Zhengzhong Liu

CONTACT INFORMATION

Email: hectorzliu@gmail.com
Personal Website: https://hunterhector.github.io/

WORKING EXPERIENCES

Head of Technology, MBZUAI Institute of Foundation Model (IFM)
Director, MBZUAI IFM Silicon Valley Lab **Nov, 2024 — Present**

- I am leading the foundation model training efforts at the MBZUAI IFM. Our key projects include Jais, the best Arabic language model, and the LLM360 model series, including K2, the largest and best fully open source models.
- I help found and direct the IFM Silicon Valley lab, a new lab focusing on advancing AI, with emphasis on the future of Foundation Models
- I also help build the IFM's technology strategies, including aligning the research efforts globally across the Silicon Valley, Paris and Abu Dhabi labs.

Head of Engineering at Petuum **May, 2022 — Nov, 2024**

- I led a team composed of staff and external contributors to conduct full cycle of Large Language Model production, from data collection, model training, finetuning to model release and marketing. We built one fo the first fully open LLM effort: LLM360.
- In the past years. I lead an Engineering team of around 20 engineering staffs and efficiently deliver several products within planned timeline, including full-fledged MLOps system, an foundation model based automation system.
- Also responsible at team building, including hiring and reorganizing the engineering team structure to align with product directions.

Research Scientist at Petuum **May, 2019 — 2022**

- Lead the research and development of advanced Natural Language Processing systems.
- I lead develop **Forte** and **Stave**, two advanced NLP systems driven by modular design principles.

Research Intern at Google **May, 2015 — Aug, 2015**

- Mentors: Edgar González Pellicer *and* Daniel Gillick
- Worked with the Google NLP team the Verb Phrase Ellipsis (VPE) problem, which is used to help improve a task-based dialogue systems.
- Study the multi-stages of the VPE problem, propose several strategies for joint-training and achieve the state-of-the-art performance on two public datasets. - We clean and contribute one of the datasets and release it to the community.

EDUCATION

Carnegie Mellon University **2014**

- Ph.D. in Language Technologies
- Thesis Topic: Diving Deep into Event Semantics

Carnegie Mellon University **2012**

- Master in Language Technologies

The Hong Kong Polytechnic University **2007**

- Bachelor of Science in Computing (with First Class Honors)
- Bachelor of Business Administration with a Major in Management

HONORS AND AWARDS

ACL Best System Demonstration Nomination **2019**

- Our system Texar is selected to be one of the top five system demonstration at ACL.

ACL Outstanding Long Paper Award **2016**

- Awarded to the 10 best paper on ACL, less than 1% of the submitted paper are selected.

Open Source

K2 and K2 Think - Large, SOTA 360-Open LLMs. https://www.llm360.ai/
- Scale up the LLM360 open source effort to 65B and 70B, best model of the same size class.

LLM360 - Fully Open Sourced LLMs https://www.llm360.ai/
- Lead the team to train a few Large Language Models, including Amber 7B and Crystal 7B. Both models show competitive performance against open LLMs released at that time.
- Advocate the LLM360 initiative to encourage fully open source and promote collaborative research.

Forte - A flexible and composable NLP toolkit https://github.com/asyml/forte
- Design a modular system architecture and flexible representations that can be integrated with deep learning systems.
- Lead the development team and implement core functions of the system.

Stave - A general purpose, extensible text annotation system https://github.com/asyml/stave
- Design the main functionalities and interfaces, use an ontology system to enable extensible text annotation.

Texar- A composable toolkit for Text Generation and NLP models https://texar.io
- Involve in project interface design and implementations of several fundamental modules.

OAQA- Open Advancement of Question Answering Systems https://oaqa.github.io/
- Questions Answering systems in High School World History Exam and Alzheimer's disease articles

DBpedia Spotlight
- Develop a collective disambiguation module in project DBpedia Spotlight using Scala and Hadoop.

Publications and Preprints

1. Zhoujun Cheng, Yutao Xie, Yuxiao Qu, Amrith Setlur, Shibo Hao, Varad Pimpalkhute, Tongtong Liang, Feng Yao, **Zhengzhong Liu**, Eric Xing, et al. IsoCompute Playbook: Optimally Scaling Sampling Compute for LLM RL. arXiv preprint arXiv:2603.12151.
2. Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, **Zhengzhong Liu**, Eric P Xing, Zhiting Hu. Vision-G1: Towards General Reasoning Vision-Language Models via Reinforcement Learning. AAAI 2026.
3. Jianshu She, Zonghang Li, Hongchao Du, Shangyu Wu, Wenhao Zheng, Eric Xing, **Zhengzhong Liu**, Huaxiu Yao, Jason Xue, Qirong Ho. PLA-Serve: A Prefill-Length-Aware LLM Serving System. arXiv preprint arXiv:2601.11589.
4. **Zhengzhong Liu**, Liping Tang, Linghao Jin, Haonan Li, Nikhil Ranjan, Desai Fan, Shaurya Rohatgi, Richard Fan, Omkar Pangarkar, Huijuan Wang, et al. K2-V2: A 360-Open, Reasoning-Enhanced LLM. arXiv preprint arXiv:2512.06201.
5. Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. PAN: A World Model for General, Interactable, and Long-Horizon World Simulation. arXiv preprint arXiv:2511.09057.
6. Jianshu She, Xinyue Li, Eric Xing, **Zhengzhong Liu**, Qirong Ho. Linear Steerability in Language Models: When It Emerges and How It Evolves. In Findings of EMNLP 2025.
7. Jianshu She, Xinyue Li, Eric Xing, **Zhengzhong Liu**, Qirong Ho. How Does Controllability Emerge in Language Models During Pretraining? arXiv preprint arXiv:2508.01892.
8. Renxi Wang, Rifo Ahmad Genadi, Bilal El Bouardi, Yongxin Wang, Fajri Koto, **Zhengzhong Liu**, Timothy Baldwin, Haonan Li. AgentFly: Extensible and Scalable Reinforcement Learning for LM Agents. arXiv preprint arXiv:2507.14897.
9. Subham Sekhar Sahoo, Zhihan Yang, Yash Akhauri, Johnna Liu, Deepansha Singh, Zhoujun Cheng, **Zhengzhong Liu**, Eric Xing, John Thickstun, Arash Vahdat. Esoteric Language Models. arXiv preprint arXiv:2506.01928.
10. Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, **Zhengzhong Liu**, Ion Stoica, Eric P Xing, Hao Zhang. Faster Video Diffusion with Trainable Sparse Attention. NeurIPS 2025.

11. Peiyuan Zhang, Haofeng Huang, Yongqi Chen, Will Lin, **Zhengzhong Liu**, Ion Stoica, Eric Xing, Hao Zhang. VSA: Faster Video Diffusion with Trainable Sparse Attention. arXiv preprint arXiv:2505.13389.
12. Yuan Li, **Zhengzhong Liu**, Eric P Xing. Data Mixing Optimization for Supervised Fine-Tuning of Large Language Models. ICML 2025.
13. Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, **Zhengzhong Liu**, Eric P Xing. MegaMath: Pushing the Limits of Open Math Corpora. arXiv preprint arXiv:2504.02807.
14. Jianshu She, Wenhao Zheng, **Zhengzhong Liu**, Hongyi Wang, Eric Xing, Huaxiu Yao, Qirong Ho. Token Level Routing Inference System for Edge Devices. arXiv preprint arXiv:2504.07878.
15. Yonghao Zhuang, Lanxiang Hu, Longfei Yun, Souvik Kundu, **Zhengzhong Liu**, Eric P Xing, Hao Zhang. Scaling Long Context Training Data by Long-Distance Referrals. ICLR 2025.
16. Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, **Zhengzhong Liu**, Hao Zhang. Fast Video Generation with Sliding Tile Attention. arXiv preprint arXiv:2502.04507.
17. Wenhao Zheng, Yixiao Chen, Weitong Zhang, Souvik Kundu, Yun Li, **Zhengzhong Liu**, Eric P Xing, Hongyi Wang, Huaxiu Yao. CITER: Collaborative Inference for Efficient Large Language Model Decoding with Token-Level Routing. arXiv preprint arXiv:2502.01976.
18. **Zhengzhong Liu**, Bowen Tan, Hongyi Wang, Willie Neiswanger, Tianhua Tao, Haonan Li, Fajri Koto, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Liqun Ma, Liping Tang, Nikhil Ranjan, Yonghao Zhuang, Guowei He, Renxi Wang, Mingkai Deng, Robin Algayres, Yuanzhi Li, Zhiqiang Shen, Preslav Nakov, Eric Xing. LLM360 K2: Building a 65B 360-Open-Source Large Language Model from Scratch. arXiv preprint arXiv:2501.07124.
19. Haonan Li, Xudong Han, Zenan Zhai, Honglin Mu, Hao Wang, Zhenxuan Zhang, Yilin Geng, Shom Lin, Renxi Wang, Artem Shelmanov, Xiangyu Qi, Yuxia Wang, Donghai Hong, Youliang Yuan, Meng Chen, Haoqin Tu, Fajri Koto, Tatsuki Kuribayashi, Cong Zeng, Rishabh Bhardwaj, Bingchen Zhao, Yawen Duan, Yi Liu, Emad A Alghamdi, Yaodong Yang, Yinpeng Dong, Soujanya Poria, Pengfei Liu, **Zhengzhong Liu**, Xuguang Ren, Eduard Hovy, Iryna Gurevych, Preslav Nakov, Monojit Choudhury, Timothy Baldwin. Libra-Leaderboard: Towards Responsible AI through a Balanced Leaderboard of Safety and Capability. In Proceedings of NAACL 2025 System Demonstrations.
20. Sukmin Yun, Rusiru Thushara, Mohammad Bhat, Yongxin Wang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, Haonan Li, Preslav Nakov, et al. Web2Code: A Large-Scale Webpage-to-Code Dataset and Evaluation Framework for Multimodal LLMs. NeurIPS 2024.
21. Liping Tang, Nikhil Ranjan, Omkar Pangarkar, Xuezhi Liang, Zhen Wang, Li An, Bhaskar Rao, Linghao Jin, Huijuan Wang, Zhoujun Cheng, Suqi Sun, Cun Mu, Victor Miller, Xuezhe Ma, Yue Peng, **Zhengzhong Liu**, Eric P Xing. TxT360: A Top-Quality LLM Pre-training Dataset Requires the Perfect Blend. https://huggingface.co/spaces/LLM360/TxT360
22. Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, **Zhengzhong Liu**, Eric P Xing, Zhiting Hu. Pandora: Towards General World Model with Natural Language Actions and Video States. arXiv preprint arXiv:2406.09455.
23. Tianhua Tao, Junbo Li, Bowen Tan, Hongyi Wang, William Marshall, Bhargav M Kanakiya, Joel Hestness, Natalia Vassilieva, Zhiqiang Shen, Eric P Xing, **Zhengzhong Liu**. Crystal: Illuminating LLM abilities on language and code. COLM 2024.
24. **Zhengzhong Liu**, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, Eric Xing. LLM360: Towards Fully Transparent Open-Source LLMs. arXiv preprint arXiv:2312.06550.

25. Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, **Zhengzhong Liu**, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. SlimPajama-DC: Understanding Data Combinations for LLM Training. arXiv preprint arXiv:2309.10818.
26. Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, **Zhengzhong Liu**, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, Eric Xing. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149.
27. Jiannan Xiang, **Zhengzhong Liu**, Yucheng Zhou, Eric Xing, Zhiting Hu. ASDOT: Any-shot data-to-text generation with pretrained language models. In EMNLP'22 Findings.
28. Mingkai Deng, Bowen Tan, **Zhengzhong Liu**, Eric P Xing, Zhiting Hu. Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation. In EMNLP'21.
29. Adithya Pratapa, **Zhengzhong Liu**, Kimihiro Hasegawa, Linwei Li, Yukari Yamakawa, Shikun Zhang, Teruko Mitamura. Cross-document Event Identity via Dense Annotation. In CoNLL'21.
30. Han Guo, Bowen Tan, **Zhengzhong Liu**, Eric P Xing, Zhiting Hu. Efficient (soft) q-learning for text generation with limited good data. In EMNLP'21 Findings.
31. **Zhengzhong Liu**, Guanxiong Ding, Avinash Bukkittu, Mansi Gupta, Pengzhi Gao, Atif Ahmed, Shikun Zhang, Xin Gao, Swapnil Singhavi, Linwei Li, Wei Wei, Zecong Hu, Haoran Shi, Xiaodan Liang, Teruko Mitamura, Eric Xing, Zhiting Hu. A Data-Centric Framework for Composable NLP Workflows. In EMNLP'20 (Demonstration).
32. Zhisong Zhang, Xiang Kong, **Zhengzhong Liu**, Xuezhe Ma, Eduard Hovy. A Two-Step Approach for Implicit Event Argument Detection. In ACL'20.
33. Zhiting Hu, Zichao Yang, Haoran Shi, Bowen Tan, Tiancheng Zhao, Junxian He, Xiaodan Liang, Wentao Wang, Xingjiang Yu, Di Wang, Lianhui Qin, Xuezhe Ma, **Zhengzhong Liu**, Devendra Singh, Wangrong Zhu, Eric P Xing. Texar: A Modularized, Versatile, and Extensible Toolkit for Text Generation. ACL'19, *best system demonstration nomination*
34. Zhiting Hu, Zichao Yang, Haoran Shi, Bowen Tan, Tiancheng Zhao, Junxian He, Xiaodan Liang, Wentao Wang, Xingjiang Yu, Di Wang, Lianhui Qin, Xuezhe Ma, **Zhengzhong Liu**, Devendra Singh, Wangrong Zhu, Eric P Xing. 2018. Texar: A Modularized, Versatile, and Extensible Toolbox for Text Generation. In *Proceedings of Workshop for NLP Open Source Software, ACL'18.*
35. **Zhengzhong Liu**, Teruko Mitamura, Eduard Hovy. 2018. Graph-Based Decoding for Event Sequencing and Coreference Resolution. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18).*
36. **Zhengzhong Liu**, Chenyan Xiong, Teruko Mitamura, Eduard Hovy. 2018. Automatic Event Salience Identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18).*
37. Hans Chalupsky, Jun Araki, Eduard Hovy, Andrew Hsi, **Zhengzhong Liu**, Xuezhe Ma, Evangelia Spiliopoulou, Shuxin Yao. 2018. Multi-lingual Extraction and Integration of Entities, Relations, Events and Sentiments into ColdStart ++ KBs with the SAFT System. In *Proceedings of the Text Analysis Conference 2017 (TAC'17).*
38. Teruko Mitamura, **Zhengzhong Liu**, Eduard Hovy. 2018. Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track. In *Proceedings of the Text Analysis Conference 2017 (TAC'17).*

39. Chenyan Xiong, **Zhengzhong Liu**, Jamie Callan, Tie-Yan Liu. 2018. Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*.

40. Chenyan Xiong, **Zhengzhong Liu**, Jamie Callan, Eduard Hovy. 2017. JointSem: Combining Query Entity Linking and Entity based Document Ranking. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM'17)*.

41. Keyang Xu, **Zhengzhong Liu**, Jamie Callan. 2017. De-duping URLs with Sequence-to-Sequence Neural Networks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*.

42. Teruko Mitamura, **Zhengzhong Liu**, Eduard Hovy. 2017. Overview of TAC-KBP 2016 Event Nugget Track. In TAC 2016.

43. Zhiting Hu, Xuezhe Ma, **Zhengzhong Liu**, Eduard Hovy, Eric Xing. 2016. Harnessing Deep Neural Networks with Logic Rules. In *Proceeding of 2016 Annual Meeting of the Association for Computational Linguistics (ACL'16)*.

44. Xuezhe Ma, **Zhengzhong Liu**, Eduard Hovy. 2016. Unsupervised Ranking Model for Entity Coreference Resolution. In *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'16)*.

45. **Zhengzhong Liu**, Edgar González Pellicer, Dan Gillick. 2016. Exploring the steps of Verb Phrase Ellipsis. In *NAACL 2016 Workshop on Coreference Resolution Beyond OntoNotes (CORBON'16)*.

46. **Zhengzhong Liu**, Jun Araki, Teruko Mitamura, Eduard Hovy. 2016. CMU-LTI at KBP 2016 Event Nugget Track. In *Proceedings of the Text Analysis Conference 2016 (TAC'16)*.

47. **Zhengzhong Liu**, Jun Araki, Dheeru Dua, Teruko Mitamura, Eduard Hovy. 2015. CMU-LTI at KBP 2015 Event Track. In *Proceedings of the Text Analysis Conference 2015 (TAC'15)*.

48. Teruko Mitamura, **Zhengzhong Liu**, Eduard Hovy. 2016. Overview of TAC KBP 2015 Event Nugget Track. In *Proceedings of the Text Analysis Conference 2015 (TAC'15)*.

49. **Zhengzhong Liu**, Teruko Mitamura, Eduard Hovy. 2015. Evaluation Algorithms for Event Nugget Detection: A Pilot Study. In *Proceedings of the 3rd Workshop on EVENTS, NAACL-HLT'15*, pages 53–57.

50. Di Wang, Leonid Boytsov, Jun Araki, Alkesh Patel, Jeff Gee, **Zhengzhong Liu**, Eric Nyberg, Teruko Mitamura. 2014. CMU Multiple-choice Question Answering System at NTCIR-11 QA-Lab. In *Proceedings of the 11th NTCIR Conference (NTCIR'14)*.

51. **Zhengzhong Liu**, Jun Araki, Eduard Hovy, Teruko Mitamura. 2014. Supervised Within-Document Event Coreference using Information Propagation. In *Proceedings of The Ninth Edition of International Conference on Language Resources and Evaluation (LREC'14)*.

52. Jun Araki, **Zhengzhong Liu**, Eduard Hovy, Teruko Mitamura. 2014. Detecting Subevent Structure for Event Coreference Resolution. In *Proceedings of The Ninth Edition of International Conference on Language Resources and Evaluation (LREC'14)*.

53. Xu, J., Lu, Q., & **Zhengzhong Liu**. 2012. Aggregating skip bigrams into key phrase-based vector space model for web person disambiguation. In *Proceedings of KONVENS'12*.

54. Xu, J., Lu, Q., **Zhengzhong Liu**, & Chai, Junyi. 2012. Topic Sequence Kernel. In H. Yuexian, N. Jian-Yun, S. Le, W. Bo, & Z. Peng (Eds.), *Lecture Notes in Computer Science*.

55. Xu, J., Lu, Q., & **Zhengzhong Liu**. 2012. PolyUCOMP: Combining semantic vectors with skip bigrams for semantic textual similarity. In *SemEval'12 Proceedings of the First Joint Conference on Lexical and Computational Semantics*.

56. Xu, J., Lu, Q., & **Zhengzhong Liu**. 2012. Combining Classification with Clustering for Web Person Disambiguation. In *Proceedings of the 21st International Conference Companion on World Wide Web (WWW'12)*. Lyon.
57. Xu Jian, **Zhengzhong Liu**, Qin Lu, Yu-lan Liu, & Chenchen Wang. 2011. PolyUCOMP in TAC 2011 Entity Linking and Slot-Filling. In *Proceedings of the Fourth Text Analysis Conference (TAC'11)*.
58. **Zhengzhong Liu**, Lu, Qin, & Xu, Jian. 2011. High Performance Clustering for Web Person Name Disambiguation Using Topic Capturing. In *The First International Workshop on Entity Orientation Search, SIGIR'11*.

Professional Activities

Co-organizer, TAC Knowledge Base Population Event Track 2015-2017
Reviewer, ACL, EMNLP, NAACL, CONLL, NLPCC

Tutorials

Modularizing Natural Language Processing	NVDIA GTC 2020, with Zecong Hu
Modularizing Natural Language Processing	AAAI 2020, with Zhiting Hu and Eric P Xing

References

Eric Xing	President, MBZUAI; Professor, Carnegie Mellon University
Eduard Hovy	Professor,Carnegie Mellon University; University of Melbourne
Teruko Mitamura	Research Professor, Carnegie Mellon University
Vincent Ng	Professor, University of Texas Dallas
Qin Lu	Professor, Computing, The Hong Kong Polytechnic University