Bhiman Kumar Baghel

Computer Science PhD · 1st Yr · University Of Pittsburgh

+1-412 844 1085 | bkb45@pitt.edu | https://bhimanbaghel.github.io/ | linkedin.com/in/bhiman-kumar-baghel | Bhiman

Summary

In my PhD, I'm eager to tackle key challenges in grounding language models for commonsense reasoning, improving their fairness and reliability for everyday use. My research background lies in NLU for Conversational AI, including multi-intent & slot classification and managing out-of-domain queries. A significant milestone in my career was leading the overhaul of the Samsung SmartThings machine learning architecture. This experience brought to light the existing model's limitations in adapting to and understanding users' context for proper reasoning. It's also crucial to address their biases for sound decision making and prevent the spread of harmful information. I am committed to pursuing these objectives throughout my doctoral research.

Education

University Of Pittsburgh — *PA, USA*

Doctor of Philosophy - PhD in Computer Science [3.75/4] — *2023 - 2028 Expected*

Indian Institute of Technology (IIT) — *Kharagpur, India*

Master of Technology - M.Tech. in Computer Science & Engineering [CGPA - 8.82/10] — *2017 - 2019*

National institute Of Technology (NIT) — *Jalandhar, India*

Bachelor of Technology - B.Tech. in Computer Science & Engineering [CGPA - 6.82/10] — *2013 - 2017*

Publications

Smart Stacking of DL Models for Granular Joint Intent-Slot Extraction for Multi-Intent SLU[Link] — *IEEE Access (Volume: 9)*

Intent Focused Semantic Parsing and zero-shot Learning for OOD detection in SLU[Link] — *IEEE Access (Volume: 9)*

Patents

The Handling Missing Slots in Single & Multi-intent cases and in Open & Closed Domain[Link] — *US, A1*

Cognitive Decline Based Wellness Management in Smart-Home — *Filling Ongoing*

ASIIST - Assisting Seamless Indirect Interaction in Smart Home — *Filling Ongoing*

Dynamic relevance time prediction in Smart Home — *Filling Ongoing*

Industry/Academic Experience

University of Pittsburgh — *PA, USA*

Teaching Assistant - Operating System — *Sep 2023 - Dec 2023*

- Conducted weekly recitations for class of 50 undergrad students. Graded quizzes and assignments.

Samsung Research Institute — *Bangalore, India*

Lead Engineer, Speech Recognition & Natural Language Processing - Voice Intelligence R&D — *June 2019 - August 2023*

- **Role:** Research, Design and Develop solutions for Bixby's (Conversational AI) industry-level NLU challenges.
- **CoSMIC - Contextual SmartThings Multi-Intent Composer:** Researched and commercialized BERT-based Multi-Intent NLU for SmartThings. Achieved **96% E2E Accuracy** with **12% performance boost**. Reduced the error rate by **67%**.
- **Conversational AI:** Worked on complex scenarios of Multi-Intent and Slot Classification, Zero-Shot Out-Of-Domain Detection, Intelligent Device Resolution service, Bixby Recall - Chat Question Answering (QA) System and Knowledge Representation.
- **Achievements: Publications (x2), Patents (x4), MBO High Performance Bonus (x3), Samsung Excellence and other awards (x5), CLab Finalist.**

IBM — *Bangalore, India*

Extreme Blue Intern - Template based Machine Learning Models for MVS — *May 2018 - July 2018*

- Designed and developed a time series forecasting, self-learning, DL system to predict, prevent or auto-resolve or assist in IT infrastructure Failure.
- **Achievement: [HA2] 2nd Runner Up - Audience Poll Category, IBM Extreme Blue Expo[Link]**

Indian Institute Of Technology (IIT) — *Kharagpur, India*

Teaching Assistant - Programming and Data Structure lab & DBMS — *July 2018 - May 2019*

- Thought lectures on Programming & DBMS concepts to a class of 80+ undergrad students. Evaluated lab assignments and tests.

Projects

Multimodal Understanding of Memes with Fair Explanations — *PhD Term Project*

Fairness Analysis of Human/AI-Generated Summaries of Student Reflections — *PhD Term Project*

DCLL: A Deep Learning Model For Simultaneous Travel Time Estimation and Traffic Congestion Prediction — *M.Tech. Thesis Project*

Automatic Concept Map generation from text-based Learning Material[Link] — *Term Project*

Twitter Sentiment Analysis[Link] — *B.Tech. Final Year Project*

Skills

Specialization	Machine Learning Engineering for Production (MLOps)[Link]
Programming	Python, C/C++, JavaScript, LaTeX, Verilog
Deep Learning	Keras, Tensorflow, Pytorch, Pytorch Lightning, MLFlow
Web Development	Flask, Django, HTML5, CSS, Bootstrap, React
Graph DB	Neo4j

Research Statement

Bhiman Kumar Baghel - 1st Yr CS PhD - University of Pittsburgh, PA, USA

In my PhD, I'm eager to tackle key challenges in grounding language models for commonsense reasoning, improving their fairness and reliability for everyday use. My background in Natural Language Understanding for Conversational AI, including multi-intent & slot classification [1] and managing out-of-domain queries [2], has prepared me well for this endeavor. A significant milestone in my career was leading the overhaul of the Samsung SmartThings machine learning architecture. This was built upon insights gleaned from my research. This experience brought to light the existing model's limitations in adapting to and understanding users' context. It spurred my deep interest in developing language models grounded in the real world, enabling them to reason and incorporate common-sense knowledge more effectively. As Conversational AIs increasingly influence our decisions, it is also crucial to address their biases and prevent the spread of harmful information. I am committed to pursuing these objectives throughout my doctoral research.

Reasoning: Compositional tasks are tasks that require multiple steps of reasoning to execute, like multiplication, or making a coffee. [3] showed that GPT-4 fails to solve multiplication when the complexity increases by increasing the number of digits. Even GPT-3.5 fails to generalize for 5-digit multiplication when exhaustively fine-tuned on 3-digit multiplication. This showed that even if the rule of reasoning remains the same, current LLMs fail to generalize when the complexity is increased.

Works have been done that try to mitigate this limitation either by using an LLM for code [4]. For tasks that require real-world grounding [5] proposed to map LLMs output to atomic actions supported by an embodied agent using a similarity measure. On the other hand, [6] proposes LLMs utilizing the value function of the embodied agent to predict the most feasible atomic action. [7] proposes to iteratively make the model generate the steps/atomic actions until successful execution.

The above-mentioned works have two common characteristics - 1) Reactive - they work on LLMs output and 2) Repetitive - the LLMs repeat the same mistake as no learning is happening. The characteristics make these work inefficient. One possible solution could be fine-tuning. However, it is resource-intensive and not targeted which makes it inefficient. Can we do a targeted learning? This is where memory editing comes into the picture as it's targeted and efficient. Memory editing [8, 9] find the layers responsible for predicting the knowledge you want to edit, and then weights to those layers are updated using gradient descent to predict new knowledge. So, currently, I am working on layer localization and updating the wrong knowledge of the computation task to the correct one.

In parallel to this, I also think the LLMs cannot generalize on compositional tasks because the correctness of the intermediate step is not evaluated. So, I am also working on providing intermediate feed to LLM. I am trying to accomplish this by tweaking the RLHF's [10] reinforcement learning algorithm i.e. PPO to enable the proposed feedback. With such a setting, I also want LLMs to be more interpretable, increasing people's trust in them.

Fairness: I am also interested in the fairness aspect of AI. My project 'FiME: Fairness in Meme Explanation' analyzes biases present in meme explanations generated by Vision Language models (LLaVA [11] & MiniGPT4 [12]). The manual analysis found 4 categories of biases. Interestingly, automatic toxicity evaluators like Perspective API were not able to identify those biased explanations in the majority of the cases. My other project deals with fairness analysis of student reflection summarization [13]. In this project, I ran topic modeling models like BERTopic [14] on student reflections and found differences between topics males and females discussed. Analyzed whether summarizations of these reflections produced by humans as well as AI contain any bias towards any gender. I understand the importance of AI fairness and will continue working in this direction.

References

[1] Niraj Kumar and **Baghel, Bhiman Kumar**. Smart stacking of deep learning models for granular joint intent-slot extraction for multi-intent slu. *IEEE Access*, 9:97582–97590, 2021.

[2] Niraj Kumar and **Baghel, Bhiman Kumar**. Intent focused semantic parsing and zero-shot learning for out-of-domain detection in spoken language understanding. *IEEE Access*, 9:165786–165794, 2021.

[3] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023.

[4] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners, 2022.

[5] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022.

[6] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J

Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022.

[7] Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. Adapt: As-needed decomposition and planning with language models, 2023.

[8] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer, 2023.

[9] Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Lorraine Li, Sarah Wiegreffe, and Niket Tandon. Editing common sense in transformers, 2023.

[10] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[12] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. 2023.

[13] Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. Coursemirror: Enhancing large classroom instructor-student interactions via mobile interfaces and natural language processing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, page 1473–1478, New York, NY, USA, 2015. Association for Computing Machinery.

[14] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.