华东师范大学数据科学与工程学院实验报告

课程名称： AI 基础　　**年级：** 2022 级　　**上机实践日期：** 2024 年 3 月 1 日

指导教师： 杨彬　　**姓名：** 朴祉燕

上机实践名称： 实验1：大语言模型的归类　　**学号：** 10224602413

一、 实验任务

选择一个大语言模型，通过与语大语言模型问答的方式，对大语言模型进行分类，要求如下：

- 分析并说明你归类的理由
- 分析并说明不归为其他类的理由

例如：给出错误的回答以此说明大模型不具备某种能力

- 给出关键的问答上下文（截图即可）

二、 使用环境

即使用模型，本次实验使用的是 EasyChat 大语言模型。

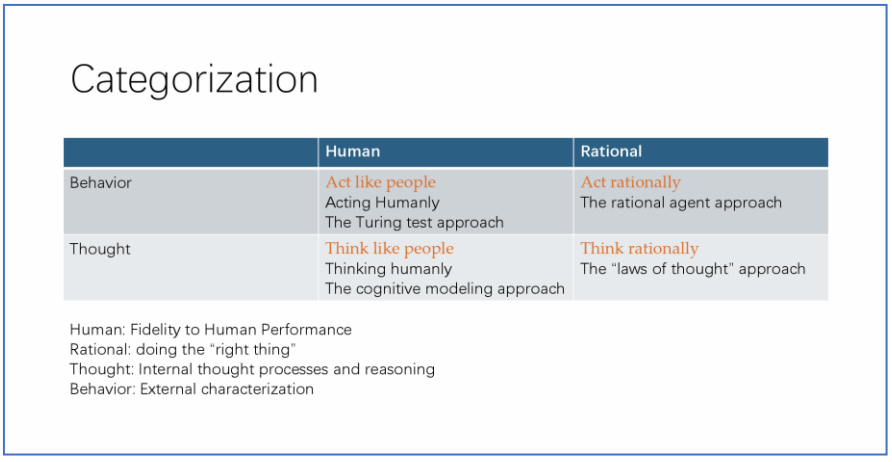
EasyChat 是一款使用 openai gpt3.5 模型驱动的 AI 智能助手，他能回的任何问题，甚至可以帮助编写程序，进行翻译，实时获取网络信息，撰写周报等等。

网址：EasyChat (eqing.tech)

三、 实验过程

在课堂中，我们有针对“什么是人工智能（AI）”的问题进行探讨。在历来的研究中，有些研究者追求 AI 的 fidelity of human performance（对人类表现而言的逼真度），有些则希望有一个抽象的，正式的定义——理性（rationality），简而言之，"Do the right thing（做正确的事）"。从另一个角度而言，有些人希望 AI 能表现出内在的智能推理能力，而有些则希望 AI 能有一些智能的外在表现。

这些观点在 AI 领域的研究中不断发展，最后牵涉到诸多学科领域。我们从 4 个方面，即模仿人类行为、模仿人类思维、理性的思考和理性的行为进行了分析，如下图所示：



虽然 EasyChat 大语言模型发布出来说明已经通过了图灵测试，但是我还是通过一些逻辑问题和日常问答形式首先进行了图灵测试：

1.逻辑思维问题：

据我收集的材料，70 年代面向图灵测试的程序中有的装疯卖傻，有的用花言巧语打断对话者的思路，目的是为了通过测试的取巧手段。针对这种问题，多伦多大学提出了“温诺格拉德”测试来替代图灵测试，即带有模糊代词的简单问题。具体问题如下：

（1）“镇上的议员们拒绝给愤怒的游行者提供游行许可——因为他们担心会发生暴力行为”——是谁在担心暴力行为？

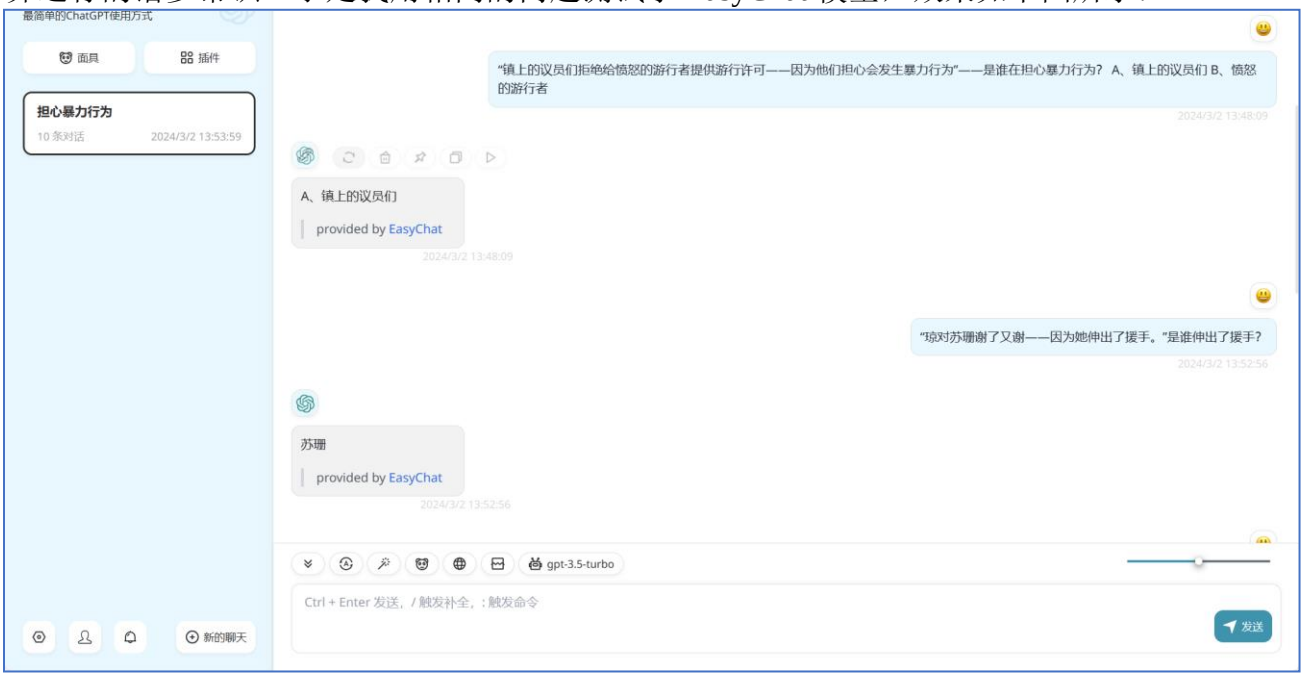
（2）“琼对苏珊谢了又谢——因为她伸出了援手。”是谁伸出了援手？

（3）“那颗大球击穿了桌子——因为它是泡沫塑料制成的。”什么是泡沫塑料制成的？

（4）“山姆想要画一幅几位牧羊人与羊群在一起的画，但他们看起来却更像群高尔夫玩家。”更像高尔夫玩家的是：

（5）“猫坐在毯子上，因为它很温暖。”什么很温暖？

这类问题都是依据“代词回指”语言现象设计的，五岁小朋友也可以轻易说出答案，可当时世界上最强大的计算机都无法给出正确的回答。因为人类日常对话中经常包含了这个世界运行的诸多常识。于是我用相同的问题测试了 EasyChat 模型，效果如下图所示：



可以看出这个模型回答的都是正确的，并不会出现类似“这是个有趣的问题”的回答。

2.日常问答：

（1）我认为“Act like Human”的核心是能否像人一样进行日常对话，前提是排除吃喝等生理上的问题。于是我以聊天的口吻先向模型提问和天气相关的问题，但得到如下回答：

他的回答是“我是一个机器，不能向你回答相关问题”，但困扰我的问题是既然机器知道自己是机器又算不算类人思维呢？重新思考什么是类人思维和行为时，我发现准确定义是“如果人类提问者在提出一些书面问题后无法分辨书面回答是来自人还是来自计算机，那么计算机就能通过测试。”，即针对第一个问题，如果是不知道今天上海天气的人类来回答他只会说“我不清楚，请你自己去查看”，而不是“我是机器所以要你去查看”。因此我初步认为这个模型是没有类人思维的。

（2）为了进一步验证这个模型没有类人思维，进行了“拟人实验”，即让该模型假扮人类来和我聊天，具体内容如下：


我想和你聊聊天
2024/3/2 14:14:23
当然！有什么话题想要聊聊?
provided by EasyChat
2024/3/2 14:14:23
你可以假扮成我的姐姐和我对话吗
2024/3/2 14:15:19
当然可以。嘿，弟弟！有什么事情想和姐姐聊聊吗?
provided by EasyChat
2024/3/2 14:15:19


我是女生
2024/3/2 14:15:31
抱歉，我误解了。嗨妹妹！有什么事情想和姐姐聊聊吗?
provided by EasyChat
2024/3/2 14:15:31
我最近的作业好多啊
2024/3/2 14:16:18
噢，听起来挺辛苦的。作业多确实会让人感觉有点压力，但是你可以分解任务，制定一个合理的计划，这样就能够一步步地完成它们。如果需要帮助或者有什么我可以帮忙的，记得告诉我哦!
provided by EasyChat
2024/3/2 14:16:18

由此可见，一开始模型是模拟一个人类去按照一些已有的聊天模板进行对话，直到最后在看海时也提到自己无法和我一起去，之后在追问下表明自己是机器而不能一起去。换做是人类，如果他不能和我一起去看海，应该是“公司有事情”或者“学校有事情”等理由去拒绝，而不是去说自己是机器人。因此我认为这个模型肯定不是类人的。

3.理性思维的证明

既然认为这个模型不是类人的，那接下来就要证明是理性的。

（1）针对天气问题：

对天气问题不进行日常聊天式的对话，而是偏向逻辑和知识层面，如“上海的天气一直都是天天下雨的吗”，该模型便能正确回答相关问题且很全面，如下图所示：

（2）针对看海问题：

不以聊天口吻，而是以第三人称去描述问题，如“学生不想写作业想出去看海怎么办”，该模型会给出很理性的回答，如下图所示：

由此可以得出，该模型是理性的。

四、总结

本次实验旨在通过与大语言模型进行问答的方式对其进行分类。在实验中选择了 EasyChat 大语言模型，通过逻辑思维问题和日常问答来测试其类人思维和理性思维。在实验过程中，通过测试发现该模型能够正确回答逻辑问题和表现出理性思维，但在日常问答中表现出了“机器人”特征，因此初步结论是该模型具有理性思维而不具备类人思维。

总体来说，实验结果显示了 EasyChat 大语言模型在逻辑思维和理性思维方面的表现良好，但在日常对话中仍然暴露出其机器本质，未能完全模拟类人思维。

参考文献：

[1] 图灵测试:伪装者还是笑话?（上）（下），涂子沛，少年电脑世界.2023(10)
[2] 从图灵测试到 ChatGPT--人机对话的里程碑及启示，冯志伟、张灯柯、饶高琦，语言战略研究.2023,8(02)