香港中文大学（深圳）李海洲/王本友教授招收大模型方向博士生（2024 FALL & 2025）、研究助理、工程师和博后

1 团队介绍

1.1 博士生导师

1.1.1 李海洲

李海洲，香港中文大学（深圳）数据科学学院执行院长、校长讲座教授，新加坡工程院院士，教育部长江学者,同时他也是新加坡国立大学客座教授和德国不来梅大学卓越讲座教授。他曾于2006年至2016年分别担任新加坡南洋理工大学和新加坡国立大学教授，于2009年担任东芬兰大学客座教授，于2011年至2016年任澳洲新南威尔士大学客座教授，于2003年至2016年担任新加坡科技研究局通信与资讯研究院首席科学家和研究总监。李教授是IEEE Fellow、ISCA Fellow、AAIA Fellow，曾任顶级期刊IEEE/ACM Transactions on Audio、Speech and Language Processing主编(2015-2018年）。他也曾是多个国际大型学术会议的大会主席，包括ACL 2012、INTERSPEECH 2014、ICASSP 2022。

主页 https://colips.org/~eleliha/

1.1.2 王本友

王本友，香港中文大学（深圳）助理教授。主要研究方向为自然语言处理（NLP）、应用机器学习、和信息检索。迄今为止，他曾获得了SIGIR 2017最佳论文提名奖、NAACL 2019最佳可解释NLP论文、NLPCC 2022最佳论文、华为火花奖、腾讯犀牛鸟项目和欧盟玛丽居里奖学金。领导开发了华佗GPT大模型和阿拉伯语大模型AceGPT，华佗GPT是首个通过当年国家药剂师考试的模型，迄今有超过40万次访问；AceGPT是发布时最好的阿拉伯语大模型。

主页 https://wabyking.github.io/old.html

2 关于研究

2.1 研究方向

- 大模型的产业化：将大数据模型融入生活的各个方面，以提升工作效率和生活质量，包括在医疗领域上的应用（参见我们的华佗GPT）和教育领域，让大模型技术实质地改善公众生活。在有大量开源大模型的基础上，如何将大模型在实际场景的应用的最后一公里打通。在医疗大模型方面，将会将其扩展到agent，构建高质量RAG数据集，可信和隐私计算。

- 大模型的平民化：将大规模语言模型变得更加可用，包括提高训练和部署效率、降低数据收集的门槛等；这部分最近将会探索端侧大模型的加速和应用。另外一方面，本课题还在做人机交互相关的研究，可以减少终端用户和大模型之间的距离。

- 多模态大模型和AI for Science的研究：语言模型处理不同模态的理解和生成任务，包括但不限于图片、音频、表格、代码以及视频等，目前已经开始在做多模态大模型，特别期待有

做CV或者语音方向的候选人可以参与到多语言大模型相关研究上来。最近希望把多模态大模型迁移到Science的场景。

- 多语言大模型的相关研究，探索大模型中的多语言处理机理，高效地将语言和对应的知识在不同模型之间迁移（参加AceGPT）。同时研究多语言里面的价值观对齐和文化多样相关的研究。

2.2 已有研究示例

我们开发有多个大模型，请扫描二维码来体验我们的模型，也可以通过在微信搜索“神仙湖”小程序体验。

(a) Medical LLM

2.3 研究哲学

本团队的科研崇尚“发现新现象、定义新问题，设计新范式”，研究哲学如下：

- 研究团队秉承“Less is more”的科研理念，做简单且有效的工作;
- 做有影响力的研究，论文作为展示我们工作的一种方式（而不是目的），做有趣或者有用的论文，力争为社区做贡献；
- 做减熵的研究，让从业人员更加豁然开朗，而不是更加困惑；
- 做通用且有深度的研究;
- 做工程、产品和科研并起的研究，用科研手段去解决工程和产品解决不了的问题。

2.4 本组代表论文

华佗**GPT** 其中公开资料显示，我们二月份发布的华佗GPT是首个国内类ChatGPT的医疗大模型，2023年2月份香港中文大学（深圳）副校长和深圳市大数据研究院院长罗智泉院士2023年2月份在中华医院信息网络大会CHINC发布的华佗GPT。目前线上已经有四十万次访问量。第二个版本的华佗GPT参数达72B，在2023年的10月份的最新药剂师考试（因为时间太新，所以很难有数据泄露问题）是唯一个及格的模型，大幅领先文心一言、GPT-4。最新华佗GPT参数达72B，也扩展到了多语言和多模态场景。

阿拉伯语大模型**AceGPT**，本团队和KAUST合作开发沙特的阿拉伯语大模型(AceGPT)，发布时是当时最好的阿拉伯语大模型，并在一个评测平台上超过ChatGPT，受到英国著名金融时报报道，论文被NAACL 2024接受。团队目前正在用上千块GPU芯片训练多语言大模型。

数学推理验证器**OVM** 我们提出的验证器模型，即"结果监督价值模型"（OVM），采用结果监督进行训练，为规划提供了一种高效直观的方法，通过优先考虑那些能导致准确结论的步骤，而非单纯关注每一步的正确性。此外，OVM摒弃了对步骤级正确性进行劳动密集型标注的需求，提升了其可扩展性。在GSM8K数据集中，我们的OVM-7B模型在不使用GPT-4或代码执行的情况下，实现了13B参数以下LLMs中的最佳性能。相关方法使用OVM将7B模型在GSM8K数据集取得了0.95的准确率，超过了GPT 4。

MathScale (ICML 2024)，这是一种利用前沿大型语言模型（如GPT-3.5）创建高质量数学推理数据的简单且可扩展方法。该方法启发于人类数学学习的认知机制，首先从种子数学问题中提取主题和知识点，再构建概念图，用以生成新的数学问题。通过这种方式，我们成功创建了一个包含两百万数学题目-答案对的数学推理数据集（MathScaleQA）。为全面评估大型语言模型在数学推理能力上的表现，我们构建了MWPBENCH基准，这是一个包括GSM8K和MATH在内的十个数据集的集合，覆盖了从K-12到大学及竞赛级别的数学问题。将MathScaleQA用于对开源大型语言模型（例如LLaMA-2和Mistral）进行微调，显著提高了它们在数学推理上的能力。在MWPBENCH上的评估显示，MathScale-7B在所有数据集上均达到了最先进的性能，其微平均准确率和宏平均准确率分别比同等大小的最佳模型高出42.9%和43.7%。

自监督学习的理论理解 **(ICLR 2024)** 研究自监督学习中的表示塌陷问题，提出一种新方法来衡量特征分布的均匀性。通过分析特征在单位超球面上的分布，发现遵循特定分布的特征表现出较好的均匀性。文章引入Wasserstein距离来量化学习到的特征与理想分布之间的差异，提出的方法不仅满足理想的均匀性标准，还能有效解决维度塌陷问题。最后，将这种均匀性作为辅助损失加入自监督学习，有效提升了下游任务的表现。

其他相关论文见：

- 面向更长上下文的多模态模型评测 MileBench: Benchmarking MLLMs in Long Context https://arxiv.org/abs/2404.18532
- 多语言医疗模型Apollo: Apollo: Lightweight Multilingual Medical LLMs towards Democratizing Medical AI to 6B People https://arxiv.org/abs/2403.03640
- 合成数据提升数学推理能力的MathScale: Scaling Instruction Tuning for Mathematical Reasoning https://arxiv.org/abs/2403.02884 (ICML 2024)
- 大模型和人同时学习的新人机交互范式： Online Training of Large Language Models: Learn while chatting https://arxiv.org/abs/2403.04790
- 自监督的学习中的Uniformity： Rethinking The Uniformity Metric in Self-Supervised Learning. ICLR 2024
- 小而好的多模态大模型 ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model https://arxiv.org/abs/2402.11684
- 人和大模型做大模型生成评估器的偏见分析: Humans or LLMs as the Judge? A Study on Judgement Biases https://arxiv.org/abs/2402.10669
- 更精细的多模态大模型的评估: MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria. https://arxiv.org/abs/2311.13951

3 招聘要求

香港中文大学（深圳）李海洲和王本友教授拟招收4名博士生（包括1名2024秋季，三名2025年）、2名研究助理、3名工程师和3名博后。研究方向如上所介绍。我们团队还招收相关研究方向的硕士生、访问学生/学者。欢迎计算机、自然语言处理、语音处理、数据科学、机器学习、信息检索等相关背景的同学申请。

3.1 全职和兼职工程师招聘

招聘对象：在读学生或在职人员，其致力于大模型落地应用，实验室支持其创业，支持参与论文发表，具体发展方向看个人兴趣。

兼职工程师和实习生，可以推荐其到国内大厂（BAT和和华为），深圳市大数据研究院就职，推荐其申请本校硕博。

可以获得千卡训练经验，实习和工作结束后可以在作为技术领导人，成为时代潮流的弄潮儿。

工作内容：算法落地、数据工程师、算法后端、UI和产品设计师。领导支持和参与一流项目的开发，孵化未来创业项目。包括但是不限于医疗大模型的产品设计、大模型落地项目和大模型端侧硬件产品开发。

全职和兼职工程师请发到 wuxiangbo@cuhk.edu.cn

3.2 博士生招聘

3.2.1 基本要求

招生对象：直博（本科生）、有意读博的硕士研究生（采用申请制，无需考研）

1. 本科或以上学历（申请制无需考研），学校背景请参考往年录取学生（https://sds.cuhk.edu.cn/student-search），具有计算机科学等相关专业背景者优先考虑；
2. 在校期间成绩优异，有扎实的数学基础，良好的英语阅读、写作能力以及编程能力；
3. 有大规模语言模型、对话系统、文本生成、多模态相关研究经验并在国内外知名会议或期刊有相应论文者优先考虑；
4. 热爱科研工作，富有责任心，良好的沟通能力、敬业精神和团队协作精神；
5. 在香港或者英语国家取得相当的学位或修学证明；托福（笔试不低于550分，机试不低于213分；网考不低于79分）；雅思（学术类不低于6.5分）。
6. 希望博士生对工作充满热情、热爱技术，并具有良好的团队协作和沟通能力，可以自我驱动。

3.2.2 博士生的实际要求

本组博士生分为下面两种:

1. 有良好的教育背景并GPA在前5%-10%的本科生
2. 有基本研究能力，例如在ICLR/NeurIPS/ICML/ACL/EMNLP/NAACL（或同等水平的期刊会议）发表有一作论文的候选人；优先本组内实习一年以上的研究助理，表现出研究潜力和领导力

3.2.3 申请程序

博士生申请流程如下：

1. 请准备好完整的中文/英文简历（请附上成绩和论文发表记录）发送至团队成员王熙栋邮箱：wangbenyou@cuhk.edu.cn，邮件标题注明：博士申请+本人姓名+感兴趣研究方向；
2. 面谈了解候选人相关背景、研究能力及研究兴趣；

3. 学院录取流程。

重要时间：

1. 招生入学时间：2024 年 9 月（春季），2025 年 1 月（春季），2025年 9月（秋季）

2. 招生截止日期：2024年秋季入学还有少量名额，对特别优秀的同学开放。2025秋季截止时间不限，越早越好。

香港中文大学（深圳）博士申请细节参见：

Ph.D. in Computer Science: https://sds.cuhk.edu.cn/en/phd-programmes-CSE
Ph.D. in Computer and Information Engineering: http://sse-mphil-phd.cuhk.edu.cn/en/basic/249

博士项目优势如下：

- 本课题组已有博士生曾去或即将去斯坦福、普林斯顿、美国创业公司和MSRA实习和访问。
- 本课题组和深圳大数据研究院罗智泉院士（中科院外籍院士）、万翔教授有联合培养博士生项目，就读期间享受丰厚的奖学金。
- 本课题组和KAUST有联合培养博士生项目，毕业可获得香港中文大学和沙特KAUST双博士学位（2024年秋季入学还有名额）。
- 本课题组还与瑞典Halmstad University的Prayag Tiwari教授有联合博士生项目，可以4-5年毕业（五年毕业可拿到当地绿卡），读博期间可以每年在牛津大学和港中深各访问两个月，开销全部由项目cover，毕业拿瑞典Halmstad University的学位。
- 学校实行与北美相同的教学体制，学制五年（有硕士学位可以四年），毕业授予香港中文大学学位，毕业可获两年IANG在港就业签证；开放的国际校园，良好便利的生活、学习设施；优秀学生可申请优厚补助；
- 由于博士生培养昂贵（每年学费和生活费接近20W），最大程度上让博士生关注学术问题，减少一些琐事干扰（专门有人负责）；因为培养经费过高且类似北美的funding机制，博士生较少参与横向项目；工程项目有工程师负责。
- 一流博士导师，前沿博士课题，科研氛围浓厚，团队内自由合作交流没有顾忌；欢迎和国内外同行保持合作，支持国内外开会交流；
- 博士生在能够独立发表一作顶会论文后，鼓励其前往国外实习访问和交流。学生不做老师的附庸，独立成为业界翘楚。

3.3 研究助理（Research Assistant）

3.3.1 基本要求

研究助理

1. 本科或以上学历（在读硕士生、在读博士生均可申请）。

2. 本科应聘者，原则上应在组内工作半年及以上；本校本科生应有充足的时间投入。

3. 硕士学历应有相关研究经验，发表有NeurIPS/ICLR/ICML/ACL/EMNLP/NAACL等会议。

4. 热爱科研工作，富有责任心，具有团队协作精神；请自我驱动。

5. 对于非本校学生，原则上倾向招聘有留下继续硕博的潜力和意愿的应聘者；本科学校应有较好的背景，其GPA在排名30%以内，且高于3.0/4；否则难以在学院通过学院的硕博选拔

6. 鼓励拿到本校硕士offer的同学入学前提前来组内实习。

7. 有论文发表记录可推荐去MSRA、腾讯、微信、华为、百度、上海AI Lab等地方实习（包括本校硕士生）。

研究助理表现优秀者，可申请攻读香港中文大学（深圳）数据科学学院硕士或博士学位；申请研究助理请发送简历到huyan@cuhk.edu.cn

3.4 博后

3.4.1 基本要求

基本要求如下：

1. 需要完成博士答辩；
2. 具有计算机科学、数学等相关专业背景者优先考虑；
3. 鼓励自由探索，做世界一流工作。

3.4.2 工作职责，薪酬及福利

工作职责，薪酬及福利如下：

1. 从事科研工作并撰写科研论文；
2. 具有竞争力的薪酬，根据个人资历和经验而定；博士后年薪30W起；
3. 博后可以协助指导团队内的本科生、硕士生和博士生，支持下一站学术生涯；
4. 留学回国人员符合条件者可申请人才计划；
5. 数据科学学院博士后出站后大多入职国内985和顶级211，不少获得终身教职。

申请博后请发送邮件到wangbenyou@cuhk.edu.cn。

3.4.3 申请程序

申请程序如下：

1. 请准备好完整的中文/英文简历（请附上成绩和论文发表记录）发送至团队成员王熙栋博士邮箱：wangxidong06@gmail.com。邮件标题注明：应聘职位+本人姓名；
2. 合适者将接受团队成员面试，了解候选人背景、能力及研究兴趣；
3. 本职位空缺有效期截止到招聘到合适人选为止。

3.5 自费研究型硕士(Mphil)招聘

要求如下

1. 本科学校应有较好的背景，其GPA在排名30%以内，且高于3.0/4；也欢迎硕士学历申请，包括跨专业硕士
2. 有良好的研究动机，建议提前来本校实习RA
3. 可与南方科技大学荆炳义教授联合培养，基本可以cover学费。
4. 优先推荐在国内外大厂实习，包括腾讯AI Lab、华为诺亚方舟实验室、百度NLP实验室（深圳），上海AI Lab等。
5. 独立发表有顶会论文，可以申请转博（Mphil修读课程可以转为博士学分）。
6. 优秀学生可以强推国外学校。

2024年秋季还有少量名额，香港中文大学（深圳）Mphil申请细节参见：

Mphil. in Computer Science: https://sds.cuhk.edu.cn/en/phd-programmes-CSE

申请博后请发送邮件到huyan@cuhk.edu.cn。

4 相关信息

4.1 学校及团队优势

- 组内有纯研究用的上千块GPU卡，可积累千卡训练经验。组内多人有千卡训练经验，可独当一面，将持续为业界和学界输送技术和学术领导人。欢迎未来的技术领导人加入本组，探索大模型的魅力。
- 与一流业界知名公司/研究机构等紧密合作，拥有丰富的计算资源，努力不让计算资源限制了想象力，充分发挥人的创造力；目前团队有充足的A100算力、海量的数据等研究资源，同时拥有大量的对外部署资源，希望把研究内容可以通过在线演示网站的形式展示出来；
- 团队优势是兼有活力的年轻PI和资深的PI，可以提供多方面多层次的指导。
- 支持毕业去学术界/工业界发展；殷切地希望培养出独立的博士生，毕业后可以独当一面，具备在一流大学成为独立PI的潜质；承办国内外相关领域的顶级会议，支持学生承担会议sevices，助力在学术界和工业界的早期事业起步。
- 重视培养学生的全方位科研能力，包括口头报告、写作、团队写作能力、领导力、研究品位等等，不锚定静态的生产力指标（比如CCF A类会议论文中稿数量），着重提升学生的创造力和潜力，力求让学生每篇论文比上一篇论文的质量更好。
- 欢迎团队成员向业界发展，并帮忙联系投资人，开展创业的尝试。
- 团队的国际视野强，和北美、欧洲、中东、新加坡、港澳联系紧密，培养面向全球的技术领导者和学术新秀。

4.2 背景介绍

香港中文大学（深圳）是一所经国家教育部批准，传承香港中文大学的办学理念和学术体系的大学。目前，来自世界各地的 8000 多名优秀学子正在港中大（深圳）求学。经过九年的发展，大学学科建设已逐步完善，已面向全球招聘引进了 400 余名国际知名优秀学者和研究人员，其中包括诺贝尔奖得主 5 名，图灵奖得主 2 名，菲尔兹奖 1 名，各国院士近 30 名（其中全职 10 名），国家级特聘专家近 60 名，ACM/IEEE 等协会会士近 40 名。目前引进的教师 100% 具有在国际一流高校执教或研究工作经验。大学已经连续六年成为广东省内院校中录取分数最高的大学，毕业生颁发香港中文大学的学位证。香港中文大学（深圳）在CS ranking中排名内地高校第八名（https://csrankings.org/#/fromyear/1970/toyear/2024/index?all&cn），鉴于学校年轻AP较多，未来发展空间巨大。

香港中文大学（深圳）数据科学学院师资团队现有62人（含兼职15人），其中校长学勤讲座教授4人、校长讲座教授10人、教授12人、副教授11人、助理教授23人、讲师1人。师资中包括加拿大皇家科学院和中国工程院外籍双院士1人、加拿大皇家科学院和加拿大工程院双院士1人、新加坡工程院院士1人、运筹学与管理学研究协会INFORMS会士1人、国际电气与电子工程学会IEEE会士4人、国际系统与控制科学院IASCYS 会士1人、国际数理统计学会IMS会士2人、美国工业与应用数学学会SIAM会士2人、国际语音通信学会ISCA会士1人、国家级高层次人才9人、省市区高层次人才7人。目前数据科学学院已录取的博士生均来自知名大学，如清华大学、北京大学、上海交通大学、浙江大学、南京大学、中国人民大学、中国科学技术大学、武汉大学、同济大学、香港中文大学（深圳）、南方科技大学等。本科专业包括数学与应用数学、计算机科学、信息科学、工业工程等。其中，42%的学生本科成绩排名为专业前5%，80%的学生本科成绩排名为专业前10%。