

Beiduo Chen

beiduochen@gmail.com | +49 15222000786 | Homepage | GitHub | Google Scholar

RESEARCH SUMMARY

- Ph.D. researcher in Natural Language Processing (NLP) and Large Language Models (LLMs), working on *understanding, evaluating, and interpreting LLM behavior beyond a single accuracy number*. My research spans Chain-of-Thought reasoning and cross-model reasoning transfer, human label variation, uncertainty- and explanation-grounded annotation, trustworthy and human-centered evaluation, and multilingual/cultural NLP. I am broadly interested in robust evaluation protocols, distributional and aggregation-aware metrics, LLM-as-a-judge evaluation, and aligning model reasoning with the diversity of human judgments.

EDUCATION

Ludwig-Maximilians-Universität München

Ph.D. in Natural Language Processing, *MainNLP Lab*, supervised by Prof. Barbara Plank (ACL 2026 President)
Research: LLM Evaluation, Trustworthiness, Human-Centered NLP, Uncertainty, and Chain-of-Thought Reasoning

Munich, Germany
2024 – 2027 expected

University of Cambridge

Exchange *ELLIS Ph.D. Student*, *Language Technology Lab*, supervised by Prof. Anna Korhonen

Cambridge, United Kingdom
2025

University of Science and Technology of China

M.Eng. in Information and Communication Engineering, supervised by Prof. Wu Guo
Thesis: *Multilingual Representation Learning and Applications Based on Pre-trained Language Models*

Anhui, China
2020 – 2023
GPA: 3.93/4.3, top 1%

University of Science and Technology of China

B.Eng. in Electronic and Information Engineering, supervised by Prof. Wu Guo
Thesis: *Speaker Recognition Based on Depth Features*

Anhui, China
2016 – 2020
GPA: 3.75/4.3, top 3%

RESEARCH EXPERIENCE

Microsoft Research Asia

Research Intern, hosted by Shaohan Huang

Beijing, China
Jun 2022 – Jan 2023

- Designed and evaluated a multi-perspective curriculum learning framework for ELECTRA-style language model pre-training, published at ACL Findings 2023.
- Conducted empirical analysis of pre-training strategies and downstream model behavior on GLUE and SQuAD 2.0 benchmarks.

iFLYTEK Research

Research Intern, hosted by Quan Liu

Anhui, China
Jun 2021 – Mar 2022

- Developed multilingual and cross-lingual transfer methods for named entity recognition across diverse languages, scripts, and domains.
- Led the first-author SemEval-2022 shared task system with strong results across 13 multilingual tracks.

SELECTED RESEARCH DIRECTIONS

LLM Reasoning and Chain-of-Thought Evaluation

Controlled evaluation of reasoning traces, model priors, and cross-model transfer

2025 – 2026

- Designed Cross-CoT experiments to isolate reasoning text from intrinsic model priors, showing that CoT can drive top-answer accuracy while failing to calibrate fine-grained distributional uncertainty. [ACL 2026 Findings-a]
- Introduced a prefix-based provider–receiver framework for cross-model CoT transfer, distinguishing answer extraction, reasoning scaffolding, receiver competence, and partial-solution accumulation. [Preprint 2026-b]
- Developed a CoT-based pipeline for extracting supporting and opposing evidence to explain human label variation, together with a rank-based evaluation framework for distributional alignment. [EMNLP 2025 [Oral](#)]

Human-Centered LLM Evaluation and Human Label Variation

Evaluation protocols for label distributions, explanations, and annotator-specific behavior

2024 – 2026

- Proposed CAPO, a cross-annotator preference optimization method, and aggregation-aware metrics for testing whether LLMs can learn stable annotator-specific label-explanation behavior. [Preprint 2026-a]
- Used explanation taxonomies and similarity analyses to decompose NLI annotation variation beyond label agreement, revealing individual preferences in explanation strategies and label choices. [ACL 2026 Findings-b] [EMNLP 2025 [SAC Highlights](#)]
- Showed that LLM-generated explanations can help approximate human judgment distributions from sparse labels/explanations, while requiring both instance-level and global distributional evaluation. [ACL 2025 Findings] [EMNLP 2024 Findings]

Scalable evaluation across languages, cultures, tasks, and model families

- Contributed to MAKIEval, a multilingual Wikidata-based framework for open-ended cultural awareness evaluation across languages, regions, topics, and model families. [EMNLP 2025 Findings]
- Built the first-author SemEval-2022 Task 11 system, ranking 1st on Chinese, Code-mixed, and Bangla tracks, and 2nd on ten additional multilingual tracks. [SemEval 2022@NAACL]
- Developed cross-lingual alignment, layer aggregation, and distillation methods for robust multilingual transfer and benchmark generalization. [ICASSP 2022] [ICPR 2022] [EMNLP 2022]

PUBLICATIONS AND PREPRINTS (*EQUAL CONTRIBUTION.)

- [Preprint 2026-a] **Beiduo Chen**, Pingjun Hong, Ziyun Zhang, Benjamin Roth, Anna Korhonen, Barbara Plank. *Human Label Variation as Stable Signal: Learning Annotator-Specific Explanation Behavior via Cross-Annotator Preference Optimization*. Proposed CAPO, a preference-optimization method and aggregation-aware evaluation framework for learning annotator-specific label-explanation behavior.
- [Preprint 2026-b] Xinyuan Cheng*, **Beiduo Chen***, Philipp Mondorf, Barbara Plank. *Reasoning that Travels: Dissecting How Chain-of-Thought Transfers Across Models*. Introduced a provider–receiver framework for analyzing whether cross-model CoT transfer reflects answer extraction, reasoning scaffolding, or receiver competence.
- [Preprint 2026-c] Benedetta Muscato, **Beiduo Chen**, Gizem Gezici, Barbara Plank, Fosca Giannotti. *Disagreeing Rationales: Rethinking Classification and Explainability Evaluation in Hate Speech Detection*. Unified classification and rationale evaluation under hard, intermediate, and soft representation spaces for subjective NLP
- [ACL 2026 Findings-a] **Beiduo Chen**, Tiancheng Hu, Caiqi Zhang, Robert Litschko, Anna Korhonen, Barbara Plank. *Decoupling the Effect of Chain-of-Thought Reasoning: A Human Label Variation Perspective*. Disentangled CoT content and model priors in distribution-based tasks, showing that CoT drives top-answer accuracy while model priors dominate distributional ranking.
- [ACL 2026 Findings-b] Pingjun Hong*, **Beiduo Chen***, Siyao Peng, Marie-Catherine de Marneffe, Benjamin Roth, Barbara Plank. *Agree, Disagree, Explain: Decomposing Human Label Variation in NLI through the Lens of Explanations*. Decomposed NLI annotation variation through label agreement, explanation similarity, taxonomy agreement, and annotator selection bias.
- [EMNLP 2025 **Oral**] **Beiduo Chen**, Yang Janet Liu, Anna Korhonen, Barbara Plank. *Threading the Needle: Reweaving Chain-of-Thought Reasoning to Explain Human Label Variation*. Extracted supporting and opposing statements from CoTs and proposed rank-based HLV evaluation for aligning LLM reasoning with human label distributions.
- [EMNLP 2025 **SAC Highlights**] Pingjun Hong*, **Beiduo Chen***, Siyao Peng, Marie-Catherine de Marneffe, Barbara Plank. *LiTeX: A Linguistic Taxonomy of Explanations for Understanding Within-Label Variation in Natural Language Inference*. Introduced a linguistic taxonomy for free-text NLI explanations and showed that taxonomy-guided generation better matches human reasoning.
- [EMNLP 2025 Findings] Raoyuan Zhao, **Beiduo Chen**, Barbara Plank, Michael A. Hedderich. *MAKIEval: A Multilingual Automatic Wikidata-based Framework for Cultural Awareness Evaluation for LLMs*. Developed scalable multilingual evaluation of cultural awareness through Wikidata-grounded metrics across languages, regions, and topics.
- [ACL 2025 Findings] **Beiduo Chen**, Siyao Peng, Anna Korhonen, Barbara Plank. *A Rose by Any Other Name: LLM-Generated Explanations Are Good Proxies for Human Explanations to Collect Label Distributions on NLI*. Showed that LLM-generated explanations, when conditioned on human labels, can approximate human judgment distributions and generalize to challenging test sets.
- [EMNLP 2024 Findings] **Beiduo Chen**, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, Barbara Plank. *“Seeing the Big through the Small”: Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations?* Demonstrated that a few expert explanations can help LLMs approximate human judgment distributions, and highlighted the need for both instance-level and global distributional metrics.
- [ACL 2023 Findings] **Beiduo Chen**, Shaohan Huang, Zihan Zhang, Wu Guo, Zhenhua Ling, Haizhen Huang, Furu Wei, Weiwei Deng, Qi Zhang. *Pre-training Language Model as a Multi-perspective Course Learner*. Proposed multi-perspective course learning for ELECTRA-style pre-training, improving GLUE and SQuAD 2.0 performance under comparable settings.
- [EMNLP 2022] Jun-Yu Ma*, **Beiduo Chen***, Jia-Chen Gu, Zhenhua Ling, Wu Guo, Quan Liu, Zhigang Chen, Cong Liu. *Wider & Closer: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition*. Proposed multi-channel distillation and parallel domain adaptation for zero-shot cross-lingual NER across nine languages.
- [ICASSP 2022] **Beiduo Chen**, Wu Guo, Bin Gu, Quan Liu, Yongchao Wang. *Multi-Level Contrastive Learning for Cross-Lingual Alignment*. Improved cross-lingual representation alignment through sentence- and word-level contrastive learning.
- [ICPR 2022] **Beiduo Chen**, Wu Guo, Quan Liu, Kun Tao. *Feature Aggregation in Zero-Shot Cross-Lingual Transfer Using Multilingual BERT*. Explored layer-wise feature aggregation in mBERT for zero-shot transfer on XNLI, PAWS-X, NER, and POS benchmarks.
- [SemEval 2022@NAACL] **Beiduo Chen**, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, Quan Liu. *USTC-NELSLIP at SemEval-2022 Task 11: Gazetteer-Adapted Integration Network for Multilingual Complex Named Entity Recognition*. Built a Wikidata-enhanced multilingual NER system ranking 1st on three tracks and 2nd on ten additional tracks.

SHARED TASKS

- 2022, Rank 1st on three tracks (Chinese, Code-mixed, and Bangla) and 2nd on the other ten tracks in the 16th International Workshop on Semantic Evaluation (SemEval-2022) Task 11: Multilingual Complex Named Entity Recognition, as the first author.

PATENT

- **Beiduo Chen**, Qingqing Huang, Jun Du. Multi-Feature Fusion Method for Neural Machine Translation Error Detection Based on Data Enhancement Training. China National Intellectual Property Administration (CNIPA). 2021.

INVITED TALKS AND PANELS

- Invited Panel Discussion, 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives) at EMNLP 2025, Suzhou, China. November 8th, 2025. (Panelist)
- Invited Talk, Dealing with Meaning Variation in NLP – 3rd Yearly Workshop at Utrecht University, Netherlands. *Explanations as a Catalyst: Leveraging Large Language Models to Embrace Human Label Variation*. October 28th, 2025.
- Invited Talk, Language Technology Lab Seminars at University of Cambridge, United Kingdom. *Explanations as a Catalyst: Leveraging Large Language Models to Embrace Human Label Variation*. October 10th, 2025.
- Invited Talk, 4th International Workshop on Dependability Modeling and Digitalization (WDMD) at the 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2025), Naples, Italy. *Understanding and Modeling Human Label Variation in LLM*. June 25th, 2025.

HONORS AND AWARDS

- 2025, Senior Area Chair Highlight Award at EMNLP 2025 (top 1%).
- 2025, Mobility Grant of European Lighthouse on Secure and Safe AI (ELSA).
- 2024, European Laboratory for Learning and Intelligent Systems (ELLIS) Ph.D. Program (5% acceptance).
- 2023, Outstanding Graduate Award of University of Science and Technology of China.
- 2023, Outstanding Graduate Award of Ordinary Colleges and Universities in Anhui Province.
- 2022, China National Scholarship.
- 2020, Suzhou Yucai Scholarship: Top 1 GPA in the class (1/120).
- 2019, Scholarship of the Institute of Electrics, Chinese Academy of Sciences.
- 2018, Third Prize (provincial) in the Contemporary Undergraduate Mathematical Contest in Modeling of China.
- 2017, Gold Award for Outstanding Student of USTC.

TEACHING

- **Lectures:** Information Retrieval (SS2025, SS2026), LLM Agents (SS2025), Symbolic Programming Language (WS2024/25, WS2025/26), Multi-modal NLP (SS2024) at LMU Munich; Signals and Systems (2021), Computer Programming A (2019), Electromagnetism C (2018) at USTC.
- **Seminars:** Explaining and Interpreting Annotations in NLP (WS2025/26), Discourse Modeling and Processing (WS2024/25), NLP for Climate Change (SS2024) at LMU Munich.

MENTORSHIP

- 2026, Xinyuan Cheng, MSc Thesis on LLM Reasoning. *Measuring the Quality of Chain-of-Thought in Reasoning-Tuned Large Language Models via Cross-Model Utility*.
- 2026, Lihui Zhu, MSc Thesis on LLM Agents. *Beyond Majority Vote: Structured Aggregation of Reasoning Trajectories in Multi-Agent LLM Debate*.
- 2025, Pingjun Hong, MSc Thesis on Within-label Variation. *Within-Label Variation in Natural Language Inference: A Linguistic Taxonomy for Explanations and Its Impact on Model Interpretation of Label Decisions*.
- 2024, Vu Thanh Trung Bui, BSc Thesis on Multilingual LLM. *Exploring Multilingual Capabilities in Large Language Models with Soft Prompt Tuning*.

SERVICE

- Program Committee/Reviewer: TPAMI, ACL, EMNLP, NAACL, ACL Rolling Review, COLM, ICASSP, ICPR, NLPOR@COLM2025.

SKILLS

- **Research:** LLM pre-training and post-training (SFT, preference optimization, reinforcement learning), Chain-of-Thought reasoning, LLM evaluation and benchmark design, distributional and uncertainty-aware evaluation, aggregation-aware metrics, LLM-as-a-judge evaluation, prompt-based evaluation, human annotation analysis, statistical analysis, multilingual and cross-lingual transfer.
- **Programming:** Python (proficient), C/C++, MATLAB, HTML (academic project level).
- **Frameworks and Tools:** PyTorch, TensorFlow, HuggingFace Transformers, vLLM, DeepSpeed, Keras, Scikit-Learn, Pandas, NumPy, Jupyter, CUDA, Git, LaTeX.