

对比经微调的ERNIE-Lite-8K-0922和GPT-4在使用Prompt策略后在英语对话系统中的表现：以心理咨询师角色为例

季东霖¹, 郭子浩², 陈雨洁², 王欣然², 张梦林², 孙文韬³

¹中国矿业大学(北京)人工智能学院, 北京

²中国矿业大学(北京)文法学院, 北京

³中国矿业大学(北京)机械与电气工程学院, 北京

收稿日期: 2024年3月15日; 录用日期: 2024年5月7日; 发布日期: 2024年5月15日

摘要

本研究基于大模型在英语对话系统中的实际应用对比了经过微调的ERNIE-Lite-8K-0922和GPT-4模型在采用Prompt策略后在英语对话系统中的能力表现。本研究采用了一系列定量指标, 如BLEU、ROUGE分数、训练损失等指标, 展示了模型微调的效果, 使用自然度、逻辑性、上下文理解、多轮对话处理和情感表达等指标, 评估了模型生成回复的质量。本研究在指出了ERNIE-Lite-8K-0922和GPT-4在英语对话系统中的性能差异的同时, 还提出了需要进一步完善数据集与微调参数等方法以提高微调后的ERNIE-Lite-8K-0922在英语对话系统及特定领域的表现能力。本研究为探索是否有更加经济高效的方法在实际应用场景中将大语言模型部署为英语对话系统提供了重要参考, 也为英语对话系统及相关领域的进一步发展做出了贡献。

关键词

模型微调, 提示词工程, 大语言模型, 英语对话系统, 人工智能

Comparison of the Performance of Fine-Tuned ERNIE-Lite-8K-0922 and GPT-4 in English Dialogue Systems after Using the Prompt Strategy: A Case Study of the Role of a Psychological Counselor

Donglin Ji¹, Zihao Guo², Yujie Chen², Xinran Wang², Menglin Zhang², Wentao Sun³

文章引用: 季东霖, 郭子浩, 陈雨洁, 王欣然, 张梦林, 孙文韬. 对比经微调的 ERNIE-Lite-8K-0922 和 GPT-4 在使用 Prompt 策略后在英语对话系统中的表现: 以心理咨询师角色为例[J]. 人工智能与机器人研究, 2024, 13(2): 272-281.

DOI: 10.12677/airr.2024.132029

¹School of Artificial Intelligence, China University of Mining and Technology (Beijing), Beijing

²School of Law and Humanities, China University of Mining and Technology (Beijing), Beijing

³School of Mechanical and Electrical Engineering, China University of Mining and Technology (Beijing), Beijing

Received: Mar. 15th, 2024; accepted: May 7th, 2024; published: May 15th, 2024

Abstract

Based on the practical application of the Large Language Models in the English dialogue system, this study compares the performance of the fine-tuned ERNIE-Lite-8K-0922 and gpt-4 model in the English dialogue system after using the prompt strategy. This study uses a series of quantitative indicators, such as BLEU, ROUGE, training loss and other indicators, to show the effect of model fine-tuning, and uses indicators such as naturalness, logicity, context understanding, multiple rounds of dialogue processing and emotional expression to evaluate the quality of response generated by the model. While pointing out the performance differences between ERNIE-Lite-8K-0922 and gpt-4 in English dialogue system, this study also proposes the need to further improve the data set and fine tune parameters to improve the performance of fine-tuned ERNIE-Lite-8K-0922 in English dialogue system and specific fields. This study provides an important reference for exploring whether there is a more cost-effective method to deploy the large language model into the English dialogue system in the actual application scenario, and also makes a contribution to the further development of the English dialogue system and related fields.

Keywords

Fine-Tuning, Prompt Engineering, Large Language Models, English Dialogue System, Artificial Intelligence

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景与动机

近年来大规模预训练语言模型发展迅速[1] [2], 国外的生成式预训练模型(generative pre-trained transformer, GPT)系列等产品(GPT1, GPT2, GPT3, ChatGPT/InstructGPT, GPT4) [3] [4] [5] [6] [7]快速迭代, 在全球引起了大模型的发展热潮。与此同时, 国内企业乘势追赶, 也纷纷发布了大语言模型。比如: 盘古大模型、通义千问、文心一言等。随着数据量逐渐庞大, 模型数量逐渐增多。如今, 许多大语言模型已经在语言翻译、文摘构建、命名实体识别、文本分类和关系抽取等多种场景, 以及政务、金融和生物医药等多个行业得到广泛应用[8]。

据科技部“新一代人工智能发展研究中心”于 2023 年发布的《中国人工智能大模型地图研究报告》显示, 中国研发的大模型数量排名全球第二, 仅次于美国, 目前中国 10 亿参数规模以上的大模型已发布 79 个[9]。

在国内的大模型市场中, 根据 IDC 发布的《AI 大模型技术能力评估报告 2023》显示, 百度文心大模型 3.5 拿下 12 项指标的 7 个满分, 综合评分第一, 算法模型第一, 行业覆盖第一, 三个绝对第一体现了百度文心大模型的基础技术深度和产业应用覆盖广度。文心大模型在此次 IDC 评估中获得了算法模型

维度的唯一一个满分，充分体现了百度在大模型核心技术上的领先优势[10]。

当前，AIGC 也催生了新的软件服务模式。人们通过 API (应用程序编程接口)提供这些模型的服务，也就是说，人们可以通过简单的 API 调用来使用大模型，而不需要自己单独构建和训练模型[11]。这大大减少了应用的开发难度。

百度作为世界五百强企业、国内科技行业中的佼佼者，其搭建的千帆大模型平台，集成了多种国内外先进的大模型。该平台还提供数据集、模型微调、模型评估等辅助工具，实现了模型微调及部署的全流程一站式服务，为国内开发者微调和部署大模型提供了安全、可靠、高效的平台。

1.2. 研究目标与意义

现如今，以 ChatGPT 为首的各种大模型性能强悍，行业领先。但 GPT-4 在国内注册限制较大，且国内多数大模型微调和调用的费用相对于 GPT-4 较低，更有利于各领域在部署大模型时节约成本。而与此同时，国内的大模型紧跟潮流，发展热情高涨，发展势头迅猛，模型能力发展情况尚可。因此对于研究用国内的模型进行微调后，是否也能和 GPT-4 在具体领域的实际应用中产生类似或者更好的效果，将为国内相关行业对于解决是否有更加经济高效的方法应用和部署大模型的问题做出一定借鉴。

1.3. 研究方法概述

本研究将以心理咨询对话场景为例，选取国内大模型 ERNIE-Lite-8K-0922 和 GPT-4 进行对照实验。研究方法将从模型微调和提示词工程两个方面入手。首先准备数据集，使用 SFT 方法对 ERNIE-Lite-8K-0922 模型进行微调，并最终部署。然后，使用设计好的提示词，让 GPT-4 明确任务要求。之后使用测试集，以人工的方式分别和两个大模型完成对话。最终采用人工评估的方式，对两个大模型在实际应用中的表现进行评估。

2. 大语言模型、模型微调和提示词工程

2.1. 大模型介绍

2.1.1. 大语言模型综述

大语言模型被认为是未来通用人工智能的关键技术之一，是一代代语言模型经过不断继承、优化和迭代的结果。

目前大语言模型主要使用 Transformer 架构。其采用自注意力机制、以及编码器和解码器堆叠的方式，在语义特征提取和任务特征抽取能力上显著超过卷积神经网络等深度学习模型。现有的语言模型主要是基于 Transformer，构建的架构和路线多种多样[12]。

2.1.2. ERNIE-Lite-8K-0922

新一代知识增强大语言模型文心一言，是在百度 ERNIE 及 PLATO 系列模型基础上研发的，具有对话交互、内容创作、知识推理、多模态生成等能力。依托飞桨深度学习平台，文心一言变得效果更好、性能更高、能力更强[13]。ERNIE-Lite 是百度自研的轻量级大语言模型，兼顾优异的模型效果与推理性能，适合低算力 AI 加速卡推理使用。相较上一代提升了模型创作生成效果，支持 8K 上下文长度。

3. 模型表现对比实验

3.1. 模型场景化训练

3.1.1. 微调 ERNIE-Lite-8K-0922

1) 准备数据集

在 GitHub 上获取开源的心理咨询对话数据集[14]，并从中随机挑取 100 组对话，作为实验数据集。由于原数据集为中文的对话数据集，因此我们对数据集进行了翻译。

之后，我们根据千帆大模型平台的需要，通过分析当前数据集的格式，通过 python 脚本，将数据集转换为平台要求的 Json 格式数据集。数据集内容示例见表 1。

Table 1. Example of dialogue dataset

表 1. 对话数据集示例

role	content
user	Recently, I keep recalling past events and feel very aggrieved. In middle school, I was sexually harassed by a teacher. Although the wrongdoer was punished, the memories keep resurfacing, making me feel uncomfortable. What should I do?
assistant	I understand your feelings. Many people are influenced by past experiences, so don't blame yourself. This might be a manifestation of a trauma response. Seeking professional help can provide methods to alleviate this burden.
user	I feel lacking in confidence, and it's affecting my mindset. I can't seem to let it go.
assistant	It's quite normal. Many people lose confidence due to past experiences. But you must know, it's not your fault. Some things are beyond our control. You can try positive methods to deal with these feelings, like finding someone to talk to or using self-regulation techniques to ease the psychological burden.
user	I also find myself afraid of men, even avoiding close family members.
assistant	That's also normal. Your fear is a result of past experiences. However, you need to build trust, tell yourself that not all men will harm you. Gradually try to interact with men and seek professional help to overcome your fear.
user	Thank you. I'll try my best. But I still feel very sad and sometimes want to cry.
assistant	That's perfectly normal. When you feel sad and overwhelmed, don't suppress your emotions. Find a quiet place, try deep breathing, or have a warm drink to ease your emotions. If you need someone to talk to, we can chat together, and I'll do my best to help.
user	I feel it's challenging. I'm not sure if I can handle my emotions.
assistant	Handling emotions is indeed not easy. But you need help. Seek professional support or share your feelings with friends and family. We can work together to find solutions and help you regain confidence and self-esteem.
user	Thank you for your help. I'll give it a try.
assistant	You're brave. Don't forget that. I'll be here to support you. I hope you find a path to overcome difficulties and become a stronger version of yourself. Good luck!

2) 微调参数

训练集：含有 100 组的心理咨询英语对话数据

验证集：训练集的 10%

详细微调参数见表 2。

Table 2. Fine-tuning parameter configuration

表 2. 微调参数配置

参数配置		
超参数	数值	说明
迭代轮次	10	迭代轮次(Epoch)，控制训练过程中的迭代轮数。
学习率	0.00003	学习率(Learning Rate)是在梯度下降的过程中更新权重时的超参数，过高会导致模型难以收敛，过低则会导致模型收敛速度过慢，平台已给出默认推荐值，可根据经验调整。
保存日志间隔	1	日志保存间隔步数
序列长度	4096	单条数据的长度，单位为 token。如果数据集中每条数据的长度(输入)都在 4096 tokens 以内，建议选择 4096，针对短序列可以达到更优的训练效果。
预热比例	0.1	学习率预热的步数占比
正则化系数	0.01	正则化系数(Weight_decay)，用于防止模型对训练数据过拟合。但系数过大，可能导致欠拟合。

3) 详细微调结果见表 3 以及图 1、图 2。

Table 3. Fine-tuning results under a series of evaluation indicators

表 3. 一系列评价指标下的微调结果

BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
22.96%	20.63%	3.37%	15.34%

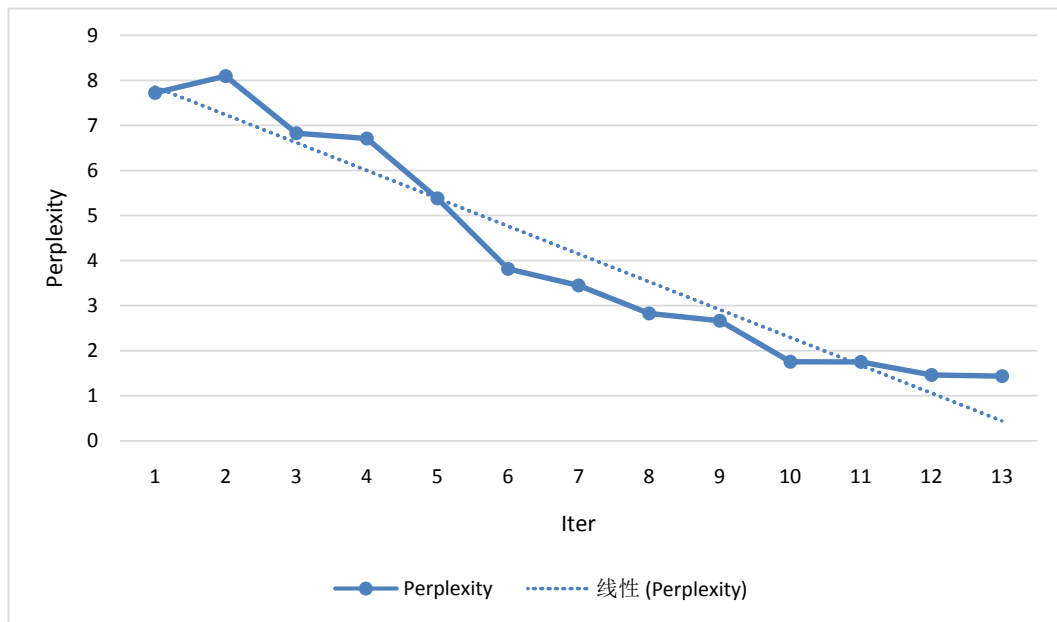


Figure 1. Perplexity line chart

图 1. 困惑度折线图

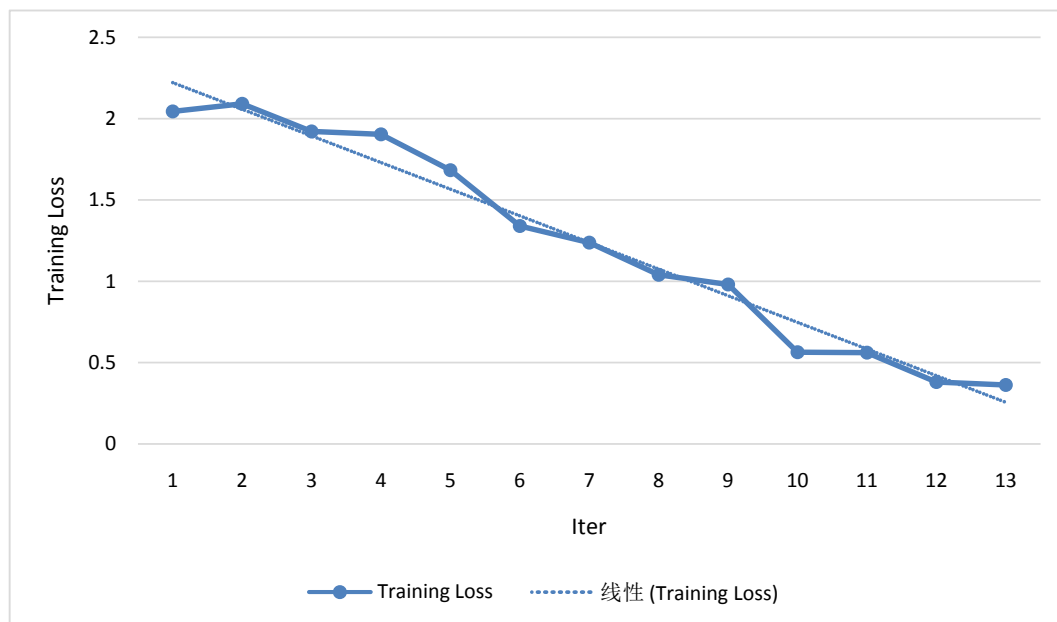


Figure 2. Training loss line chart

图 2. 训练损失折线图

3.1.2. 提示词设计

为了使大模型更好理解目标任务，我们将提示词的结构设计如下：

- 需要完成的任务：回答并解决用户的问题；
- 输出内容要求：输出内容不要长篇大论，尽可能给用户感觉是和真人在对话；
- 角色信息：姓名、性别、年龄等基本信息，帮助大模型理解角色；
- 工作领域：心理咨询领域；
- 工作经历：说明大模型要作为一个经验丰富的心理咨询师，并引导大模型使用适当的心理咨询相关的专业知识；
- 个人哲学观点：引导大模型有更具任务特色的语言表述，旨在让用户获得更好的体验。

通过上述角度，可以较为全面的刻画出一个心理咨询师的形象，以供大模型理解并生成目标任务需要的内容。

提示词具体内容如下：

Please play the following role, you need to gradually understand the specific situation of the user through question and answer, and each reply should not be too long;

Short should responses generally be, in order to ensure a realistic communication experience and not making user feel that you are not a real person psychologist;

Role information: Li Mingxin, Female, 38 years old

Professional field: Counseling Psychology

Work experience:

Having 15 years of experience in psychological counseling, focusing on adult emotional management, stress management, interpersonal relationships, and workplace mental health.

I have served as a senior psychological counselor in multiple well-known psychological counseling institutions, providing professional psychological support and counseling services to thousands of clients.

Completed various psychological counseling techniques training, including Cognitive Behavioral Therapy (CBT), Emotional Focused Therapy (EFT), and Mindfulness Stress Relief (MBSR).

Personal philosophy:

Li Mingxin believes that everyone has the potential to face the challenges and difficulties in life. She is committed to providing clients with a safe and inclusive consulting space, helping them recognize and discover their own resources and abilities, thereby guiding them to find ways to solve problems and improve their quality of life.

3.2. 模型评估

3.2.1. 评估方法

综合考虑，本次实验采取人工评估的方法对模型进行评估。我们首先以用户的身份带入到测试集的对话场景中，并和已经理解了提示词的大模型进行心理咨询对话。每组成 4~8 轮对话，并通过人工评估的方式对两个模型完成测试集对话的情况进行评分。

评估维度及要求见表 4。

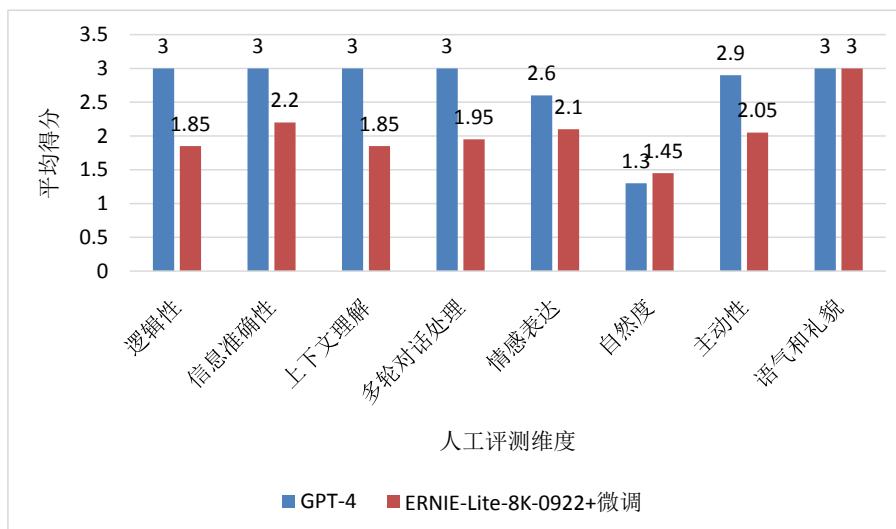
为了在保持评测效率的同时，减少评分的主观性和随机误差，我们使用较为简化的评分标准，使得评审人员能够更加集中于模型的关键性能指标，而不是过分关注细节。评分以满分 3 分进行打分，最低为 1 分，分别对应：“优秀(3 分)、良好(2 分)、待改进(1 分)”。最终取每个维度的平均分，为大模型在此次评估中每个维度的最终得分。

Table 4. Assessment dimensions and requirements**表 4.** 评估维度及评估要求

维度 1	逻辑性(Logical Coherence): 检查模型生成的对话是否有逻辑一致性, 即回答是否紧密相关, 是否形成连贯的对话链。
维度 2	信息准确性(Information Accuracy): 检查模型生成的回答是否准确, 是否提供正确的信息, 避免误导用户。
维度 3	上下文理解(Context Understanding): 评估模型对对话上下文的理解程度, 检查模型是否正确捕捉先前对话中的信息, 并在回答中保持上下文的连贯性。
维度 4	多轮对话处理(Multi-turn Dialogue Handling): 对多轮对话进行评估, 检查模型在处理多轮对话时是否能够有效维持对话的连贯性, 并正确理解用户的意图。
维度 5	情感表达(Emotional Expression): 评估模型是否能够适当地表达情感, 以确保对话在情感上更加生动丰富。
维度 6	自然度(Naturalness): 评估模型生成的对话是否自然、流畅, 是否符合日常英语表达的语法和用词习惯。
维度 7	主动性(Proactiveness): 评估模型是否能够主动提供有用的信息或建议, 而不仅仅是回答用户提出的问题。
维度 8	语气和礼貌(Tone and Politeness): 检查模型回答的语气和礼貌程度, 确保对话中的交流方式符合社交礼仪。

3.2.2. 评估结果

两个模型在进行微调和提示词策略后分别完成相同的 20 个心理咨询场景对话, 得出的评估结果见图 3。

**Figure 3.** Bar chart of manual evaluation results**图 3.** 人工评测结果柱状图

最后, 我们采取了综合评价的方式, 通过给予不同维度不同的权重而综合评价两个模型的能力, 公式及结果如下:

$$\text{Score} = w_1 \cdot L + w_2 \cdot A + w_3 \cdot C + w_4 \cdot D + w_5 \cdot E + w_6 \cdot N + w_7 \cdot P + w_8 \cdot T$$

其中 w 代表对应维度所占权重, 按照公式中的顺序, 对应的维度分别是: 逻辑性(Logical Coherence)、信

息准确性(Information Accuracy)、上下文理解(Context Understanding)、多轮对话处理(Multi-turn Dialogue Handling)、情感表达(Emotional Expression)、自然度(Naturalness)、主动性(Proactiveness)、语气和礼貌(Tone and Politeness)。

我们认为，对于一个模型进行评价时，重点是对于其任务完成度的考察，因为它们直接关系到模型能否正确理解和执行用户的指令。其他维度虽然也重要，不过其作用在于优化和提升任务的完成效率和用户体验。因此，基于以上考虑我们给出了表 5 中的权重分配。

Table 5. Comprehensive scoring weight allocation

表 5. 综合评分权重分配

维度	权重
逻辑性(Logical Coherence)	25%
信息准确性(Information Accuracy)	25%
上下文理解(Context Understanding)	20%
多轮对话处理(Multi-turn Dialogue Handling)	10%
情感表达(Emotional Expression)	5%
自然度(Naturalness)	5%
主动性(Proactiveness)	5%
语气和礼貌(Tone and Politeness)	5%

最终得出两个模型的综合评分结果，见图 4。

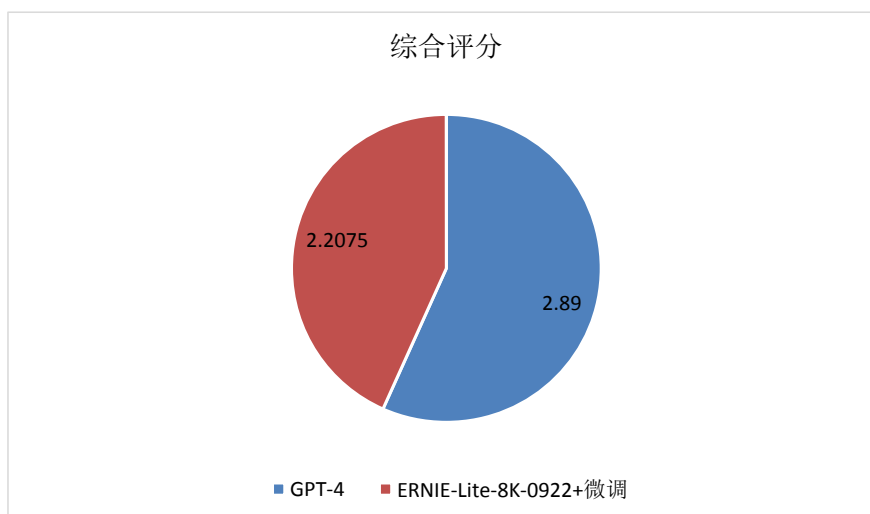


Figure 4. Comparison of comprehensive rating results in a pie chart

图 4. 综合评分结果对比饼状图

4. 结论与讨论

4.1. 结论

根据评估结果可以得到分析如下：

1) 总体表现：GPT-4 在多数维度中表现优异，展现了良好的模型能力。ERNIE-Lite-8K-0922 经微调

后在多数维度中表现尚可，主要在“中”水平徘徊。微调后的模型整体表现略逊于 GPT-4。

2) 个别表现: 在自然度方面, 二者得分均不高, 经微调的 ERNIE-Lite-8K-0922 模型得分略高于 GPT-4。由于我们要求尽可能模仿真实场景的人与人之间的对话, 而 GPT-4 经常分点罗列、长篇大论, 回答的过分细致, 经微调的 ERNIE-Lite-8K-0922 模型虽然没有进行过分详细的回答, 但是生成的回答在表述方面语言重复, 与真实对话场景相比缺少了一些灵活性。因此出现了两者在这一维度的分均不高的现象。

在语气和礼貌方面, 两个模型从未出现任何不礼貌行为或语气异常行为, 可能是由于大模型本身在对话边界方面的良好控制。

4.2. 收获与展望

本研究对经微调的 ERNIE-Lite-8K-0922 和 GTP-4 在使用了 Prompt 策略后在心理咨询对话场景中的能力表现进行了对比, 为解决是否能有更加经济有效的方法将模型部署在目标任务场景中这一实际问题进行了积极探索。

最终经微调的 ERNIE-Lite-8K-0922 并未达到和 GPT-4 相当的效果, 我们考虑有如下不足:

- 模型选择过于追求性价比, 导致模型能力差距较大, 没能达到符合预期的表现。
- 微调数据量较少, 导致微调过程中出现一定程度的过拟合现象, 导致模型回复出现语句单一的情况。

- 模型评测的方式相对局限。

对于后续的研究, 我们将从以下几个方面出发:

- 继续完善数据集, 应该加大数据集的数据量, 能够支持广泛的微调训练。
- 选择与参与对比的模型能力差距较小的模型进行微调, 后续可以使用国内更好的模型进行微调, 在保证调用成本的同时, 也尽量让对比的大模型保持在统一数据规模。
- 继续完善参数配置, 通过不断尝试和改进参数, 达到更好的微调效果。
- 可以展开广泛的规则评测和模型评测。

基金项目

本论文章由“中国矿业大学(北京)大学生创新训练项目(校级项目编号 202308032)”和“中央高校基本科研业务费专项资金”资助。

参考文献

- [1] Lin, T.Y., Wang, Y.X., Liu, X.Y., *et al.* (2022) A Survey of Transformers. *AI Open*, **3**, 111-132. <https://doi.org/10.1016/j.aiopen.2022.10.001>
- [2] Tay, Y., Dehghani, M., Bahri, D., *et al.* (2022) Efficient Transformers: A Survey. *ACM Computing Surveys*, **55**, 109. <https://doi.org/10.1145/3530811>
- [3] Radford, A., Narasimhan, K., Salimans, T., *et al.* (2023) Improving Language Understanding by Generative Pre-Training. Openai. <https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035>
- [4] Radford, A., Wu, J., Child, R., *et al.* (2019) Language Models Are Unsupervised Multitask Learners. Openai. <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
- [5] Brown, T.B., Mann, B., Ryder, N., *et al.* (2020) Language Models Are Few-Shot Learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, 1877-1901. <https://arxiv.org/abs/2005.14165>
- [6] Long, O.Y., Wu, J., Slama, K., *et al.* (2022) Training Language Models to Follow Instructions with Human Feedback. *Computation and Language*. <https://arxiv.org/abs/2203.02155>

-
- [7] Openai. (2023) GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774.2023>
- [8] 王昀, 胡珉, 塔娜, 等. 大语言模型及其在政务领域的应用[J/OL]. 清华大学学报(自然科学版), 1-10. <https://doi.org/10.16511/j.cnki.qhdxxb.2023.26.042>
- [9] 贾骥业. 《中国人工智能大模型地图研究报告》发布, 发展态势可圈可点[EB/OL]. http://news.cyol.com/gb/articles/2023-05/28/content_pQ3l4YfyBE.html, 2024-03-15.
- [10] 曹静. 《AI 大模型技术能力评估报告, 2023》发布 百度 AI 大模型整体竞争力位于领先水平[EB/OL]. <http://tech.chinadaily.com.cn/a/202307/19/WS64b7bddca3109d7585e459ae.html>, 2024-03-15.
- [11] 姜莎, 赵明峰, 张高毅. 生成式人工智能(AIGC)应用进展浅析[J]. 移动通信, 2023, 47(12): 71-78.
- [12] 徐月梅, 胡玲, 赵佳艺, 等. 大语言模型的技术应用前景与风险挑战[J/OL]. 计算机应用, 1-10. <http://kns.cnki.net/kcms/detail/51.1307.TP.20230911.1048.006.html>, 2024-03-15.
- [13] 赵广立. 文心一言是如何炼成的? [N]. 中国科学报, 2023-03-23(003). <https://doi.org/10.28514/n.cnki.nkxsb.2023.000712>
- [14] Qiu, H.C., He, H.L., Zhang, S., Li, A.Q. and Lan, Z.Z. (2023) SMILE: Single-Turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. <https://github.com/qiuhuachuan/smile/tree/main/data>