Natural language technologies have always been developed with the goal to serve humanity. Nevertheless, a reality where NLP systems (and particularly, large language models) provide measurable value to humans has only existed for a relatively short time. The abilities of modern language models (LMs) to provide assistance in impactful and pervasive ways, from writing and information seeking to summarizing and translating information, continue to improve day by day; these systems have the potential to change the way we live and work in profound ways.

However, adoption and commercialization of LMs are accelerating at a much faster rate than the basic science underpinning foundational technological innovation. There are critical reliability gaps and adaptation issues that require concerted research effort to resolve, but which industry labs have largely ignored in favor of fixating on performance. Downstream users' needs are not fully served, for example, when models output harmful content, hallucinate nonfactual information in response to fact-seeking queries, or are convincingly and confidently incorrect– all frequent issues in the current generation of LMs. However powerful models become, the nature of machine learning is one of prediction and extrapolation from data, meaning that systems are unlikely to ever be perfect or have reliability and performance *guarantees*. We must rigorously study LMs to understand their potentials and pitfalls. Because industrial research is moving away from an open and shared model of science, the onus is on academic researchers to treat LMs as a true object of scientific study.

Treating LMs as a true object of scientific study necessitates a deep understanding of how they operate. My work on **explainability and interpretability of LMs** makes progress toward this goal at the level of both specific predictions and broad system behaviors. I build off of this scientific understanding to make LMs more performant, reliable, and user-friendly. A deep understanding of LMs' behavior, from the simplest tasks up to the most complex, will enable the development of the next generation of language modeling technology that **better serves the needs of diverse downstream users**. I achieve this goal by pushing to:

1. **Build NLP systems that produce useful explanations (§1)**: Models should generate explanations that users find useful, i.e., both **faithful** to the model's underlying prediction process [1–4] and **acceptable** [5, 6], so users can know when and whether to trust systems' predictions.

2. **Make NLP systems more interpretable and performant (§2)**: Model developers should better understand models' inner workings [7–12] to prevent unanticipated outcomes. Understanding models' underlying causal mechanisms allows us to build systems that are **more performant**, for example, because we can better prompt them with textual inputs [9, 11, 13], efficiently update their weights to master a new skill [7], or update their training data to imbue new knowledge and capabilities [[8],§3].

My future work (**§3**) will continue these themes by 1) improving the utility of LM-generated explanations and reducing their propensity to mislead users in real-world settings; and 2) developing interpretability methods that provide proper data attribution, enable fine-grained localization and control, and improve safety-critical behaviors. My future work will thus, in turn, enable the next generation of LMs to more reliably provide factual, personalized, and safe information to serve users in high-stakes applications such as healthcare, education, and policy.

1 Building NLP Systems that Produce Useful Explanations

It has long been understood that explanation does not exist in a vacuum, but instead is "a three-place predicate: *someone* explains *something* to *someone*" [14]. In explainable AI, this typically means that an *AI system* provides *explanations* of its predictions to *human users*. Building explainable systems, therefore, requires solving a communication problem between AI systems and the people who use them– a problem that can be overcome by generating **useful** explanations.

What makes an explanation useful? People often fall for the illusion of explanatory depth: that if a model-generated explanation of behavior appears reasonable, then the model performing the behavior must have done so in a reasonable manner. In [1], I proposed that useful explanations must be both **acceptable** to humans *and* **faithful** to the model's underlying reasoning process (Fig. 1). This **bi-criteria view** of explanation utility has since been widely adopted and led to a significant increase in research on faithful, non-deceptive explanation of NLP systems. My research has improved the ability of NLP systems to produce useful explanations in service of two goals: 1) to determine whether trust in an AI system is warranted and 2) to build user trust when trust is warranted.

Producing Faithful Explanations of Model Behavior [1–4]: When trust in an AI system is weak, uncertain, or input-dependent, it is vital to explain model predictions so users can determine when to trust them. Faithful explanations help human users develop an **accurate mental model** of an (imperfect) AI system. Humans can then simulate the system's predictions–both correct and incorrect–with this mental model, and calibrate their trust accordingly. I have shown that effective explanations do, in fact, achieve this purpose [2].

How can we produce faithful explanations from LMs' internals? I have proposed 3 means of achieving this at the level of individual predictions (Fig. 1). Neural network attention mechanisms, a key component of most architectures, have historically been used to highlight key words and phrases in an input document (Fig. 2). But can a single component in isolation faithfully represent the entire network's behavior? I contributed a key work to the debate on whether attention can serve as a faithful explanation by developing **a suite of practical tests** [1]. With **over 1100 citations** and 50 Github stars, this work **changed the research community's understanding** of not only what makes an explanation of an NLP system meaningful, but also to what extent model internals can provide this meaning. To address my conclusion that not all attention mechanisms are faithful, I subsequently proposed an architecture– the natural language bottleneck– that produces explanations with faithfulness *guarantees* [3]. I recently gave an invited keynote at the "Big Picture" retrospective EMNLP workshop on this line of research, highlighting its continued relevance.

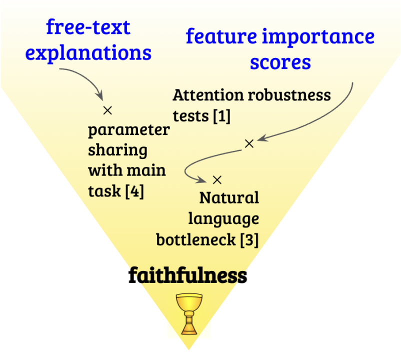

free-text explanations
feature importance scores
Attention robustness tests [1]
parameter sharing with main task [4]
Natural language bottleneck [3]
faithfulness

Figure 1: An illustrative "faithfulness spectrum". I assess and improve different types of model-generated explanations to be as faithful as possible to models' prediction processes.

As NLP systems have become more powerful, I have argued for an increased focus on generating more expressive free-form textual explanations [4, 15]. I was correct in this position– free-form explanations have since experienced a surge in popularity as "chain-of-thought prompting" [16]. But are LMs producing textual explanations using the same subsets of their parameters to both predict and explain? If they are not, the generated textual explanations will not be faithful to the model's prediction process– raising the possibility to deceive and mislead users. I was the **first to quantify how faithfully free-text explanations generated by LMs align with the LMs' predictions**, showing high parameter sharing between prediction and explanation [4]. For this work, I won the Allen Institute for AI's outstanding intern award.

Producing Explanations that are Acceptable and Useful to Users [5,6]: When trust in an AI system is warranted, I have developed explanations that facilitate users building that trust. For example, in the clinical domain, my state-of-the-art architecture for the ICD coding task was the **first to provide textual explanations for each predicted label** [5]; a physician judged these explanations to be informative (Fig. 2). With **over 700 citations**, this work has majorly influenced explainable architectures in clinical NLP.

Motivated by the success of general-purpose LMs and my prediction that collecting ground truth explanation datasets would become untenable at scale [15], I was the **first to evaluate and improve the human acceptability of few-shot free-text explanations** generated by LMs such as GPT-3 [6]. I developed an evaluation framework inspired by psychology research that has been used extensively in subsequent work (over 130 citations and 30 Github stars). I was also the **first to apply human preference modeling**, which relies on more cost-effective human annotation than explanation writing, **to train a system to select high-quality explanations from an LM**, pushing the state-of-the-art in automated explanation generation.

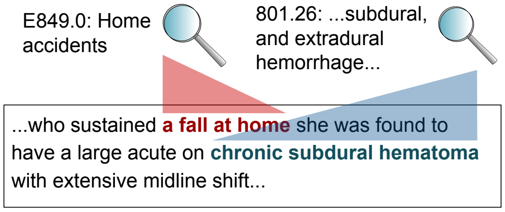

E849.0: Home accidents
801.26: ...subdural, and extradural hemorrhage...
...who sustained a fall at home she was found to have a large acute on chronic subdural hematoma with extensive midline shift...

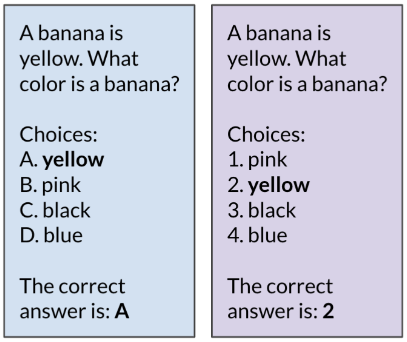
Figure 2: My work has provided informative explanations in applications like clinical coding.

2 Making NLP Systems More Interpretable and Performant

Closely aligned with my work on producing faithful explanations of LM predictions, a main tenet of my research program is understanding how LMs function more *generally* on tasks and skills of importance. Standard benchmark evaluation is often too coarse-grained to capture nuances of model behavior; unintentional and surprising behaviors subsequently emerge at deployment time due to a lack of fundamental scientific understanding of LMs' internal mechanisms. Indeed, a consistent theme in my work is discovering that models with similar performance can have vastly different internal mechanisms [1, 10]. My work in model interpretability **fills in the gaps that evaluation cannot**– not only **increasing scientific understanding of systems** [8, 10, 12], but also **leading to model improvements through efficient weight updates** [7] **and better prompt design** [9, 11, 13].

I use methods centered in causal inference to make controlled interventions (i.e., ablations or substitutions) to neural network LMs. My work leverages two primary methodologies: **behavioral** [2, 9, 11, 12] and **mechanistic** [1, 4, 7, 8, 10]. Behavioral analyses make causal interventions on model *inputs*: I carefully construct input queries to test for specific model behaviors. This does not require access to model parameters, and can thus be used on closed-weight models. On the contrary, mechanistic analyses (popularly known as "mechanistic interpretability") make causal interventions on model *weights* and *representations* when available, allowing for deeper understanding. My meta-scientific analysis of mechanistic interpretability makes the case for why causality is so important [17].

In pursuit of my goal to provide comprehensive interpretations of performant systems in widespread use, I have instantiated my interpretability research on tasks that represent both fundamental building blocks of complex skills *and* real-world use cases of LMs. I describe 3 examples below.



A banana is yellow. What color is a banana?

Choices:
A. **yellow**
B. pink
C. black
D. blue

The correct answer is: **A**

A banana is yellow. What color is a banana?

Choices:
1. pink
2. **yellow**
3. black
4. blue

The correct answer is: **2**

Figure 3: Two variations of the same multiple-choice question. LMs must first select the correct answer string, then output its corresponding symbol. I've uncovered the internal mechanisms that models use to answer these questions robustly.

How do LMs encode factual and commonsense information? Correctly answering factual queries and performing commonsense reasoning about the world are key necessary functionalities of useful LMs. However, correcting model falsehoods in these domains is still largely accomplished by *finetuning*: indiscriminately and inefficiently updating model weights while balancing under- and overfitting. My work [7, 8] provides insight into **1) where and how information is stored in model weights**, **2) how it's learned during training**, and **3) how the findings for 1) and 2) differ for correct vs. incorrect information**. The answers to these questions provide a **promising alternative avenue to fine-tuning: targeted and efficient edits**, either to a small subset of model weights [7], or to the training data [8]. For example, in the commonsense domain, I have shown that models can be efficiently edited to provide correct plausibility assessments of real-world scenarios. Targeted edits to a model's early-layer MLP weights, found using causal interpretability techniques, can outperform fine-tuning at balancing under- and overfitting [7].

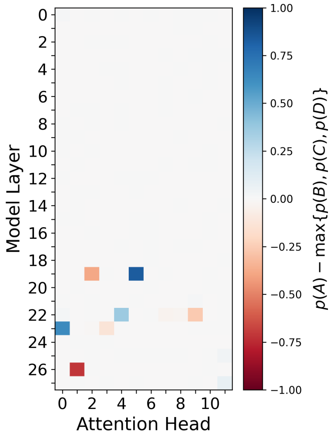


Figure 4: Specific attention heads at late layers of Transformer LMs promote the correct multiple-choice answer (blue); others either demote the correct choice or promote the incorrect choices (red) [10].

How do LMs answer multiple-choice questions? Multiple-choice question answering (MCQA) is a standard benchmark task format for evaluating LM capabilities. However, my work [9] and the works of others have shown that some state-of-the-art models are not robust to innocuous changes to the input, such as different symbols or answer choice orderings (Fig. 3). This lack of robustness is at odds with models capable of doing advanced reasoning, and can erode user trust.

I have demonstrated how performant models answer multiple-choice questions robustly and provided insight into why weaker models fail. When model weights are not accessible (such as OpenAI's), I have leveraged models' output probability distributions to determine how models prioritize certain answer choices. While prior work has argued that models are unfairly penalized when they assign high probability to invalid answer choices (termed "surface form competition"), I **demonstrated conclusive evidence that this is not the case in practice** by proposing the first mathematical formalism that could measure surface form competition's effects on model evaluation [9]. I discovered that prompting LMs to predict valid multiple-choice answer choices sometimes has surprisingly negative effects, demonstrating that in-context learning cannot always teach the symbol-mapping skill needed to do MCQA. My findings **challenged common practices in LM evaluation** and allowing me to **propose best practices** for model evaluation.

But what mechanisms exist in models that robustly answer MCQA questions (such as those in Fig. 3) regardless of the symbols used or the position of the correct answer? Throughcausal interventions on the internal functions of open-weight LMs, I found that **performant models answer questions in 3 stages**, which includes a phase for adaptation to unusual or out-of-distribution queries [10]. Notably, all of these mechanisms are driven by sparse subsets of network components, primarily individual attention heads (Fig. 4), making them good candidates for targeted intervention and control in future work.

3 Future Research

I have demonstrated that scientific understanding of LMs improves NLP systems, increases their reliability, and can calibrate human trust, but many open research questions remain that continue to grow in importance as LMs expand in reach and influence.

Improving Model Reliability by Uncovering how Models Learn Factual Information: LMs are still heavily prone to producing factually incorrect information ("hallucinating"), even when augmented with external lookup tools. Factuality is a basic requirement of trustworthy AI systems that requires, to some extent, memorization of training data. But to what extent? On the flip side, memorizing information like copyrighted data or personal information causes harm, and generalization from training data is a necessary and desirable property of machine learning models. Current research largely focuses on *either* reducing memorization *or* reducing hallucination; **any proposed solution will thus be suboptimal by increasing the other behavior**. Apart from these contradictory goals, methods in the literature to change model behavior **generally remain static**: LMs' weights are updated with computationally intensive fine-tuning and their behavior customized with finicky and ephemeral retrieval and/or few-shot prompting.

My recent work has discovered a strong correlation between interpretable linear structures for factual recall in LM representations and the frequency of these facts in the pretraining data [8]. Linearity allows for efficient model steering and control at inference time; in ongoing work, I am modifying pretraining data to better understand this causal relationship. I am pursuing a research agenda to answer many critical research questions to *both* **increase LM factuality** and **reduce verbatim reproduction of sensitive training data**: What makes data desirable vs. undesirable to memorize? How is knowledge represented internally in LM parameters? Finally, how can we efficiently and dynamically update subsets of model weights to *learn* and *unlearn* information, allowing us to simultaneously achieve both goals?

Improving Safety-Critical LM Behaviors such as Refusal: A critical safety capability of LMs is to **consistently identify and refuse to answer dangerous or inappropriate user queries and provide caveats on uncertain responses**. Yet, despite recent advances in training models to refuse to comply or provide uncertainty estimates for their predictions, models remain both overconfident and brittle to jailbreaking attacks. In recent work, I proposed a taxonomy of queries that should be refused, arguing that ideal refusal behavior should extend beyond malicious or unsafe queries to unanswerable ones, while also avoiding over-refusal that can render the system unhelpful [18]. We demonstrated that fine-tuning does not fully achieve this goal.

Recent interpretability research suggests that models **encode concepts like refusal or unanswerability in simple representational subspaces that can be easily undone at inference time**. In the same vein as my work on MCQA, factual recall, and commonsense reasoning, I will leverage my expertise in mechanistic interpretability to *first* develop a deeper understanding of the internal mechanisms underlying refusal and uncertainty expression, and *then* leverage this understanding to develop more successful and robust interventions than fine-tuning to ensure these safety-critical behaviors are deeply and reliably ingrained in LMs.

Explanations for Real-World Applications: Methods building on my work on generating explanations from large LMs (§1) have experienced a surge in popularity, alongside growing concern over the ease of generating deceptive and misleading explanations with powerful LMs. However, state-of-the-art explanation faithfulness research lacks contextualization in real-world use cases. I plan to **build and test state-of-the-art explainable NLP systems in diverse applications like healthcare, finance, and education**, where explanations that are not faithful to a model's underlying prediction process can have serious consequences. This involves answering research questions such as: how application-specific are definitions of explanation utility? How can instance-level explanations allow users to generalize their mental models of AI systems, without causing information overload? Are causal and mechanistic interpretations of LMs meaningful to lay users?

I additionally plan to research how explainable AI systems can **educate and inform policy makers** about model operations. I have advised staffers in the Senate Chamber of Commerce and the Washington State Senate's Environment, Energy and Technology Committee on explainable AI. Whether or not explanations for AI systems become a legal requirement, developing technology capable of satisfying the technical requirements of such a policy is my priority.

References

* indicates equal contribution.

[1] S. **Wiegreffe*** and Y. Pinter*, "Attention is not not Explanation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[2] K. Xie, S. **Wiegreffe**, and M. Riedl, "Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes," in *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[3] S. Jain, S. **Wiegreffe**, Y. Pinter, and B. C. Wallace, "Learning to Faithfully Rationalize by Construction," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[4] S. **Wiegreffe**, A. Marasović, and N. A. Smith, "Measuring Association Between Labels and Free-Text Rationales," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[5] J. Mullenbach, S. **Wiegreffe**, J. Duke, J. Sun, and J. Eisenstein, "Explainable Prediction of Medical Codes from Clinical Text," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[6] S. **Wiegreffe**, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, "Reframing Human-AI Collaboration for Generating Free-Text Explanations," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

[7] A. Gupta, D. Mondal, A. Sheshadri, W. Zhao, X. Li*, S. **Wiegreffe***, and N. Tandon*, "Editing Common Sense in Transformers," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[8] J. Merullo, N. A. Smith, S. **Wiegreffe***, and Y. Elazar*, "On Linear Representations and Pretraining Data Frequency in Language Models." Under review, 2024.

[9] S. **Wiegreffe**, M. Finlayson, O. Tafjord, P. Clark, and A. Sabharwal, "Increasing Probability Mass on Answer Choices Does Not Always Improve Accuracy," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[10] S. **Wiegreffe**, O. Tafjord, Y. Belinkov, H. Hajishirzi, and A. Sabharwal, "Answer, Assemble, Ace: Understanding How Transformers Answer Multiple Choice Questions." Under review, 2024.

[11] P. Hase, M. Bansal, P. Clark, and S. **Wiegreffe**, "The Unreasonable Effectiveness of Easy Training Data for Hard Tasks," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[12] Y. Elazar, B. Paranjape*, H. Peng*, S. **Wiegreffe***, K. R. Chandu, V. Srikumar, S. Singh, and N. A. Smith, "Measuring and Improving Attentiveness to Partial Inputs with Counterfactuals," in *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[13] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. **Wiegreffe**, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, "Self-Refine: Iterative Refinement with Self-Feedback," in *Neural Information Processing Systems (NeurIPS)*, 2023.

[14] D. J. Hilton, "Conversational Processes and Causal Explanation," *Psychological Bulletin*, 1990.

[15] S. **Wiegreffe*** and A. Marasović*, "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing," in *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Neural Information Processing Systems (NeurIPS)*, 2022.

[17] N. Saphra* and S. **Wiegreffe***, "Mechanistic?," in *Proceedings of the BlackboxNLP Workshop*, 2024.

[18] F. Brahman, S. Kumar, V. Balachandran*, P. Dasigi*, V. Pyatkin*, A. Ravichander*, S. **Wiegreffe***, N. Dziri, K. Chandu, J. Hessel, Y. Tsvetkov, N. A. Smith, Y. Choi, and H. Hajishirzi, "The Art of Saying No: Contextual Noncompliance in Language Models," in *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2024.