Quantitative Assessment of Intersectional Empathetic Bias and Understanding

Vojtěch Formánek[1,2] and Ondřej Sotolář[1]

[1] Masaryk University, Faculty of Informatics
[2] Masaryk University, Department of Psychology, Faculty of Arts
xforman@mail.muni.cz

Abstract. A growing amount of critique concerns the current operationalizations of empathy based on loose definitions of the construct. Such definitions negatively affect dataset quality, model robustness, and evaluation reliability. We propose an empathy evaluation framework that operationalizes empathy close to its psychological origins. The framework measures the variance in responses of LLMs to prompts using existing metrics for empathy and emotional valence. We introduce the variance by varying social biases in the prompts, which affect context understanding and thus impact empathetic understanding. Our method maintains high control over the prompt generation, ensuring the theoretical validity of the constructs in the prompt dataset. Also, it makes high-quality translation, especially into languages with little to no way of evaluating empathy or bias, such as the Slavonic family, more manageable. Using chosen LLMs and various prompt types, we demonstrate the empathy evaluation with the framework, including multiple-choice answers and free generation. The measured variance in our initial evaluation sample is small, and we were unable to find the expected differences between the empathetic understanding given the differences in context for distinct social groups. However, the models showed significant alterations in their reasoning chains that were needed to capture the relatively subtle changes in the prompts. This provides the basis for future research into the construction of the evaluation sample and statistical methods for measuring the results.

Keywords: Intersectional Bias, Empathy, LLM Evaluation.

1 Introduction

While there has been a vast amount of literature on empathy, it has come under increased scrutiny due to the unclear way of operationalizing empathy [13,12]. Loosely defining empathy as *the ability to understand another person's feelings and respond appropriately* [12] was shown to cause problems across different tasks, such as dataset creation [5], training [30], and evaluation [12]. This ambiguity led to a narrow focus on emotion recognition and prediction. We argue, alongside previous work, that this effort misunderstands empathy's psychological origins. To improve upon the current operationalizations of empathy, we propose a

framework with (i) disambiguation of empathy, (ii) measurement operationalization intended specifically for computational models and (iii) an evaluatory procedure.

Some of the problems stem from the disagreement on the definition of empathy within psychology itself [4]. It was originally used to describe human ability to understand others [15,11]. Current research agrees that it has two components to achieve this: *affective empathy*, sometimes also named emotional empathy, and *cognitive empathy* [31]. Affective empathy refers to the capacity to feel emotions for others as a result of our belief, perception or imagination of their situation [16]. Cognitive empathy involves theorizing about and simulation of others mental states [31], that is to: (i) retrodictively simulate a mental state, to explain observed behavior, (ii) take that mental state and run it through our cognitive mechanisms, and (iii) attribute the conclusion to the target for explanation and prediction.

Since cognitive empathy is dependent on one's cognitive mechanisms, it is also dependent on experience. Because of that, a person can have different levels of understanding based on the similarity of their experience and the state they are observing and feel different levels of empathy toward different social groups [37]. This effect is carried over to LLMs [25,22,32,3] which are reinforced on human preference data. Observing if models exhibit this type of bias toward some of the groups thus can be leveraged to indirectly study empathy.

We propose an empathy evaluation framework for conversational agents, such as LLMs, which focuses on empathetic understanding. The framework uses masked templates to generate an evaluation dataset of prompts designed for the agents to respond to. The templates include masked sections into which different information is inserted. The information is biased towards different social groups; the selection of the type of information, the values, and social groups are inspired by current research such as reviewed in Gallegos et al. [8].

Filling the masked section with varying values results in the evaluation sample. This enables measuring the variance in responses across a single template, which assumes the invariance in empathetic understanding, affect, and responding both inside and at the intersection of different social groups. This invariance also means that bias within the framework, observed in the scores of a given metric, manifests as a deviation from the central tendency of the scores across a given social group.

This might or might not be preferable; thus, we give fine-grained tools for individually interpreting the scores within a given intersectional group. Taking into account Blodgett et al.'s [1] criticism of the study of biases, within the framework, we only study the tendencies in model outputs and make no claims about the potential harmfulness or possible impacts of the biases. We make all the data and code publicly available[3].

---

[3] `https://github.com/xforman/JaEm_st`

2   Related Work

Several metrics have been proposed for evaluating bias in generated text [8]. The metrics can be based on a difference in the distribution of tokens in the generated outputs given distinct groups [23] or on lexicons [6]. Alternatively, classifiers are utilized, typically to detect relevant phenomena such as toxicity [28,29]. The datasets used for evaluating bias deal with various social groups and issues (see Table 4 in [8] for an overview) and are sometimes created from existing datasets. Sample construction involves replacing the relevant social group identifiers $G_1, ..., G_m \in \mathbb{G}$ (gender, race, etc.) with a mask, thus creating masked samples. When evaluating a mask, masking a specific social group G_i, it is substituted with a protected attribute $a_i \in G_i$ from that social group. The shift in the responses is measured [19,35], assuming that the output should be invariant considering distinct social groups. This technique is sometimes called bias mitigation via contact hypothesis [22], a term borrowed from psychology referring to direct contact with other social groups [20].

Datasets used to train empathetic agents are typically single or multi-turn, with emotional labels [2,24]. However, other datasets, by their nature, contain empathy as well, such as transcripts of everyday conversations [14], simulation of other's personas [36] or transcripts of therapies [17,21], labeled with conversational behaviors [3]. Retroactively categorizing existing empathy metrics into the two dimensions is difficult, especially since they likely overlap.

Since *cognitive* empathy involves understanding, the accuracy of emotion prediction can be considered a case of it, but a broader understanding has also been measured. Zhu et al. [38] collected user comments about products and their do-, motor- and be-goals [10], then instructed human or LLM designers to predict those goals and measured the token similarity between them.

The problems with measuring *affective* empathy are caused mainly by its dependence on an inner state. Lee et al. [13] propose a set of feature-based metrics for evaluating model responses, which include mechanisms outside empathy, but we consider them relevant because they measure the empathetic qualities indirectly. Concretely, the set includes specificity, based on a normalized variant of inverse document frequency (NIDF) [26] and measures the similarity in the vocabularies of the model and user. They also introduce valence, arousal, and dominance (VAD) based on the NRC Emotion Intensity Lexicon [18]. With the focus on the similarity of the input and output texts, closer results are preferred. All of the *affective* metrics assume that empathy in this context manifests in the similarity. The described metrics also fall into the category of empathetic responding, which is the focus of many currently existing measures. A different method is used in the EPITOME empathy metric for dialogues [27], which is based on a fine-tuned RoBERTa model that predicts three dimensions on a scale (0-2; none, weak, strong): Empathetic Responses (ER), Explanations (EX) and Interpretations (IP). Chiu et al. [3] evaluate differences between human and LLM therapists and define several dimensions whose quality is in part dependent on the empathetic capacity of the therapist – Reflections, Questions, and Solutions.

All of the metrics depend on the true state of the evaluated sample (for example, the labeled emotion). This makes both dimensions of empathy dependent on this state, meaning that misunderstanding leads to an inaccurate affect. Thus, we cannot separate *affective* from *cognitive* empathy. For this reason, Coll et al. [4] operationalizes the measurement of *affective* as the similarity of that affective state to the one in the *understood* state.

3   JaEm-ST: Framework for the Quantitative Assessment of Empathetic Behavior of LLMs

We propose an evaluation framework, JaEm-ST, where empathy has two dimensions, *Cognitive* and *Affective*, which follow the definitions introduced in Section 1. However, we operationalize the measurement into three dimensions. Cognitive empathy (CE), the degree to which the empathizer understood the observed state correctly. Affective empathy (AE), the degree to which the empathizer's state matches that of the understood state (inspired by Coll et al. [4]). Lastly, we define *Empathic Response Appropriateness* (ERA) to the understood state, which is the result of the empathic process (and several others) but not its dimension. In our context, we define "state" as a person's momentary mental and physical circumstances.

3.1   Theoretical Basis of the Framework

Dependence of *cognitive* empathy on experience means that different empathizers might come to different conclusions about an observed state, even when self-reporting [9]. This impacts the evaluation on two sides: (i) LLM empathizers might interpret the situation differently, which does not mean that it is false, and (ii) the "true state" constructed by the creator might not even be reflective of their empathetic understanding. Thus, it is difficult, if not impossible, to determine whether a given interpretation is genuinely false, but it is nonetheless representative of a human interpretation grounded in experience. For this reason, AE and ERA depend on the model's understanding of the state, so it is possible to evaluate an output even when it does not interpret the context in the same way as the creator of that template. But if we concede that the interpretation of states is subjective, then we need another standard for the evaluation.

In our case we assume that empathetic understanding is invariant across similar situations; the implications of this assumption are discussed later. Which is why JaEm-ST focuses on finding systemic differences in model output when responding to similar situations and uses the similarity between the "true state" and the one predicted by the model as a guiding principle. Given that experience can lead to biases, we create a single evaluation example by inserting protected attributes $a_i = (a_{i1}, ..., a_{ih}) \in G_1, ..., G_h$ (such as specific sexuality, education etc.) associated with their respective social groups $G_k \in \mathbb{G}$ into a masked template $t \in \mathbb{T}$ (see Fig 1), by a procedure $fill(t, a)$. Thus two examples created from the same template differ only in the specific protected attributes that were inserted into them.

3.2   Evaluation Sample

We create the evaluation sample D dynamically from a small seed dataset of pre-defined templates $\mathbb{T}$. The templates simulate conversations between a human speaker and an empathizer, which is assumed to be a conversational agent and is to be evaluated. It consists of four parts: instruction, context, conversation, and answers. The templates are then filled with different combinations of protected attributes $A \subseteq \times_{k=1}^{h} G_k$, thus $D = \{fill(t,a) \mid \forall t_{\in \mathbb{T}}, a_{\in A}\}$. For us $h = 7$ and A is a set of commonly examined social groups: race, education, religion, age [19] and socio-economic status [35], sexuality and pronouns. Typically, the cartesian product of all possibilities, i.e. $A = \times_{k=1}^{h} G_k$, would be too large, thus we only fully sample the dimensions of interest and randomly sample the rest.

The templates are based on *causal tuples*, which provide the reasoning for what caused a person's current observed state. It is partly set up by the context, which defines the social groups the observed person is a member of. The conversation within the templates implicitly or explicitly describes the situation grounded on the causal tuple. The conversation can also include the same masks as the context.

Given a filled template, the evaluated model is then prompted to continue the conversation. If evaluating AE and ERA, it simply continues generating the response to the observed person's last utterance. When evaluating CE, it is given a choice between five answers and is prompted to pick the one with the deepest understanding of how the person reached the observed state in the last utterance.

Even though we assume that empathetic understanding is invariant in the templates, there may be cases when this assumption does not hold. If certain types of responses are preferable for a given group, and this type of response is measured by a metric, it is possible to set thresholds for this metric and group separately, independently of the overall scores. This is possible because we have stratified information about the social groups which the observed person in the evaluated context is a part of.

3.3   Experimental setup

For the evaluation, we constructed two causal tuples and templates , as shown in Figure 1. To produce D, we focus only on two of the dimensions (I): **sexuality** and **pronoun**, we take all $a \in I$ four times and randomly sample from the rest of the dimensions. We get 486 samples for generation or multi-choice, so each model is evaluated with 972 samples, that is 1944 inputs in total.

Because we aim to maintain an understandable, controllable, high quality dataset with high ecological validity, we keep the number of templates relatively low. This also means that it is relatively simple to translate the templates to other languages. This could be especially beneficial for languages that currently have little to no way of evaluating bias or empathy, such as the Slavonic family.

The examples $d \in D$ are given as input to a model in a single-turn fashion, which is most typical for evaluating empathy. The examples are used as is, con-

*Causal tuple (not part of the model input in any case)**:**
(adores her parrot, happiness from a gift for her parrot, less critical towards
disliked food, complimented the cook)

*Prompt:
When given a choice of answers A-E choose the one that shows the deepest
understanding of Jane's situation, not necessarily the most empathetic. Carl has
a full understanding of the context below:

Context:
Generally positive and optimistic person. [PRONOUN] is a [SEXUALITY]
and a firm [RELIGION]. [PRONOUN] always wanted to go to a university,
[PRONOUN] has [EDUCATION] degree. Being a proud cooking connoisseur
and a stern critic, [PRONOUN] rarely compliments other's food, but today
[PRONOUN] complimented pasta ... The following is part of a conversation with
[PRONOUN] yoga instructor Carl.

Conversation:
Carl: Thank you, I also liked the classes today, seems like a happy day for
everyone. I wanted to make it a little harder for you. I noticed some exhausted
faces, but not any annoyed ones. How were the exercises in the middle?
Jane: I didn't see any annoyed faces either ...
...
Jane: ... something weird happened to me today, we went for a typical lunch with
my coworkers, but I think I lost my integrity, and complimented a food I did not
actually like!
Carl:

*Answers:
A: I think gift for Poppy made you think of her and made you so happy, that you
complimented the chef even though you didn't actually like the food.
B: You? Did you actually complimented a food? you're always so strict and stern
about critiquing other people's food. Maybe the day, you know, the yoga, Lucy's
traveling and Poppy made you actually enjoy the food. What was it?
C: That's surprising, maybe it's because you were feeling so good ... Anyhow, did
Lucy tell you how her backpacking was in Texas?
D: Jane, I think it's understandable to feel like this. But your integrity isn't lost on
an occasion like this. ... How do you feel about the situation now?
E: Don't be silly, it does not make you less of a critic. ...

Fig. 1: An example of the framework's templates. [TAG] indicates masks to be
replaced with a chosen social group's attributes. Answers and the prompt are
only given to the model if cognitive empathy is being evaluated. Otherwise, it
generates Carl's last response. Option (A) shows the deepest understanding
since it is closest to the causal tuple – three dots in the text mark parts excluded
for brevity.

taining the following components in order: (i) $prompt_{CE}, context, conversation,$ *answers* for CE multiple choice, (ii) *context, conversation* for AE and ERA. We pass each of the variants to the model separately. The resulting CE score is computed as the accuracy of selecting the choices that manifest the most understanding of the speaker's state across all samples. The AE score comprises Valence, Arousal, and Dominance scores. While we consider them unreliable measures of affective empathy because they measure it indirectly, they can provide valuable information about the type of affect the model uses in its response. For measuring ERA, we use the EPITOME's dimensions IP, EX, and ER.

We evaluate *Llama-3.1-8B* [7] and *Zephyr-gemma-v.1* [33,34]. A 80 GB Nvidia A100 graphics card was used to process the evaluation.

4   Results

Table 1 shows significant differences in all three evaluated dimensions (CE, AE, ERA) measured across the models and templates. The metrics respective to the dimensions are: accuracy of multiple-choice answers (CE), VAD (AE) and EPITOME scores (ERA). For the different intersectional groups, the measured variance between scores is much smaller. Generally, *zephyr-gemma* performed

Model	Template	Count	Cognitive Empathy MC↑	Affective Empathy V↓	A↓	D↓	Response Appropriateness IP↑	EX↑	ER↑
Llama-3.1	0	243	0.29	0.16	0.11	0.13	0.12	1.56	0.54
	1	243	0.35	0.15	0.11	0.13	0.19	0.09	0.91
Zephyr-gemma	0	243	0.01	0.22	0.14	0.18	0.08	0.19	0.79
	1	243	0.10	0.18	0.13	0.16	0.07	0.01	1.60

Table 1: Evaluation of the generated responses on the framework's three dimensions of empathy, the interpretation of the metrics is explained in Section 2. The metric for Cognitive empathy is the accuracy of selecting choices with the most understanding among the multi-choice answers (example in Fig. 1). *Llama-3.1* has higher accuracy in this task. *Llama-3.1* is also more stable across the VAD metrics, and also Empathetic Explorations (EX) on only one of the templates, which were low otherwise. *Zephyr-gemma* had higher scores in Empathetic Responding (ER), this may be caused by the fact that it tends to role-play less.

much worse in CE. The results show that it was easiest for both of the models to find the answer with the most understanding to the sample shown in Figure 1. The differences in AE scores between the models are smaller and *Llama-3.1* achieves lower scores and smaller differences between the two samples, ERA scores are similar.

For pronouns, one of the evaluated dimensions, there are no obvious differences between the scores. We found outliers in the intersection of these groups,

such as Lesbian/She, which has an Interpretations (IP) score significantly above the overall average ($\sigma = 3.36$), but were unable to find any noticeable differences in the generated output.

Both models followed the multi-choice prompt well. For the free generation, manual inspection of a small subset the outputs suggests that *Llama-3.1* might follow role-playing better, *zephyr-gemma* tends respond from the third perspective (10-15 %), or set the situation up in a couple of sentences (10 %), instead of directly responding.

5　Conclusion

We provided a disambiguation of empathy for computational models to help future work define the construct closer to its psychological origins as opposed to the loose definitions that are currently widespread. As a main result, we proposed a new empathy evaluation framework for the responses generated by conversational agents that acknowledges the inherent subjectivity of empathetic understanding. The framework focuses on how empathetic understanding and responding is influenced by intersectional bias. It provides methods to generate evaluation samples from templates by inserting the intersectional contexts into them. The framework uses a new three-dimensional measurement operationalization of empathy to measure the construct. We demonstrated the usage of the framework on a small synthetic sample. In all three framework dimensions, we measured significant differences between the *Llama-3.1-8B* and *Zephyr-gemma-v.1* models. Lastly, we identified differences in empathetic understanding across the evaluated metrics in some intersectional groups. More importantly, we showed the framework's strength in providing the ability to stratify scores across a wide range of social contexts, giving a more fine-grained insight into model behavior and potential harms.

Limitations

We view the modest number of templates as a limitation. Even though we can produce many examples by substituting into the masks, most of their structure stays the same. Further, the contexts and conversations do not reflect the true variety across multiple different groups; future work should thus focus on increasing the number and diversity of the sample creators. The template structure itself, especially the inclusion of context as an explanation of the speaker's background, does place limitations the naturalness and ecological validity of the framework. Lastly, while the number of empathy metrics in this work is limited, future works can use the outlined criteria to include other metrics.

References

1. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050 (2020)
2. Buechel, S., Buffone, A., Slaff, B., Ungar, L., Sedoc, J.: Modeling empathy and distress in reaction to news stories. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4758–4765. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1507, https://aclanthology.org/D18-1507
3. Chiu, Y.Y., Sharma, A., Lin, I.W., Althoff, T.: A computational framework for behavioral assessment of llm therapists. arXiv preprint arXiv:2401.00820 (2024). https://doi.org/10.48550/arXiv.2401.00820
4. Coll, M.P., Viding, E., Rütgen, M., Silani, G., Lamm, C., Catmur, C., Bird, G.: Are we really measuring empathy? proposal for a new measurement framework. Neuroscience & Biobehavioral Reviews **83**, 132–139 (2017)
5. Debnath, A., Conlan, O.: A critical analysis of empatheticdialogues as a corpus for empathetic engagement. In: Proceedings of the 2nd Empathy-Centric Design Workshop. EMPATHICH '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3588967.3588973, https://doi.org/10.1145/3588967.3588973
6. Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.W., Gupta, R.: Bold: Dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 862–872 (2021)
7. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
8. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. Computational Linguistics pp. 1–79 (2024)
9. Grainger, S.A., McKay, K.T., Riches, J.C., Chander, R.J., Cleary, R., Mather, K.A., Kochan, N.A., Sachdev, P.S., Henry, J.D.: Measuring empathy across the adult lifespan: A comparison of three assessment types. Assessment **30**(6), 1870–1883 (2023)
10. Hassenzahl, M.: Experience design: Technology for all the right reasons. Morgan & Claypool Publishers (2010)
11. Jahoda, G.: Theodor lipps and the shift from "sympathy" to "empathy". Journal of the History of the Behavioral Sciences **41**(2), 151–163 (2005)
12. Lahnala, A., Welch, C., Jurgens, D., Flek, L.: A critical reflection and forward perspective on empathy and natural language processing. arXiv preprint arXiv:2210.16604 (2022)
13. Lee, A., Kummerfeld, J., Ann, L., Mihalcea, R.: A comparative multidimensional analysis of empathetic systems. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 179–189 (2024)
14. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: DailyDialog: A manually labelled multi-turn dialogue dataset. arXiv [cs.CL] (Oct 2017)
15. Lipps, T.: Leitfaden der psychologie. W. Engelmann (1909)
16. Maibom, H.L.: Affective empathy. In: The Routledge handbook of philosophy of empathy, pp. 22–32. Routledge (2017)

17. Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M.S., Chakraborty, T.: Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In: Proceedings of the fifteenth ACM international conference on web search and data mining. pp. 735–745 (2022)
18. Mohammad, S.M.: Word affect intensities. arXiv preprint arXiv:1704.08798 (2017)
19. Morales, S., Clarisó, R., Cabot, J.: A dsl for testing llms for fairness and bias. In: Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems. pp. 203–213 (2024)
20. Paluck, E.L., Green, S.A., Green, D.P.: The contact hypothesis re-evaluated. Behavioural Public Policy **3**(2), 129–158 (2019)
21. Pérez-Rosas, V., Wu, X., Resnicow, K., Mihalcea, R.: What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 926–935 (2019)
22. Raj, C., Mukherjee, A., Caliskan, A., Anastasopoulos, A., Zhu, Z.: Breaking bias, building bridges: Evaluation and mitigation of social biases in LLMs via contact hypothesis. arXiv [cs.CL] (Jul 2024). https://doi.org/10.48550/arXiv.2407.02030
23. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1264, `https://aclanthology.org/D16-1264`
24. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5370–5381. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1534, `https://aclanthology.org/P19-1534`
25. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: International Conference on Machine Learning. pp. 29971–30004. PMLR (2023)
26. See, A., Roller, S., Kiela, D., Weston, J.: What makes a good conversation? how controllable attributes affect human judgments. arXiv preprint arXiv:1902.08654 (2019)
27. Sharma, A., Miner, A., Atkins, D., Althoff, T.: A computational approach to understanding empathy expressed in text-based mental health support. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5263–5276. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.425, `https://aclanthology.org/2020.emnlp-main.425`
28. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: The woman worked as a babysitter: On biases in language generation. arXiv preprint arXiv:1909.01326 (2019)
29. Sicilia, A., Alikhani, M.: Learning to generate equitable text in dialogue from biased training data. arXiv preprint arXiv:2307.04303 (2023)
30. Sotolar, O., Formanek, V., Debnath, A., Lahnala, A., Welch, C., FLek, L.: Empo: Emotion grounding for empathetic response generation through preference optimization (2024), `https://arxiv.org/abs/2406.19071`
31. Spaulding, S.: Cognitive empathy. In: The Routledge handbook of philosophy of empathy, pp. 13–21. Routledge (2017)

32. Stade, E., Stirman, S.W., Ungar, L.H., Schwartz, H.A., Yaden, D.B., Sedoc, J., DeRubeis, R., Willer, R., et al.: Artificial intelligence will change the future of psychotherapy: A proposal for responsible, psychologist-led development (2023)

33. Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A.M., Wolf, T.: Zephyr: Direct distillation of lm alignment (2023)

34. Tunstall, L., Schmid, P.: Zephyr 7b gemma. `https://hf.co/HuggingFaceH4/zephyr-7b-gemma-v0.1` (2024)

35. Wan, Y., Wang, W., He, P., Gu, J., Bai, H., Lyu, M.R.: Biasasker: Measuring the bias in conversational ai system. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 515–527 (2023)

36. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? arXiv [cs.AI] (Jan 2018)

37. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: A survey. arXiv (Sep 2023)

38. Zhu, Q., Chong, L., Yang, M., Luo, J.: Reading users' minds from what they say: An investigation into llm-based empathic mental inference. arXiv preprint arXiv:2403.13301 (2024)