# Rami Al-Rfou

Senior Staff Research Scientist · Waymo

*Menlo Park, CA 94025*

rami@alrfou.com | alrfou.com | ramieid | BedouinRanger

## Research Interests

◇ I am excited about developing the scaling laws theory into AGI. I have been leading a team at Waymo pioneering scaling multimodal autoregressive models (world modeling). Improving motion forecasting, planning and safety of real world robotaxis. Before that I worked on LLMs at Google Research.

## Research and Industry Experience

### OpenAI
*San Francisco, CA*

Member of Technical Staff / TLM — *August 2024 - Present*

- Leads a robot learning team to develop embodied foundational models.

### Waymo Research
*Mountain View, CA*

Senior Staff Research Scientist - TLM — *March 2021 - August 2024*

- Pioneered the development foundational models for motion forecasting and planning.
- Established the scaling laws for open-loop and closed-loop metrics.
- Designed coherent, consistent, and efficient E2E multimodal behavior models.
- Developed novel distillation methods to enable deployment of foundational models.
- Managed and led a the foundational models team at Waymo Research.

### Google Research
*Mountain View, CA*

Staff Research Scientist - TL — *May 2015 - March 2021*

- **PEFT**: Envisioned and designed prompt tuning to leverage LLMs more efficiently.
- **LLMs**: Led and developed several multilingual large language models such as mT5, ByT5.
- **Deep Retrieval**: Pioneered the study of knowledge and retrieval augmented language models.
- **SmartCompose**: Desinged a deep retrieval language model to speed up Gmail SmartCompose. Awarded the 2018 *Feats of Engineering* for my contributions.
- **SmartReply**: Designed and built multilingual and character based SmartReply models that were deployed in Gmail, Youtube, Google Docs, Google Play.

### Microsoft Research
*New York, NY*

Research Intern — *Summer 2013*

Worked with Léon Bottou on active learning and prior knowledge integration.

### Google Research
*Mountain View, CA*

Research Intern — *Summer 2012*

Designed and implemented multilingual coreference resolution of noun phrases based on word embeddings. First application of word embeddings at the time.

### Middle East Technical University
*Kalkanli, North Cyprus*

Lecturer — *Feb 2009 - June 2010*

Taught laboratory courses in computer architecture and design, circuits, and analog amplifiers.

## Foundational Models for Autonomous Vehicles

- ◇ "MoST: Multi-modality Scene Tokenization for Motion Prediction", Norman Mu, Jingwei Ji, Zhenpei Yang, Nate Harada, Haotian Tang, Kan Chen, Charles R. Qi, Runzhou Ge, Kratarth Goel, Zoey Yang, Scott Ettinger, **Rami Al-Rfou**, Dragomir Anguelov, Yin Zhou, *Proceedings of CVPR 2024*

- ◇ "WOMD-LiDAR: Raw Sensor Dataset Benchmark for Motion Forecasting", Kan Chen, Runzhou Ge, Hang Qiu, **Rami AI-Rfou**, Charles R. Qi, Xuanyu Zhou, Zoey Yang, Scott Ettinger, Pei Sun, Zhaoqi Leng, Mustafa Baniodeh, Ivan Bogun, Weiyue Wang, Mingxing Tan, Dragomir Anguelov, *Proceedings of ICRA 2024*

- ◇ "Scaling Motion Forecasting Models with Ensemble Distillation", Scott Ettinger, Kratarth Goel, Avikalp Srivastava, **Rami Al-Rfou**, *Proceedings of ICRA 2024*

- ◇ "MotionLM: Multi-Agent Motion Forecasting as Language Modeling", Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S. Refaat, **Rami Al-Rfou**, Benjamin Sapp, *Proceedings of ICCV 2023*

- ◇ "Wayformer: Motion Forecasting via Simple & Efficient Attention Networks", Nigamaa Nayakanti, **Rami Al-Rfou**, Aurick Zhou, Kratarth Goel, Khaled S. Refaat, Benjamin Sapp, *Proceedings of ICRA 2023*

- ◇ "Narrowing the coordinate-frame gap in behavior prediction models: Distillation for efficient and accurate scene-centric motion forecasting", DiJia Andy Su, Bertrand Douillard, **Rami Al-Rfou**, Cheol Park, Benjamin Sapp, *Proceedings of ICRA 2022*

## Large Language Models

- ◇ "SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer", Tu Vu, Brian Lester, Noah Constant, **Rami Al-Rfou**, Daniel Cer, *Proceedings of ACL 2022*

- ◇ "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models", Linting Xue, Aditya Barua, Noah Constant, **Rami Al-Rfou**, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel, *Proceedings of TACL 2022*

- ◇ "The Power of Scale for Parameter-Efficient Prompt Tuning", Brian Lester, **Rami Al-Rfou**, Noah Constant, *Proceedings of EMNLP 2021*

- ◇ "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer", Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, **Rami Al-Rfou**, Aditya Siddhant, Aditya Barua, Colin Raffel, *Proceedings of NAACL 2021*

- ◇ "nmT5 - Is parallel data still relevant for pre-training massively multilingual language models?", Mihir Sanjay Kale, Aditya Siddhant, **Rami Al-Rfou**, Linting Xue, Noah Constant, Melvin Johnson, *Proceedings of ACL 2021*

- ◇ "Wiki-40B: Multilingual Language Model Dataset", Mandy Guo, Zihang Dai, Denny Vrandečić, **Rami Al-Rfou**, *Proceedings of LREC 2020*

- ◇ "Bridging the Gap for Token-Free Language Models", DK Choe, **Rami Al-Rfou**, Mandy Guo, Heeyoung Lee, Noah Constant, *Proceedings of BayLearn 2019*

- ◇ "Character-Level Language Modeling with Deeper Self-Attention", **Rami Al-Rfou**, Dokook Choe, Noah Constant, Mandy Guo, Llion Jones, *Proceedings of AAAI 2019*

## Retrieval & Knowledge Augmented Generation

- ◇ "LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool", Uma Roy, Noah Constant, **Rami Al-Rfou**, Aditya Barua, Aaron Phillips, Yinfei Yang, *Proceedings of EMNLP 2020*

- ◇ "Machine Translation Aided Bilingual Data-to-Text Generation and Semantic Parsing", Oshin Agarwal, Mihir Kale, Heming Ge, Siamak Shakeri, **Rami Al-Rfou**, *Proceedings of Web Nautral Language Generation 2020*

- ◇ "Efficient Natural Language Response Suggestion for Smart Reply", Matthew Henderson, **Rami Al-Rfou**, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, Ray Kurzweil, *arXiv:1705.00652*

- ◇ "Conversational Contextual Cues: The Case of Personalization and History for Response Ranking", **Rami Al-Rfou**, Marc Pickett, Javier Snaider, Yun-hsuan Sung, Brian Strope, Ray Kurzweil, *arXiv:1606.00372*

## Education

### Stony Brook University

*Stony Brook, NY*

Ph.D. in Computer Science

- Advisor: Steven Skiena | **Thesis**: Polyglot- Massive Multilingual Natural Language Processing Pipeline.

### University of Jordan

*Amman, Jordan*

B.Sc. in Computer Engineering | GPA: 3.79

## Honors & Awards

| | |
|---|---|
| 2024 | **Test of Time Award**, KDD — Barcelona, Spain |
| 2022 | **King Abdullah II Order for Distinction**, Jordan's 75th Independence Celebration |
| 2018 | **Feats of Engineering**, Google Awards |
| 2008 | **Most Innovative Activity**, IEEE Region 8 Student Branch Conference |
| 2007 | **Representative of the Youth Delegation**, Jordan's State visit to China |
| 2004 | **Jordan's High Education Ministry Fellowship**, Ranked 15th in university entrance exams |
| 2002 | **Finalist**, Jordan's first Math & Physics Olympiads |