TONGSHUANG (SHERRY) WU **RESEARCH** STATEMENT

Research in Artificial Intelligence (AI) has advanced at an incredible pace, to the point where it is making its way into our everyday lives, explicitly and behind the scenes. However, beneath their impressive progress, many AI models hide deficiencies that amplify social biases (*e.g.,* chatbot assistants making inappropriate or unfair responses to certain questions) or even cause fatal accidents (auto-driving scenarios). The presence of these issues raises a question at the heart of my research: **How do we identify, improve, and cope with imperfect models, while still benefiting from their use?**

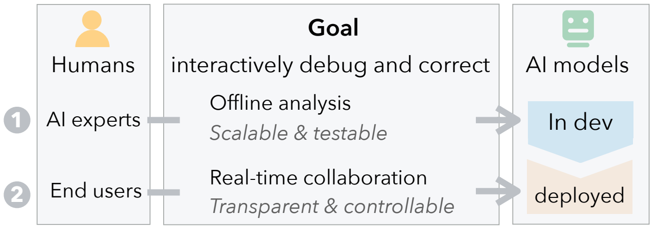


Figure 1: I empower humans to **debug and correct imperfect AI models interactively** in two scenarios with unique goals: ① I help **AI experts systematically analyze in-dev AIs** to discover and repair model failures prior to deployment. ② I support **end users** as they **collaborate with deployed AIs** so they can interpret and correct AIs *in-situ*.

I strive to **empower humans to debug and correct AI models interactively**. On the one hand, I help ① **AI experts run scalable and testable analyses on models in development** so they can diagnose and address model weaknesses before models are released. This offline analysis acts as a guardrail, making it less likely for end users to encounter unwanted model behavior. On the other hand, despite experts' best efforts, all deployed models are almost guaranteed to be imperfect due to differences in training and deployment environments. To make AIs more usable in downstream applications, I also help ② **end users collaborate with deployed AIs in a transparent and controllable manner** so they can detect and overwrite AI errors in real-time.

My research probes the intersection between Human-Computer Interaction (HCI) and Natural Language Processing (NLP), and has led to publications in top-tier conferences and journals in both areas (*e.g.,* CHI, TOCHI, ACL). I conduct *user studies* to identify pitfalls in current human-AI interaction processes [2, 6, 10], and design *interactive tools* [1, 5, 7, 11] as well as *novel models* [4, 8] that aid in AI debugging and correction. The impact of my work extends beyond academics: Several of the interfaces and frameworks I developed have been (or will soon be) integrated into open source AI libraries or deployed internally in industries; **many leading tech companies, including the Allen Institute for AI, Microsoft, Apple, and Google, have used these tools to transform their analyses of AI models and the creation of AI-infused applications.** In the future, I am eager to continue shaping how humans and AIs interact *in high-stakes, in-the-wild scenarios.* I also plan to *help humans interpret and steer AIs*, which is crucial for understanding and pushing the limits of model usability.

① EXPERT DEBUGGING OF AI IN-DEV: SYSTEMATIC, SCALABLE, AND TESTABLE ANALYSIS

To minimize malfunctions in deployment, AI experts — those who design and develop AIs — must ensure models are reliable and robust before they are released. **Throughout the whole AI development cycle** (Figure 2), **I help experts systematically analyze their models**, so they can make well-informed decisions on when a model is ready for use, and where and how to improve it.

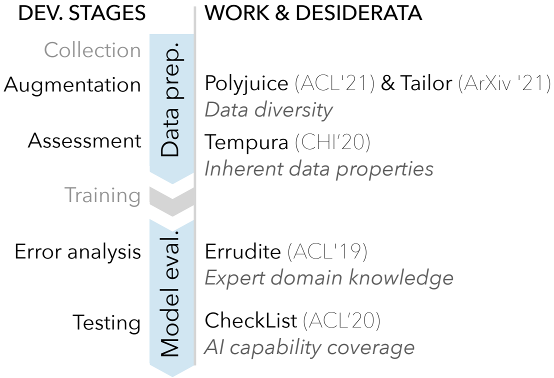
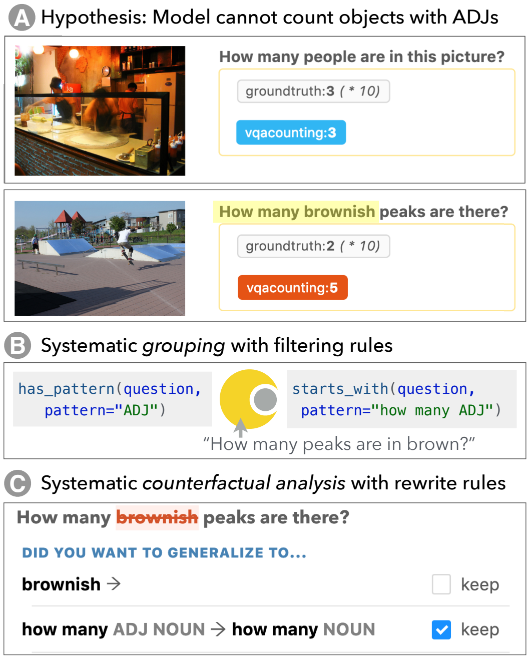


Figure 2: I support experts throughout the AI development cycle. Given a development stage (left), I distill its unique desiderata, and design NLP models and interactive tools accordingly (right).

To ground my research in the practical needs of AI experts, I first asked: **What are some pitfalls in currently applied AI analysis methodologies?** I conducted informal interviews, in-lab experiments, and long-term direct observation (*e.g.,* [11]), and found that *experts often inspect models in ad hoc and informal ways, despite their desire to find generalizable patterns.* Because NLP models are typically complex, and unstructured text data can be difficult to filter semantically, experts rarely review more than a few examples at a time. Unfortunately, **these ad hoc observations can result in confirmation bias and spurious conclusions, and hinder experts from iteratively improving model quality.** In a user study, I asked participants to iteratively select features for a sentiment analysis model after seeing related examples [10], and observed such local changes failed to improve models on average.



A Hypothesis: Model cannot count objects with ADJs

How many people are in this picture?

groundtruth:**3** *(* 10)*

vqacounting:3

How many brownish peaks are there?

groundtruth:**2** *(* 10)*

vqacounting:5

B Systematic *grouping* with filtering rules

```
has_pattern(question,
   pattern="ADJ")
```

```
starts_with(question,
 pattern="how many ADJ")
```

"How many peaks are in brown?"

C Systematic *counterfactual analysis* with rewrite rules

How many ~~brownish~~ peaks are there?

DID YOU WANT TO GENERALIZE TO...

brownish → □ keep

how many ADJ NOUN → how many NOUN ☑ keep

Figure 3: Errudite example. This tool enables scalable and testable error analysis through *systematic grouping* and *counterfactual rewriting*. (A) A visual question answering model predicts "How many people..." correctly but "How many brownish..." incorrectly. Experts suspect the adjectives make a difference. (B) They **scale up the observation**, by building groups of questions that contain `ADJ`ectives, or start with **"`How many ADJ`"**. (C) They also **test the root error cause** with rewrite rules that answer: *"If the adjectives were not there, would the model predict correctly?"*

The experiment sounded a cautionary note, and paved the way for my subsequent research: **I designed interactive tools and novel NLP models that provide experts with comprehensive and unbiased views of models.** For example, I built Errudite to facilitate analysis of when, how, and why models fail ("error analysis") [7]. I identified and implemented two essential building blocks, as shown in Figure 3: First, Errudite allows *systematic grouping* of relevant instances and thus scales observations beyond random spot checks; Second, Errudite supports *counterfactual rewriting*, which surfaces root error causes by testing what-if scenarios. Errudite transforms expert domain knowledge into actionable analysis scripts. As a result, **developers and researchers** have adopted it for their own NLP models, and **shared reproducible insights with the broader NLP community** (*e.g.*, researchers working on relation extraction[1] and an internal team at Apple doing question answering).

Recently, to further optimize for expert domain knowledge, I co-organized the NL-Augmenter challenge[2] to crowdsource counterfactual analysis strategies. The challenge received substantial interest, and I am now co-authoring a paper that **contributes 200+ creative strategies, extending far beyond a single expert's wisdom.**

The aforementioned *grouping* and *counterfactual rewriting* are also fundamental to other AI development stages. However, in those cases, relying solely on domain knowledge (as in error analysis) can introduce unnecessary bias. For example, when experts sanity check training data quality, they need groupings to detect conflicting labels between similar instances. Here, handcrafted filtering rules may be subjective and produce unrepresentative groups. To promote objective analysis, I **distilled unique requirements for each development stage** (Figure 2), and **tailored tool designs to mitigate human bias**. To this end:

First, for the aforementioned **data assessment** scenario, I developed Tempura [11], which automatically mines representative data groups, thereby **exposing inherent properties** in both academic and industrial datasets.

Second, to **unit-test model behaviors** (by crafting test cases for desired AI capabilities, *e.g.*, robustness, fairness), my collaborators and I created CheckList [1], a framework that provides suggestions on what linguistic phenomena to group and perturb. Checklist significantly **increased test coverage**: experts using it found three times as many bugs as those without it. **CheckList won the best paper award at ACL 2020** (top-1), and **has also been widely adopted**: its open source repository has received 1,500+ stars, it has been incorporated into popular NLP frameworks like AllenNLP,[3] and it will be embedded into the Papers With Code leaderboard for comparing model capabilities.

Furthermore, to disentangle spurious and robust features via **counterfactual data augmentation** (*e.g.*, to solely

[1] https://github.com/DFKI-NLP/tacrev

[2] https://gem-benchmark.com/nl_augmenter

[3] https://medium.com/ai2-blog/using-checklists-with-allennlp

highlight the impact of *negation* independent of other features using the original and perturbed sentences in Figure 4), I proposed Polyjuice [8] and Tailor [4], language-model-based generators that automatically produce **diverse counterfactuals**. In a variety of domains, not only did the two generators compensate for human omissions (Figure 4), but their generated counterfactuals successfully improved model generalization. In this way, I was able to **fix models by removing spurious correlations in data**.

② End User Debugging of AI In Situ: Transparent & Controllable Collaboration

On more occasions than we would like, models carefully developed *in the lab* still suffer *in the wild*: the development environment — with clean and static inputs — tends to oversimplify realistic challenges. As a result, end users still need to interpret AIs as they collaborate in real time: **they should follow AIs when they are correct, but identify and correct their mistakes otherwise.**

However, my collaborators and I asserted that **people tend to blindly accept AI.** In a user study [6], we saw that the human-AI team often increased accuracy on document classification when the AI system was correct but, worryingly, decreased accuracy when the AI erred, as shown in Figure 5. Even more concerning, though we might think humans could judge AI correctness more effectively if AIs explained their reasoning rationale, we found that explanations solidified beliefs and entrenched blind trust.

To mitigate blind trust, I **designed interactive tools that make human-AI collaborations more transparent** (so humans know when AIs err) **and more controllable** (so humans can guide AIs in the right direction). For example, I created AI Chaining [9] to improve interaction between end users and large language models (LLM, like GPT-3). LLMs can be flexibly tailored for a wide variety of tasks, purely through natural language descriptions. This flexibility, however, also makes them opaque. In user studies, I found that end users struggled to debug and improve their arbitrary instruction prompts. In response, I proposed *Chaining* multiple LLM runs together, *i.e.*, decomposing an overarching task into a series of highly targeted sub-tasks, mapping each to a distinct LLM step, and using the output from one step as an input to the next. These steps naturally **expose intermediate checkpoints and control knobs to end users, helping them pinpoint seemingly global errors to a local cause**. For example, thanks to the *Ideation* step in Figure 6, Chaining lets users customize which suggestions to include in the final paragraph — a function that does not exist otherwise. I am continuing research on Chaining with my collaborators, and we plan to **deploy it as an internal tool at Google** to support rapid prototyping of LLM-infused applications.

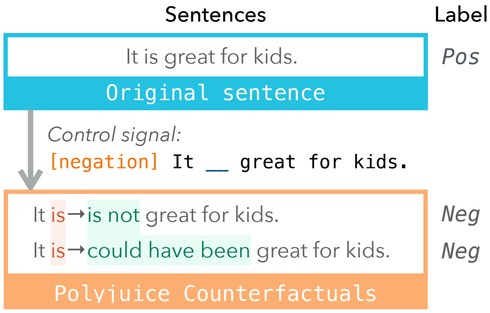

Sentences
Label
It is great for kids.
Pos
Original sentence
Control signal:
[negation] It __ great for kids.
It is→is not great for kids.
Neg
It is→could have been great for kids.
Neg
Polyjuice Counterfactuals

Figure 4: Polyjuice generates diverse counterfactuals that experts may miss. For example, to teach the model that `negation` changes sentiment, experts may augment the training data by rewriting is → is not, but they would miss teaching the model is → could have been.

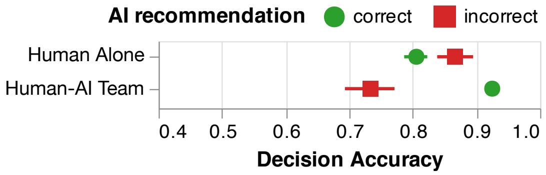
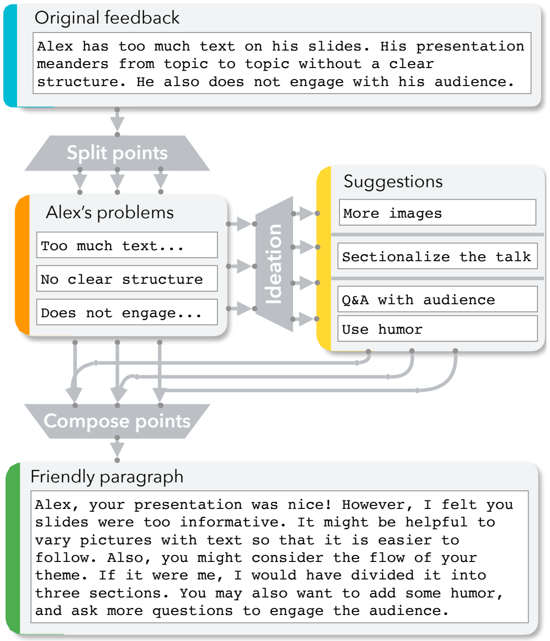


Figure 5: Humans tend to blindly agree with AIs. Compared to humans alone, the human-AI team often increased accuracy when the AI was correct, but decreased accuracy when the AI erred.


Original feedback

```
Alex has too much text on his slides. His presentation
meanders from topic to topic without a clear
structure. He also does not engage with his audience.
```

Friendly paragraph

```
Alex, your presentation was nice! However, I felt you
slides were too informative. It might be helpful to
vary pictures with text so that it is easier to
follow. Also, you might consider the flow of your
theme. If it were me, I would have divided it into
three sections. You may also want to add some humor,
and ask more questions to engage the audience.
```

Figure 6: *Chaining* makes the human-AI collaboration more transparent and controllable. For example, we decompose a peer review rewriting task into three sub-tasks: identifying each problem, providing suggestions per problem, and composing all the suggestions into one paragraph. Users can therefore inspect and steer AI in each step.

Future Research Agenda

My long-term research goal is to support humans coping with AI models that may never be perfect. In my past work, I achieved systematic analysis and transparent collaboration primarily in a *controlled environment*. Nevertheless, a single standard interaction workflow cannot support the variety of applications and people involved. How do we deal with label inconsistency for naturally subjective or controversial applications? How can we make rich yet complex models useful to humans without overwhelming them? In the coming years, I intend to answer these questions by **making human-AI interactions (1) more aware of use scenarios, and (2) more aware of human capabilities**.

To make model training and analysis more *context-sensitive* and reflective of our complex *real-world goals*, I will work with domain experts to explore **more efficient ways to collect and use benchmark datasets in the context of high-stakes applications** (*e.g.,* in medicine, law, and business). The work outlined below will naturally **close the loop on model development** in Figure 2, with a focus on more thoughtful *data collection* and *model training*.

Make data collection clean and ambiguity-aware. Data quality is crucial to AI performance, yet our existing benchmark datasets are often noisy, biased, and overly simplified. I plan to make data collection more efficient. In light of my counterfactual data augmentation work, I aim to design an active learning paradigm that generate counterfactuals for humans to label, thereby addressing the distribution gaps. To contextualize data ambiguities, I will gather labels along with annotator metadata, in order to, for example, model language toxicity based on the annotator cultural background.

Contextualize AI error severity. While we typically treat all model errors equally, the severity of errors actually varies by application area [3]. In collaboration with domain experts (*e.g.,* education practitioners [12]), I will explore users' expectations of models across domains, articulate a taxonomy of error types and severity, and extend CheckList [1] to test more societal aspects.

Consider the long-term impact of deployed models. Real-world use cases are rarely stationary. Users adjust their behavior as they interact with AIs (*e.g.,* using only short, clear commands to instruct virtual assistants). Through longitudinal studies, I aim to reweight model errors based on changes in user beliefs and actions, and design training objectives such that future model updates become compatible with users' previous experiences.

To make human-AI interaction more aware of *human capabilities*, I plan to design models and interactions that **allow people to intuitively understand and steer AIs**. These threads will advance **general human-AI interaction**, without distinguishing AI experts from end users, as I did in my prior work.

Humans understanding AIs: Explaining AI for appropriate reliance. Explanations can lead to over-reliance on AI [6]. I plan to investigate alternative explanation methods that can raise necessary suspicion. For example, instead of explaining why it believes an answer to be true, the AI might also surface evidence to the contrary — even when it agrees with the human.

Humans steering AIs: Rich controls on the model end, and intuitive interactions on the human end. My work [4, 8] has shown that language models can be fine-tuned to follow rich control codes. Still, to maintain ease of use, I plan to explore strategies for *distant control*, *i.e.,* flexibly translating intuitive user interactions into rich forms of control. Promising directions include extending the programming-by-demonstration design in Errudite [7], coupled with assisted prompt engineering, where we help users curate instructions for LLMs.

More robust, interpretable, and controllable models. For models to be understood and steered, they must be able to reason. I will develop models that are "right for the right reasons." As a first step, I will enhance the value of counterfactuals for training robust models, by adding explicit terms in the loss function that compare counterfactuals with original data, or by implementing other forms of contrastive learning.

Collaborations. As I expand my research scope to include more societal factors, I look forward to collaborating with experts in Psychology and Social Science (*e.g.,* to contextualize ethical errors). I also believe most insights from my work are transferable to AI applications beyond NLP. Through collaborations with *e.g.,* Vision, Robotics, Augmented Reality researchers, I am eager to distill shared and unique challenges in shaping human-AI interaction.

References

[1] Marco Tulio Ribeiro, **Tongshuang Wu**, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *ACL 2020*.

[2] Alison Smith-Renner, Ron Fan, Melissa Birchfield, **Tongshuang Wu**, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *ACM CHI 2020*.

[3] Jiao Sun, **Tongshuang Wu**, Yue Jiang, Ronil Awalegaonkar, Victoria Lin, and Diyi Yang. Pretty Princess vs. Successful Leader: A Quantitative Study of Gender Roles in Greeting Card Messages. In *ACM CHI 2022 (conditionally accepted)*.

[4] **Tongshuang Wu***, Alexis Ross*, Hao Peng, Matthew E. Peters, and Matt Gardner. Tailor: Generating and Perturbing Text with Semantic Controls. In *arXiv:2107.07150 (submitted to ACL Rolling Review 2021)*.

[5] **Tongshuang Wu***, Zhihang Dong*, Sicheng Song, and Mingrui Zhang. Interactive Attention Model Explorer for Natural Language Processing Tasks with Unbalanced Data Sizes. In *IEEE PacificVis 2020*.

[6] **Tongshuang Wu***, Gagan Bansal*, Joyce Zhou+, Raymond Fok+, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *ACM CHI 2021*.

[7] **Tongshuang Wu**, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. Errudite: Scalable, Reproducible, and Testable Error Analysis. In *ACL 2019*.

[8] **Tongshuang Wu**, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *ACL 2021*.

[9] **Tongshuang Wu**, Michael Terry, and Carrie J. Cai. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *ACM CHI 2022 (major revision)*.

[10] **Tongshuang Wu**, Daniel S. Weld, and Jeffrey Heer. Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis. In *ACM TOCHI 2019*.

[11] **Tongshuang Wu**, Kanit Wongsuphasawat, Donghao Ren, Kayur Patel, and Chris DuBois. Tempura: Query Analysis with Structural Templates. In *ACM CHI 2020*.

[12] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, **Tongshuang Wu**, Mo Yu, Dakuo Wang, and Jia-Jun Li. StoryBuddy: A Human-AI Collaborative Agent for Parent-Child Interactive Storytelling with Flexible Parent Involvement. In *ACM CHI 2022 (major revision)*.