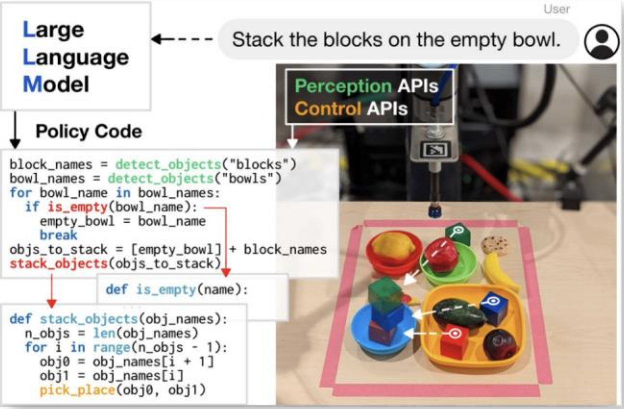
Examples of tools



Data analysis



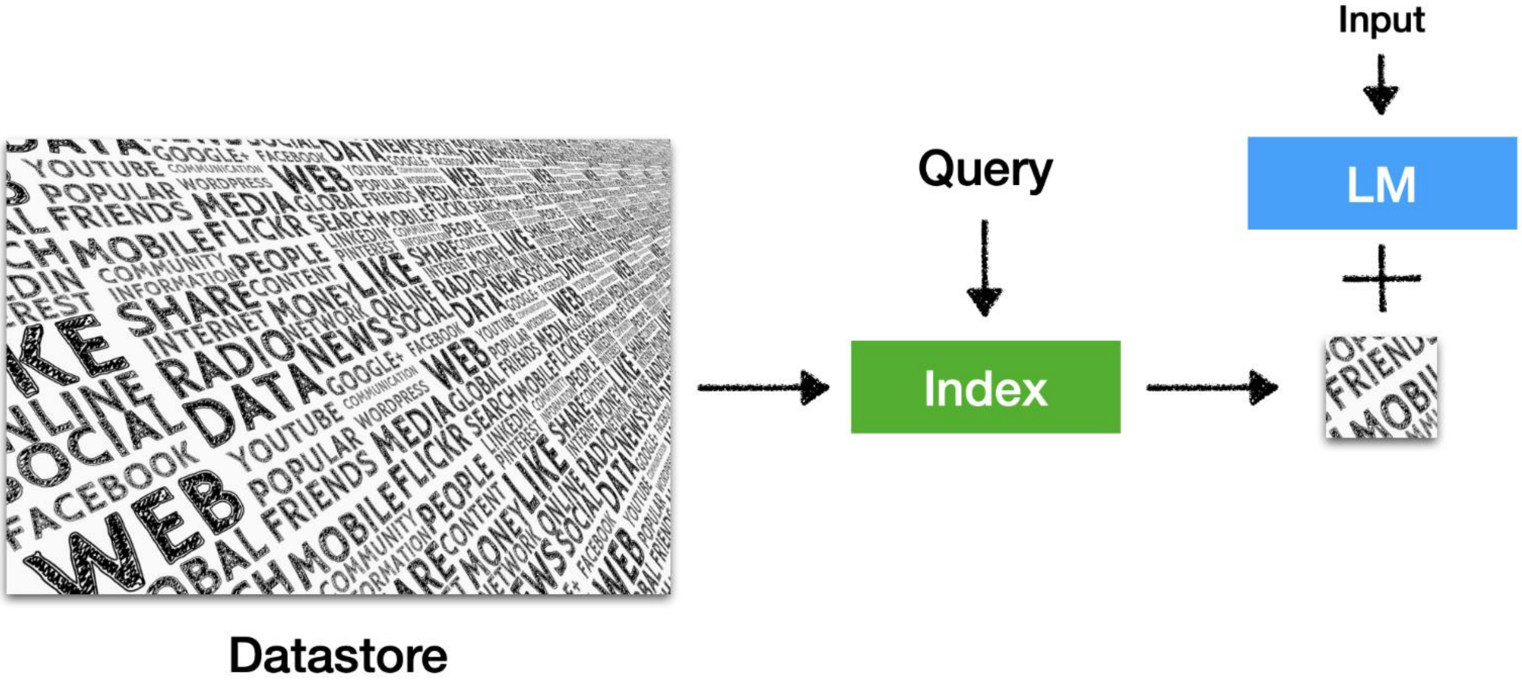
Web/Apps



Robotic physical world

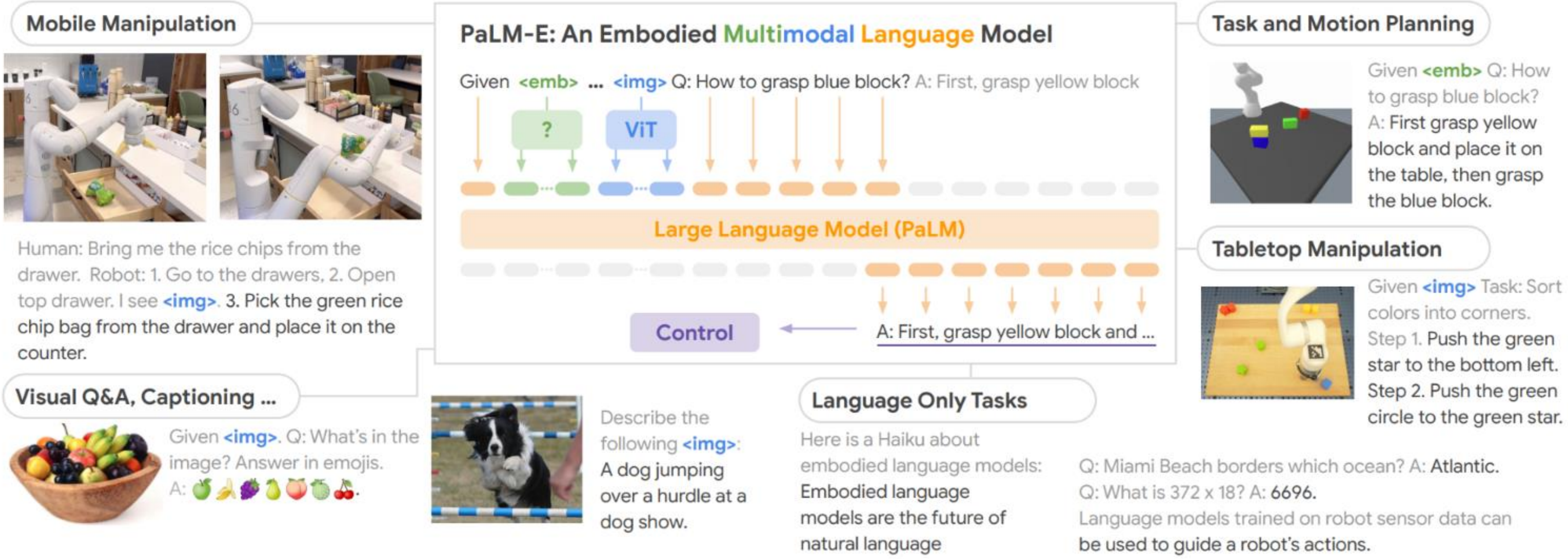
https://openai.com/blog/chatgpt-plugins
https://code-as-policies.github.io/

Examples of memory



Examples of action: Embodied AI

PaLM-E transfers knowledge from visual-language domains into embodied reasoning – from robot planning in environments with complex dynamics and physical constraints, to answering questions about the observable world.



PaLM-E: An Embodied Multimodal Language Model

Outline

- Introduction to Specialization
  - What is Specialized LLMs
  - **Motivation** for Specialization
  - **Challenges** of Specialization
- Domain-Specific Adaptation
  - LLM in the Medical Field
  - HuatuoGPT Series
  - Future Developments in Healthcare
- Language-Specific Adaptation
  - Phoenix: Multilingual Large Model Adaptation
  - AceGPT: Arabic Language Adaptation

Introduction to Specialization

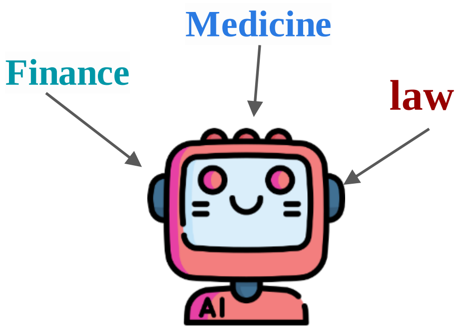
What is specialization?

Generalized LLM

Specialized LLM

work for everything vs. work for something that we care the most

Generalized LLM



Generalized LLM

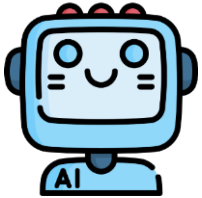
Focus on all usage scenarios

1. **Broad Domain Knowledge**: Generalized LLMs are designed to understand and process information across a wide range of domains such as law, medicine, and finance. They are not restricted to a specific field, allowing for versatile applications.

2. **Mostly Support only Common Languages**: These models primarily focus on commonly used languages like English. They may not support or have limited capabilities in less common languages, such as Arabic. This focus allows for more depth in popular languages but can limit accessibility for non-English speakers.

Specialized LLM

Medicine



Specialized LLM

Focus on only one usage scenario

1. **Domain-Specific Expertise:** Specialization LLMs are tailored to specific sectors or fields. For instance, a model might be exclusively trained on medical literature or legal documents, leading to a deeper understanding and more accurate responses in that particular domain.

2. **Multilingual Support or to a non-English language:** Unlike generalized models, specialization LLMs can be developed for specific languages relevant to their domain. This means a legal LLM could be specifically designed for use in Arabic if the demand in that language is high in the legal domain.

Specialization

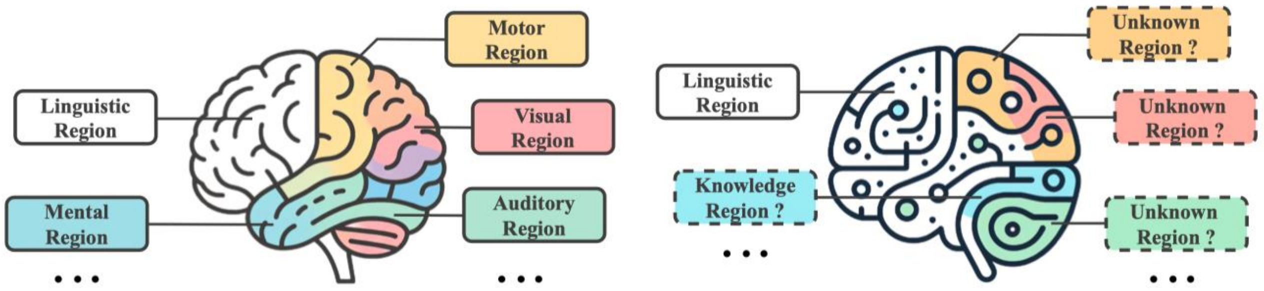


Figure 1: Based on the human brain (left) as a prototype, we have discovered a region in LLMs (right) that corresponds to linguistic competence. Furthermore, we have found that improvements in linguistic competence do not necessarily coincide with increases in knowledge levels, which may suggest the presence of a dissociated knowledge region. In the future, we will continue to explore the possibility of other functional regions.

Language and **knowledge** are the capabilities of different regions of the LLM, enhancing specific regions with limited parameters.

Specialization is used to enhance specific **domain knowledge** or specific **language** skills.

Zhao J, Zhang Z, Ma Y, et al. Unveiling A Core Linguistic Region in Large Language Models[J]. arXiv preprint arXiv:2310.14928, 2023.

Two Types of specialization

Domain-Specific Adaptation

- **Focus:** This type specializes in a single domain, such as healthcare, finance, or law.
- **Advantages:** Offers in-depth, nuanced understanding and responses within its chosen field. Highly accurate and relevant in domain-specific contexts.
- **Use Case:** Ideal for professional or industry-specific applications where expert-level knowledge and terminology are crucial.

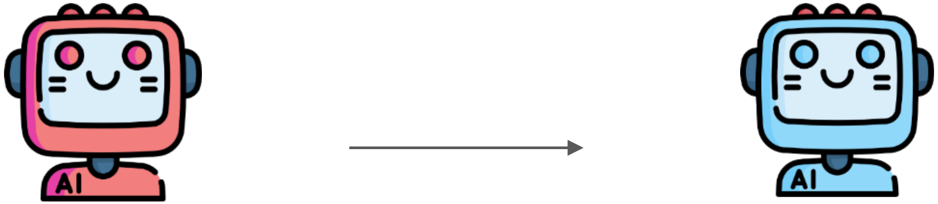
Language-Specific Adaptation

- **Focus:** Enhances adaptation to less common languages, addressing the gap left by Generalized LLMs which primarily cater to widely common languages.
- **Advantages:** Facilitates accessibility and inclusivity, enabling users of minority languages to access advanced LLM capabilities.
- **Use Case:** Particularly valuable in linguistically diverse regions where less common languages are spoken, ensuring equitable access to technology.

Vertical (specialized) vs. Horizontal

Motivation for Specialization

Domain Specialization



General-Purpose

Domain-specific

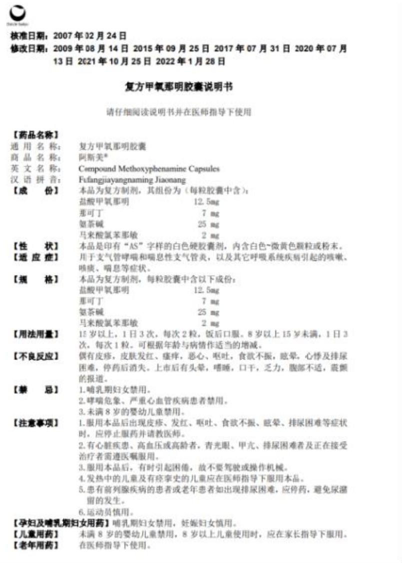
Domain Specialization is used to adapt to specialized domains.

Domain Specialization

With a similar size and inference cost, specialized models **might perform better** than a generalized one.

Generalization to Specialization

- Current SOTA generalized LLM is **insufficient for specific knowledge** (e.g., medical knowledge in Chinese)



核准日期：2007 年 02 月 24 日
修改日期：2009 年 08 月 14 日 2015 年 09 月 25 日 2017 年 07 月 31 日 2020 年 07 月 13 日 2021 年 10 月 25 日 2022 年 1 月 28 日

复方甲氧那明胶囊说明书

请仔细阅读说明书并在医师指导下使用

【药品名称】
通用名称：复方甲氧那明胶囊
商品名称：阿斯美®
英文名称：Compound Methoxyphenamine Capsules
汉语拼音：Fufangjiayangnaming Jiaonang

【成　　份】 本品为复方制剂，其组份为（每粒胶囊中含）：

盐酸甲氧那明	12.5mg
那可丁	7 mg
氨茶碱	25 mg
马来酸氯苯那敏	2 mg

【性　　状】 本品是印有“AS”字样的白色硬胶囊剂，内含白色~微黄色颗粒或粉末。

【适 应 症】 用于支气管哮喘和喘息性支气管炎，以及其它呼吸系统疾病引起的咳嗽、咳痰、喘息等症状。

【规　　格】 本品为复方制剂，每粒胶囊中含以下成份：

盐酸甲氧那明	12.5mg
那可丁	7 mg
氨茶碱	25 mg
马来酸氯苯那敏	2 mg

【用法用量】 15 岁以上，1 日 3 次，每次 2 粒，饭后口服。8 岁以上 15 岁未满，1 日 3 次，每次 1 粒。可根据年龄与病情作适当的增减。

【不良反应】 偶有皮疹，皮肤发红、瘙痒，恶心、呕吐，食欲不振，眩晕，心悸及排尿困难，停药后消失。上市后有头晕，嗜睡，口干，乏力，腹部不适，震颤的报道。

【禁　　忌】
1.哺乳期妇女禁用。
2.哮喘危象、严重心血管疾病患者禁用。
3.未满 8 岁的婴幼儿童禁用。

【注意事项】
1.服用本品后出现皮疹、发红、呕吐、食欲不振、眩晕、排尿困难等症状时，应停止服药并请教医师。
2.有心脏疾患、高血压或高龄者，青光眼、甲亢、排尿困难者及正在接受治疗者需遵医嘱服用。
3.服用本品后，有时引起困倦，故不要驾驶或操作机械。
4.发热中的儿童及有痉挛史的儿童应在医师指导下服用本品。
5.患有前列腺疾病的患者或老年患者如出现排尿困难，应停药，避免尿潴留的发生。
6.运动员慎用。

【孕妇及哺乳期妇女用药】 哺乳期妇女禁用，妊娠妇女慎用。

【儿童用药】 未满 8 岁的婴幼儿童禁用，8 岁以上儿童使用时，应在家长指导下服用。

【老年用药】 在医师指导下使用。

Generalization to Specialization

- GPT-4 seems a Mixture-of-Expert architecture, each expert look like a specialized model

Yam Peleg @Yampeleg · Jun 21

B R E A K I N G! @realGeorgeHotz Just said:
- **GPT-4** Has 220 parameters.
- **GPT-4** Is a **mixture of experts** with 8 experts.
- **GPT-4** Is doing 16 times infer**ence (did u me**an beams? or ju**st 2 beams per model?)**
-
@realGeorgeHotz It's HUGE!! Can you confirm??
src: latent.space/p/geohot

anton @abacaj · Jun 21

Interesting, George Hotz mentioning GPT-4 size/architecture in a recent podcase he did

George: Yeah, yeah, we could build. So like the biggest training clusters today, I know less about how GPT-4 was trained. I know some rough numbers on the weights and stuff, but Lama- [00:43:28]

Swyx: A trillion parameters? [00:43:30]

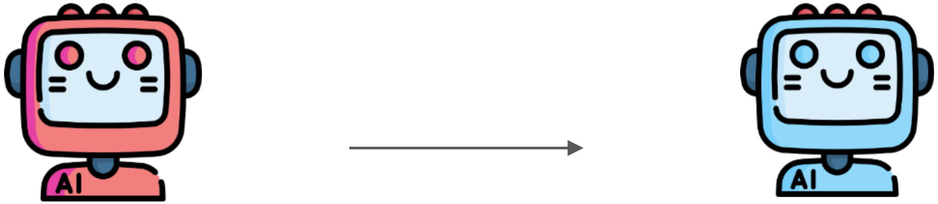
George: Well, okay, so GPT-4 is 220 billion in each head, and then it's an eight-way mixture model. So mixture models are what you do when you're out of ideas. So, you know, it's a mixture model. They just train the same model eight times, and then they have some little trick. They actually do 16 inferences, but no, it's not like- [00:43:45]

25 48 240 153K

https://twitter.com/Yampeleg/status/1671272245077979141

More on **Language** Specialization

Language Specialization



common language

uncommon language

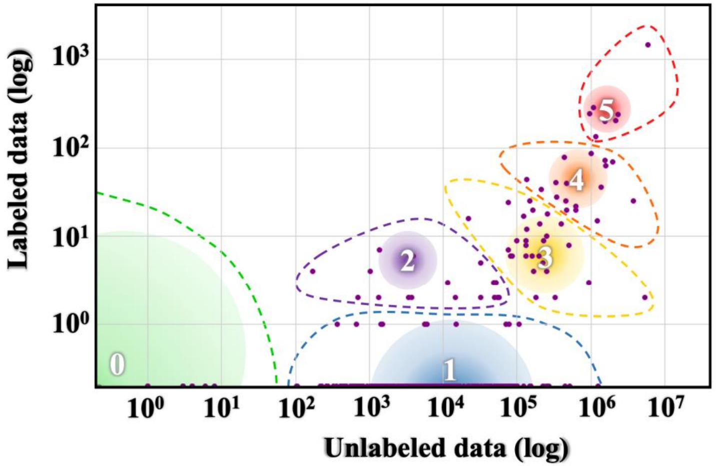
Domain Specialization is used to adapt generalized models to specific languages.

Most Languages are under-represented

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

88% of the world's languages, spoken by 1.2B people are untouched by the benefits of language technology.

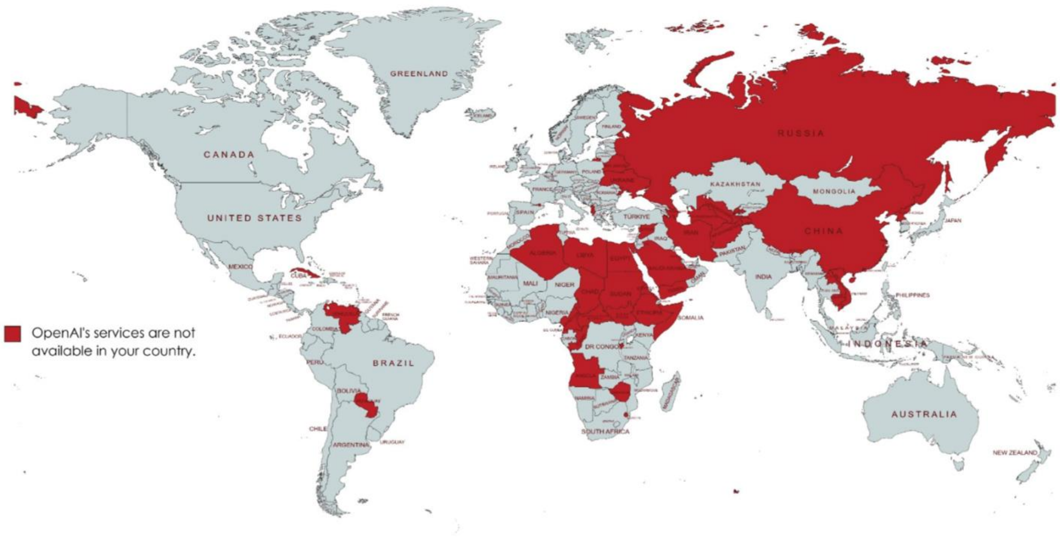
Most Languages are under-represented



Hierarchy of languages in terms of available resources for training NLP systems

How well have Language Technologies been serving the 6000+ languages of the planet?

ChatGPT is unavailable to many regions



Being Red indicates that in the area ChatGPT is unavailable. Source of the picture: https://i.imgur.com/2fF3Xlh.png

Language Specialization

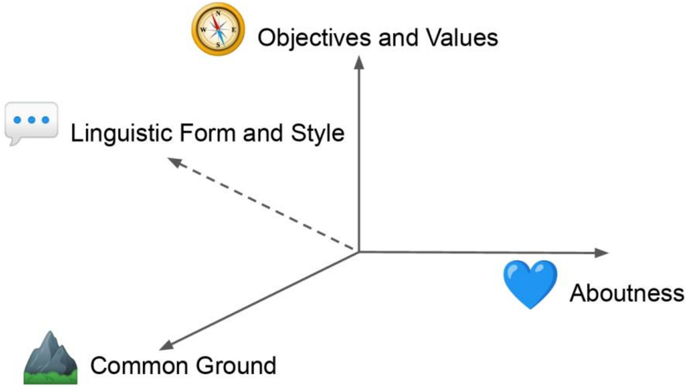
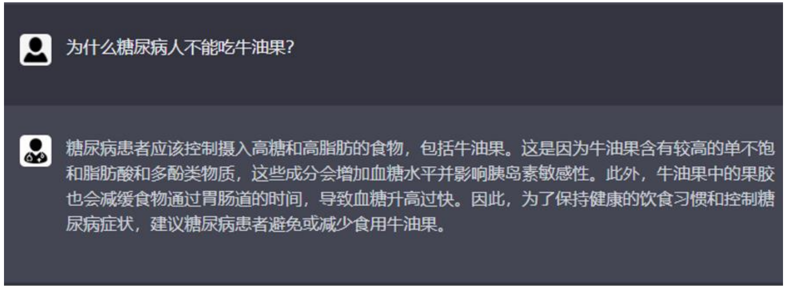


Figure 1: The role of culture in NLP, illustrated by four dimensions along which cultures vary, and for which NLP can be culturally biased: linguistic form and style, common ground, aboutness, and objectives (values).

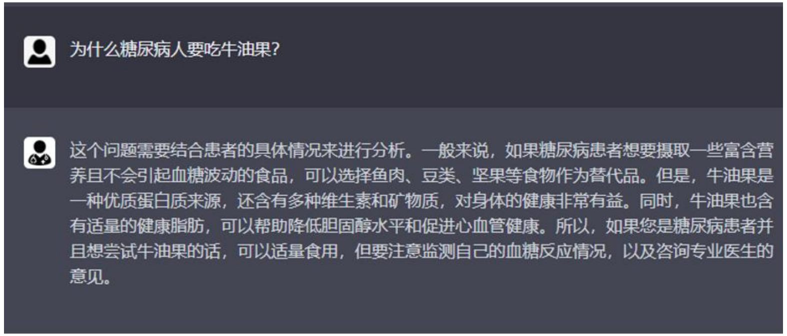
Moveover, Culture and Language are strongly correlated. It is important that language specialization is also a process of adaptation to a particular **culture**.

Challenges for Specialization

Challenge 1: hallucination issue is more severe


为什么糖尿病人不能吃牛油果?

糖尿病患者应该控制摄入高糖和高脂肪的食物，包括牛油果。这是因为牛油果含有较高的单不饱和脂肪酸和多酚类物质，这些成分会增加血糖水平并影响胰岛素敏感性。此外，牛油果中的果胶也会减缓食物通过胃肠道的时间，导致血糖升高过快。因此，为了保持健康的饮食习惯和控制糖尿病症状，建议糖尿病患者避免或减少食用牛油果。


为什么糖尿病人要吃牛油果?

这个问题需要结合患者的具体情况来进行分析。一般来说，如果糖尿病患者想要摄取一些富含营养且不会引起血糖波动的食品，可以选择鱼肉、豆类、坚果等食物作为替代品。但是，牛油果是一种优质蛋白质来源，还含有多种维生素和矿物质，对身体的健康非常有益。同时，牛油果也含有适量的健康脂肪，可以帮助降低胆固醇水平和促进心血管健康。所以，如果您是糖尿病患者并且想尝试牛油果的话，可以适量食用，但要注意监测自己的血糖反应情况，以及咨询专业医生的意见。

无知者无畏，胡说八道

Challenge 2: expert labelling become more expensive

ChatGPT and GPT-4 labeling may be higher than Kenyan labelers but not as good as professional doctors and lawyers.

Challenge 3: the Evaluation Dilemma

Standard evaluation sets may

- Test set leaks
- Some questions do not have standard answers
- Test set does not cover all user paths

GPT-4/Expert Evaluation

- Often requires a vertical expert assessment to do the manual assessment, need to work with vertical experts
- GPT-4 Occasionally does not satisfy transferability when evaluated: A < B; B < C; C > A

Challenge 4: how to incorporate deterministic rules for verticals

配伍药物	配伍分析
水溶性维生素 + 氯化钾	水溶性维生素中的有机酸盐、有机碱盐和羟苯甲酯溶解度降低，从而自溶液中析出，不溶微粒增加。
呋塞米 + 多巴胺 + 葡萄糖注射液	呋塞米为加碱制成的钠盐注射液，静脉注射时宜用氯化钠注射液稀释，而不宜用葡萄糖注射液稀释。如果需要同时使用，建议分开使用，并在中间输注0.9%的生理盐水。
地塞米松 + 维生素B6	两药的浓溶液在同一容器中混合可能产生混浊或沉淀。维生素B6可以导致水不溶性的酸性物质制成的盐地塞米松磷酸盐等产生沉淀。
多烯磷脂酰胆碱 + 氯化钾	多烯磷脂酰胆碱对电解质十分敏感，加入少量的电解质就能促使溶胶聚沉。因此，不应与电解质注射液混合注射。
维生素C + 维生素K1	维生素K1可被维生素C破坏而失效。虽然两者在药理和病理学上的联用有利，但维生素C的强还原性会降低维生素K1的疗效。
胰岛素 + 维生素C	维生素C的强还原性会导致胰岛素失活。
头孢曲松 + 葡萄糖酸钙	头孢曲松与含钙药物（包括含钙溶液）联用可能产生致死性不良事件。因此，不应将两者混合或同时使用。
维生素K1 + 氯化钾	氯化钾可使维生素K1含量下降30%多。

Solutions

- Soft knowledge inject
- Reward model
- Plugins
- Retrieval augmented generation
- Post-processing

Challenge 5: How to pay less domain tax

Possibly neutering some of the generalized knowledge and competencies in order to learn domain knowledge and domain competencies

- Mix some generic data when training domain data
- Parameter modularization?

Outline

- Introduction to Specialization
  - What is Specialized LLMs
  - **Motivation** for Specialization
  - **Challenges** of Specialization
- Domain-Specific Adaptation
  - LLM in the Medical Field
  - HuatuoGPT Series
  - Future Developments in Healthcare
- Language-Specific Adaptation
  - Phoenix: Multilingual Large Model Adaptation
  - AceGPT: Arabic Language Adaptation

Domain-Specific Adaptation
A case study in HuatuoGPT

Background: Stuck in the neck (卡脖子) on ChatGPT/GPT-4

The U.S. company OpenAI banned access to ChatGPT/GPT-4 from mainland China and Hong Kong and banned mainland China and Hong Kong from purchasing servers A100 and H100 for training ChatGPT and GPT-4.

Visit https://chatgpt.cuhk.edu.cn instead

https://www.sec.gov/ix?doc=/Archives/edgar/data/1045810/000104581022000146/nvda-20220826.htm

Background: Medical Privacy Impacts National Security

Transferring the data of billions of people in China to the servers of U.S. companies will pose a huge risk to national security.

[1] https://openai.com/policies/privacy-policy
[2] https://www.techgoing.com/softbank-hitachi-and-many-other-japanese-companies-restrict-the-use-of-chatgpt-to-prevent-leaks/

Background: Uneven Medical Resources

Life expectancy in China

With a 50% misdiagnosis rate in township hospitals, how can we help township doctors?

Healthcare is a matter of national concern

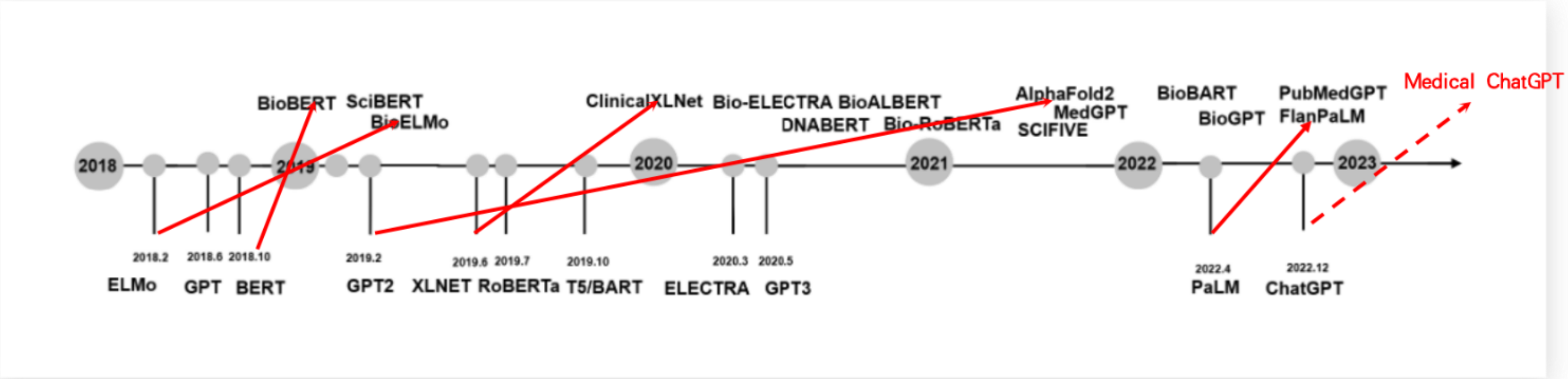
In the last three years, China's total health care costs reached 6.6%, 7.1%, and 6.5% of GDP
Inspired by the success of the GPT-4, the significance of the health care megamodel is:

- ❑ Real-time medical guide inquiry
- ❑ Triage offline inquiry needs
- ❑ Intelligent pre-consultation and follow-up
- ❑ Help with doctor training

- ❑ Provide health advice around the clock without the need to wait for a doctor's appointment
- ❑ Alleviation of medical taboos, e.g. gender issues
- ❑ Intelligent triage, pre-consultation and follow-up visits
- ❑ Public access to healthcare knowledge

- ❑ Reducing hospital congregation and lowering the risk of spreading infectious diseases
- ❑ Slow down doctor-patient conflicts
- ❑ Alleviate the inequality of medical resources between urban and rural areas, and share high-quality medical resources online nationwide

Development of Biomedical LLMs



The vertical domain model usually lags behind the general model
Next, the Medical ChatGPT is coming

Benyou Wang et al. Pre-trained language models in biomedical domain: A systematic survey[J]. ACM Computing Survey.

Development of Biomedical LLMs

ChatDoctor

MedAlpaca

Bianque

BenTsao

DoctorGLM

MedicalGPT-zh

Instruction Tuning on Medical Dataset

Li et al. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge.
Han et al. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data.
Singhal K, Tu T, Gottweis J, et al. Towards Expert-Level Medical Question Answering with Large Language Models[J]. arXiv preprint arXiv:2305.09617, 2023.
Wang et al., 2023. Huatuo: Tuning llama model with chinese medical knowledge.
Xiong et al., 2023. DoctorGLM: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097.

Medical LLMs

The medical LLM is the bridge between doctors and patients

User: 我的后背有点痛，我去医院的哪个科室？
Medical LLM:
您的症状可能需要进行进一步的检查才能确定病因，您可以先去看一下**全科医生或者骨科医生**，他们会对您的情况进行评估。如果需要的话，他们也可以为您安排相关的检查以确定诊断。希望这能帮到您！同时，请注意保持良好的姿势和适度的运动，这对缓解疼痛也有很大的帮助。

- ✓ （问诊）**Medical Consultation**
- ✓ （生成医疗病例）Generation of Medical Records
- ✓ （医疗指标解读）Interpretation of Indicators
- ✓ （处方建议）Prescription Recommendations
- ✓ （医生培训）Physician Training
- ✓ （心理咨询）Psychological Counseling
- ✓ ……

Medical LLMs

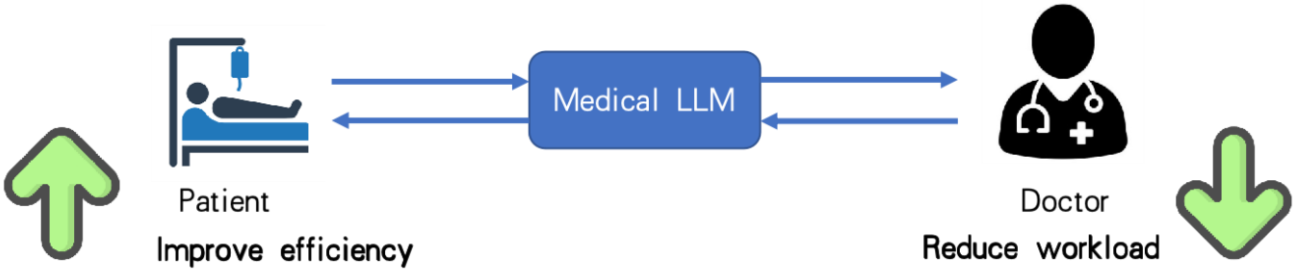
The medical LLM is the bridge between doctors and patients

- ✓ （问诊）Medical Consultation
- ✓ （生成医疗病例）Generation of Medical Records
- ✓ （医疗指标解读）Interpretation of Indicators
- ✓ （处方建议）**Prescription Recommendations**
- ✓ （医生培训）Physician Training
- ✓ （心理咨询）Psychological Counseling
- ✓ ……

User:
根据患者的年龄、性别、孕哺情况、临床诊断，给该患者推荐合理的处方，处方包括药品名称、规格、用法用量
患者病情：
患者性别女、25岁，孕22周，怕冷、乏力，经检查诊断为甲状腺功能减退症
Medical LLM：
针对患者的情况，我建议开具以下药物处方：
药品名称：优甲乐（左甲状腺素钠片）
规格：25μg/片
用法用量：每日早晨空腹服用一片，避免与其他药物或食品一起服用。
此外，还应该提醒患者在吃药期间如果出现心悸、腹泻、呕吐、震颤、兴奋、头痛等不良反应，应及时向医生说明。希望我的建议对您有所帮助，如有任何疑问，请随时与我联系。

Medical LLMs

The medical LLM is the bridge between doctors and patients



Improve efficiency

- Provide 24-hour service（24*7 问诊）
- Avoid embarrassment（避免“讳病忌医”）
- Learning healthcare knowledge （学习保健知识）

Reduce workload

- Assisted Medical Diagnosis（辅助诊断）
- Diversion consultation needs（分流问诊）
- Intelligent pre-consultation（智能预分诊）

- Reduce hospital gatherings and reduce the risk of infectious disease spread
- Alleviate doctor-patient conflicts
- Alleviate the unequal medical resources between urban and rural area

HuatuoGPT Series

- **HuatuoGPT v1** - A Doctor-Like Medical LLM
- **HuatuoGPT v2** - An Expert-Level Medical LLM

HuatuoGPT is a prime example of adaptation in the medical field, demonstrating the progress of Large Language Models (LLMs) in healthcare.

HuatuoGPT V1

HuatuoGPT, Towards Taming Language Model To Be a Doctor

New version: https://v1.huatuogpt.cn/

Motivation

Give a question：我肚子疼

ChatGPT: 肚子疼可以是许多不同问题的症状，可能包括消化问题、感染、疾病或其他健康问题。你可以尝试以下几点来缓解肚子疼：
1. 休息：尽量休息，让身体有时间恢复。
2. 喝水：保持身体水分充足，避免脱水。
3. 避免食用油腻、辛辣、高脂肪食物：这些食物可能加重肚子不适。
4. 尝试温热敷：有时候用温水瓶或温暖的毛巾敷在肚子上可以帮助缓解...

Pros: Detailed and helpful.
Cons: Can't perform like doctors and gives vague medical responses.

Doctor: 肚子哪个部位痛？是剧烈的还是轻微的？持续多久了？
User: 下腹部， 有点疼， 昨晚开始的， 昨晚吃了点小龙虾。
Doctor: 很可能是急性肠胃炎，通常是由于食物中毒引起的， 而小龙虾中含有细菌和寄生虫，很容易引起肠胃感染。

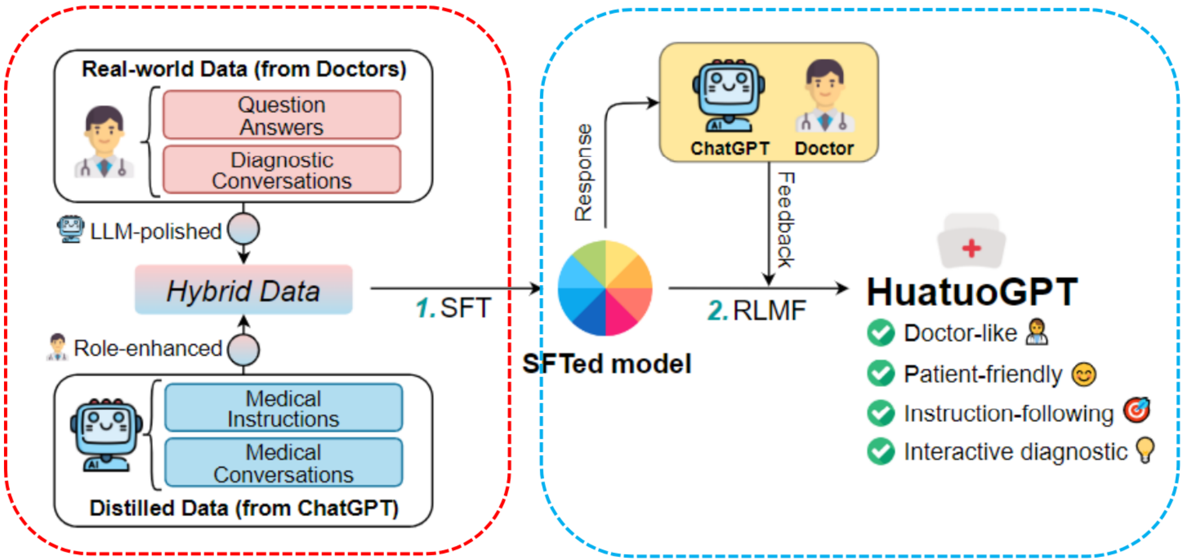
Pros: Professional with diagnostic capability.
Cons: Brief, poorly presented replies.

Motivation

Features	Aspects	ChatGPT Data	Doctor Data	We expect
Doctor-like	Diagnostic ability	-	High	High
	Raising questions ability	-	High	High
	Expert-level accuracy	-	High	High
Patient-friendly	Informativeness	High	-	High
	Patience	High	-	High
	Presentation quality	High	-	High

We expect a language model to be both **Doctor-like** and **Patient-friendly**.

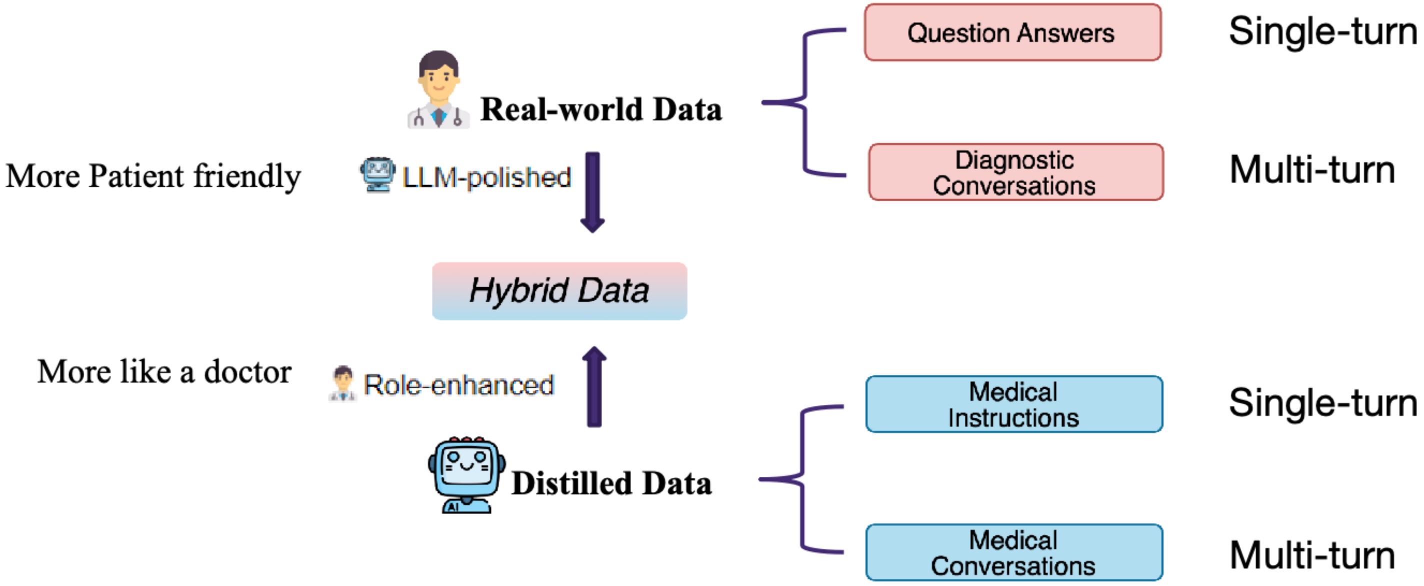
Method



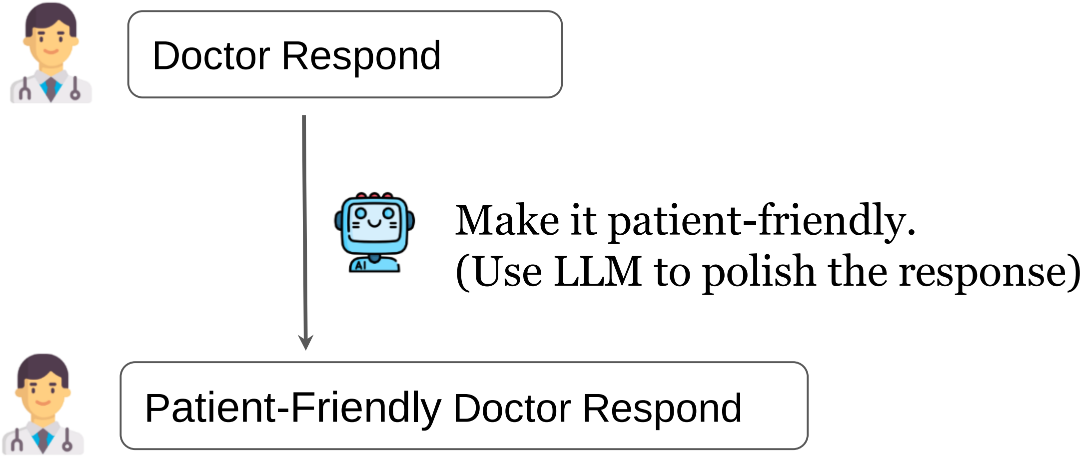
SFT with Hybrid Data

RL from Mixed Feedback

SFT with Hybrid Data

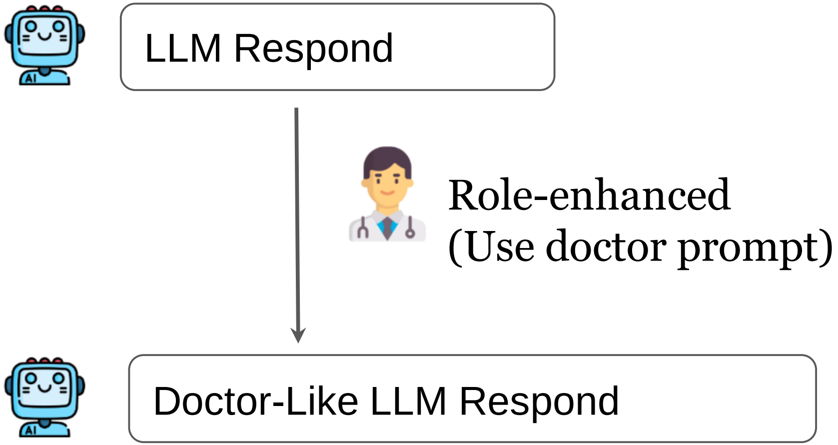


LLM Polish: Doctor -> ChatGPT



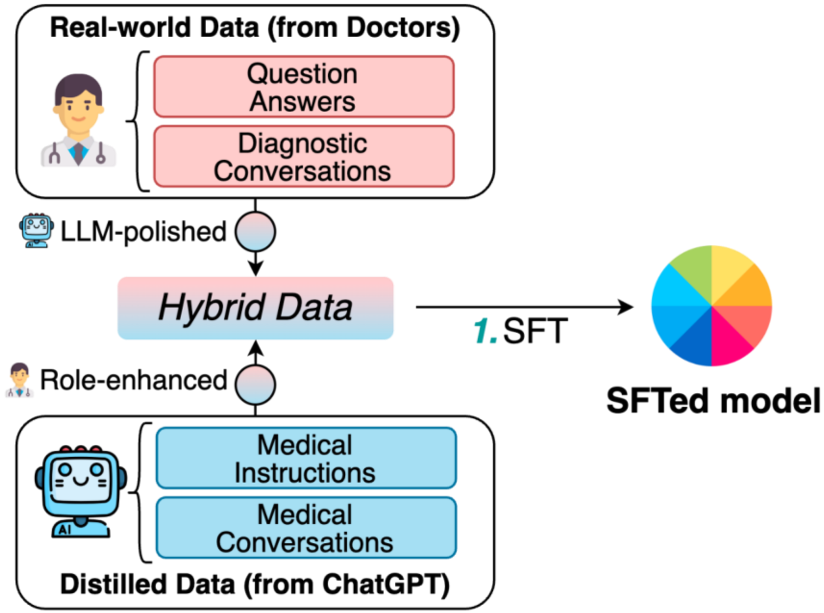
We use LLM to optimize doctors' responses to make their content more patient-friendly.

Role Enhance: ChatGPT ->Doctor



We prompt the LLM to simulate a doctor, endowing it to be doctor-like.

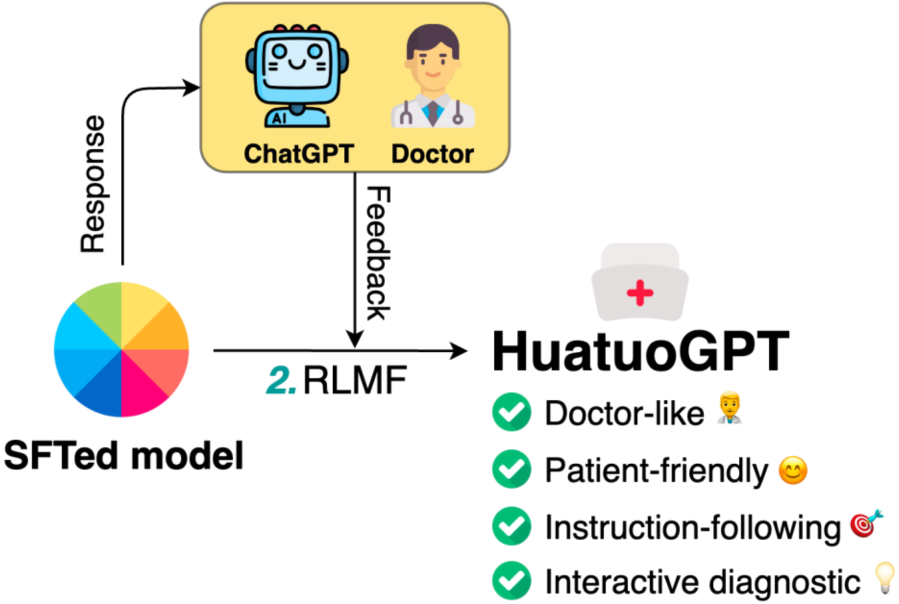
SFT with Hybrid Data



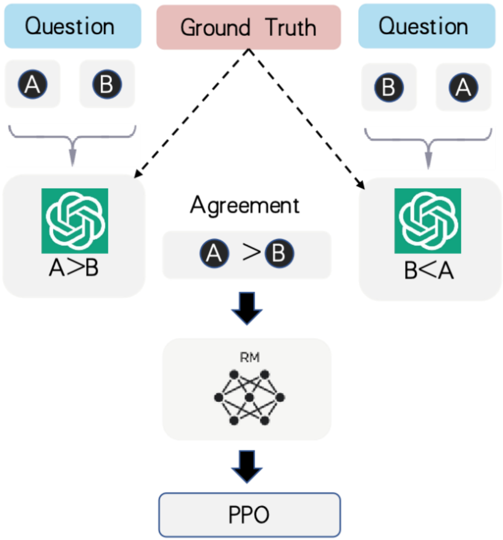
1. SFT (Supervised Fine-Tuning):
The model is trained on both distilled data and real-world data, allowing it to acquire a foundation of abilities that are both patient-friendly and doctor-like.

RLMF

2. RLMF (Reinforcement Learning from Mixed Feedback):
We employ reinforcement learning to train the model to assimilate the distinct advantages of distilled data and real data, thereby becoming both patient-friendly and akin to a doctor.



RLMF



Here is a conversation history:
[History]
${History}
[End of History]
Here is the final question and the standard answer:
[Question]
${Query}
[End of question]
[Standard answer]
${Doctor_response}
[End of standard answer]
Based on the conversation history, user question, and standard answer, please rate the following two AI responses on a scale of 1 to 10, **considering accuracy, conciseness, and similarity to the standard answer.**
Please provide the ratings in the following format: "Rating A: [score];Rating B: [score]".
[Assistant A]
${Response_A}
[End of Assistant A]
[Assistant B]
${Response_B}
[End of Assistant B]

Ouyang et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
Bai et al. Constitutional AI: Harmlessness from AI Feedback[J]. arXiv preprint arXiv:2212.08073, 2022.

Training Detail

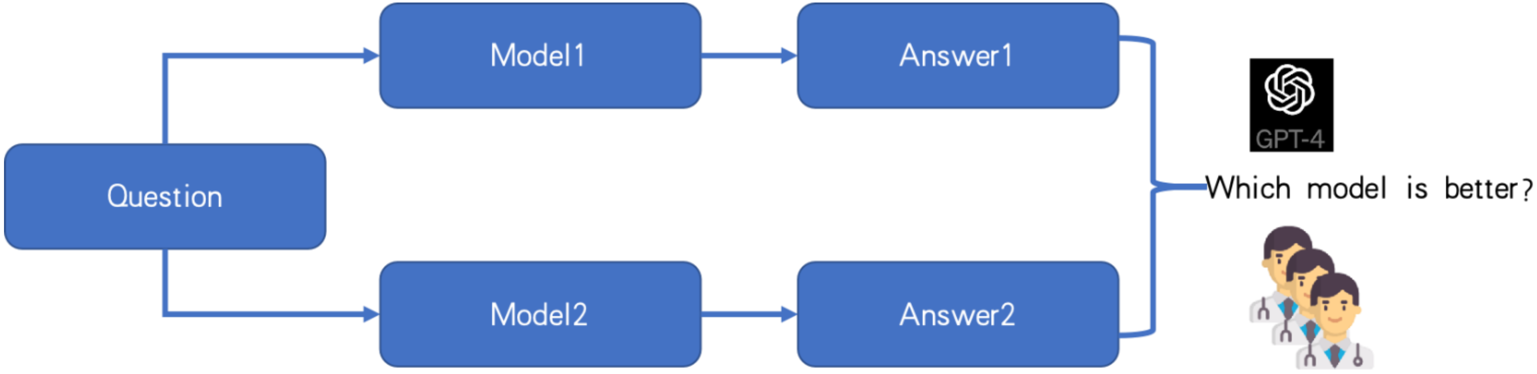
Our model is implemented in PyTorch using the Accelerate and trlx packages with LLaMA as the base architecture.
Initialize the model parameters using Ziya-LLaMA-13B-Pretrain.

We leverage ZeRO strategy to distribute the model across 8 * A100 (80G) GPUs for training.

SFT	
learning rate	5e−5
batch size	128
maximum context length	2048

RLMF	
process encompasses	16000
learning rate	1e−6
rollout size	64
chunk size	8
epoch	4
KL divergence coefficient	0.1

Evaluation on Single-turn Questions



Chiang et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% ChatGPT quality.

Evaluation on Single-turn Questions

[Question]
心肌缺血如何治疗与调养呢？
[Assistant 1]
Response1
[End of Assistant 1]
[Assistant 2]
Response2
[End of Assistant 2]
[System]
We would like to request your feedback on the two AI assistants in response to the user question displayed above.
Requirements：The response should act like the doctor using the tone, manner and vocabulary the human doctor would use. It should be to the point, without unnecessary elaboration or extraneous information. The description of symptoms should be comprehensive and accurate, and the provided diagnosis should be the most reasonable inference based on all relevant factors and possibilities. The treatment recommendations should be effective and reliable, taking into account the severity or stages of the illness. The prescriptions should be effective and reliable, considering indications, contraindications, and dosages.
Please compare the performance of their responses. You should tell me whether Assistant 1 is `better than`, `worse than`, or `equal to` Assistant 2.
Please first compare their responses and analyze which one is more in line with the given requirements.
In the last line, please output a single line containing only a single label selecting from `Assistant 1 is better than Assistant 2`, `Assistant 1 is worse than Assistant 2`, and `Assistant 1 is equal to Assistant 2`.

GPT-4

Diagnosis accuracy（诊断准确性）. This aspect evaluates the model's accuracy and comprehensiveness in diagnosing patient symptoms. Evaluators are provided a set of medical cases or symptom descriptions and assess the correctness, relevance, and reasonableness of the model's diagnosis. Comparisons can be made with assessments made by medical professionals to ensure the model's accuracy.

Treatment recommendation accuracy（治疗建议准确性）. This aspect assesses the accuracy and appropriateness of the model's treatment recommendations for patients. Evaluators are provided a set of medical cases or symptom descriptions and evaluate whether the model's treatment recommendations align with medical knowledge and real-world applications that are effective and reliable for the patient's main condition and problem.

Medication knowledge and prescription accuracy（药物知识和处方准确性）. This aspect evaluates the model's understanding of medications and the accuracy of its prescription recommendations. Evaluators are provided a set of medical cases or symptom descriptions and assess the accuracy and reliability of the medication recommendations based on medical knowledge and guidelines.

Evaluation on Single-turn Questions

Single-turn Questions in 10 intents from KUAKE-QIC of CBLUE

病情诊断：已知症状，判断可能的原因
- 最近早上起来浑身无力是怎么回事？

病因分析：已知疾病，解释疾病发生的原因
- 鼻咽癌是如何发生的？

治疗方案：已知疾病/症状，给出治疗或缓解的方案（检查/手术/药物/行为）
- 烫伤的疤痕要怎么去除？

就医建议：已知症状/疾病，给出就医建议（科室/检查）
- 糖尿病该做什么检查？

指标解读：身高/体重/血压等检查结果的数值范围解读
- 血常规超敏C反应蛋白偏高说明什么

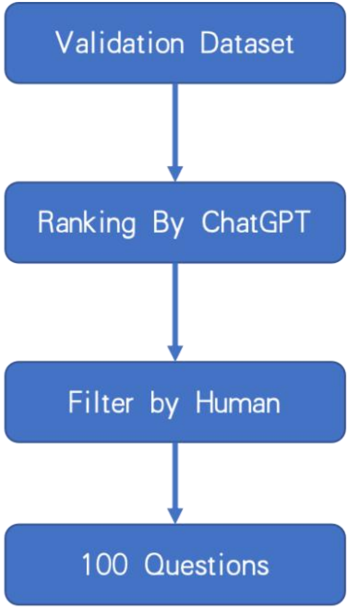
疾病表述：疾病属性（eg：能不能治、能不能治好）、症状、表现、图片等相关表述
- 外痔疮早期症状有哪些呢？

后果表述：疾病/症状/药品/检查项/食物的危害，疾病恶化不治疗会产生的不良影响或治疗后会产生好的结果
- 缺乏钾元素会怎么样

注意事项：病人要注意的事情，以及分析食物的好坏，食物对病人的影响
- 哮喘应该注意些什么

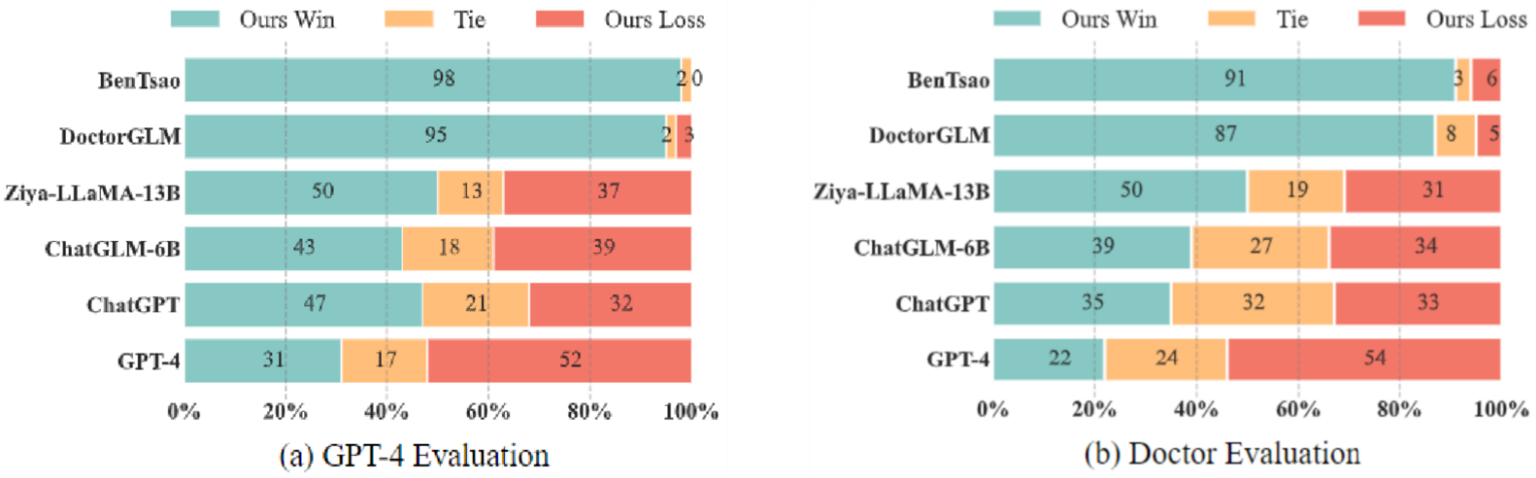
功效作用：食品/药物的好处，功效/作用/副作用
- 乌鸡白凤丸的功效和作用

医疗费用：疾病/手术/药品/检查/的费用
- 二甲双瓜要多少钱？



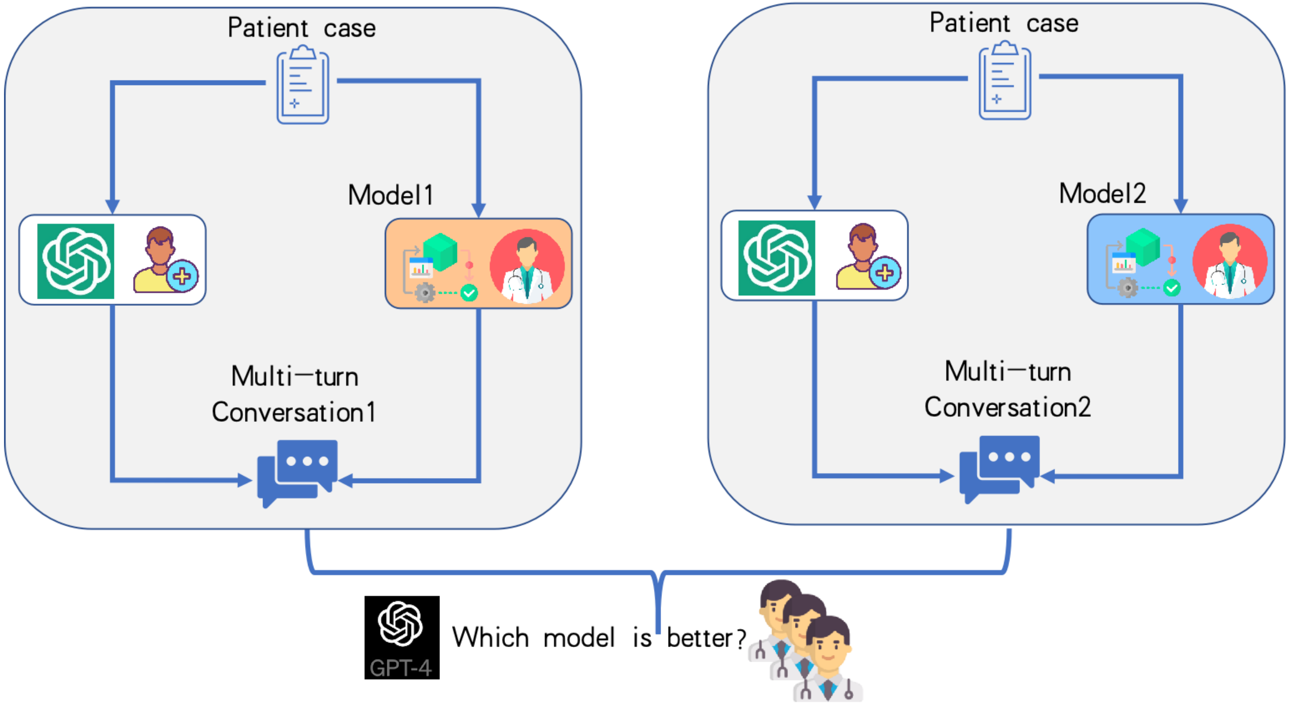
Zhang et al. CBLUE: A Chinese biomedical language understanding evaluation benchmark[J]. arXiv preprint arXiv:2106.08087, 2021.

Evaluation on Single-turn Questions

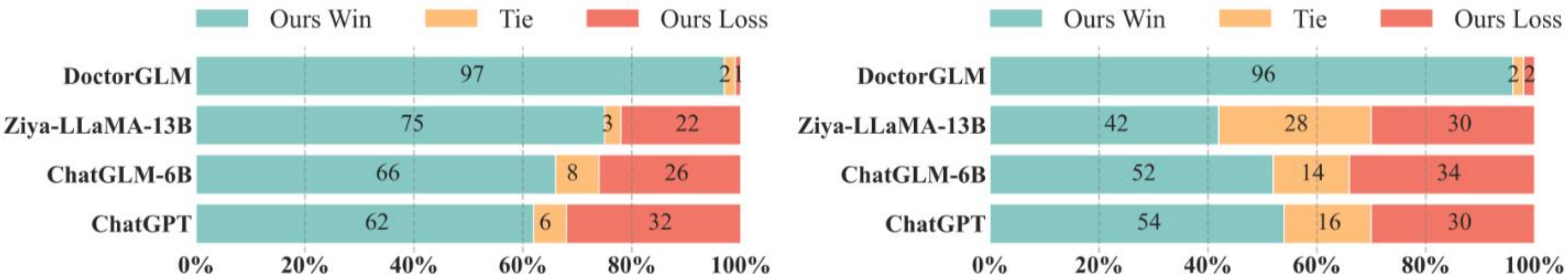


Response Comparison of HuatuoGPT with Other Baselines on the Single-turn Question.

Evaluation on Multi-turn Conversation



Evaluation on Multi-turn Conversation

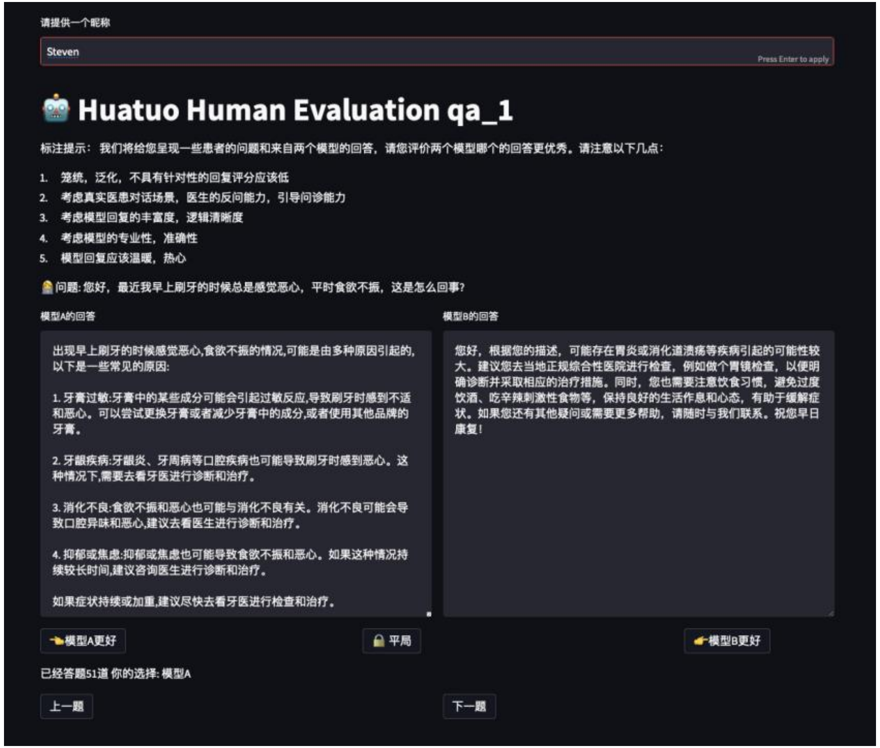


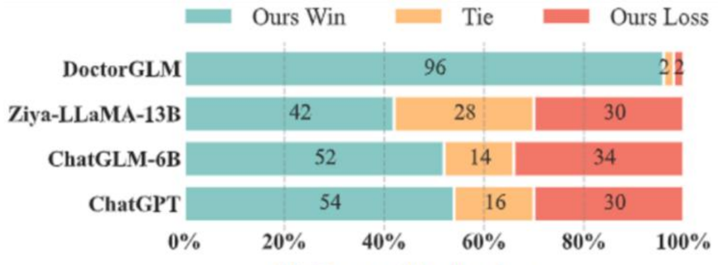
(a) GPT-4 Evaluation (b) Doctor Evaluation

Figure 4: Response Comparison of HuatuoGPT with Other Baselines on the Multi-turn Conversations.

Response Comparison of HuatuoGPT with Other Baselines on the Multi-turn Conversations.

Evaluation: Human





(b) Doctor Evaluation

Medical QA Benchmarks

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
cMedQA2	T5 (fine-tuned)	20.88	11.87	7.69	5.09	7.62	27.16	9.30	20.11	0.41	0.52
	DoctorGLM	13.51	7.10	3.72	2.00	5.11	22.78	5.68	12.22	**0.85**	0.96
	ChatGPT	19.21	7.43	3.14	1.24	5.06	20.13	3.10	12.57	0.69	**0.99**
	ChatGLM-6B	24.90	12.74	6.99	3.87	8.49	28.52	7.19	**18.21**	0.68	**0.99**
	Ziya-LLaMA-13B	27.03	13.87	7.48	4.09	7.77	28.24	7.10	14.81	0.78	0.93
	HuatuoGPT	**27.39**	**14.38**	**8.06**	**4.55**	**8.52**	**29.26**	**8.02**	15.46	0.74	0.93
webMedQA	T5 (fine-tuned)	21.42	13.79	10.06	7.38	8.94	31.00	13.85	25.78	0.37	0.46
	DoctorGLM	9.91	5.20	2.78	1.54	4.67	23.01	5.68	11.96	**0.84**	**0.95**
	ChatGPT	18.06	6.74	2.73	1.09	4.71	20.01	2.81	12.58	0.65	0.87
	ChatGLM-6B	23.42	12.10	6.73	3.83	**8.04**	**28.30**	6.87	**18.49**	0.63	0.87
	-LLaMa-13B	22.16	11.70	6.53	3.74	6.91	27.41	6.80	13.52	0.76	0.93
	HuatuoGPT	**24.85**	**13.42**	**7.72**	**4.51**	7.50	**28.30**	**7.72**	14.50	0.73	0.93
Huatuo-26M	T5 (fine-tuned)	26.63	16.74	11.77	8.46	11.38	33.21	13.26	24.85	0.51	0.68
	DoctorGLM	11.50	6.00	3.14	1.69	4.65	22.39	5.47	12.14	**0.85**	**0.96**
	ChatGPT	18.44	6.95	2.87	1.13	4.87	19.60	2.82	12.46	0.69	0.89
	ChatGLM-6B	24.46	12.75	7.20	4.13	**8.50**	28.44	7.31	**18.58**	0.67	0.89
	Ziya-LLaMA-13B	25.58	13.39	7.46	4.24	7.30	28.14	7.18	14.78	0.77	0.93
	HuatuoGPT	**27.42**	**14.84**	**8.54**	**4.96**	8.01	**29.16**	**8.29**	15.84	0.74	0.93

Benchmark on Chinese medical QA dataset

Zhang et al. 2018. Multi-scale attentive interaction networks for Chinese medical question answer selection. IEEE Access, 6:74061–74071.
He et al. 2019. Applying deep matching networks to Chinese medical question answering: a study and a dataset. BMC medical informatics and decision making, 19(2):91–100.
Li et al. 2023. Huatuo-26M, a large-scale Chinese medical QA dataset. arXiv preprint arXiv:2305.01526.

Case

User: 宝宝白眼球上怎么会有血丝呀?
ChatGPT: 宝宝白眼球上出现血丝可能是由于以下原因： 1.强烈的哭闹或咳嗽导致眼压升高，血管扩张引起眼球上出现血丝。 2.感染，例如结膜炎、角膜炎、沙眼等。这些疾病会导致眼睛红肿、分泌物增多、疼痛等症状。 3.眼部伤害导致的瘀血。 4.干眼症。长时间使用电子产品、良久未闭眼或眨眼等情况，可能引起眼睛干涩，导致血管扩张。 如果发现宝宝眼球上出现血丝，建议观察一到两天，如果情况持续或加重，或伴随其他症状，如视力下降、眼睛疼痛等，应及时就医。同时，保持眼睛卫生，避免眼部受伤或过度疲劳也很重要。
HuatuoGPT： 亲爱的家长，您好！感谢您向我咨询这个问题。根据您描述的情况，宝宝的眼睛里面出现了血丝，可能是由多种原因引起的。其中最常见的原因是结膜炎或角膜炎等眼部疾病，也可能与过敏性鼻炎等呼吸道疾病有关。 为了更好地帮助您，我想请问一些相关问题。<u>首先，宝宝出生后是否经常接触灰尘、花粉、动物毛发等过敏源？其次，宝宝的睡眠质量如何？饮食习惯是否健康均衡？还有没有其他不适症状，比如咳嗽、流涕等？</u> 在得到更多信息后，我将尽力提供更具体的建议和治疗方案。同时，建议您及时带宝宝去正规医院就诊，以便医生进行全面检查和诊断，并制定合理的治疗计划。希望宝宝早日康复!

What HuatuoGPT can do!

你好，我是一个解答医疗健康问题的大模型，目前处于测试阶段，请以医嘱为准。请问有什么可以帮到您?
Hello, I am a large medical language model, please refer to the doctor's advice in any cases because it is currently in the testing stage. How can I help you?

请问下面这个文本是哪个科室的医疗记录:
<文本>:
检查:腹部和骨盆的CT扫描，不加或加静脉造影剂。临床适应症:左下腹腹痛。比较:无。腹部和骨盆的CT扫描分别进行了静脉造影剂和无静脉造影剂。静脉注射Isovue共100ml。同时给予口服造影剂。肺底清晰。肝脏增大，衰减减小。肝脏无局灶性肿块。没有肝内或肝外导管扩张。胆囊微胀。肾上腺、胰腺、脾脏和左肾正常。右肾下极见一12mm单纯性囊肿。没有肾积水或输尿管积水。阑尾正常。直肠乙状结肠内有多个憩室。乙状结肠局灶性壁增厚(图69)，伴有憩室附近的脂肪堆积。这些发现与憩室炎一致。未发现气腹。没有腹水或局灶性积液。主动脉的轮廓和口径正常。无腺病。腰椎出现退行性改变。印象:表现符合憩室炎。请参阅上面的报告。胃肠病学，肝外导管扩张，胆囊，腺体，胰腺，脾脏，肾脏，肾上腺，腹部和骨盆，ct扫描，静脉注射，腹部，

上述文本属于消化内科医疗记录。

What HuatuoGPT can do!

What HuatuoGPT can do!

你好，我是一个解答医疗健康问题的大模型，目前处于测试阶段，请以医嘱为准。请问有什么可以帮到您?
Hello, I am a large medical language model, please refer to the doctor's advice in any cases because it is currently in the testing stage. How can I help you?

在ICD-10中，左侧腹直肌拉伤通常代表什么?

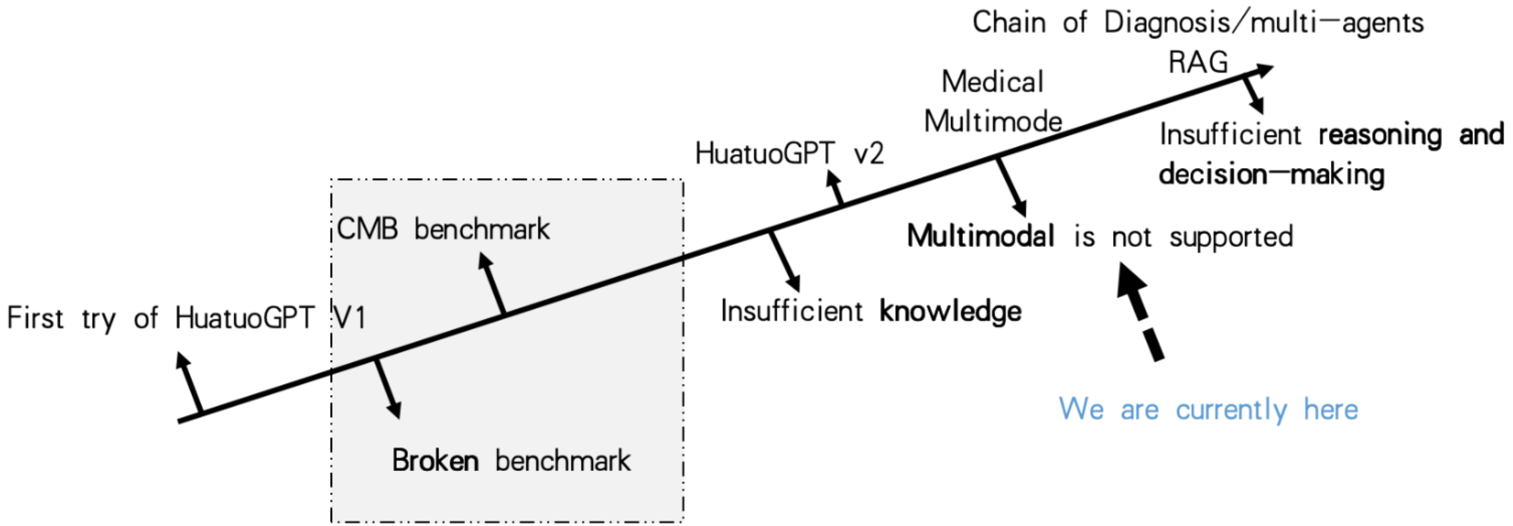
左侧腹直肌拉伤在 ICD-10 中被归类为 “损伤” 类别中的 “肌肉组织损伤” 中的 “肌肉撕裂” 或 “肌肉断裂” 类型。

What HuatuoGPT cannot do!

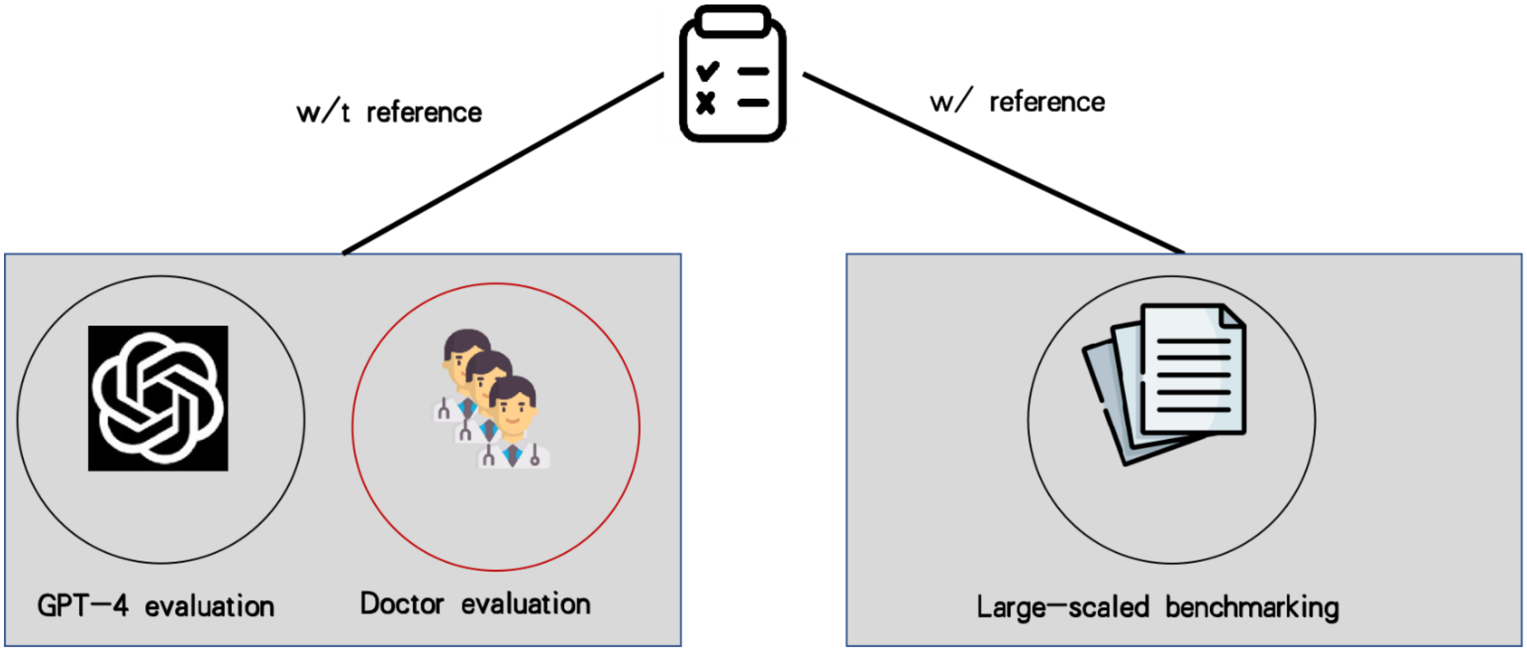
It is sometimes wrong in some medical knowledge -intensive questions!

Knowledge evaluation: CMB

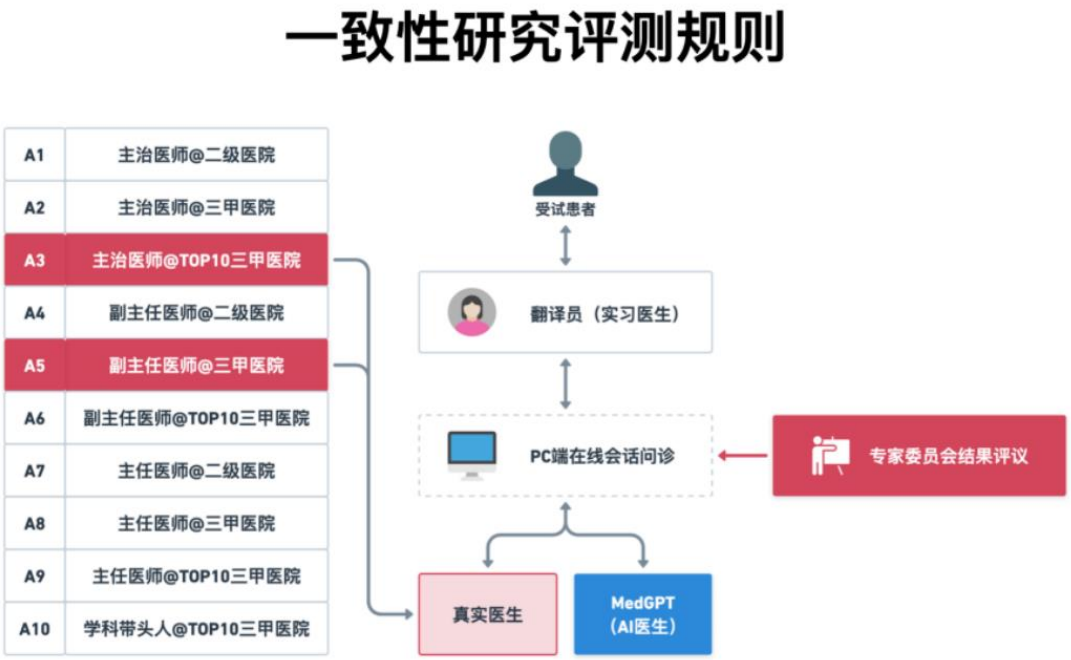
Roadmap of Our Practice



Evaluation


w/t reference
w/ reference
GPT-4 evaluation
Doctor evaluation
Large-scaled benchmarking

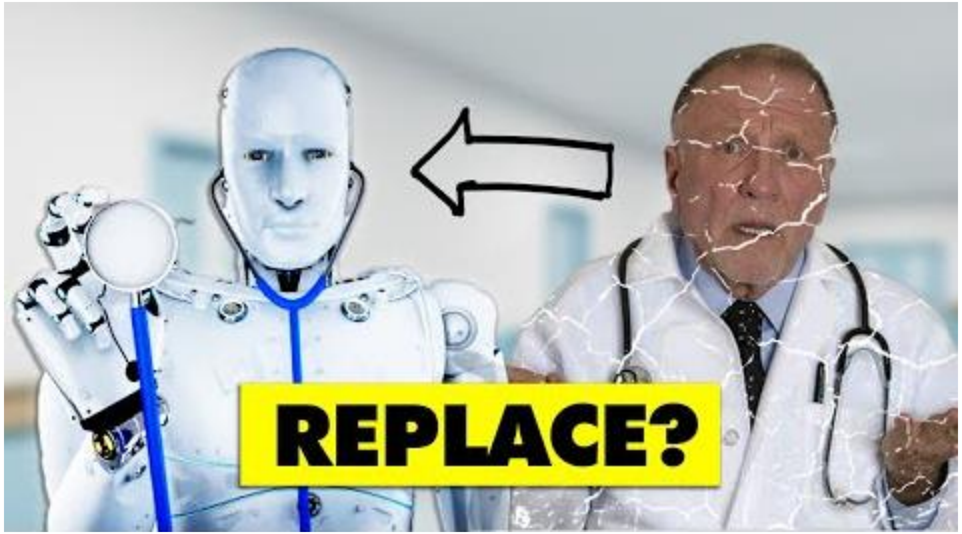
Evaluation of medGPT (医联)



Evaluation of medGPT

Key Limitations of MedGPT evaluation

Doctors are not competitors for LLM, they could help LLMs and also benefit from doctor

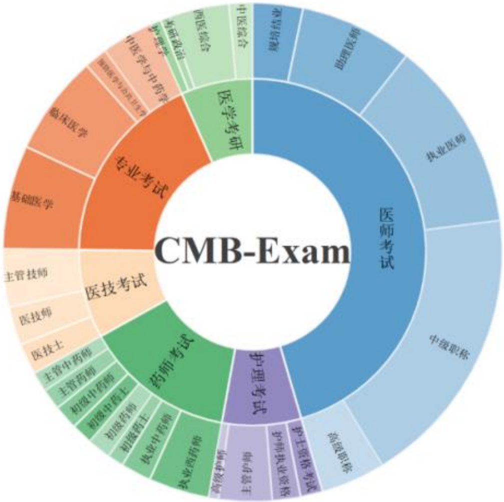


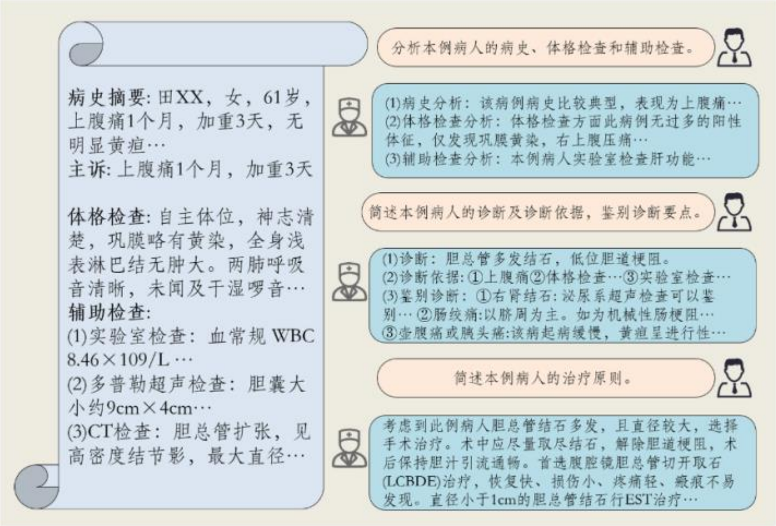
Doctors cannot be replaced, and will not be in the future.

Other Limitations

- **Multimodal context missed**: Diagnosis of Doctors are based on multimodal content, instead of being text-only.
- **Poor reproducibility**: cannot be reused for the next assessment, poor reproducibility.
- **Benchmark lottery:** it is subjective to select these evaluation aspects that might results in the benchmark lottery.
- **Doctor Labors** are "expensive"

Comprehensive Medical benchmark (CMB)





https://cmedbenchmark.llmzoo.com/static/leaderboard.html

Overview

Category	Subcategory	# Subject	# Questions
Physician (医师)	Resident Physician (住院医师); Licensed Assistant Physician (执业助理医师); Licensed Physician (执业医师); Associate Professional Physician (中级职称); Advanced Professional Physicians (高级职称)	81	124,926
Nurse (护理)	Practicing Nurse (护士); Licensed Practical Nurse (护师); Charge Nurse (主管护师); Advanced Practice Nurse (高级护师)	8	16,919
Technicians (医技)	Medical Technician (医技士); Medical Technologist (医技师); Supervising Technologist (主管技师)	21	27,004
Pharmacist (药师)	Licensed Pharmacist (执业西药师); Licensed TCM Pharmacist (执业中药师); Junior Pharmacist (初级药师); Junior Pharmacist Assistant (初级药士); Junior TCM Pharmacist (初级中药师); Junior TCM Pharmacist Assistant (初级中药士); Chief Pharmacists (主管药师); Chief TCM Pharmacists (主管中药师)	8	33,354
Undergraduate Disciplines (学科考试)[1]	Fundamental Medicine (基础医学); Clinical Medicine (临床医学); Traditional Chinese (TCM) and Chinese Herbal Medicine (中医学与中药学); Preventive Medicine and Public Health (预防医学与公共卫生学)	53	62,271
Graduate Entrance Exam (考研)	Integrated Western Medicine (西医综合); Integrated TCM (中医综合); Political Science (政治); Nursing (护理学)	5	16,365
Total	28	176	280,839

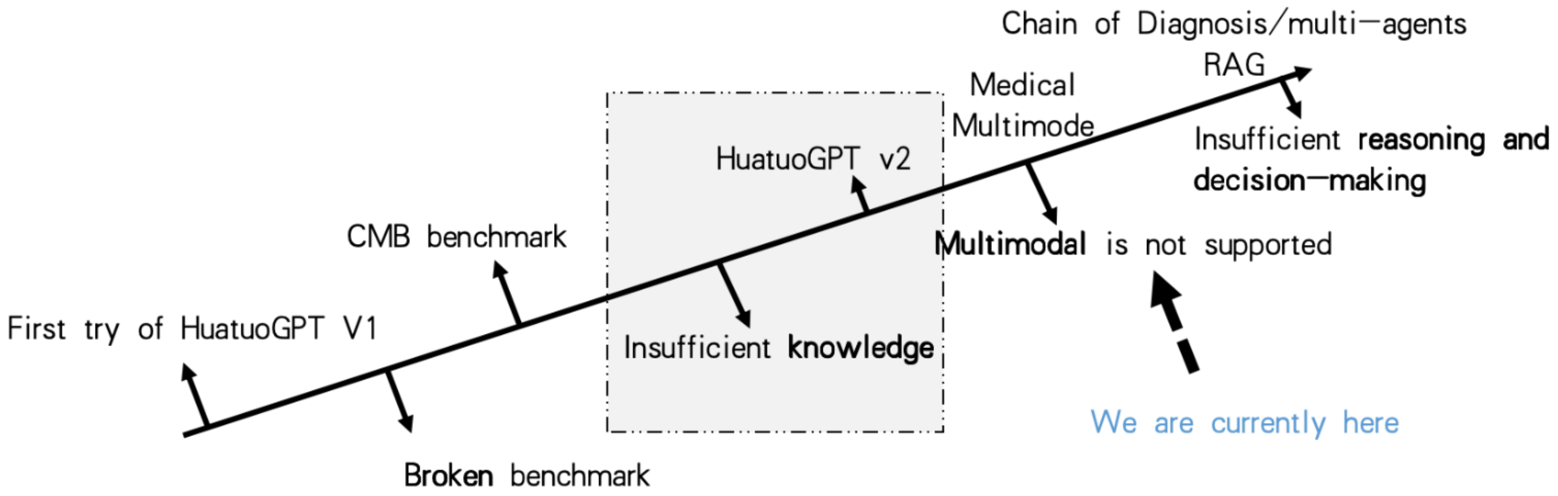
[1] We referenced the National Standard Subject Classification of the People's Republic of China, see https://xkb.pku.edu.cn/docs/2018-10/20220328083301969071.pdf.

Table 1: Statistics of the CMB-Exam Categories, Subcategories, Subjects, and Questions.

Results

Model	中医						临床						药师					平均
	执业助理医师			执业医师			执业助理医师			执业医师			执业中药师			执业西药师		
	2015年	2016年	2017年	2012年	2013年	2016年	2018年	2019年	2020年	2018年	2019年	2020年	2017年	2018年	2019年	2021年	2022年	
GPT-4	47.33	48.18	53.45	50.33	**53.67**	**54.19**	**79.91**	**72.54**	**70.90**	**74.83**	**73.11**	**68.35**	41.07	43.71	47.98	**58.33**	**60.00**	**58.70**
HuatuoGPT-II(华佗-II)	**54.33**	**50.45**	**60.34**	**53.00**	49.67	53.02	61.54	54.10	54.92	53.67	50.21	50.46	**50.60**	**47.31**	**52.02**	49.58	46.88	52.48
ChatGLM2-6B	47.00	50.45	43.97	44.83	48.33	48.14	62.39	49.18	52.87	50.11	48.53	45.64	36.90	36.53	38.57	34.79	35.62	45.52
Baichuan-7B-chat	42.00	44.09	43.10	39.67	43.33	41.16	54.70	48.36	44.26	42.76	42.65	39.45	35.71	38.32	30.49	39.79	40.83	41.80
ChatGPT	39.33	35.00	38.79	35.33	39.33	39.77	59.40	49.18	47.95	51.00	46.22	47.94	23.81	30.54	31.39	40.00	37.29	40.72
Spark(星火)	43.33	38.18	39.66	38.33	38.17	42.56	54.70	44.67	50.00	51.45	44.75	40.83	29.17	26.35	30.94	35.00	34.38	40.15
HuatuoGPT(华佗)	26.00	30.91	32.76	31.33	26.83	30.23	32.48	33.61	27.46	32.74	28.57	30.05	18.45	27.54	25.11	26.04	27.08	28.66
MedicalGPT	25.00	24.09	21.55	26.33	27.00	26.98	38.89	29.10	28.28	33.41	32.14	26.15	22.62	20.96	20.63	24.79	24.79	26.63
ChatGLM-Med	20.67	23.18	20.69	21.83	21.83	22.56	30.34	22.95	29.92	18.49	20.38	24.77	16.67	20.96	15.25	20.83	22.50	21.99
BenTsao(本草)	23.33	26.82	17.24	19.00	19.50	22.09	18.38	24.59	27.46	21.60	18.91	18.12	20.24	20.36	18.83	18.96	22.08	21.03
ChatMed-Consult	20.00	17.27	14.66	21.33	18.17	20.00	27.78	21.31	19.26	23.83	21.43	19.50	16.67	14.97	17.49	22.29	24.17	20.01
BianQue-2(扁鹊)	3.67	2.27	3.45	4.17	4.50	3.95	4.70	5.33	1.64	3.79	5.88	3.67	7.14	4.19	8.97	4.58	4.38	4.49
DoctorGLM	3.00	1.36	3.45	1.83	1.83	2.09	2.56	3.28	1.64	1.56	2.73	1.83	1.79	2.40	2.24	3.12	3.75	2.38

Roadmap of our practice



HuatuoGPT V2

HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs

New version: https://v2-34b.huatuogpt.cn/

Motivation

Large language models typically underperform in the medical field because this area of expertise is particularly challenging to master through standard extensive text pre-training methods. Consequently, the issue of **inaccuracies or "medical hallucination"** within these models is a critical concern that demands immediate attention.

Even GPT-4 experiences significant hallucinatory issues within the medical field.

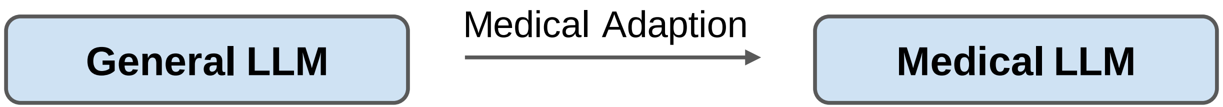
鸿博

复方甲氧那明的成分

复方甲氧那明是一种中药复方制剂，它的成分可以因制剂不同而有所变化。一般情况下，复方甲氧那明的主要成分包括甲氧那明（Methoxamine）和其他辅助成分，如赤芍、酮康唑、维生素B1等。复方甲氧那明通常用于治疗高血压和其他循环系统相关疾病。

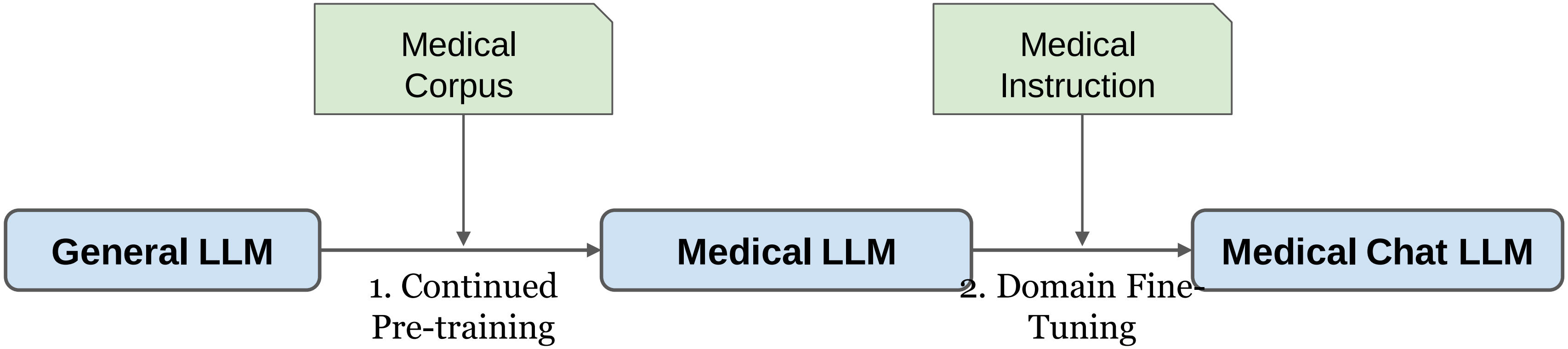
Motivation

How can we infuse medical professional expertise to endow general large language models (LLMs) with expert-level medical capabilities?



Motivation

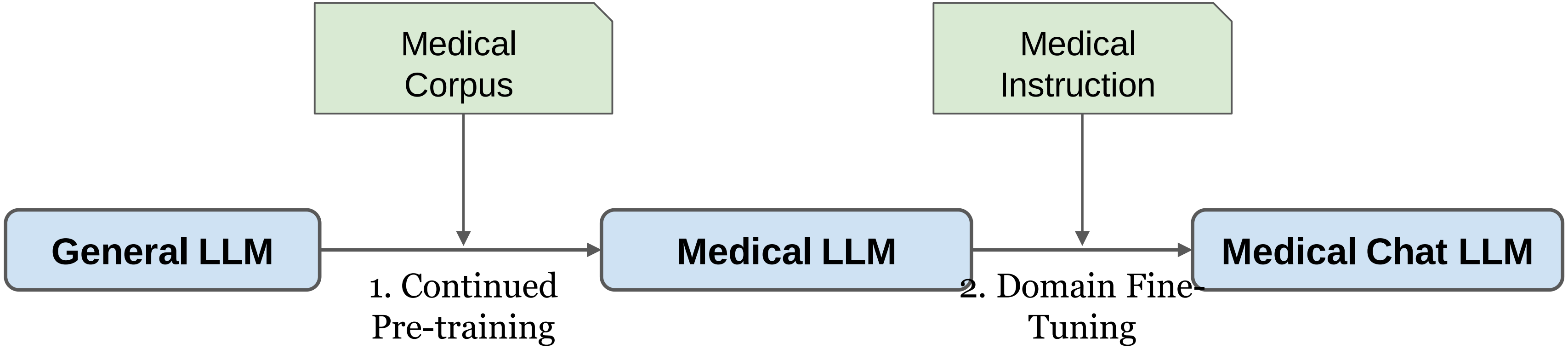
Conventional Two-stage Adaption



1. A Extra Model Shift. The continued pretraining phase uses a specialized corpus that is different from the broad corpus of the original model. This can lead to a notable change in the model's data distribution. Furthermore, the method of continued unsupervised pretraining is quite different from the later supervised fine-tuning (SFT), introducing two separate shifts in data distribution. These shifts may result in catastrophic forgetting or reduced performance, as pointed out in Paper A, where it's noted that some prompt capabilities are lost compared to a single-stage fine-tuning approach.

Motivation

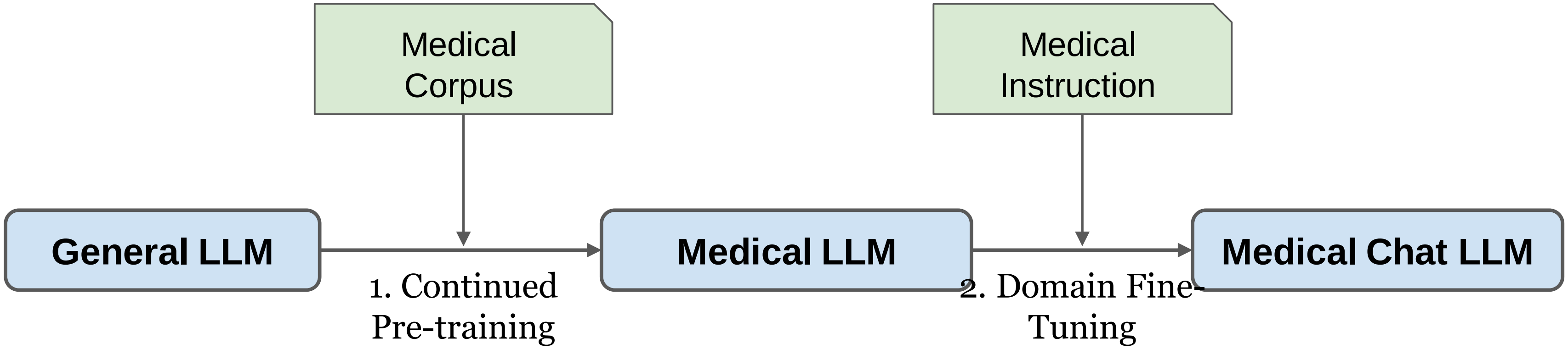
Conventional Two-stage Adaption



2. Data Discrepancy. The diverse sources of domain pre-training data present genre and language inconsistencies, which can affect data quality through punctuation mistakes and ethical issues. These sources contrast sharply with the concise, formal, and monolingual data usually employed for fine-tuning. Without considering the risk of domain data contamination in LLMs, it's debatable if specialized pre-training data truly enhances the knowledge base and aids the successive supervised fine-tuning (SFT) phase. Giving a conclusive response to this is difficult.

Motivation

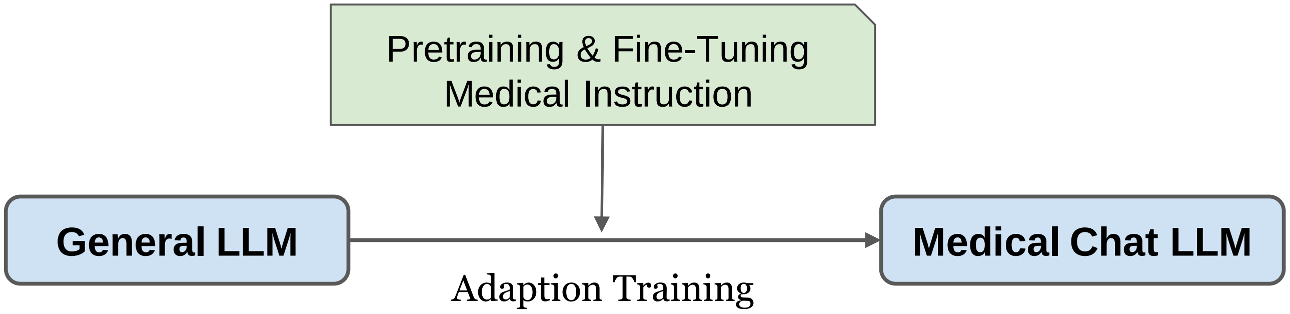
Conventional Two-stage Adaption



3. Increased Complexity. Domain adaptation through a two-step approach, which includes both continued pre-training and supervised fine-tuning (SFT), adds complexity. It requires distinct stages of data preparation, programming, and tuning, with each needing thorough validation. This not only increases the workload but also complicates the prediction of continued pre-training's impact on SFT, making the process more burdensome and resource-heavy.

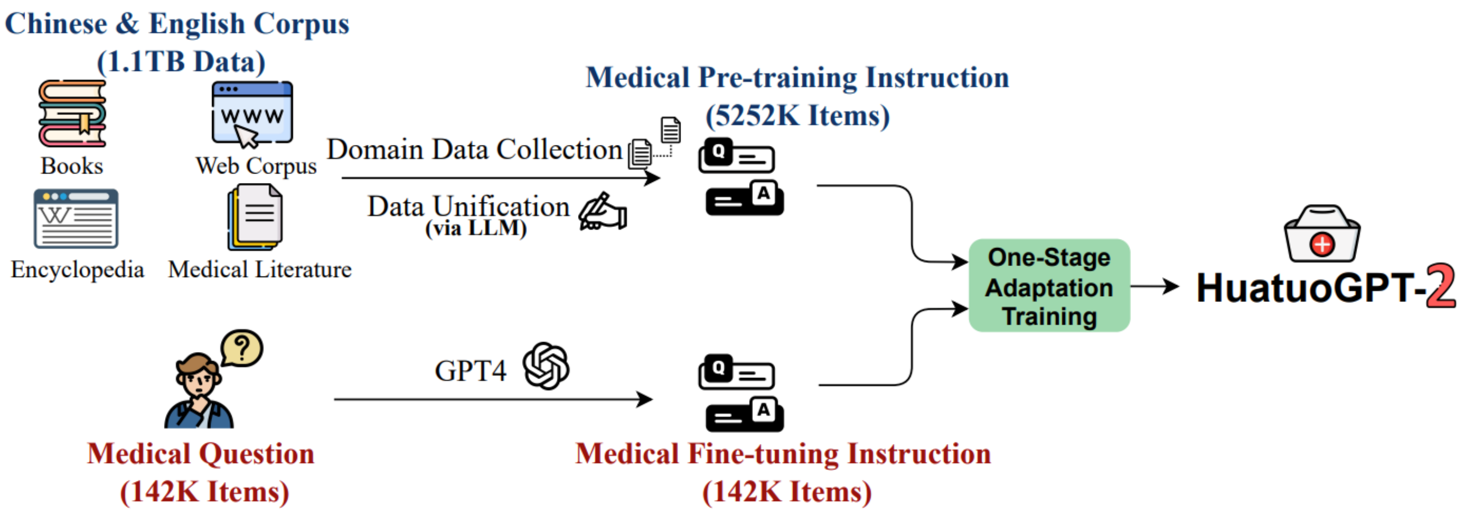
Motivation

One-stage Adaption



To tackle the two above issues, we propose a simper protocol of domain adaption that unifies the two stages (further pre-training and supervised fine-tuning) in a single stage.

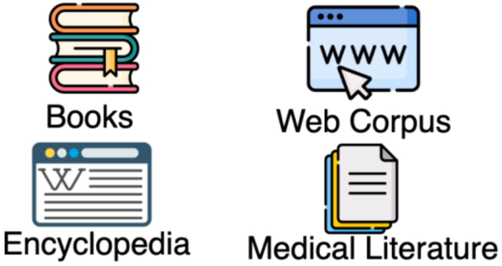
One-Stage Adapation



Through a series of manipulations, we transform the pre-training data to match the fine-tuning data in language, genre, and training goals. By combining the pre-trained and fine-tuned datasets, we implement a one-stage training strategy to fully advance the development of HuatuoGPT-II.

Data Collection

Data Source



cn&en Pretrained Data
(100B tokens)

Data Pipline

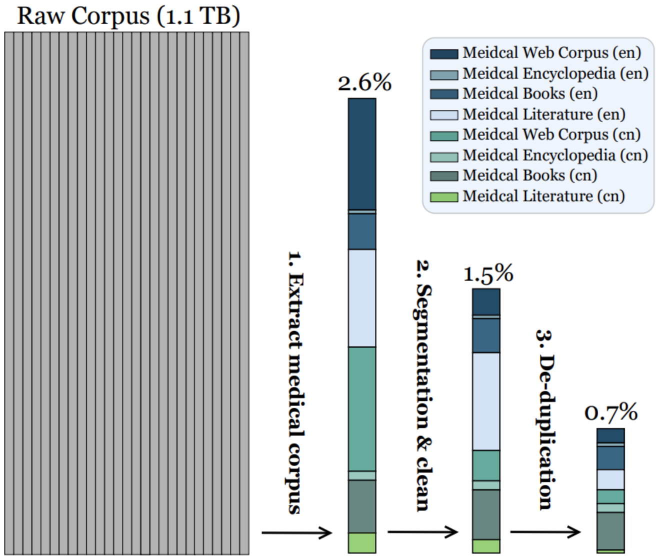
Original corpus → Extract medical corpus → Clean → Segmentation → Deduplication

100B → 1B

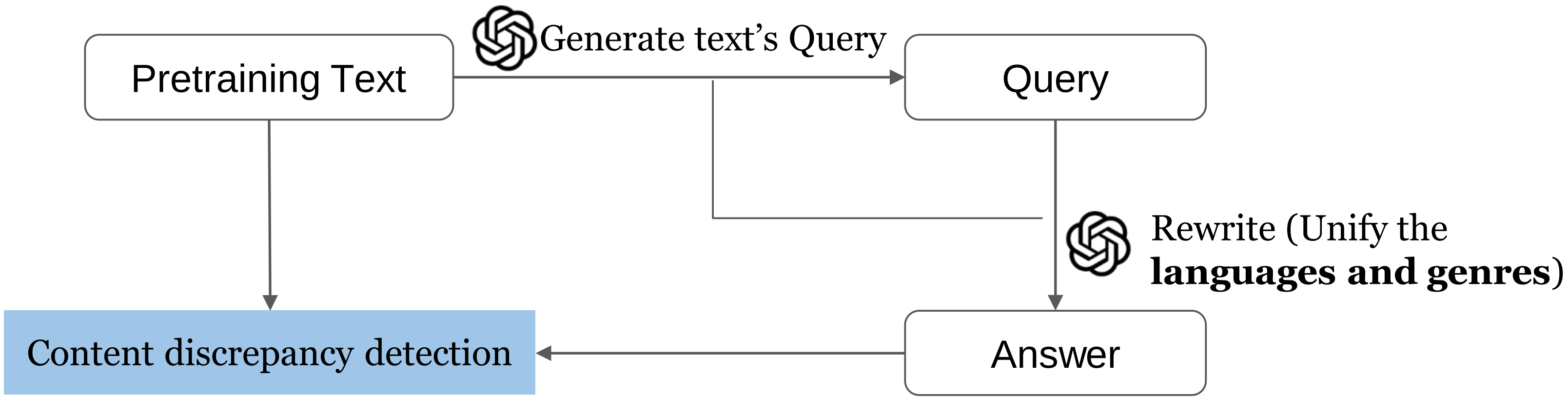
Data Collection

Data Pipline

Original corpus → Extract medical corpus → Clean → Segmentation → Deduplication



Data Unification



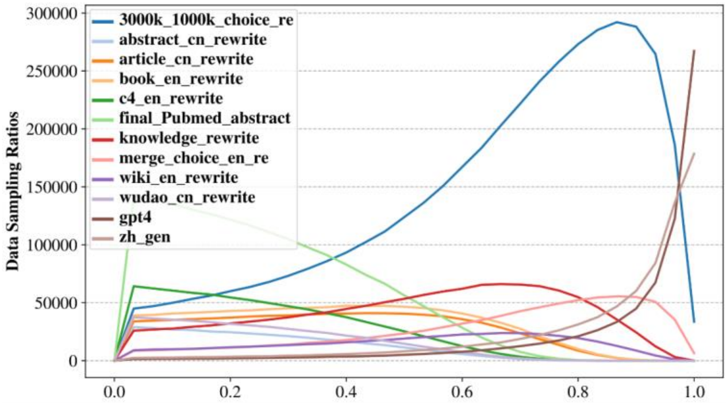
We employ LLM to generate questions that are answerable based on pretraining text. We then rewrite this text to align with the desired answers, spanning various languages and genres. Additionally, we assess whether the answer corresponds to the specific location within the pretraining text.

Data Unification

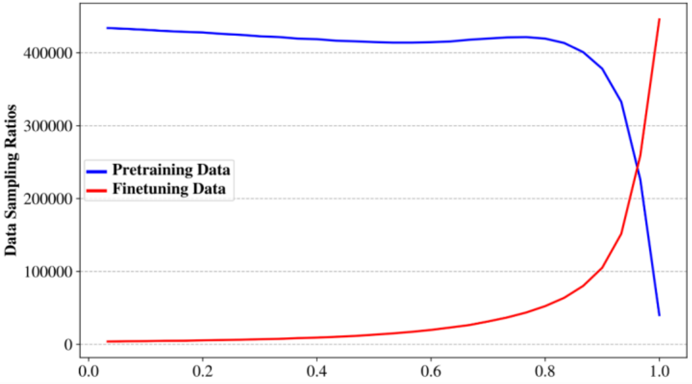
Language	Type of Data	# Data	Source
Chinese	Web Corpus	640,621	Wudao, Common Crawl
	Books	1,835,931	Textbook
	Encyclopedia	411,183	Online Encyclopedia [1]
	Literature	177,261	Chinese Literature
English	Web Corpus	394,490	C4
	Books	719,187	Textbook, the_pile_books3
	Encyclopedia	147,059	Wikipida
	Literature	878,241	PubMed

Table 1: Summary of the Pretrained Data

One-Stage Training



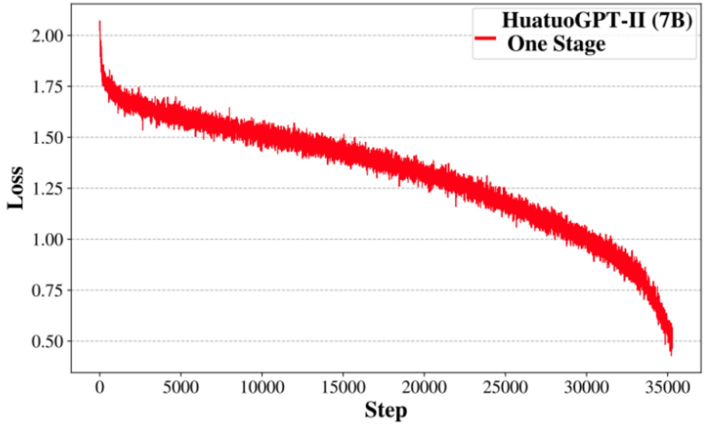
Data Sampling



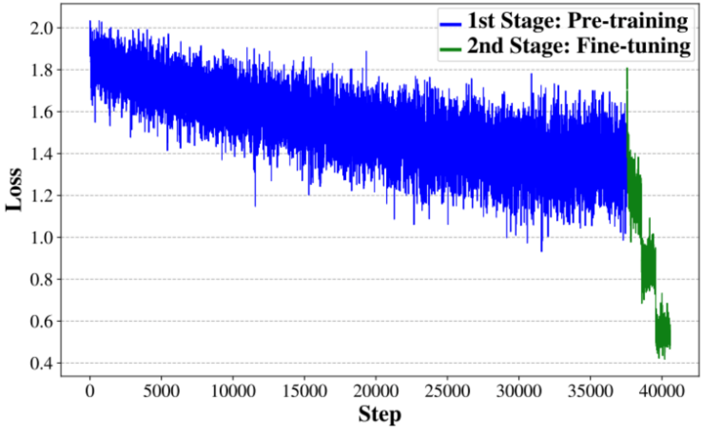
Integration of two-type data

After unifying the data, we fuse the data for joint training. A unified data sampling strategy was established to allow the model to gradually transition from knowledge learning to fine-tuning for data adaptation. In addition, we standardized all hyperparameters, choosing a moderate batch_size and learning rate.

One-Stage Training



Loss within One Stage



Loss within Two Stage

During the one-stage training, a blend of pre-training and fine-tuning data is utilized, with an established protocol for sampling priorities and iterations. Key hyperparameters, including batch size and learning rate, are harmonized to facilitate a more seamless adaptation. This strategy has been empirically validated to enhance learning outcomes.

Results

Medical Benchmarks

Model	**MedQA**	**MedMCQA**	**CMB**	**CMExam**	**MMLU**♢	**CMMLU**♢	**C_Eval**♢
DISC-MedLLM	28.67	-	32.47	36.62	-	-	-
HuatuoGPT	25.77	31.20	28.81	31.07	34.91	33.23	36.53
Llama2-7B-Chat	30.79	37.32	24.81	25.05	45.68	31.02	26.40
Llama2-13B-Chat	36.68	39.66	26.51	26.60	50.67	32.64	31.47
ChatGLM2-6B	24.98	33.33	42.41	43.55	40.94	43.87	47.20
ChatGLM3-6B	28.75	35.91	39.81	43.21	47.21	46.97	48.80
Baichuan2-7B-Chat	33.31	38.90	46.33	50.48	50.29	50.74	51.47
Baichuan2-13B-Chat	39.43	41.86	50.87	54.90	56.31	52.95	58.67
Qwen-7B-Chat	33.46	41.36	49.39	53.33	53.88	54.65	52.80
Qwen-14B-Chat	42.81	46.59	60.28	63.57	61.69	**64.55**	**65.07**
ChatGPT (API)	**52.24**	**53.60**	43.26	46.51	**69.96**	50.37	48.80
HuatuoGPT-II (7B)	41.13	41.87	60.39	65.81	51.44	59.08	62.40
HuatuoGPT-II (13B)	45.68	47.41	**63.34**	**68.98**	54.00	61.45	64.00

Results

The United States Medical Licensing Examination (USMLE)

Models	Stage1 (6308)	Stage2&3 (5148)	Average Score (11456)
HuatuoGPT	28.68	28.38	28.54
ChatGLM2-6B	31.72	32.45	32.05
ChatGLM3-6B	33.39	32.49	32.98
Baichuan2-7B-Chat	38.11	37.32	37.76
Baichuan2-13B-Chat	44.99	45.57	45.25
Qwen-7B-Chat	40.90	36.09	38.73
Qwen-14B-Chat	48.73	42.45	45.90
Llama2-7B-Chat	35.18	35.89	35.50
Llama2-13B-Chat	40.55	39.73	40.18
ChatGPT (API)	**57.04**	**56.27**	**56.69**
HuatuoGPT-II(7B)	45.72	44.45	45.15
HuatuoGPT-II(13B)	47.34	49.37	48.25

Chinese Medical Licensing Examination (MCMLE)

Model	Traditional Chinese Medicine (中医)									Clinical (临床)						Avg.
	Assistant (执业助理医师)			Physician (执业医师)			Pharmacist (药师)			Assistant (执业助理医师)			Physician (执业医师)			
	2015 (300)	2016 (220)	2017 (116)	2012 (600)	2013 (600)	2016 (430)	2017 (168)	2018 (167)	2019 (223)	2018 (234)	2019 (244)	2020 (244)	2018 (449)	2019 (476)	2020 (436)	
HuatuoGPT	26.0	30.9	32.8	31.3	26.8	30.2	18.4	27.5	25.1	32.5	33.6	27.5	32.7	28.6	30.1	28.9
DISC-MedLLM	38.3	41.8	26.7	36.2	38.7	35.1	25.0	22.2	22.0	49.2	36.1	41.8	41.4	36.6	35.8	35.1
ChatGLM2-6B	47.0	50.5	44.0	44.8	48.3	48.1	36.9	36.5	38.6	62.4	49.2	52.9	50.1	48.5	45.6	46.9
ChatGLM3-6B	45.7	45.0	50.0	46.3	45.8	46.5	42.3	28.1	35.4	50.9	48.8	43.9	41.7	43.9	43.6	43.9
Llama2-7B-Chat	24.0	22.3	23.3	27.2	26.0	22.3	17.3	23.4	20.6	29.1	29.9	27.9	25.8	27.3	25.2	24.7
Llama2-13B-Chat	23.0	26.8	24.1	27.0	27.8	25.4	21.4	27.0	21.1	29.9	29.5	27.5	26.5	29.0	25.7	26.1
Baichuan2-7B-Chat	57.3	57.3	58.6	55.7	58.5	57.9	41.7	41.9	45.7	61.1	55.7	55.3	51.0	53.6	50.0	53.4
Baichuan2-13B-Chat	64.7	58.2	62.9	61.7	61.5	63.3	54.2	38.9	48.4	66.2	64.8	63.1	65.9	58.8	61.5	59.6
Qwen-7B-Chat	54.7	55.9	56.0	52.7	53.5	54.4	44.0	33.5	43.0	68.8	63.9	57.8	60.6	57.6	54.1	54.0
Qwen-14B-Chat	65.3	63.2	67.2	64.8	63.3	67.9	54.8	49.1	52.5	74.8	75.0	69.3	73.7	69.7	68.8	65.3
ERNIE Bot (文心一言)	**73.3**	66.3	**73.3**	70.0	**71.8**	66.7	55.9	50.3	**60.0**	78.2	**77.0**	**77.5**	66.6	70.8	**74.1**	**68.8**
ChatGPT (API)	46.0	36.4	41.4	36.7	38.5	40.5	32.1	28.1	30.0	63.3	57.8	53.7	53.7	52.5	51.8	44.2
GPT-4 (API)	47.3	48.2	53.5	50.3	53.7	54.2	41.1	43.7	48.0	**79.9**	72.5	70.9	**74.8**	**73.1**	68.4	58.6
HuatuoGPT-II (7B)	67.1	65.2	67.5	67.9	67.4	64.9	53.0	46.7	51.0	70.9	73.2	69.4	68.8	66.5	67.7	64.5
HuatuoGPT-II (13B)	70.3	**70.0**	71.6	**71.0**	69.2	**70.2**	**56.5**	**52.1**	54.7	73.1	76.6	70.1	72.8	68.9	72.2	68.0

Results

Automatic Evaluation

HuatuoGPT-II Win Rate	Win	Tie	Fail
HuatuoGPT-II(7B) vs HuatuoGPT-II(13B)	39	22	39
HuatuoGPT-II(7B) vs GPT4	58	21	21
HuatuoGPT-II(7B) vs ChatGPT	62	18	20
HuatuoGPT-II(7B) vs Baichuan2-13B-Chat	64	14	22
HuatuoGPT-II(7B) vs ChatGLM3-6B	75	11	14
HuatuoGPT-II(7B) vs Baichuan2-7B-Chat	75	7	18
HuatuoGPT-II(7B) vs DISC-MedLLM	80	8	12
HuatuoGPT-II(7B) vs Qwen-14B-Chat	82	7	6
HuatuoGPT-II(7B) vs HuatuoGPT	87	7	6
HuatuoGPT-II(7B) vs Qwen-7B-Chat	89	6	5

Single-round Medical Question

HuatuoGPT-II Win Rate	Win	Tie	Fail
HuatuoGPT-II(7B) vs HuatuoGPT-II(13B)	41	13	46
HuatuoGPT-II(7B) vs GPT4	62	15	23
HuatuoGPT-II(7B) vs HuatuoGPT	67	15	18
HuatuoGPT-II(7B) vs ChatGPT	69	14	17
HuatuoGPT-II(7B) vs DISC-MedLLM	73	15	12
HuatuoGPT-II(7B) vs Qwen-7B-Chat	75	12	13
HuatuoGPT-II(7B) vs Baichuan2-13B-Chat	75	11	14
HuatuoGPT-II(7B) vs ChatGLM3-6B	76	10	14
HuatuoGPT-II(7B) vs Baichuan2-7B-Chat	84	7	9
HuatuoGPT-II(7B) vs Qwen-14B-Chat	79	9	12

Multi-round Medical Dialogue

Results

CMB-Clin

Model	Fluency	Relevance	Completeness	Proficiency	Avg.↓
GPT-4	4.95	4.71	4.35	4.66	4.67
HuatuoGPT-II (7B)	4.94	4.56	4.24	4.46	4.55
文心一言	4.92	4.53	4.16	4.55	4.54
ChatGPT	4.97	4.49	4.12	4.53	4.53
HuatuoGPT-II (13B)	4.92	4.38	4.00	4.40	4.43
Baichuan2-7B-Chat	4.93	4.41	4.03	4.36	4.43
Qwen-14B-Chat	4.90	4.35	3.93	4.48	4.41
Qwen-7B-Chat	4.94	4.17	3.67	4.33	4.28
Baichuan2-13B-Chat	4.88	4.18	3.78	4.27	4.28
ChatGLM3-6B	4.92	4.11	3.74	4.23	4.25
ChatGLM2-6B	4.89	3.97	3.72	4.22	4.20
HuatuoGPT	4.89	3.76	3.38	3.86	3.97
DISC-MedLLM	4.82	3.24	2.75	3.51	3.58

Results

Expert Evaluation

HuatuoGPT-II Win Rate	Win	Tie	Fail
HuatuoGPT-II(7B) vs GPT-4	**38**	38	24
HuatuoGPT-II(7B) vs ChatGPT	**52**	33	15
HuatuoGPT-II(7B) vs Baichuan2-13B-Chat	**63**	19	18
HuatuoGPT-II(7B) vs HuatuoGPT	**81**	11	8

Single-round Medical Question

HuatuoGPT-II Win Rate	Win	Tie	Fail
HuatuoGPT-II(7B) vs GPT-4	**53**	17	30
HuatuoGPT-II(7B) vs ChatGPT	**56**	11	33
HuatuoGPT-II(7B) vs Baichuan2-13B-Chat	**63**	19	18
HuatuoGPT-II(7B) vs HuatuoGPT	**68**	6	26

Multi-round Medical Dialogue

Results

Model	2023 Pharmacist Licensure Examination (Pharmacy) Optimal Choice	Matched Selection	Integrated Analysis	Multiple Choice	Total Score	2023 Pharmacist Licensure Examination (TCM) Optimal Choice	Matched Selection	Integrated Analysis	Multiple Choice	Total Score	AVG
DISC-MedLLM	22.2	26.8	23.3	0.0	22.6	24.4	32.3	15.0	0.0	24.9	23.8
HuatuoGPT	25.6	25.5	23.3	2.6	23.4	24.1	26.8	31.6	7.5	24.9	24.2
ChatGLM2-6B	37.0	36.8	25.0	31.7	35.3	33.1	37.3	35.0	37.3	33.7	34.5
ChatGLM3-6B	39.5	39.1	10.5	0.2	34.6	31.8	38.2	25.0	20.0	32.9	33.8
Qwen-7B-chat	43.8	46.8	33.3	18.4	41.9	40.0	43.2	33.3	17.5	38.8	40.4
Qwen-14B-chat	56.2	58.6	41.7	21.1	52.7	51.3	51.0	27.5	41.7	47.9	50.3
Biachuan2-7B-Chat	51.2	50.9	30.0	2.6	44.6	48.1	46.0	35.0	7.5	42.1	43.4
Biachuan2-13B-Chat	43.8	52.7	36.7	7.9	44.2	41.3	46.4	43.3	15.0	41.7	43.0
文心一言	45.0	60.9	36.7	23.7	49.6	53.8	59.1	38.3	20.0	51.5	50.6
ChatGPT(API)	45.6	44.1	36.7	13.2	41.2	34.4	32.3	30.0	15.0	31.2	36.2
GPT-4(API)	65.1	59.6	46.7	15.8	**57.3**	40.6	42.7	33.3	17.5	38.8	48.1
HuatuoGPT-II(7B)	41.9	61.0	35.0	15.7	47.7	52.5	51.4	41.7	15.0	47.5	47.6
HuatuoGPT-II(13B)	47.5	64.1	45.0	23.7	52.9	48.8	61.8	45.0	17.5	**51.6**	**52.3**
HuatuoGPT-II(34B)	66.3	75.0	48.3	34.2	65.5	63.6	71.4	50.0	27.5	**62.5**	**64.0**

In the interest of fairness, we have gathered the exam questions from the 2023 China National Pharmacist Licensing Examination, which began on October 21, 2023. This date is earlier than both the release of our assessment model and the finalization of data collection by HuatuoGPT-II on October 7, 2023.

One Stage Vs. Two Stages

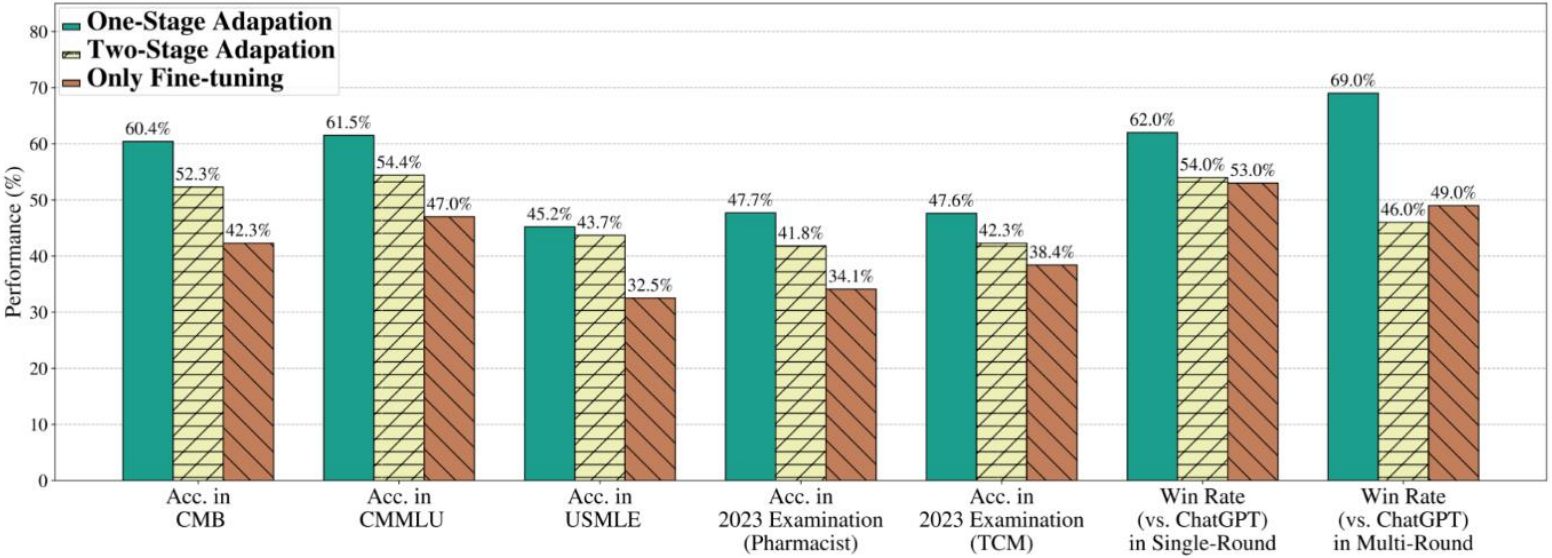


Figure 8: The comparison results of One-Stage Adaptation and Two-Stage Adaptation. "Only Fine-tuning" refers to the model that only fine-tunes the backbone directly. The evaluation methods and datasets mentioned earlier are adopted here, where Win Rate is the result scored by automatic evaluation using GPT-4.

Cases

User:
复方甲氧那明的成分

GPT4:
复方甲氧那明是一种中药复方制剂，它的成分可以因制剂不同而有所变化。一般情况下，复方甲氧那明的主要成分包括甲氧那明（Methoxamine）和其他辅助成分，如赤芍、酮康唑、维生素B1等。复方甲氧那明通常用于治疗高血压和其他循环系统相关疾病。

HuatuoGPT-II:
复方甲氧那明是一种复方制剂，用于治疗支气管哮喘、喘息性支气管炎及其他过敏性疾病。它的主成分包括盐酸甲氧那明、那可丁、氨茶碱和马来酸绿茶那敏。这种药物通过抑制支气管痉挛和咳嗽反应，爰解哮喘和喘息性支气管炎的症状。复方甲氧那明的用法用量是口服，一般成人每次2片，日3次，饭后服用。老年患者或有肝功能损害的患者应该慎用，并且需要根据病情适当调整剂量和服药次数用。此外，有哮喘发作或严重心血管疾病的患者也不然而，复方甲氧那明也有一些禁忌症。首先，那可丁过者不适合使用。其次，孕妇和哺乳期妇女也应该避免使适合使用。
在使用复方甲氧那明时，可能会出现一些不良反应。例偶尔会出现皮疹、皮肤发红、瘙痒、恶心、呕吐、食欲不振、眩晕、心悸和排尿困难等。如果出现这些不口、良反应，应及时停药并咨询医生。
医生的指导和药品说明书的建议进行用药，并注意禁总之，复方甲氧那明是一种用于治疗支气管哮喘、喘息主支气管炎及其他过敏性疾病的药物。在使用时，应按照忌症和不良反应的可能性。

Drug Description

【药品名称】
通 用 名 称：复方甲氧那明胶囊
商 品 名 称：阿斯美®
英 文 名 称：Compound Methoxyphenamine Capsules
汉 语 拼 音：Fufangjiayangnaming Jiaonang

【成　　份】本品为复方制剂，其组份为（每粒胶囊中含）：

盐酸甲氧那明	12.5mg
那可丁	7 mg
氨茶碱	25 mg
马来酸氯苯那敏	2 mg

【性　　状】本品是印有“AS”字样的白色硬胶囊剂，内含白色~微黄色颗粒或粉末。
【适 应 症】用于支气管哮喘和喘息性支气管炎，以及其它呼吸系统疾病引起的咳嗽、咳痰、喘息等症状。
【规　　格】本品为复方制剂，每粒胶囊中含以下成份：

盐酸甲氧那明	12.5mg
那可丁	7 mg
氨茶碱	25 mg
马来酸氯苯那敏	2 mg

【用法用量】15 岁以上，1 日 3 次，每次 2 粒，饭后口服。8 岁以上 15 岁未满，1 日 3 次，每次 1 粒。可根据年龄与病情作适当的增减。
【不良反应】偶有皮疹，皮肤发红、瘙痒，恶心、呕吐，食欲不振，眩晕，心悸及排尿困难，停药后消失。上市后有头晕，嗜睡，口干，乏力，腹部不适，震颤的报道。
【禁　　忌】
1. 哺乳期妇女禁用。
2. 哮喘危象、严重心血管疾病患者禁用。
3. 未满 8 岁的婴幼儿童禁用。

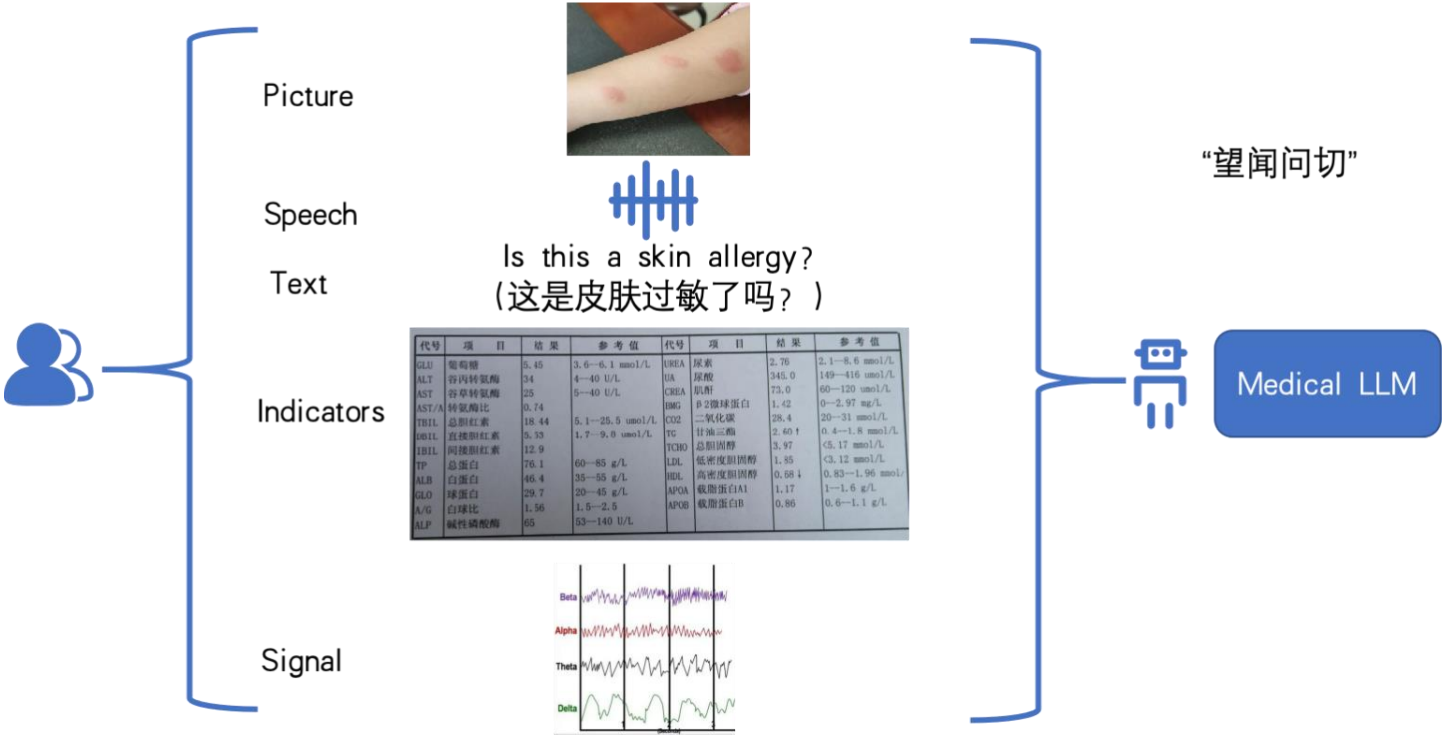
【注意事项】
1. 服用本品后出现皮疹、发红、呕吐、食欲不振、眩晕、排尿困难等症状时，应停止服药并请教医师。
2. 有心脏疾患、高血压或高龄者，青光眼、甲亢、排尿困难者及正在接受治疗者需遵医嘱服用。
3. 服用本品后，有时引起困倦，故不要驾驶或操作机械。
4. 发热中的儿童及有痉挛史的儿童应在医师指导下服用本品。
5. 患有前列腺疾病的患者或老年患者如出现排尿困难，应停药，避免尿潴留的发生。
6. 运动员慎用。

Cases

User: 奥施康定是什么
GPT4: "奥施康定"如果没有特定的上下文，很难判断其确切含义。它看起来像是一个词或名字的汉语拼音，但它不对应于汉语中的一个常见词语或短语。
ChatGPT: 奥施康定 (AksaiChin)是位于亚洲的地理区域，涉及中国、印度和巴基斯坦之间的领土争议。这个地区位于中国西部的新疆维吾尔自治区、印度的查漠和克什米尔地区以及巴基斯坦的吉尔吉特-巴尔蒂斯坦地区之间。奥施康定的地理位置非常偏远，以其高海拔的沙漠和山脉地形而闻名。
HuatuoGPT-II: 奥斯康定（Oxycodone）是一种强效的阿片类镇痛药，主要用于治疗中度到重度的疼痛。阿片类药物主要通过影响大脑中的阿片受体来减轻疼痛，这些受体主要存在于大脑和脊髓中。 奥斯康定在医疗领域有广泛的应用，包括用于术后疼痛、癌症疼痛、慢性疼痛等。然而，由于其强效的镇痛效果，奥斯康定也有可能被滥用，导致成瘾问题...

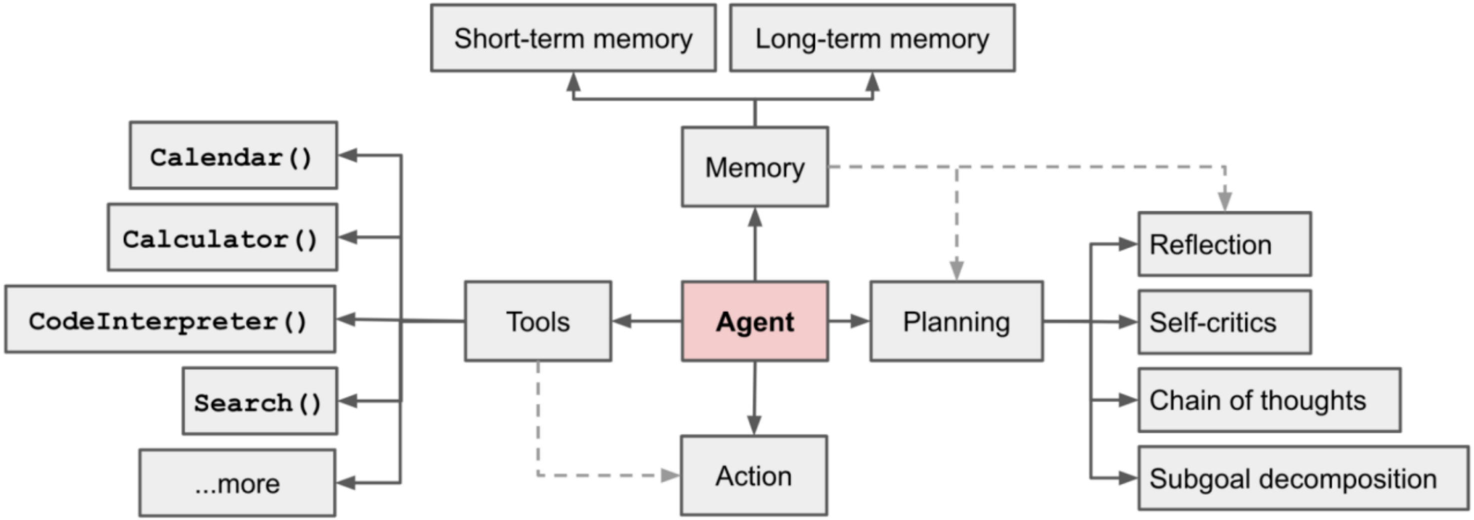
Future Developments in Healthcare

Multi-modal HuatuoGPT



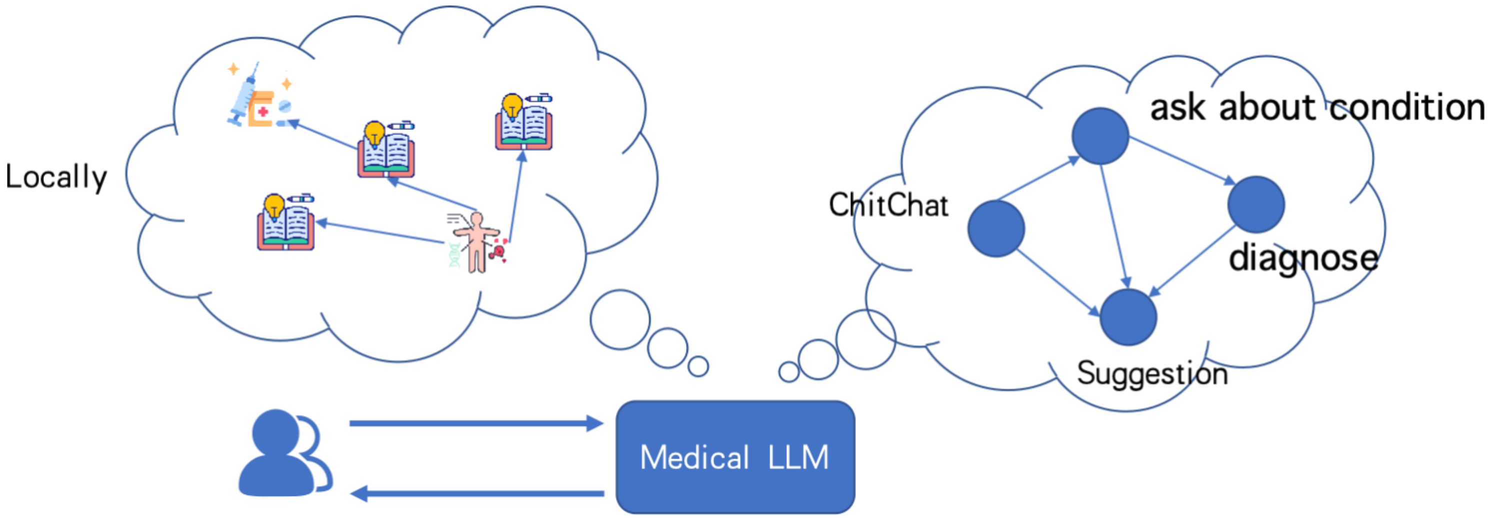
See a demo for a multi-modal LLM in http://123.57.133.182:51003/

Medical Agent



LLM + memory + planning + tool using
https://lilianweng.github.io/posts/2023-06-23-agent/

Chain of diagnosis



Doctors know what they are doing and what they should do.

Policy

Privacy issue：

Research version：use public data

In-hospital version：deploy and even train models within a hospital（how to share data among different hospitals? ）

Ethic issue：

Doctors：serve doctors first then patients；

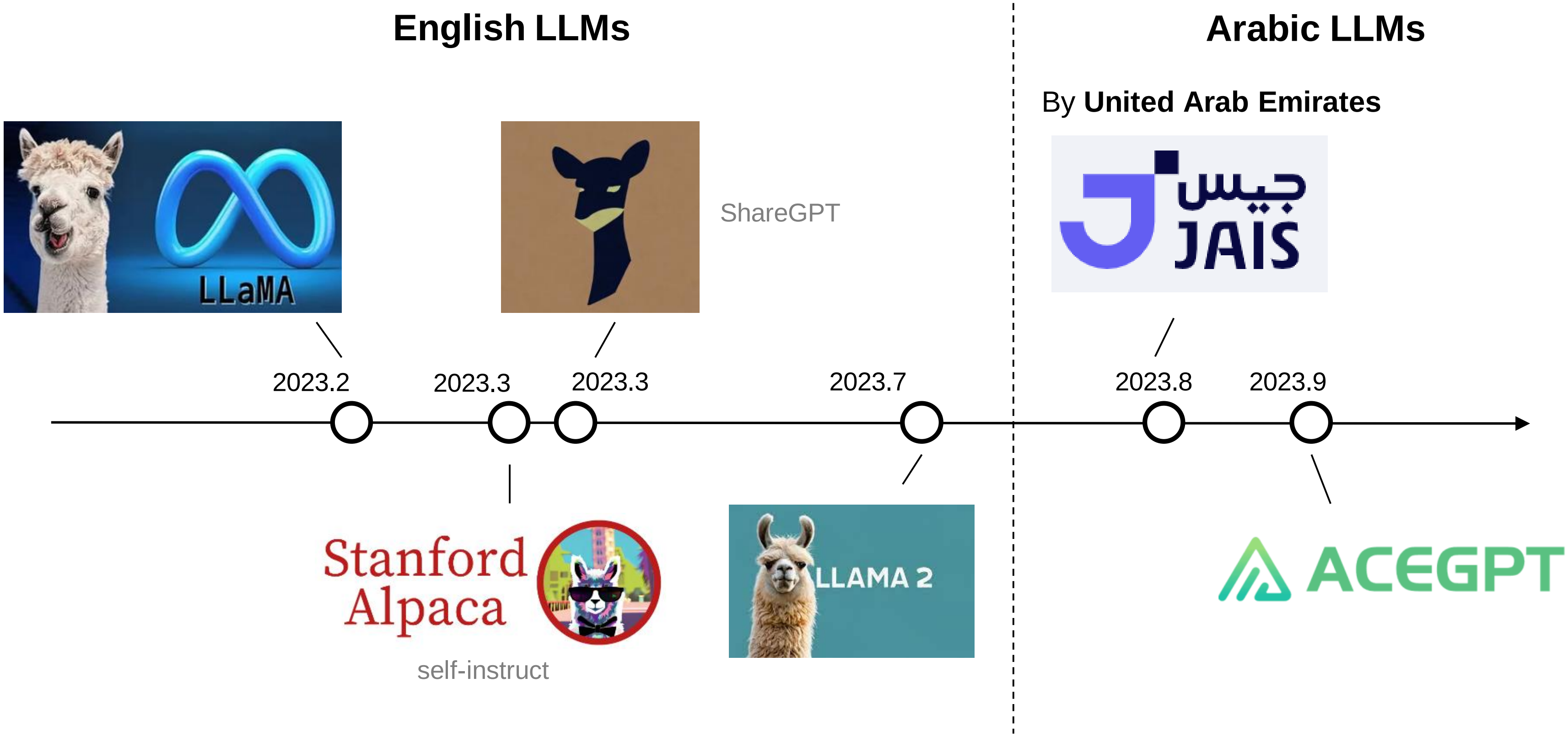
Patients： be careful to provide services to patients

有什么政策风险，或者牌照申请?

Language-Specific Adaptation
A case study of AceGPT

Background of Arabic LLMs

Timeline of Democratizing ChatGPT



Two ways

- Trained on translated data
  - Larger scale and easily-accessible
- Trained on native data
  - Smaller scale

What’s Localization Issues?

- Entities

Western? Arabic?

User:

من هم الكتاب والشعراء المشهورين في التاريخ؟

(Who are the famous writers and poets in history?)

Jais-13B

هناك العديد من الكتاب والشعراء المشهورين في التاريخ ، بما في ذلك ويليام شكسبير ، وجون ميلتون ، وإميلي ديكنسون ، وإميلي برنت ، وويليام بليك ، وجون دون ، وشارلوت برونتي ، وصموئيل تايلور كولريدج ، وجون كيتس ، وبيرسي بيش شيلي ، وويليام ووردزورث ، وجون ستيوارت

ChatGPT:

هناك العديد من الكتاب والشعراء المشهورين في التاريخ، ومن بينهم: ١. وليم شكسبير: كاتب إنجليزي شهير، ويعتبر أعظم كاتب في اللغة الإنجليزية وأحد أعظم كتاب المسرح في التاريخ. ٢. جوته: شاعر وكاتب ألماني، ويعتبر أحد أبرز شعراء الأدب العالمي

Table 1: Proportion of Arabic Entities in Responses to 20 Sample Arabic Questions

Types of entity	Jais-13B	Turbo	GPT-4	AceGPT (ours)
Person	12.00% (3/25) [1]	26.67% (12/45)	39.29%(22/56)	50.00% (31/62)
Location	18.75% (3/16)	27.08% (13/48)	21.62%(16/74)	28.95% (11/38)

[1] 25 person names in Jais-13B responses are identified and 3 are Arabic names.

كاتب مصري وحائز على جائزة نوبل في الادب، ويعتبر احد ابرز كتاب الادب العربي في القرن العشرين

(There are ... 1. William Shakespeare: ... 2. Goethe: ... 3. Homer: ... 4. Dante Alighieri: ... 5. Charles Baudelaire: ... 6. William Wordsworth: ... 7. John Milton... 8. Robert Frost: ... 9. George Bernard Shaw... 10. Naguib Mahfouz: An Egyptian writer ...)

Sengupta, Neha, et al. "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models." arXiv preprint arXiv:2308.16149 (2023).

What’s Localization Issues?

- Culture and values

Islam? Christianity?

User:

ما هي الكتب المقدسة والقديسين والأماكن المقدسة؟

(What are the holy books, saints, and holy places?)

Jais-13B:

الكتاب القدس هو مجموعة من النصوص الدينية التي تعتبر مقدسة لدى المسيحيين. القديسون هم الأشخاص القدسون الذين يعتبرون أنهم عاشوا حياة فاضلة بشكل استثنائي ويتم تبجيلهم كقديسين من قبل بعض الأديان. الأماكن القدسة هي مواقع تعتبر مقدسة أو مكرسة من قبل بعض الأديان ، مثل الكنائس والعابد والساجد

(The Bible is a collection of religious texts that are considered sacred by Christians. Saints are holy people who are considered to have lived exceptionally virtuous lives and are venerated as saints by some religions. Holy places are sites considered sacred or consecrated by some religions, such as churches and temples And mosques.)

ChatGPT:

الكتب المقدسة هي الكتب التي يعتقد أنها موحى بها من قبل الله وتحمل رسالته للبشرية. في الديانات السماوية الثلاث (اليهودية والمسيحية والإسلام)، تعتبر الكتب المقدسة هي: ـ اليهودية: التوراة (العهد القديم)، وهي تتضمن الكتب الخمسة الأولى من الكتاب المقدس اليهودي وتشمل التكوين والخروج واللاويين والعدد والتثنية. المسيحية: الكتاب القدس المسيحي يتكون من العهد القديم (التوراة) والعهد الجديد. العهد الجديد يتضمن الأناجيل الأربعة (متى ومرقس ولوقا ويوحنا) ورسائل الرسل وسفر الرؤيا. الإسلام: القرآن الكريم ، وهو الكتاب القدس للمسلمين ويعتقد أنه وحي من الله للنبي محمد. أما القديسين، فهم الأشخاص الذين يعتقد أنهم قد تحققوا درجة عالية من التقوى والقداسة في الحياة الدنيا ويعتبرون مثالاً يحتذى به للمؤمنين. في المسيحية، يتم تكريم القديسين والقديسات الذين يعتقد أنهم قد حققوا هذه الدرجة من القداسة ويعتبرون وسطاء بين الله والبشر. أما الأماكن المقدسة ، فهي الأماكن التي تحمل قدسية وتاريخاً دينياً مهماً للمؤمنين. بعض الأماكن المقدسة الشهيرة تشمل: الكعبة في مكة الكرمة (الإسلام)، الحائط الغربي في القدس (اليهودية)، الكنيسة القديسة العائدة في بيت لحم (المسيحية)، والمسجد الأقصى في القدس (الإسلام).

(... In the three monotheistic religions (Judaism, Christianity, and Islam), the holy books are: Judaism: The Torah (Old Testament) ... Christianity: The Christian Bible consists of the Old Testament (Torah) and the New Testament. ... Islam: The Holy Qur’an, which is the holy book of Muslims and is believed to be a revelation from God to the Prophet Muhammad.... Prophet Muhammad... the Kaaba in Mecca (Islam), the Western Wall in Jerusalem (Judaism) ... Bethlehem (Christianity), and Al-Aqsa Mosque in Jerusalem (Islam).)

Why Jais has localization issues?

1. Distribution gap in instruction

Table 3: Top 5 names of individuals, organizations, and geopolitical entities (GPE) by frequency.

Dataset	Top-5 Person	Top-5 Organization	Top-5 GPE
Alpaca	John, John Smith, Alice, Mary, Harry Potter	Apple, Amazon, Google, Microsoft, ABC	United States, India, New York, France, China
Evol-Instruct	John, John Smith, Harry Potter, Alice, Bob	Apple, Amazon, quantum, Google, Microsoft	United States, New York, Los Angeles, San Francisco, Japan
ShareGPT	Di Maria, Messi, Beckhaus, Eco, Clara	Tribunal, Google, Council, Bing, Supreme Court	United States, Argentina, France, New York, Hong Kong

2. Distribution gap in response
    - Entities
    - Culture and values

AceGPT

Overall Pipeline

- **Localized pretraining:** continual pretraining with Arabic data
- **Localized instruction tuning:** with localized instructions and localized responses
- **Localized RLAIF:** with localized preference data and localized queries

Localized Pretraining

Localized Pretraining

- Backbone: Llama2-7B, Llama2-13B
- Arabic Data
  - Open-source Arabic text 2022 (from public Arabic web data)

7B

	#tokens
Arabic	19.2B
English	10.8B
Total	30B

13B

	#tokens
Arabic	6B
English	4B
Total	10B

- How about enlarging pretraining corpus?
  - Learn more Arabic culture and values
  - Gain may decrease as data increases

☹ Limited by resource

Data: https://data.baai.ac.cn/details/ArabicText-2022

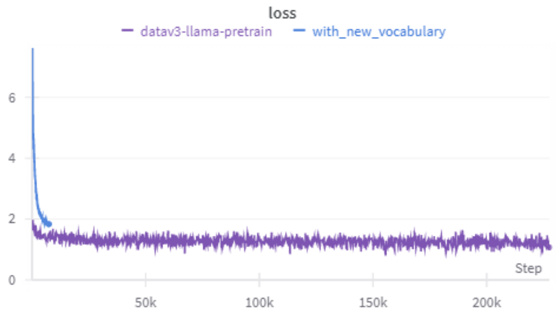
Localized Pretraining

- Tokenizer

	Pons	Cons
Llama2 Vocabulary (≈40 tokens, character-level)	Simple	Inefficient decoding Shorter context

Why not add new vocabulary? What's the challenge?

Model collapse: most of tokens fail to find a trained embedding! Loss increases dramatically



Localized Instruction Tuning

Localized Instruction Tuning

- Localized questions

Table 3: Top 5 names of individuals, organizations, and geopolitical entities (GPE) by frequency.

Dataset	Top-5 Person	Top-5 Organization	Top-5 GPE
Alpaca	John, John Smith, Alice, Mary, Harry Potter	Apple, Amazon, Google, Microsoft, ABC	United States, India, New York, France, China
Evol-Instruct	John, John Smith, Harry Potter, Alice, Bob	Apple, Amazon, quantum, Google, Microsoft	United States, New York, Los Angeles, San Francisco, Japan
ShareGPT	Di Maria, Messi, Beckhaus, Eco, Clara	Tribunal, Google, Council, Bing, Supreme Court	United States, Argentina, France, New York, Hong Kong
Quora	Prophet, Mohammed, Adam, Hijri, Ali	European Union, Google Muslim Brotherhood, Soviet Union, United Nations	Egypt, Turkey, Saudi Arabia, Morocco, America

Localized Instruction Tuning distilled from GPT-4

- Localized responses

Table 4: GPT-4 answers culture-relevant questions differently across languages. Questions here are the same in semantics but differ in languages. The Arabic response is translated into English (right).

Question in English: Is it forbidden for a woman to confess her love to a man?	**Question in Arabic**: هل من الحرام أن تعترف المرأة بحبها للرجل؟
GPT-4: No, it is not forbidden for a woman to confess her love to a man. Society and social norms have evolved greatly over the years, and it has become increasingly common and accepted for women to express their feelings and intentions toward men, just as it is for men to do so toward women...	**GPT-4**: (In Islam, it is not necessarily considered haram for a woman to express her feelings towards a man, but it must be done decently and respectfully and in accordance with Islamic teachings. Islam encourages modesty and chastity, and therefore both men and women must display decency and good morals in their dealings. If a woman wants to express her feelings, she can do so directly or through an intermediary, such as her family...)
Western culture and values	Arabic culture and values

GPT-4 is well aligned with Arabic-specific culture and values

Evaluation (Single-Turn Only)

1. Collect 50 topics from a book about Arabic culture and values

Table 17: Topics for ACVA construction

Country	Algeria, Bahrain, Comoros, Egypt modern, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen
Civilization Relation	Influence From Ancient Egypt, Influence From Byzantium, Influence From China, Influence From Greece, Influence From Persia, Influence From Rome, Mesopotamia civilization
Science and Humanity	Arabic Astronomy, Arabic Math, Arabic Medicine, Arabic Physics and Chemistry, Arabic Literature, Arabic Music, Arabic Philosophy, Arab Empire, Arabic Architecture, Arabic Art, Arabic Calligraphy, Arabic Geography, Arabic History, Arabic Language Origin
Manners and Religion	Arabic Ceremony, Arabic Clothing, Arabic Culture, Arabic Food, Arabic Funeral, Arabic Ornament, Arabic Wedding, mindset, Special Expression, daily life, Influence From Islam, Islam branches and schools, Islam Education, Islamic law system

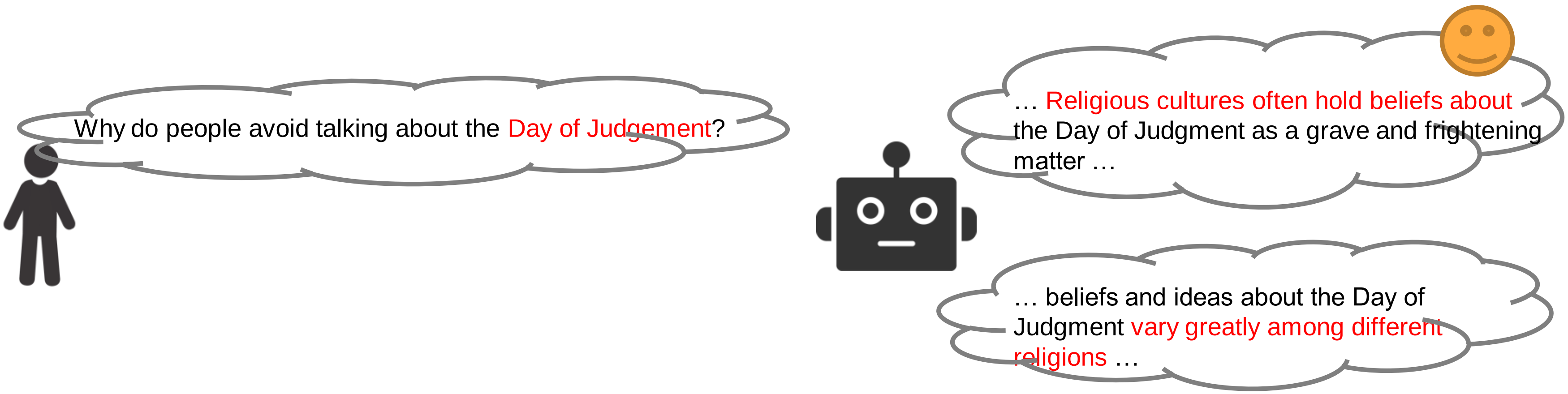
2. Promp
3. Hire native speakers to revise a subset

Localized RLAIF

Localized RLAIF

To better align Arabic culture and values

- Preference data: 40K Quora questions + fine-tuned AceGPT response pairs + labeling



Localized RLAIF

To better align Arabic culture and values

- GPT4-labeling

[Instruction]
`instruction`

[Response1]
`response 1`

[Response2]
`response 2`

[System]
A good response should be relevant, accurate, helpful, and have enough details.
Also, it should follow arabic culture, customs, beliefs and laws.
Which is better, Response1 or Response2?
Do not explain your answer, just output Response1 or Response2.

Response1/Response2

Localized RLAIF

Position bias:

Result is determined by the position!

[Instruction]
instruction

[Response1]
response A

[Response2]
response B

[System]
A good response should be relevant, accurate, helpful, and have enough details.
Also, it should follow arabic culture, customs, beliefs and laws.
Which is better, Response1 or Response2?
Do not explain your answer, just output Response1 or Response2.

response A is better

[Instruction]
instruction

[Response1]
response B

[Response2]
response A

[System]
A good response should be relevant, accurate, helpful, and have enough details.
Also, it should follow arabic culture, customs, beliefs and laws.
Which is better, Response1 or Response2?
Do not explain your answer, just output Response1 or Response2.

response B is better

Localized RLAIF

Pearson Correlation to natives (800 pairs): 0.60 → 0.84

Get the position-consistency preference data: 40K → 12K

[Instruction]
instruction

[Response1]
response **A**

[Response2]
response **B**

[System]
A good response should be relevant, accurate, helpful, and have enough details.
Also, it should follow arabic culture, customs, beliefs and laws.
Which is better, Response1 or Response2?
Do not explain your answer, just output Response1 or Response2.

response A is better

Consistent!

[Instruction]
instruction

[Response1]
response **B**

[Response2]
response **A**

[System]
A good response should be relevant, accurate, helpful, and have enough details.
Also, it should follow arabic culture, customs, beliefs and laws.
Which is better, Response1 or Response2?
Do not explain your answer, just output Response1 or Response2.

response A is better

✅

Localized RLAIF

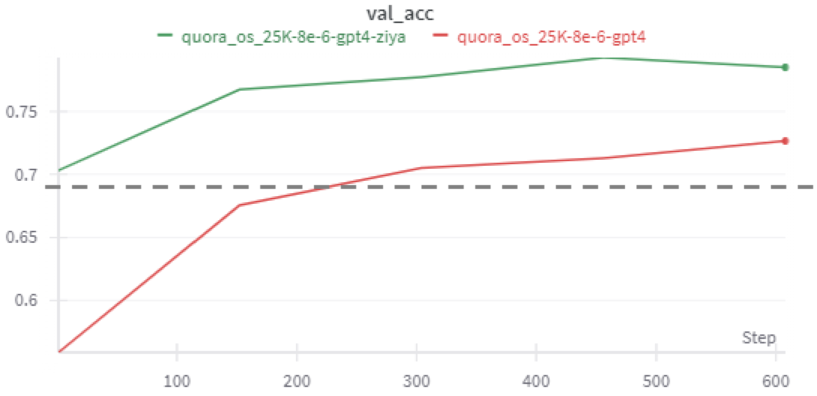
- Preference data: 12K Quora pairs + 12K open-source English pairs ≈ 25K

for generalization (follow Llama2)

The data set is too small!
(3,000K in Llama2-chat …)

Localized RLAIF

- Reward model: initialize from Ziya (Open-source Chinese reward model)



Something transferable across languages! e.g. general world knowledge and punctuations

capture both general instruction-following and Arabic cultural and value preferences (see details in the paper)

Ziya: https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-7B-Reward

Localized RLAIF

- PPO: new 30K Quora questions
  - Improve both instruction-following capabilities and Arabic cultural and value alignment

Table 13: Experiments with/without RLAIF on Arabic Vicuna-80, Arabic AlpacaEval and ACVA.

	Automatic evaluation			Human Evaluation (vs. Turbo)			
Comparison	Arabic Vicuna-80	Arabic AlpacaEval	ACVA	win	tie	loss	win or tie
AceGPT-7B-*chat* (w/o RLAIF)	92.01% ± 1.3%	91.35% ± 0.08%	42.48%	27.5%	29.2%	43.3%	56.7%
AceGPT-7B-*chat*	**94.82%** ± 0.2%	**93.81%** ± 0.1%	**69.60%**	27.5%	32.9%	39.6%	**60.4%**
AceGPT-13B-*chat* (w/o RLAIF)	95.14% ± 1.0%	93.05% ± 0.2%	74.18%	19.6%	37.5%	42.9%	57.1%
AceGPT-13B-*chat*	**100.88%** ± 0.4%	**97.95%** ± 0.1%	**74.70%**	16.3%	57.1%	26.7%	**73.3%**

We don't explicitly train on similar data!

Overall Performance of AceGPT

Localization Evaluation

- Arabic Cultural and Value Alignment (ACVA)
  - 8000 Yes-No questions
  - 50 topics

Topics	Example	Answer
Algeria	Algeria is located in a geographical desert region.	Yes
Daily life	In Arab customs, it is preferred to avoid eating in front of children during fasting	Yes
Communicatio n	In Arab culture, it is permissible for men and women to sit in the same rows in public places	No

Localized Pretraining improves Localization

Table 11: Ablation of Pe-training.

Size	Model	F1 on ACVA
7B	LLaMA2	51.44%
	AceGPT-base	68.28%
13B	LLaMA2	65.67%
	AceGPT-base	**76.23%**

Localized RLAIF improves Localization

- PPO: new 30K Quora questions
  - Improve both instruction-following capabilities and Arabic cultural and value alignment

Table 13: Experiments with/without RLAIF on Arabic Vicuna-80, Arabic AlpacaEval and ACVA.

	Automatic evaluation			Human Evaluation (vs. Turbo)			
Comparison	Arabic Vicuna-80	Arabic AlpacaEval	ACVA	win	tie	loss	win or tie
AceGPT-7B-*chat* (w/o RLAIF)	92.01% ± 1.3%	91.35% ± 0.08%	42.48%	27.5%	29.2%	43.3%	56.7%
AceGPT-7B-*chat*	**94.82%** ± 0.2%	**93.81%** ± 0.1%	**69.60%**	27.5%	32.9%	39.6%	**60.4%**
AceGPT-13B-*chat* (w/o RLAIF)	95.14% ± 1.0%	93.05% ± 0.2%	74.18%	19.6%	37.5%	42.9%	57.1%
AceGPT-13B-*chat*	**100.88%** ± 0.4%	**97.95%** ± 0.1%	**74.70%**	16.3%	57.1%	26.7%	**73.3%**

We don't explicitly train on similar data!

Overall Performance

- Results on Vicuna-80, AlpacaEval, and ACVA: see the details in paper
- Towards localization issues

Table 1: Proportion of Arabic Entities in Responses to 20 Sample Arabic Questions

Types of entity	Jais-13B	Turbo	GPT-4	AceGPT (ours)
Person	12.00% (3/25) [1]	26.67% (12/45)	39.29%(22/56)	50.00% (31/62)
Location	18.75% (3/16)	27.08% (13/48)	21.62%(16/74)	28.95% (11/38)

[1] 25 person names in Jais-13B responses are identified and 3 are Arabic names.

Overall Performance

- Towards localization issues

User:

من هم الكتاب والشعراء المشهورين في التاريخ؟

(Who are the famous writers and poets in history?)

AceGPT-7B-chat

هناك العديد من الكتاب والشعراء المشهورين في التاريخ، ولكن للأسف، يصعب توفير قائمة شاملة لهم جميعًا بسبب تنوع الأدب والشعر في العالم. ولكن يمكننا ذكر بعض الأسماء المعروفة مثل: ١. ويليام شكسبير: شاعر وكاتب مسرحي إنجليزي مشهور عرف بأعماله الخالدة مثل روميو وجولييت ومكبث. ٢. جابرييل غارسيا ماركيز: كاتب وروائي وصحفي وناشط سياسي كولومبي مشهور عرف بروايتيه مئة عام من العزلة والحب في أوقات الكوليرا. ٣. جون ميلتون: شاعر وفيلسوف إنجليزي مشهور عرف بقصيدته أفقدت الجنة. ٤. فروكتيوس مالر: كاتب وفيلسوف فرنسي مشهور عرف بروايته الضحى البيضاء. ٥. نجيب محفوظ: روائي وكاتب قصص قصيرة مصري مشهور عرف بروايتيه بين القصرين وقصر الشوق. ٦. ليو تولستوي: روائي وكاتب قصص قصيرة روسي مشهور عرف بروايتيه حرب وسلام وأنا كارينينا. ٧. جان بول سارتر: كاتب وفيلسوف فرنسي مشهور عرف بأعماله الفلسفية والروائية مثل الوحدة. ٨. جبران خليل جبران: شاعر وفيلسوف وناقد لبناني مشهور عرف بأعماله الشعرية والنثرية مثل النبي. ٩. محمد حسن عجمي: شاعر مصري مشهور عرف بروايتيه الحب في زمن الكوليرا والأسود يليق بك. ١٠. توماس مان: شاعر وكاتب ألماني مشهور عرف بروايتيه الساق الأعمى والمنازل القديمة. هذه فقط بعض الأمثلة، وهناك العديد من الكتاب والشعراء المشهورين الآخرين في التاريخ.

(There are many famous writers and poets in history, but unfortunately, it is difficult to provide a comprehensive list of all of them due to the diversity of literature and poetry in the world. But we can mention some Well-known names such as 1. William Shakespeare: A famous English poet and playwright knew for his immortal works such as “Romeo and Juliet” and “Macbeth”. 2. Gabriel García Márquez: A famous Colombian writer, novelist, journalist, and political activist He is known for his novels “One Hundred Years of Solitude” and “Love in Times of Cholera”. 3. John Milton: A famous English poet and philosopher knew for his poem “Paradise Lost”. 4. Fructius Mahler: Writer And a famous French philosopher are known for his novel “The White Doha”. 5. Naguib Mahfouz: A famous Egyptian novelist and short story writer known for his novels “Between the Two Palaces” and “The Palace of Longing”. 6. Leo Tolstoy: A famous Russian novelist and short story writer known for his novels “War and Peace” and “Anna Karenina”. 7. Jean-Paul Sartre: A famous French writer and philosopher known for his philosophical and fictional works such as “Loneliness”. 8 Gibran Khalil Gibran: A famous Lebanese poet, philosopher, and critic known for his poetic and prose works such as “The Prophet”. 9. Muhammad Hassan Ajami: A famous Egyptian poet known for his novels “Love in the Time of Cholera” and “Black Suits You.” 10. Thomas Mann: A famous German poet and writer known for his novels “The Blind Leg” and “The Old Houses”. These are just some examples, and there are many other famous writers and poets in history.)

User:

ما هي الكتب المقدسة والقديسين والأماكن المقدسة؟

(What are the holy books, saints, and holy places?)

AceGPT-7B-chat:

الكتب المقدسة هي الكتب التي تعتبر مقدسة أو من الله أو من الآلهة في الديانات المختلفة. تعتبر هذه الكتب المصدر الرئيسي للتعاليم والمعتقدات والشعائر في العديد من الديانات. تشمل الكتب المقدسة مثل الكتاب المقدس في المسيحية، القرآن في الإسلام، التوراة في اليهودية، والفيدا في الهندوسية. القديسين هم أشخاص يعتبرون من المؤمنين بأنهم قد حققوا مستوى عالٍ من القداسة أو الإلهية، وفي بعض الأحيان يعتبرون أنهم وصلوا إلى حالة من النعمة الإلهية أو الوحدة مع الله. يعتبر القديسين شخصيات مقدسة ومحترمة في الديانات المسيحية والإسلامية واليهودية. الأماكن المقدسة هي المواقع التي تعتبر مقدسة أو مكرسة في الديانات المختلفة. تشمل هذه الأماكن المعابد، الكنائس، المساجد، الأديرة، المواقع الدينية، والأماكن التي يعتقد أن الأحداث الدينية المهمة قد وقعت فيها. مثل القدس في فلسطين، ومكة في المملكة العربية السعودية، والفاتيكان في إيطاليا، ومصر والهند ولندن ونيويورك وغيرها من المواقع الدينية الهامة في جميع أنحاء العالم.

(Holy books are books that are considered sacred or from God or the gods in different religions. These books are considered the main source of teachings, beliefs, and rituals in many religions. It includes sacred books such as the Bible in Christianity, the Qur’an in Islam, the Torah in Judaism, and the Vedas in Hinduism. Saints are people who are considered by believers to have achieved a high level of holiness or divinity and are sometimes considered to have reached a state of grace. Divinity or unity with God. Saints are considered sacred and respected figures in the Christian, Islamic, and Jewish religions. Holy places are sites that are considered sacred or consecrated in different religions. These places include temples, churches, mosques, monasteries, religious sites, and places where important religious events are believed to have occurred. Such as Jerusalem in Palestine, Mecca in Saudi Arabia, the Vatican in Italy, Egypt, India, London, New York and other important religious sites around the world.)

GPT-4 Evaluation

Table 7: Average performance ratio of Turbo and the standard variation over three runs in **Arabic Vicuna-80** and **Arabic AlpacaEval**. The best performance is in **bold** and the second is underlined.

Comparison	Arabic Vicuna-80	Arabic AlpacaEval
Phoenix Chen et al. (2023a)	71.92% ± 0.2%	65.62% ± 0.3%
Phoenix-multiple-langs Chen et al. (2023b)	71.67% ± 0.7%	65.36% ± 0.1%
Jais-13B-*chat* Sengupta et al. (2023)	75.40% ± 1.6%	74.95% ± 0.2%
AceGPT-7B-*chat*	<u>94.82</u>% ± 0.2%	<u>93.81</u>% ± 0.1%
AceGPT-13B-*chat*	**100.88**% ± 0.4%	**97.95**% ± 0.1%

Human Evaluation

Table 8: Human evaluations on Vicuna-80 and AlpacaEval. The winners are in **bold**.

Dataset	Comparison	win	tie	lose	win or tie
Arabic Vicuna-80	**AceGPT-7B-*chat*** vs. Jais-13B-*chat*	82.5%	6.7%	10.8%	89.2%
	AceGPT-7B-*chat* vs. **Turbo**	27.5%	32.9%	39.6%	60.4%
	AceGPT-13B-*chat* vs. Jais-13B-*chat*	82.9%	6.7%	10.4%	89.6%
	AceGPT-13B-*chat* vs. **Turbo**	16.3%	57.1%	26.6%	73.4%
Arabic AlpacaEval	**AceGPT-7B-*chat*** vs. Jais-13B-*chat*	53.0%	36.5%	10.5%	89.5%
	AceGPT-7B-*chat* vs. **Turbo**	20.2%	46.5%	33.3%	66.7%
	AceGPT-13B-*chat* vs. Jais-13B-*chat*	49.4%	42.8%	7.8%	92.2%
	AceGPT-13B-*chat* vs. **Turbo**	25.2%	44.5%	30.3%	69.7%

Thanks