Lifetime Achievement Award

Retrospective and Future Views

Kathleen McKeown

Columbia University
kathy@cs.columbia.edu

I was honored to receive the Association for Computational Linguistics Lifetime Achievement Award in 2025. I especially want to thank the people who nominated me for the award as I know nominations require time and effort. This retrospective is a rough transcript of the speech I gave accepting the award at the conference in Vienna, Austria. In the talk, I look back at my research at early stages of my career and then look at the arc that research takes and how it relates to work that I still carry out today. I look at the trajectories of four areas of my research: language generation, text summarization, social media analysis, and multimodal analysis of artwork. In the talk, I featured videos of my current students speaking about their research and where they think the field is heading. I dedicate the talk and this article to the amazing students I have had the honor to work with over the years.

In this talk, I have decided to focus both on retrospective views of research over my career as well as a forward-looking component.

This award is particularly meaningful to me because the Association for Computational Linguistics (ACL) has been my home since the start of my career. I've participated in many ACL activities over the years. I point out reviewing because I hope to encourage many of you to be as committed to participating in the reviewing process in a strong way.

I want to focus on three events. The first is my first paper which happened at an ACL conference in San Diego. There were a hundred people in the audience and, thus, the conference was very different from today. It had a single session where we got to know everyone and talked to everyone. A second event was my first membership on an ACL Program Committee (PC). At the time there were about twenty members on the committee. We each reviewed around twenty to thirty papers and I have memories of sitting on the sofa every night for the two to three weeks coming up to the PC meeting reading and writing reviews. The PC met together as a whole and we discussed each paper before making a decision. Again, this was very different from today. Finally, I'll mention my time as ACL president. One of the tasks for the ACL president at that point in time was to give a humorous after-dinner talk at the ACL banquet that everyone attended. This was the most frightening talk of my career. For that talk, I learned a trick that I've often used since then, which is that I gave part of the talk through video. At that time, I had three young girls and I was watching many children's videos—which turned out to be relevant to our work in natural language. Recall that this was 1992 and it was the start of the Linguistic Data Consortium, which made data available at a cost. They were essentially selling corpora and dictionaries. Today, I think this is still relevant

https://doi.org/10.1162/COLI.a.605

if we think about the recent sale of Scale AI. I showed a clip in my dinner talk from the movie *The Phantom Toll Booth*, produced by MGM Studios, showing when Milo and his dog Tok enter Dictionopoulos where vendors at stands are hawking their wares—words of different kinds. If you haven't seen the movie, I highly recommend watching it in full.

Today I will provide future perspectives through video clips of my PhD students and postdoc.

I'll focus in the talk on research strategies and if you take nothing else away, it could be on research strategies that can be helpful in leading to success. I'll also be talking about people I want to celebrate—in particular, the students that I have worked with over the years. I'm going to be talking about four areas of my research. Language generation; text summarization; social media analysis; and generation from multimedia, in particular, artwork.

1. Language Generation

We'll start with language generation, which I began at the time of my PhD. I primarily focused on language generation and, of course, I continue to work on it today. I was tremendously influenced at this time by my advisor Aravind Joshi, who was a professor of computer and cognitive science at the University of Pennsylvania. He was tremendously helpful in giving me advice during the PhD as well as throughout my career.

At the time of my dissertation, everybody was working on natural language interfaces to databases and because of that, an important research topic was parsing. Systems had to parse the natural language query into the language of the underlying API of the database, such as SQL. In contrast, I was doing research on generation. This was a topic that was given to me by my advisor for the first part of my dissertation. I viewed it as a very unpopular topic and that I was somewhat on the edge of the field. I began by working on paraphrasing, rephrasing questions that are posed to databases so that the user could see whether the system had understood what they were asking. I moved from there to text generation. I worked on generation of paragraph length texts where discourse was important. These texts were generated from a system I developed to generate long answers to questions. I viewed language generation as one of choice, in contrast to parsing where the form of the sentence is given as input. In language generation, the system was given some meaning to convey, but it had to choose the syntactic structure to realize that meaning. It had to choose the words to use in the sentence, and for longer discourse, it had to determine what the ordering of the sentence would be.

Aravind Joshi, Henry Salvatori Professor of Computer and Cognitive Science, University of Pennsylvania.

The Denver Nuggets *beat* the Boston Celtics with a narrow margin, 102-101. (*game result in verb*, manner as PP).

The Denver Nuggets *edged out* the Boston Celtics 102-101. (*game result and manner in verb*)

Figure 1
Semantics of the basketball game result and manner is realized differently.

My favorite paper of ours from that period of time was a paper that we wrote on floating constraints on lexical choice. It was viewed as a difficult problem because constraints on word choice came from many different sources and where the word appeared in the sentence could vary quite a bit. Here I consider two examples from a corpus of basketball reports that we were working on at that time. The student analyzed the entire corpus manually, while today we would use statistical methods. Shown in Figure 1 are two sentences from the basketball corpus. In the first, we see that the semantics (the game result) is realized as the verb of the sentence while the manner "with a narrow margin" is realized as the prepositional phrase. In the second sentence, these two pieces of information are merged and realized as one in the main verb "edged out". In this work, we looked at how to represent the variety of constraints that determine lexical choice, whether syntactic, semantic, other word choices, or pragmatic constraints. We showed how to make use of these constraints in the functional unification formalism, a type of grammar formalism.

This was an early form of what we now refer to as **controllable generation**. In my ACL 2020 keynote, I noted that people choose words and form sentence structures intentionally to convey meaning. I said that language models do not. Of course, now we could argue that intention could be conveyed in the prompt to the system, but we would still have to ask whether the model itself had that intention. Choice of individual words, especially when we're generating a long text, is still generally not well controlled.

In our work now, we are looking at methods for controlling word choice. In particular, we're looking at the problem of style transfer, asking whether we can choose words that a particular author would use and looking at this in the context of small models, such as text diffusion models (Horvitz et al. 2024a,b). Shown in Figure 2, we see that we can choose words along a continuum from formal to informal. At the higher right, in the blue, more formal words such as "enamored" are chosen, while all the way down on the lower left, informal words, like "stoked", are chosen. In collaboration with Raghav Singhal and Rajesh Ranganath at New York University (Singhal et al. 2025), we have been looking at a general framework for any kind of control using small models such as diffusion models. My student Zachary Horvitz is working on this. We'll hear from him.

"My name is Zachary Horvitz. I am a third year PhD student at Columbia. And like many PhD students, I have a few different parallel threads. On the NLP methods

Columbia PhD student.

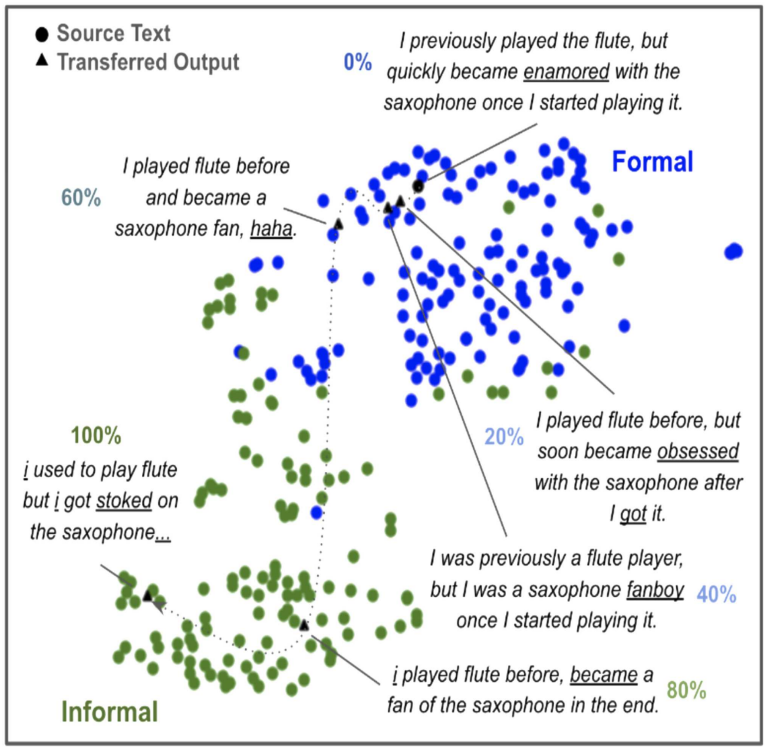


Figure 2
Interpolating in stylistic embedding space to generate texts with varying levels of formality and different word choices.

side, I'm very excited about controllable text generation, in particular with text diffusion models. On the application side, I'm very excited about creative text generation—things like humor. For text diffusion models, there are many different ways of actually controlling outputs. For example, if you're working with continuous text diffusion models that are denoising text in continuous space, you can take gradients with respect to classifiers or reward functions to guide the output iteratively towards certain attributes or certain properties. That's one way you can use diffusion models for control. What's also exciting about discrete diffusion models, which aren't operating in continuous space, is that you can estimate future tokens that you haven't fully decoded yet. You can also perform error correction and potentially fix tokens you've already generated. So there are many different ways you can control the output of a text diffusion model—that you don't necessarily have with a traditional autoregressive language model."

We are also working on methods for controlling content. For example, we are considering how to generate answers to questions or news from different perspectives. We have collected a large corpus of news from different countries that we can use to help us present the same event from the view of different countries. Yunfan Zhang will talk about this.

"My name is Yunfan Zhang and I'm a rising fourth year PhD student in Computer Science. I have been working with Professor McKeown and Professor Muresan for two years now. The topic I'm working on right now is perspective-based generation with reinforcement learning. I think this is really interesting for two reasons. The first is

Columbia PhD student.

that today's language models are trained with a fixed given perspective that is usually aligned with the perspective of model developers (Zhang, McKeown, and Muresan 2025). In real life, we want these models to process different perspectives, writing styles, and personalities. The second reason is that, for reinforcement learning, we have seen a lot of success for tasks that are easy to verify such as math and coding. But for tasks like perspective and natural language generation, good output is not easily verifiable. So it is very challenging and I enjoy working on challenging problems."

In this area of research, generation, I thought I was taking a very unpopular path and somehow that did not feel good to me. Yet as time went on, I realized that taking the less popular path could lead to seminal research, since others were not working on the problem. In my case, this led to seminal research in text generation. Today, controlling choice in language generation is still an issue. We also can see from Zach's video that taking a less traveled path in research is still an option today.

2. Text Summarization

I turn now to my work on text summarization. In this area I was tremendously influenced by Karen Spärck Jones, a professor of Computers and Information at Cambridge University. Karen really served as a mentor for me and we worked together quite a bit on summarization, in particular advising on the setup of the tasks for the Document Understanding Conference (DUC) workshops on evaluation of summarization.


Karen Spärck Jones, Professor of Computers and Information, Cambridge University.

In 1996, nobody was doing text summarization, but I felt that the time was right to begin to work on it. The field had seen a lot of success in language generation and also parsing and information extraction and I thought we could put both of these together to form a summarization system. My first proposal to the National Science Foundation

(NSF) on text summarization was rejected; the reviewers said "Too ambitious! Not feasible!".

At that point, Dragamir Radov began the PhD program at Columbia and we were determined to prove them wrong. We began to work on the task to show that it was feasible and we were essentially joined at the hip working together. This leads to another research strategy which I ascribe to: "Never take no for an answer!". A year later, we successfully got two NSF grants on summarization and this allowed us to pursue different approaches. We looked at pairing information extraction with text generation, my initial idea. We also looked at statistical techniques, one of which we called **cut and paste**, where we extracted phrases from an input document or took a sentence and compressed it, removing information. This also led to our work on text fusion for multi-document news summarization, where we took phrases from different articles in the input to generate a novel sentence. I think this was the beginning of abstractive summarization.

We developed a system called NewsBlaster which we launched on the Web and ran every day (Figure 3). We launched this right after 9/11 (2001) because we knew there would be a lot of news available on this event that we could track. NewsBlaster worked by scanning many news sites, clustering together news on the same event and for each event, and generating a paragraph length summary. This platform enabled us to pursue many research themes. In particular, we looked at discourse coherence and how to refer to entities when they appeared in a context in the summary that was different

Figure 3
NewsBlaster summarizes a cluster of related articles on collaboration with Saudis.

from the context in which they appeared in the input article. We also looked at the problem of generating summary updates—that is, how to generate a summary on day N that talked about what was new since day N-1. We developed a multilingual version of NewsBlaster where we could generate English summaries from news around the world. Each summary was an English paragraph that linked into multiple news articles in different languages.

Following NewBlaster, we began work on a patient-specific medical library that was funded from NSF as a center grant. Here we looked at patient-specific summarization of medical articles; these were personalized summaries generated from an article for a physician that focused on the patients under their care. We also worked on generating summaries of relevant literature for patients and their families in language that they can understand.

I still work on summarization today. Several years ago, in collaboration with Stanford, we showed that single-document news summarization including the faithfulness problem is solved with instruction tuned models (Zhang et al. 2024). There are nonetheless other genres for which summarization can be more difficult.

Today, in our group, we're looking at summarization of narrative, which has quite a bit of subjectivity in the input (Subbiah et al. 2024a,2025). It can be long input, which makes the task more difficult and even evaluation is harder in this setting (Subbiah et al. 2024b). We are also looking at other multi-document settings. In particular, we are studying summarizing text according to different perspectives (Deas and McKeown 2025). We'll hear from Melanie Subbiah who's working on summarization of narrative.

"I'm Melanie Subbiah. I'm a final year PhD student at Columbia and I'm studying subtext in writing. Specifically, I'm looking at understanding subtext through the context of summarizing narrative. What we mean by subtext is cases where there's implicit meaning beyond the explicit words on the page. Narrative text has a lot of subtext. For example, two characters might be having a conversation about one thing, but you are trying to interpret what they are feeling beneath this. What are the motivators for each of these characters? And so, I look at summarizing narrative in different contexts, from collaborating directly with creative writers to working with psychologists and thinking about people's real life stories. I find this line of work really exciting in part because of my interests across different fields. I've always loved creative writing and I really enjoy drawing from experts in other fields. Studying this problem has allowed me to work with writers directly as well as psychologists and to think about these problems in a real world context. I also find it really exciting because as a field we focus a lot on problems that we can evaluate easily, because they have a clear right and wrong answer and we can assess that automatically. I think dealing with subtext, especially in cases where there might be multiple right answers, is important. Summarization often involves multiple valid summaries because you can always summarize text in

Melanie Subbiah, Columbia PhD student.

many different ways. It depends who you are summarizing for whether it is a good summary. Studying problems like this really challenges us to look beyond black and white. We have an exciting collaboration with a clinical psychology lab at Northwestern University and they have an amazing data set of people telling their life story through an interview format. The psychologists go back and ask the same questions each year over a period of ten years. Each year, people are asked the same questions and each of these interviews is quite long, creating a rich data set that introduces problems not only around long input text, but also assessing change over time."

In this research area, summarization, I made a choice to look at a problem that had not been addressed and to strike out in new directions. I would encourage every researcher never to take no as the final answer. If you believe in yourself, you can always move ahead and get a positive result the next time around. I do think that a shared research platform like NewsBlaster is useful for research. It drew us together as a group. People had a shared goal but it allowed us to pursue many different threads of research. As we go forward in the field of summarization, I think evaluation and experimentation for genres that are not well studied is an area where we can find challenges.

3. Social Media Analysis

I'll turn now to my work on social media analysis. While I have done a lot of work in this area, I'm going to focus today on work that we've done to serve Black communities. I was influenced by Desmond Patton, who arrived at Columbia as a professor in social work around 2012. I was Director of Columbia's Data Science Institute at that time and he approached me looking for people who would be interested in working with him and I was intrigued. He was working on a case study of Gakirah Barnes, a young teen who had joined a gang at age fourteen when her close friend Tyquan was shot and killed, only to be shot and killed herself at age seventeen. Her childhood was complicated. It was filled with deep love from her mother, but on the other hand, from birth, it was also filled with loss, trauma, and gang-related violence. She was extremely prolific on social media and Desmond was studying her posts. His goal was to enable intervention by social workers in order to avert violence. Her posts expressed a variety of emotions, from aggression to loss.

This was the start of a ten-year collaboration between myself and Desmond Patton. We were joined by Owen Rambow for the first part of our work, and in that early part, we developed tools to automatically label posts as aggression or loss (Blevins et al. 2016; Chang et al. 2018). This work has inspired a field in social science that is called **social media–related gun violence**. We didn't have funding at the time, so the work was done by an amazing team of undergraduate students who went on further in the field. We

Desmond Patton, PIK University Professor, School of Social Policy and Practice, Annenberg School for Communication, University of Pennsylvania.

You not gonna never find your happiness cause they gonna always be a step ahead of you and you gonna always be upset about that.

Figure 4
Naturally occurring AAL containing only a few features of AAL (e.g., negative concord in "not gonna never").

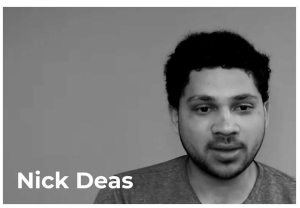
subsequently began focusing on just loss, developing an approach to identify emotional distress and the events that trigger them. We noted that in order to understand such posts, language models must be able to understand African American language. So we were joined at this point by Jessi Grieser, a professor in linguistics who specializes in African American language.

This work was funded by NSF, although it was cancelled in the spring because it didn't align with the new administration's orders.

This, like NewsBlaster, is a project that has enabled many research themes. In one line of work, we investigate language models and African American language (AAL). We've done work both on characterizing how well language models understand African American language (Deas et al. 2023) as well as mitigating their errors with the use of phonology (Deas et al. 2024a). We also study the kind of AAL data on which language models are pretrained and how that contributes to issues. We reported on this in a paper presented at ACL this year (Deas et al. 2025). I would note that in contrast to the performative text we saw in Gakirah's posts, what we see lacking in pre-training data is more naturally occurring AAL, which has fewer features. An example is shown in Figure 4. We've also done work on detecting emotions, the events that cause them, and explanations that indicate why (Turcan 2024) and subsequently have identified an unbounded set of affective states in both English and Spanish (Deas et al. 2024b). And we're currently working on being able to summarize different perspectives. We want to enable unbiased summarization of opinions from vulnerable groups. To date, we have evaluated language models for perspective summarization (Deas and McKeown 2025) and we've developed a new perspective summarization approach using reranking and direct preference optimization, which appeared in this ACL (2026) meeting's findings (Ri, Deas, and McKeown 2025).

Nick Deas is involved in all of this work and we hear from him now.

"My name is Nick Deas and I'm a third-year PhD candidate. A lot of my work focuses on evaluating how well LLMs are able to understand attitudes both from an emotion detection perspective, as well as how well they can summarize different political arguments and beliefs. Our work especially focuses on trying to understand more nuanced attitudes and emotions that people actually use when they talk about their



Columbia PhD student.

feelings, as well as the kind of language that people use when describing their political beliefs and opinions. There are lots of differences in how different cultures or speakers of different dialects express their opinions, emotions, and beliefs. And so a lot of our work also focuses on understanding how well LLMs can understand African American language as a dialect of English spoken by African Americans in the U.S. and studying whether there are encoded stereotypes or biases that prevent this kind of understanding of more nuanced expressions of attitudes. Our work has mostly shown that the models really struggle to understand lots of different features of African American language. They kind of understand the basic patterns of some features that are more common. But especially when it comes to features that are really only evident on social media or more camouflaged features that aren't very well distinguished from mainstream English, the models really struggle to understand the nuances of how those features are used and how they contribute to kind of the overall meaning of what the speaker is intending to say."

In this line of work, I was moved to join the project because of the work Desmond was doing on Gakirah Barnes. Here I found that if I followed my heart in selecting research, I suddenly found myself in many new unexpected and thrilling research projects. I also found that a socially impactful project is very attractive to students. It draws in students, enabling the building of big teams where people really care about what they're working on. In this sort of project, ethical issues abound. So, for example, we realize that creating language models that understand African American language will enable important applications in mental health and medicine, enabling them to be used by a much broader segment of the population. At the same time, it will also enable unintended uses such as for surveillance by police. Thinking about the impact requires an interdisciplinary team. There are many issues here that I would not be able to identify myself without the help of Desmond or Jessi.

4. Multimodal Analysis of Artwork

I'll turn now to the last topic, multimodal generation from artwork. This is a very recent project. We are only beginning to get results. I was influenced by one of my students, Amith Ananthram, who came to me and was very interested in switching his topic to work on the generation of text about art. I was intrigued and worked hard to secure funding so we could pursue this topic.

The idea here is to generate detailed image descriptions. Similar to *alt-text*, the descriptions are used by museums to help individuals who are blind or have low vision access to visual art. We've done this work in part in collaboration with the National Gallery of Art. We've used one of their datasets, which pairs images of artwork with very detailed alt-text that are written by PhDs in art history for the National Gallery of Art. Looking closely at an example (see Figure 5), we can see that the language is rich with detail about angles, relative positions, colors, and so forth: "Here we look slightly down onto a woman dressed in golden yellows sitting in a pale green chair with a nude child sitting in her lap."

So far, we have characterized the cultural perspective of vision language models (VLMs), showing that VLMs adopt a Western viewpoint even when prompted in non-Western languages because of the predominance of English in their pretraining (Ananthram et al. 2025). This was joint work done with the University of North Carolina. Amith is now working on a new metric to evaluate detailed image descriptions, including alt-text for artwork, called PoSh, and this will be published in ICLR 2026 (Ananthram et al. 2026). More recently, I was approached by a professor of art history,

Woman with a Sunflower, Mary Cassatt, 1905

We look slightly down onto a woman dressed in golden yellows, sitting in a pale green chair, with a nude child sitting in her lap as they both gaze into a mirror in this vertical portrait painting. Both the people have pale, peachy skin. The chair is angled to our left so the woman's knees and child cant down toward the lower left corner of the composition, and the woman leans onto the arm closer to us. The chair is painted mint green and the rose-pink upholstery is visible on the seat and a corner behind the woman's shoulder. To our right, the woman's vibrant, copper-colored hair is pulled loosely to the back of her head. She has a rounded nose, flushed cheeks, and her full, coral-pink lips are closed. Her long dress has a low, U-shaped neckline. The fabric shimmers from pale, cucumber green to light sunshine yellow. The sleeves of the dress split over the shoulder and a second long, goldenrod-yellow sleeve falls from her elbow off the bottom edge of the canvas. An oversized sunflower, larger than the woman's face, is affixed to her dress near her left shoulder, closer to us. She looks with dark eyes down toward the small, gold-rimmed mirror she holds in her right hand, farther from us. The child also holds the handle of the mirror with both hands, and in the reflection, the child looks back at us with dark eyes, a button nose, and pink lips. The child's hair in the reflection is the same copper color as the woman's, but the child on her lap has blond, shoulder-length hair. The woman rests one hand on the child's left shoulder, closer to us. The child has a rounded belly and smooth, rosy limbs. The woman and child are reflected in a second mirror hanging on the wall alongside them, opposite us. Their reflections are very loosely painted. The wall behind the pair is sage green across the top and it shifts to fawn brown across the bottom. Brushstrokes are visible throughout, especially in the woman's dress and hair, and are more blended in the bodies and faces. The artist signed the painting in the lower right corner, "Mary Cassatt."

Figure 5
A detailed image description from the National Gallery of Art.

Noam Elcott, to extend our work on interpretability (Alshomary et al. 2025) to vision models. They were very interested in understanding how a VLM analyzes art. We have begun looking at whether vision language models interpret architectural style using the same visual features as art historians or whether they rely instead on the large body of text in their pretraining data that describes this work (Bin et al. 2024). This work is very interdisciplinary. In addition to art history, we have participation from law (Kate Crawford), vision (Carl Vondrick), and an NLP group.

And so in Amith's words:

"My name is Amith Ananthram. I'm a fifth year PhD student. I work on vision language models—in particular, trying to characterize both the potential and the limitations of leveraging language for image understanding. A recent work of ours looked at VLMs and tried to understand whose perspective they model when prompted in English versus another language. We were motivated by some nice results in the cognitive sciences that show that visual perception is to a degree culturally mediated. We were able to show experimentally that the amount of multilingual language that a model sees during language-only pretraining actually affects the perspective that downstream VLMs have when they're making sense of images. A particular area of interest for us in the lab is working with visual art. We are interested in building models that are able to describe and interpret visual art. I think visual art is quite an exciting area to work in—good art for most of us is surprising or interesting in some way. So definitionally, it's slightly out of distribution. And a lot of AI systems are, you know,

Columbia PhD student.

statistical machines under the hood. So seeing how they perform on visual art, which typically contains interesting visual features or objects that are not often seen together can stress the generalization capabilities of these systems."

And now we turn to Milad Alshomary, who is working on the interpretation of LLMs.

Columbia postdoctoral fellow.

"My name is Milad Alshomary. I'm a post-doctoral research scientist at Columbia University for about a year and a half now, and I'm working on the interpretability of deep neural models. Here, the task is to interpret why and how AI makes certain decisions, and the core purpose of this is to gain user trust. Here, interpretability is geared towards the idea of how can we enable models to provide explanations. Recently, we joined a new project to examine how AI analyzes the style of an art piece, collaborating with art historians. The task is to see how art historians analyze art and how they examine the visual component of a picture to classify it as Baroque or Gothic, and whether this process aligns with vision language models. These powerful large language models, basically, you give them an image, and they can do an analysis of its style, but the question that we are interested in is whether they do this in the same way as art historians. Do they really look at the same aspects of an image in order to come up with their prediction?"

In this research project, I was open to following the research interests of my students and again this led me in directions that I hadn't expected and opened up a number of collaborations that I had not been involved in before. Again, interdisciplinary collaborations are a key component of the work, and we're right at the beginning. There are still a lot of challenges and much work to be done.

5. The Past and the Future Converge

Looking to the future, I think being in academia offers a range of research directions. It enables the building of interdisciplinary teams drawing from many schools across the university. It enables focusing on socially impactful projects that draw the attention of students. In academia, there is a lot of choice in how you do research: choice in the paths taken. I do think going forward that ethical issues should be central to our work. We need to take responsibility for some of the unintended consequences that we see coming out of the work with large language models. For example, the use of models by students to do coursework. The use of models by professors to grade coursework. The use of models in reviewing to judge papers and the use of models that engage teenagers in unexpected ways that is harmful to their health.

Let's hear what the younger generation has to say.

Zachary Horvitz: I think another piece is that academia can explore these wilder, very different approaches from what we have. We're at a point where industry has largely taken one general solution: large scale training of large scale models that are

trained in a GPT style, and those work really well. I think a role of academia is to explore these wilder ideas, like text diffusion models or state space models. There are all these exciting alternative approaches that people in industry, because of incentives, aren't engaging with, and we can actually explore them.

Milad Alshomary: Interpretability is very important; a lot of work has been done on that, but the biggest challenge is ensuring that the interpretations provided for a model's prediction are faithful to the prediction and to communicate these explanations effectively to different audiences.

Yunfan Zhang: How can we apply reinforcement learning to things that are not easily verifiable, such as natural language generation and summarization, and how can we evaluate those abilities?

Amith Ananthram: Historically, the field has focused on topline benchmarks, and there have been many efforts with model developers to push numbers on certain benchmarks and get state of the art. Over the last year or two, we've seen the limits of that, and evaluation broadly feels a little broken.

Nick Deas: Models perform poorly or are not trained on enough data representing different groups. There will be many issues where there are two sides of the coin. We will trade off problems where models don't work well and issues such as surveillance and malicious uses that might be enabled when models perform better and have fewer biases and stereotypes.

Melanie Subbiah: The biggest challenge for the field is presenting a positive vision for how this technology will be used and how it will advance. We can say there will be benefits, but it's not clear how this improves quality of life and people's overall well-being. What does it mean to replace jobs or tasks, and how do we do this in a way that is a smooth economic transition?

6. Mentoring

I'd like to close by talking a little bit about mentoring. Mentoring has been a big part of my life and it's the part of my work that I enjoy the most. On one side of the coin, I think one of my contributions has been training and sending into the world these very talented students. But on the other side of the coin, I have had the honor of my life in working with such amazingly talented students who have gone on themselves to become so influential and I learn every day from the students I work with.

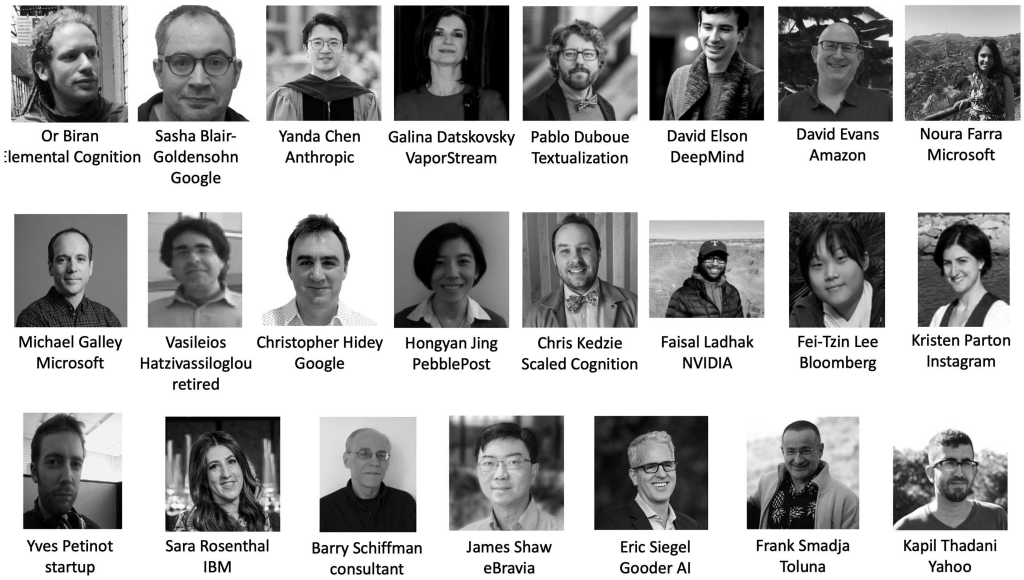
So I want to credit the students that I featured in the talk. First the academic PhD students: First and foremost, I would like to acknowledge Cecile Paris, Director of the Collaborative Intelligence Research Programme, at CSIRO (Commonwealth Scientific and Industrial Research Organisation), the national science agency of Australia. As my first graduating PhD, she has stood by my side over many years. Next, Regina Barzilay, who is now the distinguished AI and Health Professor at MIT; Michael Elhadad, who is now Professor and Chair of the Computer Science Department at Ben Gurion University. Noemie Elhadad, his sister, who is now Professor and Chair of the Biomedical Informatics Department at Colombia. Min-Yen Kan, who is Professor and Dean of Undergraduate Students at the National University of Singapore. Dragamir Radev, who held a named chair professorship at Yale. Jacques Robin, who worked as a researcher at University of Paris 6, and Ani Nenkova, who was at University of Pennsylvania and moved to Adobe. I have many other PhD students who have gone into academia. I would like to mention my most recent graduate who has gone into a faculty position, Emily Allaway, who is at the University of Edinburgh. I particularly value

Graduated PhD students who are now in academia or research institutes.

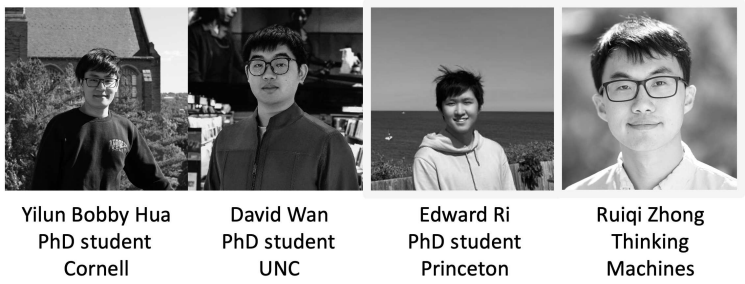
when students go into academia because then they have advised students themselves who become my academic grandchildren, and the family grows.

I will also mention the students I featured in this talk who went on to industry: Yanda Chen, who is now at Anthropic; David Evans, who went at the beginning to Amazon and is still there; Faisal Ladhak, who recently joined Nvidia; and Hongyan Jing, Barry Schiffman, and Frank Smadja, who all went to startups. Many more past PhD students have gone to industry to companies like DeepMind, Microsoft, Google, Nvidia, Bloomberg, IBM, and Yahoo, as well as many startups.

I will call out the undergraduate students I mentioned who worked on projects: Terra Blevins, who is now faculty at Northeastern; Serina Chang, now faculty at Berkeley; Edward Ri, who is going on to become a PhD student at Princeton; and Ruiqi Zhong, who just graduated from Berkeley and is at Thinking Machines, as well as many other distinguished students.



Graduated PhD students who are now in industry.



Undergraduate students who carried out research with us.

Acknowledgments

I would like to acknowledge funding from various grants: Amazon, the Defense Advanced Research Projects Agency, the Knight Foundation, the Intelligence Advanced Research Projects Agency, and the National Science Foundation supported many aspects of this research over the course of my career. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. Many thanks to the Columbia Video Network for help with the videos. Finally, thank you to my husband, who keeps me grounded in the real world and reminds me that there are things I can enjoy in life other than work. Without his support, I could never have worked as hard as I did or traveled to as many conferences as I did with three young daughters. Thank you!

References

Alshomary, Milad, Narutatsu Ri, Marianna Apidianaki, Ajay Patel, Smaranda Muresan, and Kathleen McKeown. 2025. Latent space interpretation for stylistic analysis and explainable authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1124–1135.

Ananthram, Amith, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. See it from my perspective: How language affects cultural bias in image understanding. In *The Thirteenth International Conference on Learning Representations*.

Ananthram, Amith, Elias Stengel-Eskin, Lorena A. Bradford, Julia Demarest, Adam Purvis, Keith Krut, Robert Stein, Rina Elster Pantalony, Mohit Bansal, and Kathleen McKeown. 2026. PoSH: Using scene graphs to guide LLMs-as-a-judge for detailed image descriptions. In *The Fourteenth International Conference on Learning Representations*.

Bin, Yi, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. GalleryGPT: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7734–7743. https://doi.org/10.1145/3664647.3681656

Blevins, Terra, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. 2016. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. Matsumoto, Yuji and Rashmi Prasad, editors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206.

Chang, Serina, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathy McKeown. 2018. Detecting gang-involved escalation on social media using context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 46–56. https://doi.org/10.18653/v1/D18-1005

Deas, Nicholas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824. https://doi.org/10.18653/v1/2023.emnlp-main.421

Deas, Nicholas, Jessica A. Grieser, Xinmeng Hou, Shana Kleiner, Tajh Martin, Sreya Nandanampati, Desmond U. Patton, and Kathleen McKeown. 2024a. PhonATe: Impact of type-written phonological features of African American language on generative language modeling tasks. In *First Conference on Language Modeling*.

Deas, Nicholas and Kathleen McKeown. 2025. Summarization of opinionated political documents with varied perspectives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8088–8108.

Deas, Nicholas, Elsbeth Turcan, Ivan Ernesto Perez Mejia, and Kathleen McKeown. 2024b. MASIVE: Open-ended affective state identification in English and Spanish. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20467–20485. https://doi.org/10.18653/v1/2024.emnlp-main.1139

Deas, Nicholas, Blake Vente, Amith Ananthram, Jessica A. Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard Iii, and Kathleen McKeown. 2025. Data caricatures: On the representation of African American language in pretraining corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 29192–29217. https://doi.org/10.18653/v1/2025.acl-long.1416

Horvitz, Zachary, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024a. ParaGuide: Guided diffusion paraphrasers for plug-and-play textual style transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18216–18224. https://doi.org/10.1609/aaai.v38i16.29780

Horvitz, Zachary, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024b. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390. https://doi.org/10.18653/v1/2024.findings-emnlp.781

Ri, Narutatsu, Nicholas Deas, and Kathleen McKeown. 2025. Reranking-based generation for unbiased perspective summarization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24701–24723. https://doi.org/10.18653/v1/2025.findings-acl.1268

Singhal, Raghav, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. 2025. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 55810–55827.

Subbiah, Melanie, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024a. STORYSUMM: Evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005. https://doi.org/10.18653/v1/2024.emnlp-main.557

Subbiah, Melanie, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024b. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310. https://doi.org/10.1162/tacl_a_00702

Subbiah, Melanie, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. 2025. Is the top still spinning? Evaluating subjectivity in narrative understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 185–203.

Turcan, Elsbeth. 2024. *Detecting and explaining emotional reactions in personal narrative*. Ph.D. thesis. https://doi.org/10.7916/FXXX-WX64

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57. https://doi.org/10.1162/tacl_a_00632

Zhang, Yunfan, Kathleen McKeown, and Smaranda Muresan. 2025. Exploring chain-of-thought reasoning for steerable pluralistic alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25636–25649. https://doi.org/10.18653/v1/2025.emnlp-main.1301