Masahiro Kaneko

Postdoctoral Researcher, MBZUAI (Prof. Timothy Baldwin's team)

Research Interests

- AI Safety: harms caused by AI
- AI Security: attacks against AI

masahiro-kaneko.com Google Scholar github.com/kanekomasahiro

> Selected Publications

2026

Poisoning-Based Backdoors Are Not Cryptographic
Masahiro Kaneko
NeurIPS

A Multilingual Social Bias Benchmark Incorporating Thinking Processes
Masahiro Kaneko, Danushka Bollegala, Timothy Baldwin
ACL

JailNewsBench: Multi-Lingual and Regional Benchmark for Fake News Generation under Jailbreak Attacks
Masahiro Kaneko, Ayana Niwa, Timothy Baldwin
ICLR

On the Alignment of Large Language Models with Global Human Opinion
Yang Liu, Masahiro Kaneko, Chenhui Chu
AAAI Best Paper Award (Alignment Track)

2025

Bits Leaked per Query: Information-Theoretic Bounds on Adversarial Attacks against LLMs
Masahiro Kaneko, Timothy Baldwin
NeurIPS Spotlight

> Experience

Hakuhodo DY Holdings, Human-Centered AI Institute Japan
2026.07 – Present Research Advisor

Simulacra Inc. Japan
2026.01 – Present Principal Researcher

Research and Development Center for Large Language Models, National Institute of Informatics Japan
2025.11 – Present Project Researcher (AI Safety Group)

MBZUAI UAE
2023.07 – Present Postdoctoral Researcher (Prof. Timothy Baldwin's team)

Tokyo Institute of Technology — Japan

2023.07 – 2026.03 Research Fellow (Prof. Naoaki Okazaki's lab)

2021.04 – 2023.05 Postdoctoral Researcher

Tokyo Metropolitan University — Japan

2021.04 – 2023.03 Visiting Researcher (Prof. Mamoru Komachi's lab)

Carnegie Mellon University — USA

2022.09 – 2022.11 Visiting Researcher (Prof. Graham Neubig's team)

University of Liverpool — UK

2018.10 – 2019.02 Visiting Researcher (Prof. Danushka Bollegala's team)

RIKEN Center for Advanced Intelligence Project — Japan

2018.08 – 2021.03 Research Assistant (Prof. Kentaro Inui's team)

> Education

Tokyo Metropolitan University, Graduate School of Systems Design — Japan

2018.04 – 2021.03 Ph.D. in Computer Science, Advisor: Prof. Mamoru Komachi

2016.04 – 2018.03 M.S. in Computer Science

Kitami Institute of Technology, Faculty of Engineering — Japan

2012.04 – 2016.03 B.Eng., Advisor: Prof. Hiroshi Masui

> Awards

2026.03 **Best Paper Award** 3/789
The 32nd Annual Meeting of the Association for Natural Language Processing

2026.03 **Committee Special Award** 32/748
The 32nd Annual Meeting of the Association for Natural Language Processing

2026.03 **Committee Special Award** 32/748
The 32nd Annual Meeting of the Association for Natural Language Processing

2026.03 **ELYZA Award and CyberAgent Award**
The 32nd Annual Meeting of the Association for Natural Language Processing

2026.03 **Best Paper Award**
Journal of Natural Language Processing (2025)

2026.01 **Best Paper Award (Alignment Track)**
AAAI 2026

2025.09 **Polaris.ai Award**
The 20th Young Researchers Symposium on Natural Language Processing

2025.05 **JSAI Annual Conference Award**
The 39th Annual Conference of the Japanese Society for Artificial Intelligence

2025.05 **JSAI Annual Conference Award**
The 39th Annual Conference of the Japanese Society for Artificial Intelligence

2025.03 **Committee Special Award** 32/726
The 31st Annual Meeting of the Association for Natural Language Processing

2025.03 **Committee Special Award** 32/726
The 31st Annual Meeting of the Association for Natural Language Processing

2024.09 **Encouragement Award and CyberAgent Award**
The 19th Young Researchers Symposium on Natural Language Processing

2023.08 **PKSHA Technology Award and Hakuhodo Technologies Award**
The 18th Young Researchers Symposium on Natural Language Processing

2023.03 **Committee Special Award** 26/579
The 29th Annual Meeting of the Association for Natural Language Processing

2022.03 **Outstanding Paper Award** 7/386
The 28th Annual Meeting of the Association for Natural Language Processing

2021.08 **Recommended Doctoral Thesis**
Information Processing Society of Japan

2017.03 **Young Researcher Award**
The 23rd Annual Meeting of the Association for Natural Language Processing

> Fellowships

2021.04 – 2024.03 **JSPS Research Fellowship for Young Scientists (PD), declined**

2019.04 – 2021.03 **JSPS Research Fellowship for Young Scientists (DC2)**

2018.07.02 – 07.27 **Deep Learning Camp Jeju**
Selective research camp

> Grants

2019.04 – 2021.03 **Automatic Grammatical Error Correction Considering Learners' Writing Proficiency**
JSPS KAKENHI Grant-in-Aid for JSPS Fellows (19J14084), Principal Investigator, JPY 900,000

> Publications

■ Conferences and Workshops

2027

Sparse-to-Sparse Feature Learning for Cross-Layer Analysis and Steering
Masahiro Kaneko, Ayana Niwa, Timothy Baldwin
EACL

Tracing How Language Models Integrate Temporal Anchors into Factual Recall
Ayana Niwa, Masahiro Kaneko, Kentaro Inui
EACL

2026

Poisoning-Based Backdoors Are Not Cryptographic
Masahiro Kaneko
NeurIPS

Measuring the Impact of Big Five Personality Traits on Suicide Severity Level Detection with Large Language Models
Kit May Shum, Yusuke Sakai, Masahiro Kaneko, Taro Watanabe
AACL-IJCNLP Findings

DyMT-ESB: Dynamic Multi-Turn Evaluation of Social Bias in User-LLM Interactions
Rem Hida, Masahiro Kaneko, Daisuke Oba, Danushka Bollegala, Naoaki Okazaki
EMNLP Findings

In-Context Learning as Implicit Policy Gradient
Masahiro Kaneko, Timothy Baldwin
COLM

Autoregressive Direct Preference Optimization
Masanari Oi, Mahiro Ukai, Masahiro Kaneko, Naoaki Okazaki, Nakamasa Inoue
ICML

Disentangling the Effects of Unlearning in Measuring Parametric Faithfulness of Chain-of-Thought
Ryo Mitsuhashi, Gaku Morio, Ayana Niwa, Masahiro Kaneko, Kentaro Inui, Terufumi Morishita, Yuta Koreeda, Yasuhiro Sogawa
ACL SRW

A Multilingual Social Bias Benchmark Incorporating Thinking Processes
Masahiro Kaneko, Danushka Bollegala, Timothy Baldwin
ACL

Is Human-like Text Liked by Human? Multilingual Human Detection and Preference Against AI
Yuxia Wang, Rui Xing, Jonibek Mansurov, Giovanni Puccetti, Zhuohan Xie, Minh Ngoc Ta, Jiahui Geng, Jinyan Su, Mervat Abassy, Saad El Dine Ahmed, Kareem Elozeiri, Nurkhan Laiyk, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Ryuto Koike, Masahiro Kaneko, Artem Shelmanov, Ekaterina Artemova, Vladislav Mikhailov, Akim Tsvigun, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, Preslav Nakov

ACL

From Heard to Lived Opinions: Simulating Opinion Dynamics with Grounded LLM Agents in Economic Environments
Ryuji Hashimoto, Masahiro Kaneko, Ryosuke Takata, Takehiro Takayanagi, Kiyoshi Izumi
ACL Findings

ExaGPT: Example-Based Machine-Generated Text Detection for Improving Interpretability
Ryuto Koike, Masahiro Kaneko, Ayana Niwa, Preslav Nakov, Naoaki Okazaki
ACL Findings

JailNewsBench: Multi-Lingual and Regional Benchmark for Fake News Generation under Jailbreak Attacks
Masahiro Kaneko, Ayana Niwa, Timothy Baldwin
ICLR

Stopping Computation for Converged Tokens in Masked Diffusion-LM Decoding
Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, Naoaki Okazaki
ICLR

HarmonicEval: Multi-modal, Multi-task, Multi-criteria Automatic Evaluation Using a Vision Language Model
Masanari Ohi, Masahiro Kaneko, Naoaki Okazaki, Nakamasa Inoue
LREC

On the Alignment of Large Language Models with Global Human Opinion
Yang Liu, Masahiro Kaneko, Chenhui Chu
AAAI **Best Paper Award (Alignment Track)**

2025

Bits Leaked per Query: Information-Theoretic Bounds on Adversarial Attacks against LLMs
Masahiro Kaneko, Timothy Baldwin
NeurIPS **Spotlight**

Online Learning Defense against Iterative Jailbreak Attacks via Prompt Optimization
Masahiro Kaneko, Zeerak Talat, Timothy Baldwin
IJCNLP-AACL

Investigating How Pre-training Data Leakage Affects Models' Reproduction and Detection Capabilities
Masahiro Kaneko, Timothy Baldwin
EMNLP

Balanced Multi-Factor In-Context Learning for Multilingual Large Language Models
Masahiro Kaneko, Alham Fikri Aji, Timothy Baldwin
EMNLP

Social Bias Evaluation for Large Language Models Requires Prompt Variations
Rem Hida, Masahiro Kaneko, Naoaki Okazaki
EMNLP Findings

Sampling-based Pseudo-Likelihood for Membership Inference Attacks
Masahiro Kaneko, Youmi Ma, Yuki Wata, Naoaki Okazaki
ACL Findings

Rectifying Belief Space via Unlearning to Harness LLMs' Reasoning
Ayana Niwa, Masahiro Kaneko, Kentaro Inui
ACL Findings

An Ethical Dataset from Real-World Interactions Between Users and Large Language Models
Masahiro Kaneko, Danushka Bollegala, Timothy Baldwin
IJCAI

The Gaps between Fine Tuning and In-context Learning in Bias Evaluation and Debiasing
Masahiro Kaneko, Danushka Bollegala, Timothy Baldwin
COLING

GenAI Content Detection Task 1: English and Multilingual Machine-generated Text Detection: AI vs. Human
Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, Preslav Nakov
GenAIC

2024

How You Prompt Matters! Even Task-Oriented Constraints in Instructions Affect LLM-Generated Text Detection
Ryuto Koike, Masahiro Kaneko, Naoaki Okazaki
EMNLP Findings

SAIE Framework: Support Alone Isn't Enough - Advancing LLM Training with Adversarial Remarks
Mengsay Loem, Masahiro Kaneko, Naoaki Okazaki
ECAI

Likelihood-based Mitigation of Evaluation Bias in Large Language Models
Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, Naoaki Okazaki
ACL Findings

Controlled Generation with Prompt Insertion for Natural Language Explanations in Grammatical Error Correction
Masahiro Kaneko, Naoaki Okazaki
LREC-COLING

Evaluating Gender Bias of Pre-trained Language Models in Natural Language Inference by Considering All Labels
Panatchakorn Anantaprayoon, Masahiro Kaneko, Naoaki Okazaki
LREC-COLING

Solving NLP Problems through Human-System Collaboration: A Discussion-based Approach
Masahiro Kaneko, Graham Neubig, Naoaki Okazaki
EACL Findings

In-Contextual Bias Suppression for Large Language Models
Daisuke Oba, Masahiro Kaneko, Danushka Bollegala
EACL Findings

OUTFOX: LLM-generated Essay Detection through In-context Learning with Adversarially Generated Examples
Ryuto Koike, Masahiro Kaneko, Naoaki Okazaki
AAAI

2023

Reducing Sequence Length by Predicting Edit Operations with Large Language Models
Masahiro Kaneko, Naoaki Okazaki
EMNLP

The Impact of Debiasing on the Performance of Language Models in Downstream Tasks is Underestimated
Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki
IJCNLP-AACL

Exploring Effectiveness of GPT-3 in Grammatical Error Correction: A Study on Performance and Controllability in Prompt-Based Methods
Mengsay Loem, Masahiro Kaneko, Sho Takase, Naoaki Okazaki
BEA

Comparing Intrinsic Gender Bias Evaluation Measures without using Human Annotated Examples
Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki
EACL

2022

Gender Bias in Meta-Embeddings
Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki
EMNLP Findings

NAIST-NICT-TIT WMT22 General MT Task Submission
Hiroyuki Deguchi, Kenji Imamura, Masahiro Kaneko, Yuto Nishida, Yusuke Sakai, Justin Vasselli, Huy Hien Vu, Taro Watanabe
WMT

Debiasing isn't enough! - On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks
Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki
COLING

IMPARA: Impact-based Metric for GEC using Parallel Data
Koki Maeda, Masahiro Kaneko, Naoaki Okazaki
COLING

Gender Bias in Masked Language Models for Multiple Languages
Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, Naoaki Okazaki

NAACL

ExtraPhrase: Efficient Data Augmentation for Abstractive Summarization
Mengsay Loem, Sho Takase, Masahiro Kaneko, Naoaki Okazaki
NAACL SRW

Proficiency Matters Quality Estimation in Grammatical Error Correction
Yujin Takahashi, Masahiro Kaneko, Masato Mita, Mamoru Komachi
LREC

Interpretability for Language Learners Using Example-Based Grammatical Error Correction
Masahiro Kaneko, Sho Takase, Ayana Niwa, Naoaki Okazaki
ACL

Sense Embeddings are also Biased -- Evaluating Social Biases in Static and Contextualised Sense Embeddings
Yi Zhou, Masahiro Kaneko, Danushka Bollegala
ACL

Unmasking the Mask -- Evaluating Social Biases in Masked Language Models
Masahiro Kaneko, Danushka Bollegala
AAAI

2021

Studying The Impact Of Document-level Context On Simultaneous Neural Machine Translation
Raj Dabre, Aizhan Imankulova, Masahiro Kaneko
MT Summit

Comparison of Grammatical Error Correction Using Back-Translation Models
Aomi Koyama, Kengo Hotate, Masahiro Kaneko, Mamoru Komachi
NAACL SRW

Sentence Concatenation Approach to Data Augmentation for Neural Machine Translation
Seiichiro Kondo, Kengo Hotate, Tosho Hirasawa, Masahiro Kaneko, Mamoru Komachi
NAACL SRW

Debiasing Pre-trained Contextualised Embeddings
Masahiro Kaneko, Danushka Bollegala
EACL

Dictionary-based Debiasing of Pre-trained Word Embeddings
Masahiro Kaneko, Danushka Bollegala
EACL

2020

Autoencoding Improves Pre-trained Word Embeddings
Masahiro Kaneko, Danushka Bollegala
COLING

Cross-lingual Transfer Learning for Grammatical Error Correction
Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, Mamoru Komachi

COLING

Generating Diverse Corrections with Local Beam Search for Grammatical Error Correction
Kengo Hotate, Masahiro Kaneko, Mamoru Komachi
COLING

SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction
Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, Mamoru Komachi
COLING

Towards Multimodal Simultaneous Neural Machine Translation
Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, Mamoru Komachi
WMT

A Self-Refinement Strategy for Noise Reduction in Grammatical Error Correction
Masato Mita, Shun Kiyono, Masahiro Kaneko, Jun Suzuki, Kentaro Inui
EMNLP Findings

Translation of New Named Entities from English to Chinese
Zizheng Zhang, Tosho Hirasawa, Wei Houjing, Masahiro Kaneko, Mamoru Komachi
WAT

Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction
Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, Kentaro Inui
ACL

2019

Gender-preserving Debiasing for Pre-trained Word Embeddings
Masahiro Kaneko, Danushka Bollegala
ACL

Controlling Grammatical Error Correction Using Word Edit Rate
Kengo Hotate, Masahiro Kaneko, Satoru Katsumata, Mamoru Komachi
ACL SRW

Grammatical-Error-Aware Incorrect Example Retrieval System for Learners of Japanese as a Second Language
Mio Arai, Masahiro Kaneko, Mamoru Komachi
BEA

Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection
Masahiro Kaneko, Mamoru Komachi
CICLing

Cross-Corpora Evaluation and Analysis of Grammatical Error Correction Models — Is Single-Corpus Evaluation Enough?
Masato Mita, Tomoya Mizumoto, Masahiro Kaneko, Ryo Nagata, Kentaro Inui
NAACL

2017 **Grammatical Error Detection Using Error- and Grammaticality-Specific Word Embeddings**
Masahiro Kaneko, Yuya Sakaizawa, Mamoru Komachi
IJCNLP

Journals

2022 **Pre-trained Word Embedding and Language Model Improve Multimodal Machine Translation: A Case Study in Multi30K**
Tosho Hirasawa, Masahiro Kaneko, Aizhan Imankulova, Mamoru Komachi
IEEE Access

2019 **Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection**
Masahiro Kaneko, Mamoru Komachi
Computación y Sistemas

Preprints

2026 **How Does Post-Training Shape LLM Value Preferences?**
Yang Liu, Shaoyang Xu, Masahiro Kaneko, Chengzhi Zhong, Wenxuan Zhang, Chenhui Chu
arXiv

Shaping the Document, Not the Model: Optimized Suffixes for Context Following
Gia-Bao Dinh Ho, Sekh Mainul Islam, Masahiro Kaneko, Timothy Baldwin, Isabelle Augenstein
arXiv

Rethinking the Evaluation of LLM Fact-Checkers
Ayana Niwa, Miyu Yamada, Kouta Nakayama, Masahiro Kaneko, Hisami Suzuki, Satoshi Sekine
arXiv

Silicon Samples Match the Answers but Not the Respondents
Yusuke Kumagae, Masahiro Kaneko
arXiv

Speak vs. Know: Balancing Behavior and Knowledge Distributions of Large Language Models
Yang Liu, Masahiro Kaneko, Chenhui Chu
arXiv

Disclosure Begets Disclosure: A Self-Reinforcing Privacy Loop in LLM Agents with Persistent Memory
Masahiro Kaneko, Hiromi Arai, Timothy Baldwin
arXiv

A Japanese Benchmark for Evaluating Social Bias in Reasoning Based on Attribution Theory
Taihei Shiotani, Masahiro Kaneko, Naoaki Okazaki
arXiv

Paraphrasing Adversarial Attack on LLM-as-a-Reviewer
Masahiro Kaneko
arXiv

2025 **Machine Text Detectors are Membership Inference Attacks**
Ryuto Koike, Liam Dugan, Masahiro Kaneko, Chris Callison-Burch, Naoaki Okazaki
arXiv

Intent-Aware Self-Correction for Mitigating Social Biases in Large Language Models
Panatchakorn Anantaprayoon, Masahiro Kaneko, Naoaki Okazaki
arXiv

> Talks

2026.10 **AI Safety Through Its Questions**
Large Language Model Course 2, Matsuo-Iwasawa Lab, The University of Tokyo (invited lecture)

2026.10 **Talk at the Working Group Meeting**
Research and Development Center for Large Language Models, National Institute of Informatics

2026.04 **Offensive AI Safety: N Must-Read Papers**
NLP Colloquium

2026.03 **A Remedy for the International Conference Slump**
Workshop ”Aiming for International Conferences with Ease”, The 32nd Annual Meeting of the Association for Natural Language Processing

2025.08 **Talk at the Tuning, Evaluation, and Safety Working Group**
LLM-jp

2025.07 **Panelist, Panel Discussion 2**
LLM Safety Symposium 2025

2025.07 **Toward Guardrails for Japanese LLMs**
LLM Safety Symposium 2025

2023.08 **Fairness of Language Models**
The 18th Young Researchers Symposium on Natural Language Processing (invited)

2022.07 **Interpretability for Language Learners Using Example-Based Grammatical Error Correction**
NLP Colloquium

2021.11 **Fairness in Pre-trained Word Embeddings**
Nagoya NLP Seminar

2020.07 **Grammatical Error Detection and Correction, and Discriminatory Bias**
Okazaki Lab, Tokyo Institute of Technology

2019.10 **Gender-preserving Debiasing for Pre-trained Word Embeddings**
Academia Meets Amazon

2019.06 **My Research at the University of Liverpool**
The 17th NLP Tokyo D Meetup

2018.07 **Document-level Re-ranking for NMT**
TFCON 2018

2018.06 **My Master's Research and Work at BEA**
The 14th NLP Tokyo D Meetup

2017.12 **English Grammatical Error Detection Using Word Embeddings That Consider Correctness Information and Error Patterns**
The 9th Workshop on Input Method (IM 2017)

> Organizing

2026.03 **Workshop organizer**
The 32nd Annual Meeting of the Association for Natural Language Processing

2025.05 **Organized session "Safety Measures for Large Language Models" (lead organizer)**
The 39th Annual Conference of the Japanese Society for Artificial Intelligence

2021.03 – 2023.03 **Editorial Committee**
Journal of Natural Language Processing

2018 – 2020 **Organizing Committee**
Young Researchers Association for NLP Studies

2018 – 2020 **Organizer**
NLP Tokyo D Meetup

> Reviewing

2027 **Area Chair**: AAAI, ARR (EACL, NAACL)

Reviewer: ICLR

2026 **Area Chair**: ARR (EACL, ACL, EMNLP, AACL)

Reviewer: AAAI, ICLR, ICML, Computational Linguistics

2025 **Area Chair**: ARR (NAACL, ACL, EMNLP, AACL)

Reviewer: NeurIPS, AAAI, TACL

2024 **Area Chair**: *SEM, COLING, ARR (EMNLP)

Reviewer: AAAI, IJCAI, KDD, ECAI, ARR (NAACL, ACL)

2023 **Area Chair**: EMNLP

	Reviewer: ACL, AAAI, BEA
2022	**Reviewer**: AAAI, COLING, ARR (EMNLP)
2021	**Reviewer**: NAACL, ACL, EMNLP, ARR
2020	**Reviewer**: ACL, EMNLP
	Secondary Reviewer: ACL, EMNLP, CIKM
2019	**Reviewer**: PACLIC
	Secondary Reviewer: NAACL, ACL, ACL Demo, *SEM, BEA
2018	**Secondary Reviewer**: NAACL, ACL Demo, EMNLP Demo
2017	**Secondary Reviewer**: NLPTEA