www.gkovacs.com

Geza Kovacs

geza@cs.stanford.edu

Education

Ph.D, Computer Science	Stanford University	GPA: 4.0/4.0	2019
B.S. and **M.Eng**, Computer Science	Massachusetts Institute of Technology	GPA: 5.0/5.0	2013

Work Experience

Member of Technical Staff Anthropic San Francisco Aug 2025 - present
Post-training research with a focus on RL with rubrics, reward models, automatic evaluation, and synthetic data.

Senior Research Scientist Google Mountain View Mar 2022 - Aug 2025
Worked on Gemini post-training and evals, focusing on improving Gemini's multilingual capabilities by training models, running post-training experiments, generating synthetic datasets, and building reward models and evals.
Developed LLM inference techniques to improve quality via MBR decoding with metric ensembles (WMT 2024).
Trained specialized Gemini and Gemma models via distillation and post-training (TranslateGemma, WMT 2025).
Built Gemini autoraters and reward models such as Specialist AutoMQM (ICML 2025) and MetricX-25 (WMT 2025).
Built LLM evaluations and datasets, such as WMT24++ (ACL 2025) and SMOL (WMT 2025).
Led post-training and built synthetic datasets for the Gemini-based personal health assistant LLM.
Developed LLM agents for health question answering (Nature 2026), and health QA benchmarks for LLMs (paper).

Principal Research Scientist LILT San Francisco Aug 2019 - Mar 2022
I headed a team of researchers to evaluate and improve Lilt's interactive neural machine translation system.
Developed metrics for evaluating interactive machine translation systems (WMT 2023, WMT 2022, AMTA 2020).
Developed corpus and system for automatic translation error detection and correction (NAACL 2022 best paper).
Evaluated effects of human vs machine translations on website engagement (AMTA 2022 best presentation).
Developed novel architectures and techniques to improve interactive machine translation inference speed (Patent).

Graduate Researcher (Ph.D) Stanford University Advisor: Michael Bernstein Sep 2013 - July 2019
Created HabitLab, a reinforcement learning powered adaptive behavior change system with 12,000+ active users.
Published papers on adaptive interventions (CHI 2021, CHI 2019, CSCW 2018), crowdsourcing (UIST 2017), large-scale interaction data (L@S 2016), NLP for language learning (CHI 2014) and education (CSCW 2017).

Research Intern Microsoft Research Redmond and Beijing Summer 2015 and 2014
Software Engineering Intern Google Mountain View Summer 2013, 2012, and 2011
Software Development Engineer Intern Microsoft Redmond Summer 2010

Select Awards and Honors

Best Paper Award (Best New Task) and Best Paper Award (Best New Resource), NAACL 2022	2022
Outstanding Paper Award (contribution to special theme on human-centered NLP), NAACL 2022	2022
Best Presentation Award, AMTA 2022	2022
National Defense Science and Engineering Graduate Fellowship	2013
National Science Foundation Graduate Research Fellowship	2013
Phi Beta Kappa (top 10% of students at MIT), Tau Beta Pi (top 12.5% of Engineering students at MIT)	2012

Patents

Generative Model Based Health and Activity Recommendations. US Patent US20250378934A1.
Partial execution of translation in browser. US Patent US11900073B2.
Apparatus and method for accurate translation reviews and consistency across multiple translators. US Patent US11361170B1.
Automated formation of specialized dictionaries. US Patent US9483460B2.

Select Publications.

Transforming Wearable Data into Health Insights using Large Language Model Agents. Nature, 2026.
From Jack of All Trades to Master of One: Specializing LLM-based Autoraters to a Test Set. ICML 2025.
WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. ACL 2025.
Google Translate's Research Submission to WMT2025. WMT 2025.
MetricX-25 and GemSpanEval: Google Translate Submissions to the WMT25 Evaluation Shared Task. WMT 2025.
SMOL: Professionally translated parallel data for 115 under-represented languages. WMT 2025.
Mitigating Metric Bias in Minimum Bayes Risk Decoding. WMT 2024.

TranslateGemma Technical Report. Google Deepmind Technical Report, 2026.
Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Google Deepmind Technical Report, 2025.
Automatic Correction of Human Translations. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. 2022*. **Best Paper Award**, Best New Task, and Best New Resource.
Large Language Models are Few-Shot Health Learners. Under review, 2025.
Findings of the Word-Level AutoCompletion Shared Task in WMT 2023. *Proceedings of the Seventh Conference on Machine Translation (WMT). 2023.*
Measuring the Effects of Human and Machine Translation on Website Engagement. *Proceedings of the 15th Conference of the Association for Machine Translation in the Americas (Research Track). 2022*. **Best Presentation Award.**
Findings of the Word-Level AutoCompletion Shared Task in WMT 2022. *Proceedings of the Seventh Conference on Machine Translation (WMT). 2022.*
The Impact of Text Presentation on Translator Performance. *Target: International Journal of Translation Studies, 2021*.
Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Expect To Re-Strengthen It Imminently. *ACM annual conference on Human Factors in Computing Systems (CHI) 2021*. Acceptance rate: 23%.
Conservation of Procrastination: Do Productivity Interventions Save Time Or Just Redistribute It? *ACM annual conference on Human Factors in Computing Systems (CHI) 2019*. Acceptance rate: 23.8%.
Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. *ACM annual conference on Computer-Supported Cooperative Work and Social Computing (CSCW) 2018*. Acceptance rate: 26%.
Crowd Research: Open and Scalable University Laboratories. *ACM Symposium on User Interface Software and Technology (UIST) 2017*. Acceptance rate: 22%.
EduFeed: A Social Feed to Engage Preliterate Children in Educational Activities. *ACM annual conference on Computer-Supported Cooperative Work and Social Computing (CSCW) 2017*. Acceptance rate: 35%.
Effects of In-Video Quizzes on MOOC Lecture Viewing. *ACM annual conference on Learning at Scale (L@S) 2016*. Acceptance rate: 22%.
Smart Subtitles for Vocabulary Learning. *ACM annual conference on Human Factors in Computing Systems (CHI) 2014*. Acceptance rate: 23%.

Open Source Projects

UNetbootin (LiveUSB Creator) https://en.wikipedia.org/wiki/UNetbootin
40 million downloads. UNetbootin creates bootable USB flash drives for various (50+) Linux distributions.

Wubi (Ubuntu Installer for Windows) https://en.wikipedia.org/wiki/Wubi_(software)
Now part of Ubuntu. Built the first versions of Wubi, which allows Ubuntu to be installed from Windows.

HabitLab (In-the-wild Behavior Change Research Platform) https://habitlab.stanford.edu
12,000+ daily active users. I built HabitLab over my Ph.D, and it is still used for research at Stanford Medical School.

Academic Conference Reviewing and Committees

Organizing Committee, WMT Shared Task on Word-Level Auto-Completion	2022-2023
Program Committee, EACL 2021 Bridging HCI and NLP Workshop	2021
Reviewer, Association of Computational Linguistics (ACL)	2024-2025
Reviewer, ACM Conference on Human Factors in Computing Systems (CHI)	2015, 2018-2019, 2021-2024
Reviewer, ACM Conference on Designing Interactive Systems (DIS)	2023
Reviewer, ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)	2021
Reviewer, ACM Transactions on Computer-Human Interaction (TOCHI)	2022
Reviewer, ACM Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)	2019
Reviewer, ACM Symposium on User Interface Software and Technology (UIST)	2017-2018

Researchers Managed

Sai Gouravajhala, Senior Research Scientist at Lilt.	August 2020 – March 2022
Hannah Yan, Senior Data Scientist at Lilt.	September 2020 – March 2022
Jordan Huffaker, Research Intern at Lilt. Now a Ph.D student at University of Michigan.	Summer 2021
Jessy Lin, Research Engineer at Lilt. Now a Ph.D student at UC Berkeley.	August 2019 – August 2020
Ming-Chang Chiu, Data Science Intern at Lilt. Now a Ph.D student at USC.	Summer 2020

Skills

LLM Development: Model training, evaluation, distillation, synthetic data generation, data ablations, data filtering.
Programming Languages and Tools: Python, JAX, Tensorflow, PyTorch, Numpy, Scipy, Jupyter, Apache Beam.
Languages: Fluent: English, Chinese (Mandarin), Hungarian. Intermediate: Japanese, Vietnamese, Spanish.