Jiahuan Pei

ppsunrise99@gmail.com
https://jiahuan-pei.github.io/
https://scholar.google.com/citations?user=cnhyEW0AAAAJ&hl=en
https://github.com/Jiahuan-Pei

I specialize in natural language processing (NLP) and information retrieval (IR) challenges, with experience spanning both industry and academia. I am working as a researcher at Centrum Wiskunde & Informatica (CWI), the national research institute for mathematics and computer science in the Netherlands. I focus on generative dialogue agents in extended reality (XR). Formerly, I contributed as an applied scientist at Amazon, focusing on essential NLP/IR aspects for Amazon product search. I pursued my Ph.D. with the IRLab at University of Amsterdam, supervised by Prof. dr. Maarten de Rijke. My expertise encompasses diverse NLP aspects (e.g., large language models, dialogue systems, word embedding, parsing, summarization), IR aspects (e.g., query understanding, recommender system, matching algorithm embedding, faithfulness evaluation), XR aspects (e.g., point cloud assessment and multimodal immersive systems), and foundational machine learning (e.g., classification, regression, uncertain estimation). I am eligible to self-sponsor my work with a long-term EU residence permit.

Skills & Certificate

Knowledge	Natural language processing (NLP), large language models (LLMs), information retrieval (IR), machine learning (ML), deep learning (DL), artificial intelligence (AI), extended reality (XR).
Coding	Python, Pytorch, Tensorflow, Java, C, C++, C#, Ruby, PHP, HTML, JavaScript, SQL, LaTeX, Linux, AWS (Athena, Sagemaker, S3, Stepfunctions), Git, Docker, MLOps, Langchain.
Languages	Mandarin Chinese (First language), English (IELTS-6.5), Dutch (A2), Japanese (JLPT-N2).
Teaching	TLC Training Certificate (Teaching and Learning Centre, University of Amsterdam).

Education

Oct. 2017 – Dec. 2022

Ph.D., University of Amsterdam, Science of Informatics.
Thesis title: *Collaborative agents for task-oriented dialogue systems.* [link]
Supervisor: Prof. dr. Maarten de Rijke

Sept. 2014 – Jun. 2017

M.Sc., Dalian University of Technology, Computer Applied Technology.
Thesis title: *Research on Chinese word semantic similarity computation.* [link]

Sept. 2010 – Jul. 2014

B.Sc., Dalian Maritime University, Software Engineering (Major), International Economics and Trade (Minor).
Thesis title: *Chinese functional maximal-length noun phrase recognition.* [link]

Experience

Mar. 2023 – Now

Researcher, CWI, Amsterdam, Netherlands.
My role is to define the infrastructure and lead the development of generative dialogue agents in extended reality (XR), which cooperates with multiple partners in the European project VOXReality (No. 101070521). This work combines the fields of NLP, CV, and XR to create immersive, interactive, and engaging conversational experiences in XR environments. By leveraging LLM powered autonomous dialogue agents, we develop intelligent agents capable of understanding and responding to user input in a dynamic and contextually appropriate manner.

Experience (continued)

Dec. 2021 – Mar. 2023

Applied Scientist, Amazon, Berlin, Germany.
My role is to explore practical NLP/IR problems related to customer obsession with Amazon search and propose scientifically grounded solutions. The main responsibilities include: Refining and revising the project plan to learn matcher embeddings for improving Amazon's product search engine; Creating datasets for training and evaluating machine learning models for this task; Evaluating the models, analyzing their performance, and reporting the findings to the wider team.

Mar. 2021 – May 2021

Applied Scientist Intern, Amazon, Berlin, Germany.
My role is to explore stochastic Transformers architecture and enhance it with uncertainty estimation with Monte Carlo method.

Mar. 2020 – Jun. 2020

Applied Scientist Intern, Amazon Alexa AI, Aachen, Germany.
My role is to develop a lightweight adaptive Transformers model for natural language understanding in Alexa, which is able to remain more than 95% of the effectiveness while using less than 50% of the computational cost.

Teaching

Sep. 2018 – Now

Research Supervisor, University of Amsterdam and Shandong University.
I have supervised 5 Ph.D. students, and 7 master and bachelor students for their research papers and thesis. My role is to help with weekly discussions, model design, result analysis, and paper writing. More details at [student supervision].

Sep. 2018 – Oct. 2020

Teaching Assistant, Bachelor's IR course, University of Amsterdam.
My role is to design the final search engine project, prepare pipeline code, grade, and answer students' questions.

Nov. 2018 – Jan. 2019

Teaching Assistant, Master's IR course, University of Amsterdam.
My role is to supervise two groups of students to wrap up their research projects, which includes weekly meetings, model implementation and paper writing.

Activities

2020 – Now

Reviewer and PC member, Journals (e.g., TOIS, TOMM, IPM, TOIT), conferences (e.g., ACL, SIGIR, AAAI, WWW, IJCAI).

2021 – 2022

Amazon campus ambassador, Amazon Campus Intern and New Grad Ambassador Program.

2019 – 2020

Soos talk chair, IRLab, University of Amsterdam.

Sept. 2015 – Mar. 2016

Research grant application: Sentence similarity computation based on deep Learning. My role is the lead author in the application for the National Natural Science Foundation of China (No. 61672127).

May. 2014 – Sept. 2015

Research grant participation: English functional noun phrases identification. My role is research assistant, taking charge of model design and implementation in National Social Science Foundation of China (No.15BYY175).

Talks

Feb. 2024 Multimodal Conversational AI Agents Towards Smarter Assistant, Invited talk, Microsoft Research, Cambridge, United Kingdom.

Dec. 2023 Multimodal Dialogue Agents Towards Intelligent XR, Invited talk, Utrecht University, Utrecht, Netherlands.

Nov. 2023 Maximising efficiency in NLP model training and XR environments, invited panel talk, Immersive Tech Week, Rotterdam, Netherlands. [link].

Sep. 2023 Language model-powered dialogue agents and virtual knowledge base, Invited talk, Sony AI, Barcelona, Spain.

Jul. 2023 Advancements and prospects in dialogue agents and LLMs, Invited talk, Bosch Center of Artificial Intelligence (BCAI), Renningen, Germany. [slides].

Apr. 2023 Generative AI towards scientific discovery, Invited talk, Microsoft Research AI4Science, Amsterdam, Netherlands. [slides].

Oct. 2022 Learning embeddings to represent information retrieval systems, conference talk, Amazon Machine Learning Conference (AMLC), Berlin, Germany.

Jul. 2022 Frontiers of collaborative task-oriented dialogue systems, Invited talk, University College London (UCL), London, United Kingdom. [slides].

Apr. 2022 Transformer uncertainty estimation with hierarchical stochastic attention, Invited talk, Search Engine Amsterdam (SEA), Amsterdam, Netherlands. [link].

Apr. 2021 A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles, conference talk, WWW' 21, Ljubljana, Slovenia. [slides]. [vedio].

Sep. 2020 Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation, conference talk, ECAI' 20, Santiago de Compostela, Spain. [slides]. [video].

Jun. 2020 Adaptive Transformers for efficient natural language understanding, Invited talk, Amazon, Germany.

Aug. 2019 SEntNet: Source-aware recurrent entity network for dialogue response selection, conference talk, IJCAI' 19 SCAI Workshop, Macao, China. [slides].

Dec. 2016 Combining word embedding and semantic lexicon for Chinese word similarity computation, conference talk, NLPCC' 16, Kunming, China. [slides].

Jun. 2016 DUT-NLP-CH @ NTCIR-12 temporalia temporal intent disambiguation subtask, conference talk, NTCIR' 16, Tokyo, Japan. [slides].

Publications

Conference Papers

1. **Pei, Jiahuan** et al. "Autonomous Workflow for Multimodal Fine-Grained Training Assistants Towards Mixed Reality". In: ***ACL*** (***Class A***). 2024.

2. Ren, Pengjie, Chengshun Shi, Shiguang Wu, Mengqi Zhang, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and **Jiahuan Pei**. "Mini-Ensemble Low-Rank Adapter for Parameter-Efficient Fine-Tuning". In: ***ACL*** (***Class A, Corresponding author***). 2024.

3. Sun, Xin, **Jiahuan Pei**, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. "Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study". In: ***COLING*** (***Class B, Corresponding author***). 2024.

4. **Pei, Jiahuan**, Cheng Wang, and György Szarvas. "Transformer Uncertainty Estimation with Hierarchical Stochastic Attentions". In: ***AAAI*** (***Class A***). 2022.

5. **Pei, Jiahuan**, Pengjie Ren, and Maarten de Rijke. "A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles". In: *TheWebConf* (***WWW, Class A***). 2021.

6. **Pei, Jiahuan**, Pengjie Ren, Christof Monz, and Maarten de Rijke. "Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation". In: ***ECAI*** *(**Class B**)*. 2019.

7. **Pei, Jiahuan**, Pengjie Ren, and Maarten de Rijke. "A Modular Task-oriented Dialogue System Using a Neural Mixture-of-Experts". In: ***SIGIR*** *Workshop on Conversational Interaction Systems*. 2019.

8. **Pei, Jiahuan**, Arent Stienstra, Julia Kiseleva, and Maarten de Rijke. "SEntNet: Source-aware Recurrent Entity Network for Dialogue Response Selection". In: ***IJCAI*** *Workshop on Search-Oriented Conversational AI*. 2019.

9. **Pei, Jiahuan**, Danushka Bollegala, and Omar Zaidan. "Learning Embeddings to Represent Information Retrieval Systems". In: *Amazon Machine Learning Conference (**AMLC**)*. 2022.

10. **Pei, Jiahuan** and Cheng Wang. "A Simple Way to Estimate Uncertainty in Transformers". In: *Amazon Machine Learning Conference (**AMLC**)*. 2021.

11. **Pei, Jiahuan**, Cong Zhang, Degen Huang, and Jianjun Ma. "Combining Word Embedding and Semantic Lexicon for Chinese Word Similarity Computation". In: ***NLPCC*** *(**Class C**)*. 2016.

12. **Pei, Jiahuan**, Degen Huang, Jianjun Ma, Dingxin Song, and Leyuan Sang. "DUT-NLP-CH @ NTCIR-12 Temporalia Temporal Intent Disambiguation Subtask". In: ***NTCIR***. 2016.

13. Deng, Wentao, **Jiahuan Pei**, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. "Syllogistic Reasoning for Legal Judgment Analysis". In: ***EMNLP*** *(**Class B**)*. 2023.

14. Yan, Guojun, **Jiahuan Pei**, Pengjie Ren, Zhaochun Ren, and Maarten de Rijke. "ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues". In: ***SIGIR*** *(**Class A**)*. 2022.

15. Zhang, Weijia, Mohammad Aliannejadi, **Jiahuan Pei**, Yifei Yuan, Jia-Hong Huang, and Evangelos Kanoulas. "Towards Fine-Grained Citation Evaluation in Generated Text: A Comparative Analysis of Faithfulness Metrics and Humans in Citation Evaluation". In: ***SIGIR*** *Workshop on Large Language Models for Evaluation in Information Retrieval*. 2024.

Journal Articles

1. **Pei, Jiahuan**, Guojun Yan, Pengjie Ren, and Maarten de Rijke. "Mixture-of-languages Routing for Multilingual Dialogues". In: (2024). ACM Transactions on Information Systems (**TOIS**, **Class A**, In Production).

2. Deng, Wentao, **Jiahuan Pei**, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. "Intent-calibrated Self-training for Answer Selection in Open-domain Dialogues". In: *Transactions of the Association for Computational Linguistics (**TACL**, **Class B**)* (May 2023).

3. Huang, Degen, **Jiahuan Pei**, Cong Zhang, Kaiyu Huang, and Jianjun Ma. "Incorporating Prior Knowledge into Word Embedding for Chinese Word Similarity Measurement". In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP' 18, **JCR-Q2**, SCI)* (2018).

4. Wang, Benyou, Qianqian Xie, **Jiahuan Pei**, Prayag Tiwari, Zhao Li, and Jie Fu. "Pre-trained Language Models in Biomedical Domain: A Systematic Survey". In: *ACM Computing Surveys (**JCR-Q1**)* (2023).

5. Ma, Jianjun, **Jiahuan Pei**, Degen Huang, and Dingxin Song. "Syntactic parsing of clause constituents for statistical machine translation". In: *International Journal of Computational Science and Engineering (IJCSE'18, **JCR-Q3**)* (2018).

6. Ma, Jianjun, Mulangma Zhu, Degen Huang, and **Jiahuan Pei**. "Recognition of Syntactic Relationship between Clauses Using CRFs". In: *International Journal of Advanced Intelligence* (2018).

7. Ma, Jianjun, **Jiahuan Pei**, and Degen Huang. "Identification of English Functional Noun Phrases by CRFs and the Semantic Information". In: *Journal of Chinese Information Processing (Class B, Chinese Journal)* (2016).

8 Zhang, Cong, **Jiahuan Pei**, Kaiyu Huang, Degen Huang, and Zhangzhi Yin. “Semantic graph optimization algorithm based Chinese microblog opinion summarization”. In: *Journal Of Shandong University (Natural Science)* (2017).

Released Datasets

1 *ReMeDi Dataset for Multi-domain, Multi-service, Medical Dialogues*. (It contains 96,965 conversations between doctors and patients, including 1,557 conversations with fine-grained labels. It covers 843 types of diseases, 5,228 medical entities, and 3 specialties of medical services across 40 domains.) 2022. 🔗 URL: `https://github.com/yanguojun123/Medical-Dialogue/tree/main`.

2 *Syllogistic Reasoning Dataset for Legal Judgment Analysis*. (It contains 11,239 criminal cases which cover 4 criminal elements, 80 charges, and 124 articles.) 2023. 🔗 URL: `https://github.com/dengwentao99/SLJA`.

3 *Bilingual Motivational Interviewing Dialogue Dataset with Skill Codes*. (It consists of 80 conversations with 8,572 utterances between 18 clients and therapists in real motivational interviewing counseling sessions conducted in Dutch, along with their translated counterparts in English. These sessions explore a diverse array of 26 skill codes.) 2024. 🔗 URL: `https://github.com/XIN-von-SUN/Eliciting-MISC-in-Psychotherapy-with-LLMs`.

4 *LEGO-MRTA Dataset for LEGO Assembly Training in XR*. (It contains 65 LEGO instruction manuals as the grounding for 1,423 human-human natural conversations between trainers and trainees, generated via GPT-3.5 API calls. We created 26,405 context-response pairs from these conversations and vision-language pairs, which serve as data samples for instruction-following fine-tuning.) 2024. 🔗 URL: `https://github.com/Jiahuan-Pei/AutonomousDialogAgent4AugmentedReality`.

Ongoing projects

1 **Pei, Jiahuan** *, Haochen Huang*, Irene Viola, Mohammad Aliannejadi, Moonisa Ahsan, Zhaochun Ren, Chuang Yu, Junxiao Wang, and Pablo Cesar. “Fine-Grained Vision-Language Modeling for Multimodal Training Assistants in Augmented Reality”. ACM Multimedia (**MM**, Under Review). 2024.

2 Deng, Wentao, **Jiahuan Pei**, and Pengjie Ren. “Autonomous Task-Dependency Planning via Uncertainty-aware Bidirectional Reasoning”. Neural Information Processing Systems (**NeurIPS**, Under Review). 2024.

3 Liu, Yuanxing, **Jiahuan Pei**, Ming Li, Weinan Zhang, Wanxiang Che, and Maarten de Rijke. “Augmentation with Neighboring Information for Conversational Recommendation”. ACM Transactions on Information Systems (**TOIS**, Under Review). 2023.

4 Wang, Xiao, **Jiahuan Pei**, Diancheng Shui, Zhiguang Han, Xin Sun, Dawei Zhu, and Xiaoyu Shen. “Should multiple defendants and charges be treated separately in legal judgment prediction: An exploratory study and dataset”. **EMNLP** (Under Review). 2024.

5 Zhou, Xuemei, Irene Viola, Yunlu Chen, **Jiahuan Pei**, and Pablo Cesar. “Deciphering Perceptual Quality in Colored Point Clouds: Prioritizing Geometry or Texture Distortion?” ACM Multimedia (**MM**, Corresponding author, Under Review). 2024.

6 Zhang, Wenhao, Mengqi Zhang, Shiguang Wu, **Jiahuan Pei**, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Pengjie Ren. “ExcluIR: Exclusionary Neural Information Retrieval”. **EMNLP** (Under Review). 2024. 🔗 URL: `https://arxiv.org/pdf/2404.17288`.

7 Sun, Xin, Xiao tang, Abdallah El Ali, Zhuying Li, Xiaoyu Shen, Pengjie Ren, Jan de Wit, **Jiahuan Pei**, and Jos Bosch. “Chain-of-Code Planning with LLMs: Aligning the Generation of Psychotherapy Dialogue with Behavior Codes in Motivational Interviewing”. **EMNLP** (Under Review). 2024.

8 Shen, Xiaoyu, Rexhina Blloshmi, Dawei Zhu, **Jiahuan Pei**, and Wei Zhang. “Assessing “Implicit” Retrieval Robustness of Large Language Models”. **EMNLP** (Under Review). 2024.

9. Zhang, Erhang, **Jiahuan Pei**, and Junxiao Wang. “Instructable Agents With Retrieval-augmented Generation in Virtual Reality”. Ongoing. 2024.

Awards

- 2016 The 2nd ranked team in NTCIR-12 Temporalia TID task.
- National scholarship for outstanding master students (top 1%).
- Excellent graduate student in Dalian city (top 5%).
- Outstanding master’s thesis award (top 5%).
- 2011–2014 Outstanding student scholarship for bachelor students (top 3%).
- 2013 Outstanding winner for MCM/ICM media contest (International competition).