Adaptive Explanations via Direct Preference Optimization

Jacopo Teneggi[1], **Zhenzhen Wang**[1], **Paul H. Yi**[2], **Tianmin Shu**[1], **Jeremias Sulam**[1]

[1]Johns Hopkins University
[2]St. Jude Children's Research Hospital
{jtenegg1,zwang218,tianmin.shu,jsulam1}@jhu.edu, paul.yi@stjude.org

Abstract

Machine learning explainability aims to make the decision-making process of black-box models more transparent by finding the most important input features for a given prediction task. Recent works have proposed composing explanations from semantic concepts (e.g., colors, patterns, shapes) that are inherently interpretable to the user of a model. However, these methods generally ignore the communicative context of explanation—the ability of the user to understand the prediction of the model from the explanation. For example, while a medical doctor might understand an explanation in terms of clinical markers, a patient may need a more accessible explanation to make sense of the same diagnosis. In this work, we address this gap with listener-adaptive explanations. We propose an iterative procedure grounded in principles of pragmatic reasoning and the rational speech act to generate explanations that maximize communicative utility, and we evaluate our method on classification of lung X-rays. Our procedure only needs access to pairwise preferences between candidate explanations, relevant in real-world scenarios where a listener model may not be available.

Introduction

Understanding the decision-making process of modern machine learning systems is critical for their responsible use. This need has motivated considerable research efforts on the safety, fairness, and trustworthiness of black-box predictors. A notable approach is that of *explaining* predictions in a post-hoc fashion, i.e. finding the features that contributed the most to the output of a model, either globally over a population or locally for a given input (Covert, Lundberg, and Lee 2020). Instead of using input features (e.g., pixels for images or words for text), recent works have proposed to compose explanations with human-interpretable concepts, such as the presence of certain objects, colors, or patterns in images (Kim et al. 2018; Koh et al. 2020; Yuksekgonul, Wang, and Zou 2022; Teneggi and Sulam 2024). Although closer to the way humans articulate explanations compared to input features, these methods ignore the *context* of explanation (i.e., its *pragmatics*), which is a fundamental aspect of human reasoning (Grice 1975; Hobbs et al. 1987; Carston et al. 1993; Wilson and Sperber 2012; Recanati 1989; Gibbs Jr and Moise 1997). For example, a pragmatic explainer should first reason about its context, taking into consideration whom it is communicating with (e.g., a medical doctor, a nurse, or a patient), and adapt its output to maximize communicative utility.

In this work, we introduce a framework that tailors explanations to the communicative needs of different users without requiring an explicit listener model. We follow the rational speech act (RSA) (Goodman and Frank 2016) to formalize pragmatic reasoning as probabilistic inference over explanations. We leverage recent advances in reinforcement learning from human feedback (RLHF) (Christiano et al. 2017; Bai et al. 2022; Ouyang et al. 2022) and direct preference optimization (DPO) (Rafailov et al. 2023; Rosset et al. 2024) to study cases where the listener model is unknown but preference data can be collected. We focus on image classification tasks, and we propose an iterative procedure that, given a fixed predictor, jointly trains a speaker and a listener in a pragmatic reference game (a cooperative signaling game (Sobel 2020)). The goal of the game is for the listener to guess the prediction of the classifier (the signal) without seeing the input image but the utterance of the speaker only (i.e., the explanation), as depicted in Figure 1. We evaluate our method on CheXpert (Irvin et al. 2019), simulating listeners with different levels of technical knowledge.

Pragmatic Explanations

Recall that the goal of post-hoc explainability is to find the features that contribute the most towards the output of a fixed predictor on a particular input. Existing research has focused on developing tools to investigate the internal mechanisms of black-box models, for example with gradients (Selvaraju et al. 2016; Wang et al. 2024; Kolek et al. 2022), game- and information-theory (Lundberg and Lee 2017; Kolek et al. 2020, 2022; Teneggi, Luster, and Sulam 2022; Covert, Lundberg, and Lee 2021), conditional independence (Teneggi et al. 2022; Burns, Thomason, and Tansey 2020; Teneggi and Sulam 2024; Shi et al. 2024; Bharti, Yi, and Sulam 2024), and mechanistic interpretability (Nanda et al. 2023; Conmy et al. 2023; Bereska and Gavves 2024), among other strategies. These methods, however, overlook the context of such explanations (i.e., pragmatics). Following the RSA, we observe that an explanation is useful if a listener can correctly infer the prediction made by the machine learning

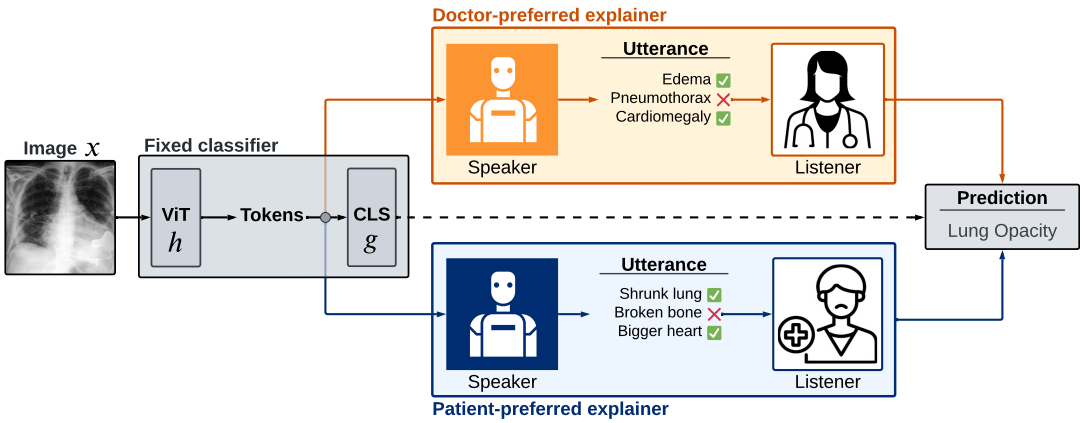

Figure 1: Illustration of our listener-adaptive explanation framework: a speaker generates utterances to help a listener infer model predictions without access to the input image.

model.

More formally, let $f : \mathcal{X} \to \mathcal{Y}$ with $\mathcal{Y} = [k] := \{1, \dots, k\}$ be a fixed predictor that maps an input $x \in \mathcal{X}$ to one of k classes. We assume f can be seen as the composition of a feature encoder $h : \mathcal{X} \to \mathbb{R}^d$ with a downstream classifier $g : \mathbb{R}^d \to \mathcal{Y}$, i.e. $f(x) = g(h(x))$, and that we have access to the embeddings $h(x)$. Denote $S : \mathbb{R}^d \to \mathcal{U}$ a speaker model (i.e., the explainer) that takes embeddings as input and generates utterances in $\mathcal{U}$. We consider utterances (i.e., explanations) composed of at most l claims from a vocabulary $\mathcal{C} = \{c_1, \dots, c_m\}$, where each claim can be labeled $+1$ or -1 depending on whether it is true or false for input x, i.e. $\mathcal{U} \subseteq (\mathcal{C} \times \{-1, +1\})^{\leq l}$. For example, in chest X-ray classification, f may predict whether a scan shows signs of lung opacity and $\mathcal{C}$ is a vocabulary of medical findings such as *edema*, *consolidation*, or *pneumothorax*. Then, a plausible utterance for the prediction *"signs of lung opacity"* could be $u = [(\texttt{edema}, +1), (\texttt{pneumothorax}, -1)]$. To connect machine learning explanations with pragmatic reasoning, let $L : \mathcal{U} \to \mathcal{Y}$ be a listener that receives an utterance and outputs one of the classes. The RSA implies that for an input x and prediction $\hat{y} = f(x)$, the distribution of utterances induced by a pragmatic explainer should satisfy

$$\log P_S(u \mid h(x)) \propto \underbrace{\log P(u \mid x)}_{\text{fidelity}} + \underbrace{\log P_L(\hat{y} \mid u)}_{\text{utility}}, \quad (1)$$

where the first term represents the *fidelity* of the utterance to the input x, and the second is the *utility* of the utterance for the listener. Then, our objective becomes to jointly optimize both the explainer S and listener L to maximize communicative reward, i.e.

$$\arg\max_{S,L} \mathop{\mathbb{E}}_{\substack{(x,\hat{y})\sim\mathcal{D}_f \\ u\sim P_S(h(x))}} \left[\log P(u \mid x) + \alpha \log P_L(\hat{y} \mid u)\right], \quad (2)$$

where $\mathcal{D}_f$ is the joint distribution of $(x, \hat{y} = f(x))$ and $\alpha \geq 0$ is a hyperparameter that controls the tradeoff between fidelity and utility (when $\alpha = 0$ the listener is ignored and the speaker is literal, maximizing fidelity only).

The Training Procedure

Our training procedure follows an alternating optimization approach where, at each step $t = 1, \dots, T$, the speaker and the listener are sequentially updated. Note that, given a dataset $D = \{(x_i, \hat{y}_i)\}_{i=1}^n$ containing inputs and their respective predictions by a fixed classifier f:

- Given a fixed listener $L^{(t-1)}$, the new speaker $S^{(t)}$ can be found via DPO on $D_{\text{pref}}^{(t)} = \{(x_i, \{(u_j^+, u_j^-)\}_{j=1}^{n_{\text{pref}}})\}$, where $n_{\text{pref}} = b(b-1)/2$ is the number of pairwise preferences constructed by ranking b utterances from $S^{(t-1)}$ according to Equation (1). Note that, in this step, $S^{(t-1)}$ plays the role of reference model in the DPO objective.
- Given a fixed speaker $S^{(t)}$, its optimal listener $L^{(t)}$ can be estimated from $L^{(t-1)}$ via cross-entropy minimization over a dataset of explanations $D_{\text{expl}}^{(t)} = \{(\hat{y}_i, (u_{i,1}, \dots, u_{i,n_{\text{expl}}}))\}_{i=1}^n$, where n_{expl} is the number of utterances for each input x_i, i.e. $u_{i,j} \sim P_{S^{(t)}}(\cdot \mid h(x_i))$.

Using DPO to update the speaker allows our method to handle scenarios with human subjects, whose fidelity and utility might be difficult to explicitly learn from data but for whom preferences can be collected. Within the scope of this work, we simulate listeners with neural networks to study the effectiveness of our proposed method.

Grounding Explanations. Recall that Equation 1 is used to rank utterances only, and that the fidelity term should ground them in a knowledge base, such as image captions, expert annotations, or the likelihood of a pretrained vision-language model. Here, we let $P(u \mid x) = \exp(\texttt{fidelity}(u, x)) / \sum_{u'} \exp(\texttt{fidelity}(u', x))$, where $\texttt{fidelity}(u, x) \in [0, 1]$ is an unnormalized score. This way, the pairwise ranks of candidate utterances do not depend on the normalizing constant $Z(x) = \sum_{u'} \exp(\texttt{fidelity}(u', x))$ but on $\texttt{fidelity}(u, x) + \alpha P_L(\hat{y} \mid u)$ only. Given the binary nature of the claims—and as done in previous works (Koh et al. 2020; Kim et al. 2018; Chattopadhyay et al. 2023; Chattopadhyay, Chan, and Vidal 2024)—

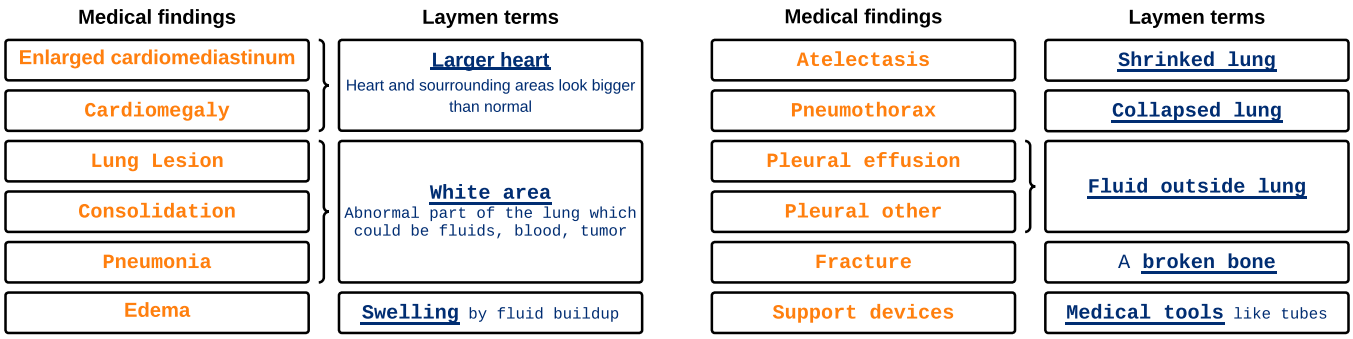

Figure 2: Mapping of medical claims in the CheXpert dataset to laymen friendly descriptions. The underlined terms are the abbreviations used in the figures in the main text of the manuscript.

we assume access to *semantics* $z : \mathcal{X} \to \{-1, 0, +1\}^m$ in the form of ground-truth annotations or predictions from a vision-language model (e.g., CLIP (Radford et al. 2021), Concept-QA (Chattopadhyay, Chan, and Vidal 2024)). These semantics $z(x) \in \{-1, 0, +1\}^m$ indicate which claims are true for x (i.e., $z(x)_j = +1$), which are false (i.e., $z(x)_j = -1$), and which are unknown (i.e, $z(x)_j = 0$). Then, we define `fidelity` as the weighted average of the number of true positive claims with the number of true negative claims, i.e. $\text{TP} = |\{(c_j, \hat{z}_j) \in u : \hat{z}_j = 1 \wedge z(x)_j = 1\}|$, $\text{TN} = |\{(c_j, \hat{z}_j) \in u : \hat{z}_j = -1 \wedge z(x)_j = -1\}|$, and

$$\texttt{fidelity}(u, x) = \frac{\text{TP} + \gamma \text{TN}}{|u|}, \ \gamma \in [0, 1]. \tag{3}$$

As $\gamma \to 0$, utterances with negative claims are downvoted, and the opposite as $\gamma \to 1$. Varying γ accounts for the fact that as the vocabulary $\mathcal{C}$ grows, it is easier for a claim $c \in \mathcal{C}$ to be false, making the ground-truth semantics imbalanced.

Simulating listener preferences. There are several ways to simulate listeners with different preferences over utterances, such as preference over certain topics, categories, or styles of presentation (e.g. technical nomenclature vs. common parlance). Here, we use temperature scaling (Guo et al. 2017) with respect to a prior distribution over *groups* of claims. That is, we assume access to a family of groups $\mathcal{G} = \{g_j : \mathcal{C} \to \{0, 1\}\}$ such that $g_j(c) = 1$ if c belongs to the j^{th} group (and groups may intersect). Then, a listener L_π has a prior distribution over groups $\pi \in \Delta^{|\mathcal{G}|}$ and

$$P_{L_\pi}(\hat{y} \mid u) = \texttt{softmax}\left(\frac{\xi(u)}{\tau \cdot \text{KL}(g(u) || \pi) + 1}\right), \tag{4}$$

where $\xi : \mathcal{U} \to \mathbb{R}^k$ are the unnormalized outputs (the logits) of the neural network simulating the listener, $g : \mathcal{U} \to \Delta^{|\mathcal{G}|}$ is the distribution of groups in utterance u, and $\tau \geq 0$ is a temperature parameter that strengthens the effect of the prior. Intuitively, the larger the distance between the observed and prior distributions, the more uniform $P_{L_\pi}(\hat{y} \mid u)$, making it harder for the listener to predict $\hat{y}$ confidently. We note that we use the reverse KL to penalize utterances with claims belonging to groups that have zero mass in the prior (i.e., $\pi_j = 0$ and $g(u)_j > 0$), which would not have an effect on the forward KL.

Lung Opacity Detection in Chest X-Rays

We evaluate our method on lung opacity detection with the CheXpert dataset (Irvin et al. 2019), which contains 224,316 chest X-rays labeled for the presence of 14 clinical findings from their textual reports. We predict whether a scan contains signs of lung opacity with the zero-shot classifier *"No signs of lung opacity"* and *"Findings suggesting lung opacity"* of BiomedVLP (Bannur et al. 2023), which has a base accuracy of 78,21%. We train on the first 10,000 patients in the dataset (40,702 images total), and keep the original validation set of 234 scans. We note that we use the VisualCheXbert (Jain et al. 2021) labels, and that concept annotations in the validation set are verified by trained radiologists. We include 12 of the 14 clinical findings in our explanation vocabulary, excluding "no findings" and "lung opacity". When simulating a non-technical listener, we expand this vocabulary to 20 terms by aggregating and mapping medical claims to laymen friendly descriptions based on an expert radiologist's input (see Figure 2).

We implement our speaker models with the CoCa architecture (Yu et al. 2022) and we simulate listeners with bidirectional transformers (Devlin et al. 2019). We train all models with AdamW (Loshchilov and Hutter 2017) with learning rate of 0.0001, weight decay of 0.01, gradient norm clipping at 1.0, and cosine annealing after each iteration of the procedure. We use 4 claims per utterance, $\alpha = 0.2$ when training pragmatic speakers, true negative weight $\gamma = 0.4$, DPO regularization strength $\beta = 0.6$, number of utterances $n_{\text{expl}} = 8$, and number of candidate utterances $b = 4$ (i.e., $n_{\text{pref}} = b(b-1)/2 = 6$). We now present and discuss the main findings of our experiment.

Pragmatic Explanations Improve Listener Accuracy

Recall that if the pragmatic strength $\alpha = 0$, the reward ignores utility, and the speaker model S is *literal* in the sense that it maximizes fidelity only. We found that pragmatic listeners (i.e., listeners trained jointly with a pragmatic speaker), have higher performance than literal ones (98.75% vs 84.62%). This confirms that pragmatic explanations communicate findings more effectively. Furthermore, we note listener accuracy is higher than the base classifier's.

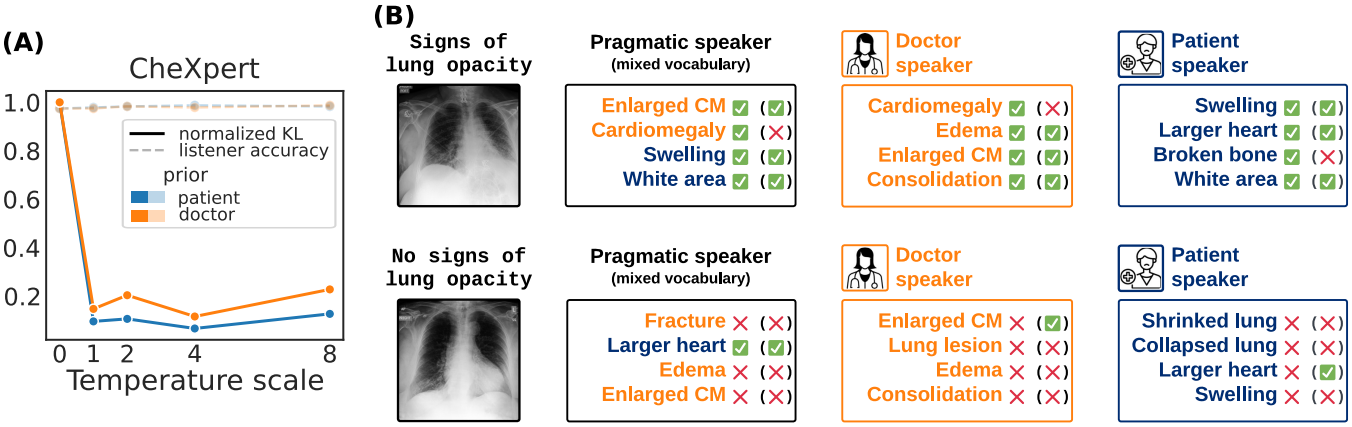

Figure 3: **(A)** Preference alignment (KL divergence) and listener accuracy as a function of temperature scale τ. **(B)** Example results for preference adaptation. Blue claims are clinical, and orange ones their meaning in laymen terms. Predicted semantics are included without parentheses, and ground-truth ones in parentheses.

We remark that the former is measured with respects to predictions, and the latter with ground-truth labels, hence why listener models can have higher accuracy than the base classifier. More importantly, this means that the speaker model is sometimes able to communicate wrong predictions. This highlights the importance of including safeguards such as explaining confident predictions only or restricting explanations to using only the claims that were statistically significant for the prediction of the base classifier (Teneggi and Sulam 2024) in order to avoid unintended consequences in real-world scenarios.

Pragmatic Explanations Adapt to Listener Preferences

We consider two listeners: a *doctor* listener who prefers medical terminology, and a *patient* listener favoring laymen descriptions. We remark that the laymen vocabulary contains fewer terms, such that adapting is not equivalent to finding a one-to-one mapping of the claims. Figure (3) (Panel A) reports listener accuracy and preference alignment (evaluated as the KL divergence between the distribution of groups in the utterances and the prior) as a function of temperature scale τ. We normalize KL divergence to a pragmatic speaker with no preference adaptation at $\tau = 0$. We found that a temperature parameter greater than 0 leads to a significant reduction of KL divergence, i.e. the speaker successfully aligns with the simulated prior, and that alignment does not reduce classification performance. Figure 3 (Panel B) includes a representative example for a true positive scan and a true negative scan. We can see that a pragmatic speaker with no adaptation (i.e., $\tau = 0$) mixes medical findings with laymen terms (orange and blue, respectively), and that the doctor and patient speakers successfully restrict claims to their respective vocabularies. Furthermore, we note that the claims in the utterances of the doctor and patient speaker do not necessarily match, showing that adaptation is not simply a translation of utterances.

Limitations and Future Work

In this work, we put forth a framework to explain predictions of black-box models grounded in principles of pragmatic reasoning, the rational speech act, and direct preference optimization. While prior works have focused on technical advancements to inspect the inner mechanisms of these complex systems, we address the gap between machine learning explanations and human communication, providing empirical evidence that the pragmatics of explanations are important for real-world applications with several different stakeholders.

Naturally, our work has limitations. First, our training procedure relies on iterative reinforcement of the speaker and listener models, which may be expensive for large datasets. In particular, the space of utterances grows exponentially with the number of claims in the vocabulary $\mathcal{C}$, which leads to a *exploration-exploitation* tradeoff (Sutton, Barto et al. 1998) when sampling the preference and explanation datasets ($D^{(t)}_{\text{pref}}$ and $D^{(t)}_{\text{expl}}$). Possible strategies to achieve better coverage of the space of utterance when the vocabulary $\mathcal{C}$ is large may be to retain candidate utterances from previous iterations (Rosset et al. 2024), or to use an exponential moving average as the reference model (Llama Team 2024). Second, free-form text explanations may be more flexible than claim-based ones as vocabularies evolve over time. This requires claim decomposition and verification techniques to prevent hallucinations (Wanner, Van Durme, and Dredze 2024; Kamoi et al. 2023) and generation of misleading explanations. Finally, we consider training speaker models that can adapt to multiple listeners at once as future work.

Acknowledgments

This research was supported by NSF CAREER Award CCF 2239787.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon,

C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15016–15027.

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety–A Review. *arXiv preprint arXiv:2404.14082*.

Bharti, B.; Yi, P.; and Sulam, J. 2024. Sufficient and necessary explanations (and what lies in between). *arXiv preprint arXiv:2409.20427*.

Burns, C.; Thomason, J.; and Tansey, W. 2020. Interpreting black box models via hypothesis testing. In *Proceedings of the 2020 ACM-IMS on foundations of data science conference*, 47–57.

Carston, R.; et al. 1993. Conjunction, explanation and relevance.

Chattopadhyay, A.; Chan, K. H. R.; Haeffele, B. D.; Geman, D.; and Vidal, R. 2023. Variational information pursuit for interpretable predictions. *arXiv preprint arXiv:2302.02876*.

Chattopadhyay, A.; Chan, K. H. R.; and Vidal, R. 2024. Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In *The Twelfth International Conference on Learning Representations*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Conmy, A.; Mavor-Parker, A.; Lynch, A.; Heimersheim, S.; and Garriga-Alonso, A. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36: 16318–16352.

Covert, I.; Lundberg, S.; and Lee, S.-I. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209): 1–90.

Covert, I.; Lundberg, S. M.; and Lee, S.-I. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33: 17212–17223.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Gibbs Jr, R. W.; and Moise, J. F. 1997. Pragmatics in understanding what is said. *Cognition*, 62(1): 51–74.

Goodman, N. D.; and Frank, M. C. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11): 818–829.

Grice, H. P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Hobbs, J. R.; et al. 1987. *Implicature and definite reference*. CSLI.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Jain, S.; Smit, A.; Truong, S. Q.; Nguyen, C. D.; Huynh, M.-T.; Jain, M.; Young, V. A.; Ng, A. Y.; Lungren, M. P.; and Rajpurkar, P. 2021. VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, 105–115.

Kamoi, R.; Goyal, T.; Rodriguez, J. D.; and Durrett, G. 2023. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Kolek, S.; Nguyen, D. A.; Levie, R.; Bruna, J.; and Kutyniok, G. 2020. A rate-distortion framework for explaining black-box model decisions. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 91–115. Springer.

Kolek, S.; Nguyen, D. A.; Levie, R.; Bruna, J.; and Kutyniok, G. 2022. Cartoon explanations of image classifiers. In *European Conference on Computer Vision*, 443–458. Springer.

Llama Team, A. . M. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Nanda, N.; Chan, L.; Lieberum, T.; Smith, J.; and Steinhardt, J. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Recanati, F. 1989. The pragmatics of what is said.

Rosset, C.; Cheng, C.-A.; Mitra, A.; Santacroce, M.; Awadallah, A.; and Xie, T. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

Selvaraju, R. R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; and Batra, D. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Shi, C.; Beltran Velez, N.; Nazaret, A.; Zheng, C.; Garriga-Alonso, A.; Jesson, A.; Makar, M.; and Blei, D. 2024. Hypothesis testing the circuit hypothesis in LLMs. *Advances in Neural Information Processing Systems*, 37: 94539–94567.

Sobel, J. 2020. Signaling games. *Complex social and behavioral systems: Game theory and agent-based models*, 251–268.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Teneggi, J.; Bharti, B.; Romano, Y.; and Sulam, J. 2022. Shap-xrt: The shapley value meets conditional independence testing. *arXiv preprint arXiv:2207.07038*.

Teneggi, J.; Luster, A.; and Sulam, J. 2022. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4494–4503.

Teneggi, J.; and Sulam, J. 2024. Testing Semantic Importance via Betting. *Advances in Neural Information Processing Systems*, 37: 76450–76499.

Wang, Y.; Zhang, T.; Guo, X.; and Shen, Z. 2024. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*.

Wanner, M.; Van Durme, B.; and Dredze, M. 2024. DnD-Score: Decontextualization and Decomposition for Factuality Verification in Long-Form Text Generation. *arXiv preprint arXiv:2412.13175*.

Wilson, D.; and Sperber, D. 2012. Linguistic form and relevance. *Wilson & Sperber (Eds.), Meaning and Relevance*, 149–168.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*.