Wittawat Jitkrittum

New York wittawat@google.com wittaw.at in wittawat-jitkrittum

Education

PhD in Machine Learning *2013 - 2017*
Gatsby Unit, University College London

- **Dissertation:** Kernel-based distribution features for statistical tests and Bayesian inference
- **Advisor** Arthur Gretton

MEng in Computer Science *2010 - 2012*
Tokyo Institute of Technology, Japan

BSc in Computer Science *2005 - 2009*
Sirindhorn International Institute of Technology, Thammasat University, Thailand

Research Experience

Senior Research Scientist *Aug 2025 – present*
Google DeepMind

Senior Research Scientist *May 2024 – Aug 2025*
Google Research, New York

Research Scientist *May 2020 – April 2024*
Google, New York

- Developed and published machine learning techniques for efficient inference of large language models.

Postdoctoral Researcher *Jan 2020 – April 2020*
Max Planck Institute for Intelligent Systems, Tübingen, Germany

- Developed statistical techniques for comparing intractable generative models (e.g., unnormalized models, GANs).
- Published at NeurIPS 2018, ICML 2019, 2× NeurIPS 2019, 2× UAI 2020, NeurIPS 2020, and AISTATS 2020.

Awards and Honors

ICLR 2025 Outstanding Paper Honorable Mention *2019*

- For work titled ”Faster Cascades via Speculative Decoding” co-authored with colleagues at Google.
- Awarded to 6 out of over 11000 total submissions.

Google Tech Impact Award 2024 *2024*

- For product impact from our work on model routing.
- Awarded to 10 most impactful projects in 2024.

Google Research Tech Impact Award Winner 2023 *2023*

- For product impact from our work on cascade models

ELLIS PhD Award *2019*

- For outstanding research achievements during the PhD dissertation phase. Awarded to 6 recipients in 2016-2018. https://ellis.eu/news/ellis-phd-award.

NeurIPS 2017 Best Paper Award *2017*

- For work titled ”A Linear-Time Kernel Goodness-of-Fit Test”.
- Awarded to 3 out of 3240 submissions to NeurIPS 2017.

Gatsby Unit Studentship (PhD study) *2013 – 2017*

- Full scholarship with stipend for PhD study at Gatsby Unit, University College London. Awarded to 2-5 students globally per year.

Gatsby Unit Studentship (PhD study) *2013 – 2017*

- Awarded to 3 out of 3240 submissions to NeurIPS 2017.

Okazaki Kaheita Scholarship (master's degree) *2010 – 2012*

- Awarded to one Thai student once every 2-3 years.

Runner-up at National Software Contest (NSC), Thailand *2010*

- Designed a character-level word tokenizer with a binary classifier. Written in Java. Accuracy: 95.5%.

Runner-up at National Software Contest (NSC), Thailand *2009*

- One of the first factoid Thai Q&A systems based on Thai Wikipedia.

Filed Patents

- US20240311405A1: Dynamic selection from among multiple candidate generative models with differing computational efficiencies (2024)
- US20240135254A1: Performing classification tasks using post-hoc estimators for expert deferral (2023)

Professional Activities

- Volunteer **arXiv moderator** for Computation and Language (since Oct 2024)
- **Area chair** for ICML 2025, NeurIPS 2024, 2025, ICLR 2023, ACML 2020-2021, 2024, AISTATS 2025
- **Workflow Chair** for AISTATS 2021
- **Publicity Chair** for AISTATS 2016
- **Reviewer**
  - NeurIPS 2015-2022 [Top 10% reviewer of NeurIPS 2020].
  - ICML 2016-2019, 2021-2023 [Top 5% reviewer of ICML 2019].
  - UAI 2021 emergency reviewer
  - AISTATS 2017-2019
  - Asian Conference on Machine Learning (ACML) 2017
  - International Conference on Learning Representations (ICLR) 2017, 2025
  - NeurIPS Workshop on Advances in Approximate Bayesian Inference 2015-2017.
- **Co-organizer of regional machine learning summer schools:**
  - The 2nd MLRS in Bangkok, Thailand, 2-9 August 2023. https://www.mlrs.ai
  - The 2nd OAMLS 2022. https://www.acml-conf.org/2022/oamls.html.
  - The 1st Online Asian Machine Learning School (OAMLS) 2021. https://www.acml-conf.org/2021/school/.
  - The Machine Learning Summer School (MLSS) 2020 (virtual). http://mlss.tuebingen.mpg.de/2020/.
  - The Southeast Asia Machine Learning School, Indonesia, 8-12 July 2019. https://www.seamls.ai.
  - The 1st Machine Learning Research School (MLRS) in Bangkok, Thailand, 4-11 August 2019.

Invited Talks

Chulalongkorn University, Thailand	Jan 2024
Stein's Method: The Golden Anniversary, National University of Singapore	July 2022
Deep Learning and Artificial Intelligence Summer School 2021, Thailand	May 2021
IBM Research, NY, USA. Virtual talk on model comparison of generative models.	Dec 2020
EURECOM, France (virtual talk)	Nov 2020
NeurIPS 2019 Tutorial (audience size: 3000+).	Dec 2019
Swiss Data Science Center, Zürich	Oct 2019

Data, Learning and Inference (DALI) 2019, Spain (http://dalimeeting.org)	Sep 2019
Vidyasirimedhi Institute of Science and Technology (VISTEC), Thailand	Dec 2018
Vidyasirimedhi Institute of Science and Technology (VISTEC), Thailand	Mar 2018
Chulalongkorn University, Thailand	Mar 2018
Bangkok Machine Learning Meetup	Mar 2018
Workshop on Functional Inference and Machine Intelligence, Japan	Feb 2018
Department of Computer Science, University of Bristol	Dec 2017
MLTrain Workshop: Learn How to Code a Paper at NeurIPS 2017	Dec 2017
Probabilistic Graphical Model Workshop II, The Institute of Statistical Mathematics, Japan	Feb 2017
Sugiyama-Sato Lab, University of Tokyo	April 2016
Probabilistic Graphical Model Workshop, The Institute of Statistical Mathematics, Japan	Mar 2016

Mentorship

Name	Affiliation	Context
Sizhe Li	London School of Economics and Political Science	Master's thesis (2023)
Yujie Zhang	London School of Economics and Political Science	Master's thesis (2023)
Zhendong Wang	London School of Economics and Political Science	Master's thesis (2023)
Jingyan Lu	London School of Economics and Political Science	Master's thesis (2023)
Jen Ning Lim	Max Planck Institute for Intelligent Systems	Pre-doctoral internship (2019)

Selected Publications

For the complete list of publications, please see https://wittaw.at.

Preprints

1. Jitkrittum, W., Narasimhan, H., Rawat, A. S., Juneja, J., Wang, Z., Lee, C.-Y., Shenoy, P., Panigrahy, R., Menon, A. K., & Kumar, S. (2025). Universal model routing for efficient LLM inference. https://arxiv.org/abs/2502.08773
2. Li, Q., Chen, K., Su, C., Jitkrittum, W., Sun, Q., & Sangkloy, P. (2025). Cost-aware routing for efficient text-to-image generation. https://arxiv.org/abs/2506.14753
3. Rawat, A. S., Sadhanala, V., Rostamizadeh, A., Chakrabarti, A., Jitkrittum, W., Feinberg, V., Kim, S., Harutyunyan, H., Saunshi, N., Nado, Z., Shivanna, R., Reddi, S. J., Menon, A. K., Anil, R., & Kumar, S. (2024). A little help goes a long way: Efficient LLM training by leveraging small LMs. https://arxiv.org/abs/2410.18779
4. Wang, C., Augenstein, S., Rush, K., Jitkrittum, W., Narasimhan, H., Rawat, A. S., Menon, A. K., & Go, A. (2024). Cascade-aware training of language models. https://arxiv.org/abs/2406.00060

Peer-reviewed articles

1. Narasimhan, H., Jitkrittum, W., Rawat, A. S., Kim, S., Gupta, N., Menon, A. K., & Kumar, S. (2025). Faster cascades via speculative decoding. *ICLR*
2. Kim, S., Rawat, A. S., Zaheer, M., Jitkrittum, W., Sadhanala, V., Jayasumana, S., Menon, A. K., Fergus, R., & Kumar, S. (2024). USTAD: Unified single-model training achieving diverse scores for information retrieval. *ICML*
3. Gupta, N., Narasimhan, H., Jitkrittum, W., Rawat, A. S., Menon, A. K., & Kumar, S. (2024). Language model cascades: Token-level uncertainty and beyond. *ICLR*
4. Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., & Gretton, A. (2023). A Kernel Stein Test for Comparing Latent Variable Models. *Journal of the Royal Statistical Society*, *85*(3), 986–1011
5. Jitkrittum, W., Gupta, N., Menon, A. K., Narasimhan, H., Rawat, A. S., & Kumar, S. (2023). When does confidence-based cascade deferral suffice? *NeurIPS*

6. Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A. S., & Kumar, S. (2022). Post-hoc estimators for learning to defer to an expert. *NeurIPS*
7. Sangkloy, P., Jitkrittum, W., Yang, D., & Hays, J. (2022). A sketch is worth a thousand words: Image retrieval with text and sketch. *European Conference on Computer Vision*
8. Schrab, A., Jitkrittum, W., Szabó, Z., Sejdinovic, D., & Gretton, A. (2022). Discussion of multiscale fisher's independence test for multivariate dependence. *Biometrika*
9. Park, M., Vinaroz, M., & Jitkrittum, W. (2021). ABCDP: Approximate Bayesian computation with differential privacy. *Entropy*, *23*(8)
10. Rawat, A. S., Menon, A. K., Jitkrittum, W., Jayasumana, S., Yu, F. X., Reddi, S., & Kumar, S. (2021). Disentangling sampling and labeling bias for learning in large-output spaces. *ICML*
11. Jitkrittum, W., Kanagawa, H., & Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. *UAI*
12. Lim, J. N., Yamada, M., Schölkopf, B., & Jitkrittum, W. (2019). Kernel Stein tests for multiple model comparison. *NeurIPS*
13. Jitkrittum*, W., Sangkloy*, P., Gondal, M. W., Raj, A., Hays, J., & Schölkopf, B. (2019). Kernel mean matching for content addressability of GANs [**Equal contribution. Long oral presentation.*]. *ICML*
14. Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., & Gretton, A. (2017). A linear-time kernel goodness-of-fit test [Best paper award, 3 out of 3240 submissions]. *NeurIPS*
15. Jitkrittum, W., Szabó, Z., Chwialkowski, K., & Gretton, A. (2016). Interpretable distribution features with maximum testing power [(Oral presentation, 1.8%)]. *NeurIPS*
16. Park*, M., Jitkrittum*, W., & Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings [**Equal contribution. Oral presentation, 6.5%*]. *AISTATS*
17. Jitkrittum, W., Gretton, A., Heess, N., Eslami, S. M. A., Lakshminarayanan, B., Sejdinovic, D., & Szabó, Z. (2015). Kernel-based just-in-time learning for passing expectation propagation messages. *UAI*
18. Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, *26*(1)

(Last update: September 2025)