深圳技术大学本科毕业论文（设计）

开题报告

题　　目	大语言模型评估平台的对比与评测				
学生姓名	黄思研	学号	202001020107	专业	计算机科学与技术
学　　院	大数据与互联网学院	指导教师	黄炳顶		

本选题的意义及国内外发展状况：

1、本选题意义：

大语言模型，英文名 Large Language Model[1]，常缩写为 LLM，是一种人工智能模型。这种模型的主要目标是理解和生成人类语言。为了实现这个目标，大语言模型需要在大量的文本数据上进行训练，以学习自然语言的各种模式和语义。根据 2023 年 11 月新华社研究院中国企业发展研究中心的大语言模型报告显示，OpenAI 在 2022 年 11 月发布了 GPT3.5，仅过了三个月就推出了 GPT4，参数规模大幅增长。百度也在 2023 年 6 月发布了文心一言 3.5，仅四个月后又推出了文心一言 4.0，实现了基础模型的全面升级。其他厂商的产品也在不断升级迭代。

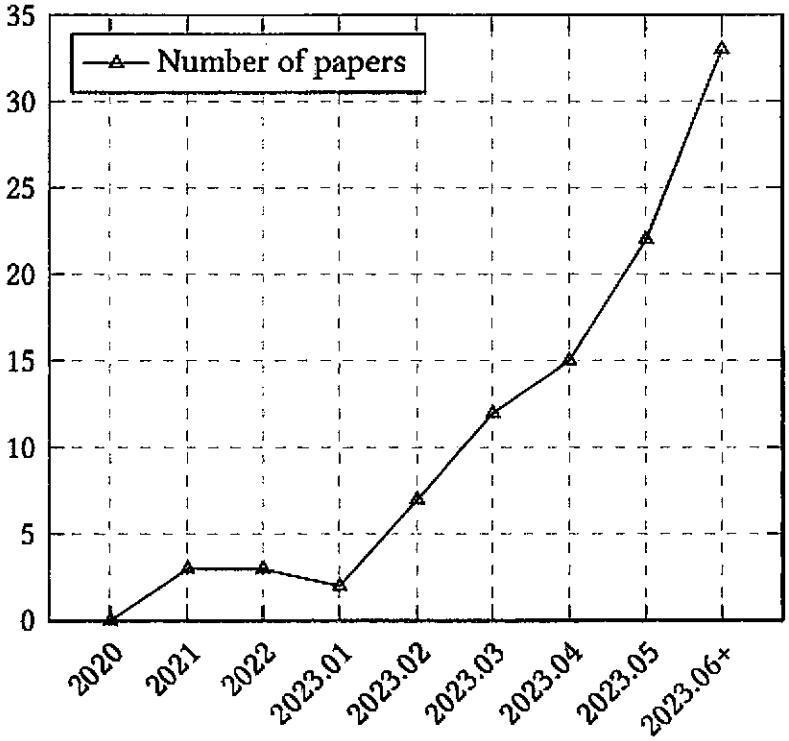
但大语言模型在快速发展的同时也面临着一些问题。首先，大语言模型存在不稳定性[2]，对于相同的提示词，其回答可能存在前后不一致的情况。其次，大语言模型仍然具有一定的局限性，会出现“幻觉”事实并犯推理错误。最后，大语言模型在模型训练、线上推理服务、安全测试等方面需要持续加强安全合规能力。同时市场需求不断变化，ChatGPT 的功能不断地得到深度挖掘，如今在各个领域的内容创作和行业细分领域深度结合的应用越来越多。市场对于大语言模型的需求也随着用户所在行业的变化而不断变化。因此大语言模型评估测试集和平台对于检测、优化、解决大语言模型的不稳定性与个性化微调大语言模型以满足市场需求的不断变化至关重要。

大语言模型产品的初期评测阶段，主要基于小规模问题集进行评测[3]。然而，这种方式也存在一些明显的缺陷。包括但不限于有限的覆盖范围，小规模问题集通常只包含少量的样本，无法全面覆盖大语言模型的各种应用场景和任务类型；可靠性问题，大语言模型有时会犯事实性错误，也就是所谓的“幻觉”，这种可靠性问题在小规模问题集中可能不容易被发现；无法全面评估，小规模问题集评测主要关注模型的逻辑推理，翻译和沟通能力等方面，无法全面评估模型的高级知识和推理能力；缺乏与下游任务的融合，大语言模型的评估不仅仅是一个终点，而小规模问题集评测可能无法满足这一需求。

Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

图 1 早期大语言模型评测方法，大多为判断题和选择题等[4]

也正因大语言模型的发展，而评测方法难以跟进，因此市面上的大语言模型评估数据集和平台开始逐年增加，鉴于不同数据集和平台的侧重点不同以及大语言模型本身存在的不稳定性，对比不同评估数据集和评估平台在不同大语言模型和任务的优劣性是很有必要的。



图二 大语言模型测评方面文章数量逐年增加[5]

大语言模型的评估数据集和评估平台是最近兴起的热门话题，市面上不同侧重的评估方法也较为混杂。现如今出现的评估方法也难以精确满足用户不断变化的需求，因此收集和整理大语言模型评估数据集和评估平台，对比不同评估数据集和评估平台在不同任务上的利弊以便用户可以自由选择和评估大语言模型，更好的满足用户需求。

2、国内外发展状况：

大语言模型的测评是随着大语言模型的不断发展而不断变化的，而语言模型总体来说可以分为三个阶段：基于规则阶段、基于统计学阶段、基于神经网络阶段[6]，其评测方法也同样分为这三个阶段。基于规则的语言模型评估通常涉及对其在特定任务上表现的分析，包括几个关键方面：语法正确性、词汇使用、句法分析准确性、信息提取效果等。如 Roche[7]在 1997 年提出的有限状态自动机在自然语言处理中对这些模型的评估和比较。到了基于统计学阶段，语言模型的评估将结果进行了量化，在评估方法上进行了改进，主要在困惑度、准确率、召回率、语言模型生成、词性标注评估、数据集划分和交叉验证上进行了创新和改进。比如 Joshua T. Goodman[8]在 2001 年提出的 N-gram 语言模型中的平滑技术，并探讨了用困惑度作为评估模型性能的指标。而到了基于神经网络阶段的语言模型，比如 GPT[9]等的出现，也对其评估方法做出了相应的改进。主要体现在人工评估文本生成任务中的质量评估、机器翻译任务中的 BLEU[10]分数评估、文本相似度评估、生成多样性评估和专家或志愿者对模型生成的文本的正确性进行评估。如由 Alec Radford[11]在 2018 年提出的对 GPT 模型的性能评估方法主要聚焦于模型在语言建模任务和下游自然语言处理任务上的表现。困惑度作为语言建模的评价指标，反映了模型对文本的理解程度，而在下游任务上的性能评估则更直接地衡量了模型在实际应用中的效果。虽然该篇论文主要集中于展示预训练模型相对

于传统方法的性能提升，但其评估方法也对后续的大语言模型开发和评估都展现了不小的启示。

而关于评估大语言模型的数据集，目前使用较多的包括 G-Eval、M3KE、MMCU、CMExam、Panda LLM 和 SuperCLUE 等数据集。其中，G-Eval 是一个全面评估大型中文语言模型在各个学科领域生成能力的评估，模型的性能是根据它们对六个学科中不同类型问题的准确和相关回应的能力来评估的；M3KE 是一个大规模的多级且多主题的知识评估基准，通过测试模型在零和少数设置中的多任务精度来衡量其知识获取能力；MMCU 数据集首先使用了 3331 个高考多选题来衡量模型对世界的基本理解，然后使用了 2819 个、3695 个和 2001 个多选题来评估模型在医学、法律和心理学等专业领域的专业知识；CMExam 是一个全面的中文医学考试数据集，用于评估大型语言模型；Panda LLM 项目专注于通过指令调整来增强开源大型语言模型的性能，并提供全面评估；SuperCLUE 是一个全面的中文大语言模型基准测试，包括开放式问题、多选问题和与人类评估并行的比较。

研究内容：

1、数据集方面，拟收集目前大语言模型评估方法使用较多或较为经典的数据集。目前拟采用 G-Eval、M3KE、MMCU、CMExam、Panda LLM 和 SuperCLUE 等数据集。

2、大语言模型评估平台（方法）；拟据 7 种较为常用的 LLM 评估方法：困惑评估、BLEU（双语评估替补）、ROUGE、METEOR（显式排序翻译评估指标）、人体评估、多样性评估、零样本评估，并比较不同方法的区别和注重。

3、安装并部署较为经典的大语言类模型，并借助不同数据集和评估方法对模型进行评估和比对。

4、测试不同开源大语言模型在不同评估数据集和评估平台的性能参数，比对模型的准确性、流畅性、一致性、困惑性及人工评估。

5、编写 Linux 的 shell 脚本，实现对不同评估数据集和评估平台的测评的自动化。

研究方法、手段及步骤：

研究方法：

1、收集大语言模型评估的资料并阅读相关文献；

2、搜索学习不同的大语言模型评估方法，对比传统方法的优劣；

3、部署大语言模型并且使用不同的数据集和评估方法对开源大语言模型进行评测；

4、探索大语言模型中的评估方法的创新及运用。

研究手段：主要包括前期调研、选定评估实现方案、搭建实验环境等。

1、通过阅读文献完成对深度学习、大语言模型评估等理论知识的准备。

2、与导师交流，选定实验环境，安排实验日程。

3、搭建实验环境，完成，使用 PyCharm 等软件编写 Python 的代码。

4、收集并统计测评结果，测试不同开源大语言模型在不同评估数据集和评估平台的性能参数

5、总结研究进展和问题，解决构建和评估模型过程中碰到的问题。

研究步骤：

起止时间	工作规划
2023.12.01-2023.12.31	收集整理大语言模型评估数据集和评估平台
2024.01.01-2024.01.15	部署不同开源的大语言模型在服务器上

2024.01.16-2024.01.31	测试大语言模型在不同评估数据集和评估平台的性能参数
2024.02.01-2024.02.29	建立起对不同评估数据集和评估平台的评测方法，撰写论文初稿
2024.03.01-2024.04.01	撰写并修改毕业论文，制作答辩 PPT

参考文献：

[1] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A.S. (2023). A Comprehensive Overview of Large Language Models. ArXiv, abs/2307.06435.

[2] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023). A Survey of Large Language Models. ArXiv, abs/2303.18223.

[3] Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., & He, J. (2023). C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In Advances in Neural Information Processing Systems.

[4] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In T. Linzen, G. Chrupa, & A. Alishahi (Eds.), Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (pp. 353-355). Association for Computational Linguistics.

[5] Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. ArXiv, abs/2307.03109.

[6] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and Applications of Large Language Models. ArXiv, abs/2307.10169.

[7] YLI-JYRÄ, A., KORNAI, A., & SAKAROVITCH, J. (2011). Finite-state methods and models in natural language processing. Natural Language Engineering, 17(2), 141-144. doi:10.1017/S1351324911000015

[8] Joshua T. Goodman,A bit of progress in language modeling,Computer Speech & Language,Volume 15, Issue 4,2001,Pages 403-434,ISSN 0885-2308,https://doi.org/10.1006/csla.2001.0174.

[9] Brown, T.B., Mann, B.,Ryder,N,Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A.,Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv, abs/2005.14165.

[10] Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[11] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

学生签名：黄思研

2023年 12月 6日

指导教师意见：

同意开题！

签名：

学院领导意见：

同意

签名：

2023年 12月 7日