# Kaibo Liu

*Curriculum Vitæ*

☎ *(541) 368-8197*
✉ *kaiboliu.me@gmail.com*
*kaiboliu.github.io*
*Google Scholar*
*LinkedIn*

## Education

2016 – 2018 **Oregon State University**, 🎓 *M.Sc.*, Computer Science [GPA: 4.0/4.0].

2010 – 2013 **Peking University, China**, 🎓 *M.Sc.*, Electronics Engineering [GPA: 3.71/4.0].

2006 – 2010 **Peking University, China**, 🎓 *B.Sc.*, Physics [GPA: 3.49/4.0].

## Work Experience

Oct 2022 – Now **Senor Research Scientist**.
Bytedance, San Jose, CA
- Focus on Code LLM and multi-modal pretrained models

A1 **Seed-Coder base: pretraining dataset construction for base code LLM**.
- Built a data deduplication and decontamination pipeline to minimize human involvement in constructing code LLM pretraining data.
- Designed a high-quality code data filtering pipeline using a scoring model (distilled from a $33B$ model across 10 programming languages, generating $150k$ labeled samples for filter training), achieved outstanding $F1$ and $RMSE$, enabling scalable filtering of $1.2TB$ curated code data for pretraining.
- Pretrained a $7B$ model on $2T$ tokens of high-quality code corpus, along with long-context training by supporting sequences of up to $32K$ tokens during the annealing phase, outperformed same-scaled $DeepSeek$-$Coder$ model on $HumanEval$ and $MBPP$, demonstrating superior data quality impact.

A2 **Seed-Coder instruct and reasoning: post-training for code LLM**.
- Developed supervised fine-tuning (SFT) frameworks, to support flexible deployment and to adapt our base models to various real-world tasks, followed by Direct Preference Optimization (DPO) to enhance specific capabilities such as code generation and reasoning.
- Conducted diversity-focused data construction, difficulty & quality-focused data filtering, sandbox verification along with DPO training; achieved significant improvement over same-scaled Coder-Instruct models from $DeepSeek$ and $Qwen2.5$ in terms of $HumanEval$, $NaturalCodeBench$, $LiveCodeBench$, etc.

A3 **TnS (Trust-and-Safety) Moderation: Multi-Modal Pretraining**.
- Optimized a multi-platform data pipeline with efficient data packing and frame extraction, achieving 2× throughput over legacy setups.
- Applied incremental multi-modal pretraining and tri-modal (video+text+audio) models, improving Violation $AUC$ and $bF1@1$.
- Evaluated generalization across dataset domains; enhancing $CLIP$ model as SOTA on full TnS data.

A4 **Tiktok Shop E-commerce AI: Multilingual & Multimodal Pretraining for Price Comparison**.
- Scaled multilingual-multimodal dataset to $1B$ samples, training $CLIP$ with $mBERT$-$Swin$ and $XLM$-$R$-$Swin$ text tower, obtained significant gains on downstream tasks.
- Boosted *recall* by +5% for multilingual retrieval at 100M data scale, and improved multimodal retrieval $P80$ by +4.5%.

Jun 2018 – Oct 2022 **Research Scientist - Staff Research Scientist**.
Baidu Research, Sunnyvale, CA
- **1** business product delivered, **2** applications deployed, **10** conference and journal papers on NLP and Bioinformatics published (including a featured paper accepted by **Nature**), and **3** patents licensed.

B1 **STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency**.
- A novel prefix-to-prefix framework for simultaneous translation that implicitly learns to anticipate in a single translation model.
- A simple yet surprisingly effective wait-k policy was trained to generate the target sentence concurrently with the source sentence, but always k words behind.
- Received many reports from influential media worldwide, e.g., CNBC, MIT tech review, FORTUNE.
- Paper published at ACL 2019 and patented. Demo and code can be found simultrans-demo.github.io

B2 **Incremental Text-to-Speech Synthesis with Prefix-to-Prefix Framework**.

- The first neural incremental TTS approach based on prefix-to-prefix framework. Speech is synthesized in an online fashion, playing a segment of audio while generating the next, O(1) over O(n) latency.
- Experiments show similar speech naturalness compared to full sentence method, but only with a fraction of time and a constant (1-2 words) latency.
- Paper published at EMNLP 2020. Synthesized demo audios can be found on inctts.github.io

B3 **LinearDesign: an efficient algorithms for Optimized mRNA Sequence Design**.

- A surprisingly high efficient solution from computational linguistics to jointly optimize Messenger RNA (mRNA) vaccines' stability and codon usage, to tackle the critical issue of mRNA instability and degradation.
- LinearDesign takes only 11 minutes for the COVID-19 Spike protein. The design substantially improve mRNA half-life and protein expression in vitro, and dramatically increase antibody response by up to 23× in vivo.
- **1** business product was delivered, commercialization achieved, 3 business contracts signed, 1 patent licensed and 1 paper reviewed by Science, open access demo web server is available at rna.baidu.com

B4 **CoV-Seq: a New Tool for SARS-CoV-2 Genome Analysis and Visualization**.

- Developed an integrated web service for fast and easy analysis of custom SARS-CoV-2 sequences. CoV-Seq automatically predicts gene boundaries and identifies genetic variants, which are displayed in an interactive genome visualizer and are downloadable for further analysis. A weekly updated database of genetic variants of all publicly accessible SARS-CoV-2 sequences is also provided
- The method paper was accepted by JMIR and patented, the web service is available covseq.baidu.com

B5 **LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search**.

- LinearFold is the first approximate algorithm in RNA folding to achieve linear runtime (and linear space) without imposing constraints on the output structure such as base-pair distance.
- Merged status of intermediate statuses to compress the size of stacks, and eliminated redundant statuses by beam size.
- Live demo and pre-computed results deployed, demo web server is available at linearfold.org, visualized results are available here

B6 **AutoSimTrans: 3-straight-year workshop on Automatic Simultaneous Translation**.

- Served in Program Committee to host the Workshop and Challenge parts
- In charge of data preparation, submission pipeline and judgement system for Shared Task Challenge
- The workshops were hosted at ACL 2020, NAACL 2021 and NAACL 2022

July 2013 – June 2016 **Engineer & Research program executive**.
China Electric Power Research Institute (CEPRI), State Grid Cooperation of China

D1 **Smart substation network and reliability research**.

- Automatic redundant network path generating technology with high reliability for substation.
- Large scale online network test for smart substation (latency, synchronous signal, network stress and packet loss test).
- **2** conference papers published, and **3** patents authorized.

## Publications

[1] He Zhang, Liang Zhang, Ang Lin, Congc Xu, Ziyu Li, **Kaibo Liu**, David H. Mathews, and Liang Huang, *Algorithm for optimized mRNA design improves stability and immunogenicity [J]*, Nature, 2023.

[2] He Zhang, Liang Zhang, Ziyu Li, **Kaibo Liu**, Boxiang Liu, David H. Mathews, and Liang Huang, *LinearDesign: Efficient Algorithms for Optimized mRNA Sequence Design*, arXiv Preprint, 2022.

[3] Sizhen Li, He Zhang, Liang Zhang, **Kaibo Liu**, Boxiang Liu, et al, *LinearTurboFold: Linear-time global prediction of conserved structures for RNA homologs with applications to SARS-CoV-2 [J]*, PNAS, 2021.

[4] **Kaibo Liu**, Boxiang Liu, He Zhang, Liang Zhang, and Liang Huang, *CoV-Seq: SARS-CoV-2 Genome Analysis and Visualization [C]*, JMIR, 2020.

[5] Baigong Zheng, **Kaibo Liu**, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang, *Simultaneous Translation Policies: From Fixed to Adaptive [C]*, ACL, 2020.

[6] Renjie Zheng, Mingbo Ma, Baigong Zheng, **Kaibo Liu**, and Liang Huang, *Opportunistic Decoding with Timely Correction for Simultaneous Translation [C]*, ACL, 2020.

[7] Renjie Zheng, Mingbo Ma, Baigong Zheng, **Kaibo Liu**, et al, *Fluent and Low-latency Simultaneous Speech-to-Speech Translation with Self-adaptive Training [C]*, ACL, 2020.

[8] Mingbo Ma, Baigong Zheng, **Kaibo Liu**, Renjie Zheng, et al, *Incremental Text-to-Speech Synthesis with Prefix-to-Prefix Framework [C]*, EMNLP, 2020.

[9] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, **Kaibo Liu**, et al, *STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework [C]*, In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3025-3036. 2019.

[10] Liang Huang, Liang, He Zhang, Dezhong Deng, Kai Zhao, **Kaibo Liu**, David Hendrix, and David Mathews, *LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search [C]*, Bioinformatics, 35, no. 14 (2019).

[11] **Kaibo Liu**, Hang Lu, Zhongqing Li, et al, *Application of High Sampling Rate Data in Merging Unit for Relay Protection [C]*, 5th IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies, 2015, 1099-1104.

[12] Zhijuan Tu, **Kaibo Liu**, Huaxiang Yi, et al, *A compact evanescently-coupled germanium PIN waveguide photodetector [C]*, Proceedings of SPIE- The International Society for Optical Engineering, 2012, 8564(19): 425-430.

[13] Zhongqing Li, **Kaibo Liu**, Xiao Li, et al, *Sampled data synchronization scheme for relay protection in smart substation [C]*, Power System Technology (POWERCON), International Conference on IEEE, 2015, 1778-1784.

[P1] Boxiang Liu, **Kaibo Liu**, He Zhang, etc., *Systems and methods for genome analysis and visualization [IP]*, 20220108773, 2020.

[P2] He Zhang, Liang Zhang, Ziyu Li, **Kaibo Liu**, Boxiang Liu, Liang Huang, *Systems and methods for sequence design* , 28888-2424, BN200427USN1, 2020.

[P3] Mingbo Ma, Liang Huang, Hao Xiong, **Kaibo Liu**, etc., *Systems and methods for simultaneous translation with integrated anticipation and controllable latency (STACL) [IP]*, 11126800, 2019.

[P4] **Kaibo Liu**, Huanzhang Liu, Zhongqing Li, et al, *The criterion for the polarization of a single ended distance protection [IP]*, CN201510955373.9, 2015.

[P5] Zhongqing Li, Zexin Zhou, Yongli Li, **Kaibo Liu**, et al, *A fault diagnosis method of circuit breaker operating mechanism based on least squares vector [IP]*, CN201510214317.X, 2015.

[P6] Zhongqing Li, Botong Li, Xianguo Jiang, **Kaibo Liu**, et al, *A fault location method for hybrid line of overhead line and high voltage cable [IP]*, CN201510316122.6, 2015.

## Awards

| | |
|---|---|
| Mar 2022 | **General TC Technology Incentive Award 2021**, Baidu, CHINA & USA. |
| Jan 2022 | **Star of Q4 2021**, Baidu Research, USA. |
| Jan 2021 | **Baidu Pride Best Team Award 2020**, Baidu, USA. |
| Jul 2020 | **AIG-TC Technology Incentive Award 2020-H1**, Baidu, USA. |
| Dec 2018 | **AIG-TC Technology Innovation Award 2018-H2**, Baidu, USA. |
| May 2015 | **First prize for scientific and technological progress in CEPRI**, State Grid Co. of China. |
| 2010 – 2013 | **National Second-order Scholarship of China**, Peking University, CHINA. |
| Nov 2009 | **Sumitomo Mitsui Bank(JP) Global Foundation Scholarship**, Peking University, CHINA. |

## Skills & Abilities

| | |
|---|---|
| Field | LLM, NLP, Machine Translation, Computational biology, Computer vision, Deep learning, Data analysis |
| Programming Language | ♥ Python,♥ C/C++, MySql, Matlab, LaTeX |
| Frame | Pytorch, PaddlePaddle, TensorFlow, Torch, Keras, Caffe, OpenCV |
| Web | Flask, Django, Node.js, JavaScript, HTML5 |
| Deep love | in algorithm |