Introduction to Department & Strategic Plans

DEPARTMENT OF Computer Science and Engineering (CSE)

Prof Xiaofang Zhou
5 April 2024
HKUST – CSE rankings

World’s Top 700 Universities in Computer Science
(#1 in Hong Kong for 11 times in 13 years)
QS World University Rankings 2023

No. 40

Times Higher Education World University Rankings (2024)

(retrained the No. 1 place in Hong Kong for 8 years in a row)
No. 31

CS Rankings (2024)

No. 18
CSE FACTS AND FIGURES

Established in 1991

**Total student enrolment** (as of Sep 2023)
- Undergraduate (CS majors + joint programs): 1,058
- Postgraduate Taught: 259
- Postgraduate Research (PhD: 398): 480

**Total regular faculty** (as of Jan 2024)
- Professor: 30
- Associate Professor: 9
- Assistant Professor: 11
- Teaching focused: 7
- Research Assistant Professor: 6
- Postdoc/Research Associate: 29
- Research Assistant: 37
- Non-academic: 24
CSE Faculty

Current: 50
Target: 65
Female: 5 (10%)
HK background: 12 (24%)
Mainland background: 27 (54%)
Other background: 11 (22%)
CSE Research Themes and Areas

Artificial Intelligence (AI)
- Artificial Intelligence (AI)
- Vision and Graphics (VG)
- Natural Language Processing (NLP)

Data Science (DS)
- Data, Knowledge and Information Management (DKIM)
- Human-Computer Interaction (HCI)
- Theoretical Computer Science (TH)

Systems (SY)
- Cybersecurity (SEC)
- Software Engineering and Programming Languages (SEPL)
- Networking and Computer Systems (NE)
HKUST CSE
UG and PG Programme

Undergraduate Programs

Major Programs
- BEng in Computer Science (COMP)
- BSc in Computer Science (COSC)
- BEng in Computer Engineering (CPEG)
- BSc in Data Science and Technology (DSCT)
- BSc in Risk Management and Business Intelligence (RMBI)
- Dual Degree Program in Technology and Management
- Extended Major in Artificial Intelligence (AI)
- Extended Major in Digital Media and Creative Arts (DMCA)
- Additional Major(s)

Minor Programs
- Minor Program in Big Data Technology (for non-IT Minor students)
- Minor Program in Information Technology (for non-CSE students)

Postgraduate Programs

Research Programs
- Master of Philosophy (MPhil)
- Doctor of Philosophy (PhD) Programs

Taught Programs
- Master of Science (MSc) Program in Big Data Technology
- Master of Science (MSc) Program in Information Technology

UG
- From school-based admission to department-based admission
- New BEng in AI program from Fall 2025

TPG
- Increased number of students in MSc IT and MSc BDT from Fall 2024
- New MSc IT concentration in Cybersecurity from Fall 2024
- New MSc in AI from 2025

(2022)
CSE Computing Infrastructure

Individual Research Group Servers pool for PG Students:

- Total no. of servers: 124 nodes
- Total no. of CPU: 248 nodes
- Total no. of GPUs: 700 nodes
- Total storage: 3 PB
- Total memory: 40 TB

Departmental Servers:

1. GPU Cluster for teaching:
   - 7-node cluster
   - 2 x 10 cores Intel Xeon Gold 5115 (2.4GHz) processors
   - 256GB physical memory
   - Inter-connected with 10Gb Ethernet
   - Nvidia RTX 2080Ti GPUs x 56
   - OS: Linux

2. General GPU server for UG students:
   - 7 servers with a total of 32 GPUs
   - (Nvidia GeForce RTX 2080/3090/3090Ti)

CSE Labs
- UG teaching labs x 4
- PG Lab. x 7
- PG server room
- PG Communal Lab x 2
- Special lab.
- RPg Hub

CSE Computing Infrastructure
Many Area-based Labs

GPU server (Dell DSS 8440) x (5+2)
- 2 CPU, 8 GPU A100 80GB, 512GB RAM, 54TB SSD
- 2x8 RTX4090 with 24GB

CPU server (Dell PowerEdge R940) x 6
- 4 CPU, 1.5TB RAM, 26TB HHD

Storage server (Dell PowerEdge R740xd2) x 2
- 2 CPU, 256GB RAM, 1.6TB SSD + 480TB HDD

* New strategies to establish centralized computing infrastructure at HKUST
HK Generative AI (HKGAI)

An InnoHK project led by HKUST

- With HKU, CUHK, PolyU, CityU, plus NUS
- To build open-source foundation models
- Vertical applications in legal, medical and creative arts areas, plus X (finance, telecommunications...)
HKGAI Establishment

Our Establishment

• Established in October 2023 with government funding ~USD100 million
• A new InnoHK R&D Center, one of the important initiatives in AI committed by HK Government
• Included in the Chief Executive’s 2023 Policy Address Policy Measure

Our Position

• Hong Kong as an international innovation and technology hub
• A cornerstone to support Hong Kong AI ecosystem
• Promote innovation & collaboration & application
• Shape future efficiency, intelligence and interconnection
• Talent cultivation and retention

The Chief Executive's 2023 Policy Address

Policy Measures

2023.10.25

Promoting R&D

• Establish the Hong Kong Microelectronics Research and Development Institute within 2024 to lead and facilitate the collaboration among universities, R&D centres and the industry on the R&D and application of microelectronics, including joint exploration of the third-generation semi-conductor core technology, as well as fully leveraging the well-developed manufacturing industry chains and enormous market in the GBA. (ITIB)

• Make preparations for the establishment of the third InnoHK research cluster which focuses on advanced manufacturing, materials, energy and sustainable development, with a view to expanding our world-class R&D collaboration and enhancing the R&D development of Hong Kong. (ITIB)

• Establish a new InnoHK R&D Centre specialising in R&D in generative AI technology, and conduct studies on the appropriate rules and guidelines for the application of AI technology. (ITIB)
Timing is Everything

**International Dynamics**
- **Aug, 2022**: US banned NVIDIA from selling of A100 and H100 GPUs to China-based customers without a license
- **Feb - Mar, 2023**: ChatGPT first reached 1 billion visits in 2 months
- **Apr, 2023**: GPT4 released by OpenAI and LLM
  - Became a Phenomena
- **May, 2023**: News on US further ban on NVIDIA’s sale of A800 and H800 GPUs to China
- **June, 2023**: July 13, 2023

**Project Interview**

**Our Work**
- **Kickoff Meeting**
- **Summary Submission**
- **Further Discussion**
- **Proposal Preparation**
- **Developing HKGPT v0.1**
US Government restricts exports of NVIDIA’s China-Exclusive H800 & A800 AI GPUs To China

OpenAI released text-to-video model Sora in Feb, 2024.

Master Agreement
- Signed Master Agreement with ITC, project approval
- NVIDIA H800 GPUs arrived in Hong Kong

Budget Approved
- ITC approved the budget for HKGAI
- NVIDIA H800 GPU Cluster installed
- LLM training project started while building the learning infra
- First 7B Model delivered
- 7B Based MoE will finish training

HKGPT v1.0 will be released
- Applications to release:
  - HK GPTs Store (with end user applications)
  - HK Gov Co-Pilot
  - HK Legal AI

Our Work
- HKUST ordered 1016 NVIDIA H800 GPUs
What we have done

- **Data**: 3T token + HK govt data + official media data, etc
- **Infra**: 1000+ NVIDIA H800 fully connected and operational
- **LLM**: finished the whole training process of delivery a validation model
- **Human capital**: 70+ members on-site + remote
- **Application**: GPTs Store, Govt co-pilot, etc
- **AI Compliance**: advising HK Government on AI Governance

- Open-source state-of-the-art FM
- Meet diverse local and national needs
### Alpha Version

<table>
<thead>
<tr>
<th>Language</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>600</td>
</tr>
<tr>
<td>Chinese</td>
<td>300</td>
</tr>
<tr>
<td>Code</td>
<td>300</td>
</tr>
<tr>
<td>Other</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1300</strong></td>
</tr>
</tbody>
</table>

#### B tokens

- **English**: red-pajama v2/slimpajama/dolma/refinedweb
- **Chinese**: CWP/CCI/skyPile/Wanjuan-text/Yayi/Wudao/ChineseWeb/TeleChat
- **Code**: Starcoder1
- **Other**: Math (Math23K, etc)
B tokens

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>1800</td>
</tr>
<tr>
<td>Chinese</td>
<td>600</td>
</tr>
<tr>
<td>Code</td>
<td>900</td>
</tr>
<tr>
<td>Other</td>
<td>500</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3800</td>
</tr>
</tbody>
</table>

English: red-pajama v2/slimpajama/dolma/refinedweb
Chinese: CWP/CCI/skyPile/Wanjuan-text/Yayi/Wudao/ChineseWeb/TeleChat
Code: Starcoder-2
Other: Papers, math, Arxiv, wiki, Math23K, NSSD (after OCR), FLAN, COIG
Efficient Edge LLM

➢ The Research Goal:
  i. Meeting the Size, Weight, Power Constraints (SWaP) of edge AI for LLMs

➢ The Research Topic:
  i. Quantization of LLM
  ii. Efficient finetuning of LLM
  iii. Context pruning for LLM reasoning
  iv. Model/hardware architecture co-design and co-optimization

➢ Some of these research tasks are part of his InnoHK ACCESS centre’s research agenda
Prof. Tim Cheng’s Research

Specific Projects on LLM

LLM Quantization (published in EMNLP 2023):
- Introduce LLM-FP4, a post-training quantization framework which for the first time is capable of quantizing both the activation and weight of LLM to 4 bits without substantial loss in accuracy, outperforming previous methods by up to 13.1%.

Context Pruning for LLM Reasoning (submitted for publication):
- Propose CoT-Influx, a context pruner trained with reinforcement learning which automatically removes redundant tokens to incorporate more high-quality examples in prompt.

Efficient Finetuning of LLM (submitted for publication):
- Introduce DoRA, a new parameter-efficient fine-tuning approach, which consistently outperforms LoRA in fine-tuning LLM without incurring additional inference cost.
Significant progress and found wide application in various fields, like ChatGPT

The success and efficiency of DL models depend on **proper data management**

Training deep learning-based image classifiers is challenging

i. Without properly labeled data

ii. Efficiency is hindered by large datasets, complex models, and numerous hyperparameters

iii. Lack of validation and explanation limits model applicability

**In the later session — Prof. Chen will discuss more:**

1) Effective data preparation for DL, including extraction, integration, labeling and selection

2) DL training optimization, involving data compression and computation graph optimization

3) Importance of model explanation for robustness and transparency

4) Demonstrate the important industry collaboration and the future research directions
Large Language Models (LLM)

- Enterprises are exploring the use of these models to help detect software faults and vulnerabilities in their systems.
- LLMs have demonstrated promising results in popular coding benchmarks like HumanEval.

**In the later session — Prof. Cheung will discuss more:**

- Using LLMs for test generation and briefly introduce recent research efforts.
Dr. Junxian He’s Research

Large Language Models (LLM)

- Such as: Reasoning, Data selection, and Evaluation
- He works on:
  1. Improving the complex reasoning abilities of LLMs
  2. how to select the best data to train LLMs
  3. how to evaluate LLMs in general

C-Eval - Very first evaluation benchmark for Chinese LLMs

- Developed by Dr. He and his students in 2023 May
- The most commonly used Chinese benchmark for LLMs
- Dataset has been downloaded over 1 millions times in less than a year
- Played an important role during the explosive development of LLMs in China over the past year
Highlights of CSE Faculty and Alumni

Prof Qiang YANG
Co-founder
Chair Professor
MPhil, CSE

Mr Jianxiong XIAO
Founder, CEO
MPhil, CSE

Mr Tat LEE
Chief Executive
and Executive Director
BEng (Comp Sci)

Welcome to Dr. Bingsheng He's

Professor,
Vice Dean (Research), NUS School of Computing,
Department of Computer Science,
National University of Singapore
Office: 205-02-12, COM3 Building, 11 Research Link, NUS,
Singapore 119076 (Texas)
Ph.D., NUS, 2009
Email: hes@comp.nus.edu.sg

Jianliang Xu
Head and Professor
Department of Computer Science (Mac) Hong Kong Baptist University
Kowloon Tong, Hong Kong
Tel: +852 3411 9800  Fax: +852 3411 7992
Email: jlxu@comp.hkbu.edu.hk

Highlights of CSE Faculty and Alumni

Professor Qiang YANG
Co-founder
Chair Professor
MPhil, CSE

Mr Jianxiong XIAO
Founder, CEO
MPhil, CSE

Mr Tat LEE
Chief Executive
and Executive Director
BEng (Comp Sci)

Welcome to Dr. Bingsheng He's

Professor,
Vice Dean (Research), NUS School of Computing,
Department of Computer Science,
National University of Singapore
Office: 205-02-12, COM3 Building, 11 Research Link, NUS,
Singapore 119076 (Texas)
Ph.D., NUS, 2009
Email: hes@comp.nus.edu.sg

Jianliang Xu
Head and Professor
Department of Computer Science (Mac) Hong Kong Baptist University
Kowloon Tong, Hong Kong
Tel: +852 3411 9800  Fax: +852 3411 7992
Email: jlxu@comp.hkbu.edu.hk

Barbara ZHONG
PhD 2005
Associate Professor, Singapore Management University

Nicole Yat-Ming SHIP
PhD 2010
Professor, Chinese University of Hong Kong

Jianliang Xu
Head and Professor
Department of Computer Science (Mac) Hong Kong Baptist University
Kowloon Tong, Hong Kong
Tel: +852 3411 9800  Fax: +852 3411 7992
Email: jlxu@comp.hkbu.edu.hk

12 university professors from this group
(4 this year, UC Davis, Minnesota x 2, Texas A&M)
## Research Funding in the Past Five Years

<table>
<thead>
<tr>
<th>Category</th>
<th>#Grants</th>
<th>HK$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGC-GRF</td>
<td>119</td>
<td>96.6M</td>
</tr>
<tr>
<td>RGC/UGC Others</td>
<td>6</td>
<td>20M</td>
</tr>
<tr>
<td>ITF/External Industry</td>
<td>54</td>
<td>129M</td>
</tr>
<tr>
<td>Contract research</td>
<td>103</td>
<td>82.5M</td>
</tr>
<tr>
<td>International funding bodies</td>
<td>54</td>
<td>26.5M</td>
</tr>
<tr>
<td>Internal</td>
<td>256</td>
<td>39M</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>393.6M</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Large-scale, collaborative research:

- **Systems (SY):** One RGC - Theme-based Research Scheme grant (iSING Lab, 2020), and one RIF - Research Impact Fund (2021).
- **Data Science (DS):** One CRF - RGC - Collaborative Research Fund (2019-2022) and has one ongoing RIF project (2020).
- **Artificial Intelligence (AI):** One RIF project (2022) and one NSFC-RGC grant (2021).

* New strategies for more large-scale projects

(August 2023)
Joint Research Labs with Companies
What’s Important to CSE

1) Research
   • RAE (next round 2026) and impact stories
   • Funding, especially RGC large grants
   • New research areas
   • Collaboration – (i) inside CSE/HKUST, (ii) internationally, (iii) with industry

2) Teaching
   • New UG and PG programs: BEng in AI and MSc in AI
   • Enhanced TPG programs: more students, more industry collaboration, and a new cybersecurity concentration
   • Student quality
   • Student satisfaction

3) Faculty
   • Quality, size, and diversity, in all areas with priority in AI and cybersecurity
   • More teaching-track faculty and industry lecturers