Predict the Next Word: <*Humans exhibit uncertainty in this task and language models _____*>

Anonymous ACL submission

Abstract

Language models (LMs) are statistical models trained to assign probability to human-generated text. As such, it is reasonable to question whether they approximate linguistic variability exhibited by humans well. This form of statistical assessment is difficult to perform at the passage level, for it requires acceptability judgments (*i.e.*, human evaluation) or a robust automated proxy (which is non-trivial). At the word level, however, given some context, samples from an LM can be assessed via exact matching against a prerecorded dataset of alternative single-word continuations of the available context. We exploit this fact and evaluate the LM's ability to reproduce variability that humans (in particular, a population of English speakers) exhibit in the 'next word prediction' task. This can be seen as assessing a form of calibration, which, in the context of text classification, Baan et al. (2022) termed *calibration to human uncertainty*. We assess GPT2, BLOOM and ChatGPT and find that they exhibit fairly low calibration to human uncertainty. We also verify the failure of expected calibration error (ECE) to reflect this, and as such, advise the community against relying on it in this setting.

1 Introduction

Language models (LMs) are trained to assign probability to human-generated text. The typical LM treats a piece of text as a sequence of tokens whose joint probability it factorises autoregressively, with conditional token probabilities predicted from the available context by a neural network (Mikolov et al., 2010; Radford et al., 2019; Scao et al., 2022). An LM can be viewed as a representation of uncertainty about human linguistic production (Serrano et al., 2009; Takahashi and Tanaka-Ishii, 2019; Meister and Cotterell, 2021; Giulianelli et al., 2023), specifically, one that reflects the production variability exhibited by the population(s) who generated the training data. Despite how plausible this

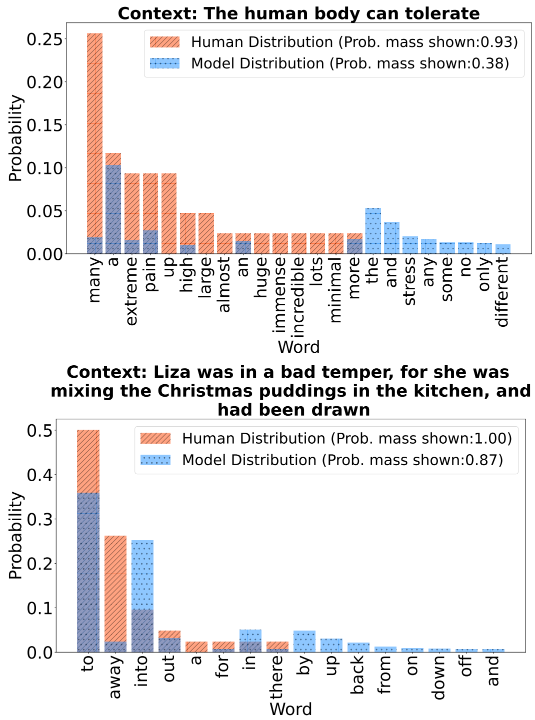


Figure 1: Estimated human and model distributions for contexts (15 most probable words of each distribution).

variability is, LMs are not consistently exposed to it at the level of individual contexts (*i.e.*, due to data sparsity, most contexts are unique) leading us to investigate their ability to predict it well.

One way to appreciate plausible variability is to ask humans to perform *next word prediction*: show multiple participants the same prefix of a passage and ask each of them to contribute a word that plausibly extends it. An LM that assigns probability to any next-word candidate similar to the proportion of the human population contributing it as the next word serves as a good proxy to the production variability of that human population—a desideratum Baan et al. (2022) termed *calibration to human uncertainty*.[1] Studying different notions of cal-

[1]Such calibration might be assessed against any population

ibration of text classifiers, Baan et al. (2022) show that the very popular expected calibration error (ECE; Guo et al., 2017) is flawed in the presence of data uncertainty (*e.g.*, due to the task's inherent ambiguity (Plank, 2022)). As data uncertainty is hardly avoidable in language modelling, we must entertain the possibility that ECE is not a reliable tool to assess the predictive distributions of an LM, despite its widespread use (Kumar and Sarawagi, 2019; Wang et al., 2020; Tian et al., 2023).

To assess calibration to human uncertainty, we compare the uncertainty exhibited by LMs to the uncertainty exhibited by humans in the next word prediction task—for which we use Provo Corpus (Luke and Christianson, 2018), a dataset (in English) with multiple human responses per available context. We analyse three pretrained LMs of different sizes and training objectives (*i.e.*, GPT2 (Radford et al., 2019), BLOOM (Scao et al., 2022) and ChatGPT (OpenAI, 2022)) and find that they exhibit low calibration to human uncertainty. We verify ECE's unreliability in this setting and advise the community against relying on it as a meaningful notion of calibration of generative models.

2 Background

Given context, an autoregressive LM predicts a conditional probability distribution (cpd) over the model's vocabulary of known tokens (*i.e.,* subword units). Hence, at this level, an LM can be regarded as a probabilistic multi-class classifier. This motivates research (Müller et al., 2019; Kumar and Sarawagi, 2019; Wang et al., 2020) assessing the extent to which probabilities predicted by LMs are interpretable as 'rate of correctness', a property referred to as calibration (Niculescu-Mizil and Caruana, 2005; Naeini et al., 2015; Guo et al., 2017).

A multi-class classifier is said to be *confidence-calibrated* if its probabilities predict the classifier's accuracy, specifically, if $(100 \times q)\%$ of its predictions made with probability (close to) q are judged to be correct. The ECE estimator (Guo et al., 2017) is the average absolute difference between average confidence and frequency of correctness across confidence bins.[2] Baan et al. (2022) uncovered a logical flaw in measuring ECE under data uncertainty—settings in which human disagreement is a plausible property of the task and hence not to be dismissed as error (Aroyo et al., 2019; Plank, 2022).[3] They show this in theory and empirically, and propose to assess predicted probabilities against estimates of *target probabilities*. The idea is to exploit multiple judgments per input to obtain the maximum likelihood estimate (MLE) of the target cpd and compare that to the model cpd at the instance level.

3 Methodology

We compare the uncertainty that LMs and humans exhibit in next word prediction. For that, we must represent their uncertainty over a shared space.

Human distributions. Given some context c, we assume that human uncertainty is captured by a single underlying cpd and, hence, regard human responses to the next word prediction task as i.i.d. draws from it. Then, given multiple responses, the MLE for this cpd assigns probability $p(w|c)$ to word w given c equal to the relative frequency with which humans predict w to follow c.

Model distributions. LMs decompose sentences as sequences of subword units, rather than words. However, humans predict complete words, hence, we establish a process for re-expressing the model cpds over the space of complete words.[4] For a given context c, we sample complete words from the model and use an empirical estimate of their probabilities; a word w drawn given c is assigned probability $q(w|c)$ equal to its relative frequency in the sample. To generate complete words, we (i) generate a token sequence generally long enough to include a word boundary; (ii) merge subword units and slice the first complete word from each generation (using a basic tokeniser); and, finally, (iii) reject samples that failed to generate a full word. This procedure samples potentially different segmentations of the same word(s) approximately marginalising out tokenisation ambiguity—which Cao and Rimell (2021) show to be an important and unduly neglected aspect of LM evaluation.

of interest, *e.g.* a specific target audience in a human-machine interaction setting (*e.g.* Williams and Reiter (2008)).

[2]Correctness is determined by comparing the mode of the predicted cpd to the target label (as pre-recorded in a dataset); the mode's probability is regarded as the classifier's confidence; closeness to q is determined via a binning scheme.

[3]There are many variants of ECE in the literature (Kumar et al., 2018; Widmann et al., 2019; Gupta et al., 2021; Si et al., 2022; Dawkins and Nejadgholi, 2022). Some variants, in particular, evaluate all probabilities of a cpd (not only the mode probability; *e.g.*, class-wise (Vaicenavicius et al., 2019; Kull et al., 2019), static and adaptive (Nixon et al., 2019)), these still assume no aleatoric uncertainty in the data generating process and, hence, remain inadequate tools for our setting. Besides, they are not common in language generation literature.

[4]Though artificial, one could tokenise the human data and analyse cpds over subword units, we do that in Appendix D.

4 Experiments

Data. Provo Corpus (Luke and Christianson, 2018) contains 55 passages (50 words long on average) in English from various sources *e.g.* news, fiction, science. Each prefix sequence of all passages (2686 prefixes) is given as context to 40 humans, on average, who predict a one-word completion. We use this corpus to estimate target cpds.

Models. For each context, we estimate cpds for different models. First, GPT2 Small (Radford et al., 2019), for which we use 1000 samples per context. To investigate whether a potential mismatch of training and test domain has an effect on our analysis, we fine-tune GPT2 on a subset of the original passages from Provo; we call this setting GPT2_{FT} (the complete experimental setup is described in Appendix F). To test the effect of scale on calibration to human uncertainty, we also analyse BLOOM 176B (Scao et al., 2022). Due to its high computational costs, we opt for sampling 40 generations per context (we motivate this choice empirically in Appendix C). Due to limited access to the API, we use a random subset of 669 Provo contexts. We are also interested in the effect of reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ibarz et al., 2018), hence we analyse ChatGPT (OpenAI, 2022). As with BLOOM, we draw 40 samples per context and use a random subset of 500 Provo contexts. In one setting we prompt ChatGPT 40 independent times, in another setting ($\text{ChatGPT}_{\text{D}}$) we prompt it once to generate a list with 40 options (prompt and additional details in Appendix C).[5] For each context, we also have a 'control cpd' formed by splitting the human annotation in two disjoint parts from which we estimate two cpds, one regarded as target, one regarded as an oracle model; this allows us to form an expectation about realistic levels of calibration.

Metrics. For each context, we compare a pair of cpds (a model vs the target for that context) in terms of their total variation distance (TVD).[6] To study a whole dataset, we plot TVD's distribution across contexts; for a numerical summary, following Baan et al. (2022), we report *expected TVD* (average TVD for all contexts) as a measure of calibration to human uncertainty. Finally, we compute ECE by comparing the mode of each model cpd to the original corpus word and ECE variants that use as targets the human or oracle majority per context.

[5]We will share all generations with the community.

[6]$\text{TVD}_c(p, q) = \frac{1}{2}\sum_w |p(w|c) - q(w|c)|$, where the sum is over the union of model- and human-generated words.

Gold Label	ECE ↓						
	Human	Oracle_2	GPT2	GPT2_{F}	Bloom	ChatGPT	$\text{ChatGPT}_{\text{D}}$
Original	0.14	0.11	0.02	0.03	0.07	0.45	0.10
Human Maj.	0.60	0.57	0.21	0.22	0.09	0.37	0.08
Oracle_1 Maj.	0.30	0.32	0.19	0.19	0.07	0.37	0.08
Avg TVD ↓	-	0.42	0.64	0.66	0.61	0.76	0.82

Table 1: ECE (the row indicates the target, the column indicates the system) and Expected TVD results. We resample the disjoint oracles 20 times and report the mean ECE (standard deviations < 0.1).

5 Results

Table 1 presents ECE and Expected TVD results. As predicted, ECE ranks most models as better calibrated than human oracles, confirming that it cannot be trusted in this setting. Figure 2 illustrates kernel density estimate (KDE) plots of instance-level TVD values between our models' cpds and the target (human) cpds, along with the KDE plot of TVD values between two disjoint oracles. We observe how the distributions of all models are skewed towards higher TVD values, with ChatGPT performing the worst. The inability of models to reproduce variability is not due to population mismatches (as GPT2_{FT} displays similar trends to GPT2) and persists in larger models, while RLHF worsens the issue (for both sampling strategies).

We measure a difference of around 0.2 TVD units between GPT2's and oracles' means, but, we lack understanding of its practical significance. That is, we do not how much worse than an oracle cpd a system that scores 0.2 TVD units more really is. To gain some insight, we conduct a controlled experiment. We artificially improve k% of the model's cpds by replacing them by an oracle estimate. We then measure TVD between this artificial improvement and a disjoint oracle allowing us to associate units of TVD with an interpretable rate of improvement (*i.e.*, percentage of plausible cpds). We find that we need to replace about 60% of GPT2's cpds to achieve TVDs that distribute similarly to human performance.[7]

For further insight, we analyse GPT2's inability to reliably reproduce human variability. We perform Bayesian regression with automatic relevance determination (ARD; Neal, 2012) using,

[7]In Appendix E, we verify that our findings a robust to choices of k, random seed and sample size.

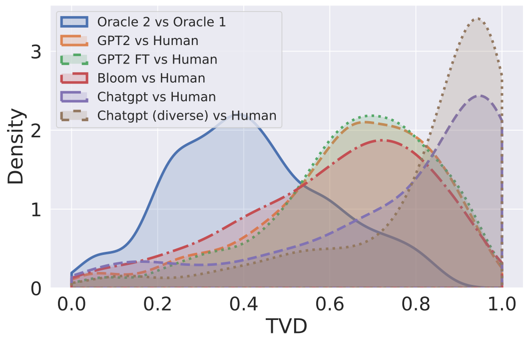


Figure 2: KDE plot of TVD values between a model and the estimated human target cpd, and between oracles.

for each context, TVD between GPT2 and the oracle cpd as the regression target, and predictors that are indicative of how constraining a context is (TVD between oracles, entropy of target cpd), as well as context length and the entropy of the model cpd. The former two are high precisely for contexts that admit more plausible variability. ARD ranked length as least important and TVD between oracles as most important, confirming that GPT2 struggles precisely in those cases of higher plausible variability (details in Appendix B). In Figure 1, we visualise target cpds and GPT2's (for the top-15 highest probability words) for two contexts; Appendix G lists a full passage. We choose the distributions of Figure 1 to demonstrate some observations; (1) GPT2's cpd fails to align with the human one, in samples where the outcome is barely constrained (true for the majority of the many instances we examined), and (2) when the outcome is fairly constrained, such as when completing a prepositional verb, GPT2 performs much better.

6 Related Work

There has been work that exploits predictive distributions of LMs in various ways. LeBrun et al. (2022) analyses such distributions and finds that they overestimate the probability of ill-formed sequences. Others investigate alternative training signals that minimise the distance between the data and model distributions (Ji et al., 2023; Labeau and Cohen, 2019; Zhang et al., 2023). Our work exploits predictive distributions as an uncertainty representation of human linguistic production and study their calibration. Several works study how well-calibrated LMs are and how to alleviate miscalibration (He et al., 2023; Lee et al., 2022; Xiao et al., 2022; Ahuja et al., 2022; Chen et al., 2022; Kumar and Sarawagi, 2019; Li et al., 2022; Xiao and Wang, 2021) — the majority using ECE to substantiate their findings, whose inadequacy makes us believe that a new round of studies is needed to assess this matter; our work being an example.

There is a line of work that stresses the value of obtaining multiple human labels per input (Plank, 2022; Basile et al., 2020; Grossmann et al., 2022; Prabhakaran et al., 2021), embracing data uncertainty in classification; Baan et al. (2022) propose calibration metrics that accommodate label variability in natural language inference (NLI; Bowman et al., 2015). In concurrent work, Lee et al. (2023) measure the calibration of LM-based classifiers to human uncertainty on ChaosNLI (Nie et al., 2020), also using Baan et al.'s expected TVD.

Other work further investigates uncertainty in an NLG setting. Zhou et al. (2023) and Kadavath et al. (2022) prompt LMs to output uncertainty linguistically. Kuhn et al. (2023a) prompt LMs to ask for clarifying questions when faced with ambiguous inputs. Similarly, Cole et al. (2023) sample repeatedly from LMs to assess whether they are able to answer ambiguous questions. Giulianelli et al. (2023) analyse various NLG tasks, their variability, and the ability of LMs to capture it. Additionally, Kuhn et al. (2023b) introduce semantic entropy, which incorporates linguistic invariances such as meaning equivalence, while Santurkar et al. (2023) prompt LMs to assess whether they represent the political views of US Americans from different demographics. Finally, Eisape et al. (2020) analyse the miscalibration of LMs from a psycho-linguistic lens, and fine-tune an LSTM model using multiple labels. Our work is an addition to this line of work.

7 Conclusion

Our work joins a stream of work acknowledging and better incorporating data uncertainty into evaluation protocols (Baan et al., 2022; Giulianelli et al., 2023). In particular, we find empirical evidence for ECE's unreliability in this setting and advise further research into calibration of LMs not to use it. With a more appropriate tool, we analyse three modern pretrained LMs and find that they are not well calibrated to human uncertainty, unlike ECE might suggest. We believe that this inability stems from models not being consistently subjected to human production variability during training, and plan to investigate this further in future work.

Limitations

The assessment of calibration to human uncertainty we have conducted is only one aspect of a system's quality and is not meant to de-emphasise the importance of any other sound form of evaluation, but rather to offer a complementary tool that supports an insightful set of observations about modern LMs. The computational costs of generating a large amount of continuations can be restrictive; as well as the cost of multiple annotations for each context. However, we believe that the benefits of obtaining such data and measuring uncertainty with more reliable methods, outweigh these costs. To foster research, we share the generations that supported this research. The high cost of obtaining data with multiple references per prompt results in another limitation: the limited availability of such labelled data. The limited number of human annotations per context is another limitation which is hard to alleviate. We considered all human annotations to be draws from the same underlying distribution, which is an assumption we cannot verify easily (*e.g.* we do not know if all participants had similar perspectives and backgrounds). Lastly, we only studied models trained for English. For less resourced languages, data-scarcity is expected to have worse effects on LMs' calibration. Simultaneously, English has a relatively fixed word order and simple morphology. Other languages might exhibit even greater variability due to their own typological features. In turn, we might be required to annotate larger datasets or study the phenomenon at a different level of granularity.

References

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. *arXiv e-prints*, pages arXiv–2210.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Kris Cao and Laura Rimell. 2021. You should evaluate your language model on marginal likelihood over tokenisations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2104–2114, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2022. A close look into the calibration of pre-trained language models. *arXiv e-prints*, pages arXiv–2211.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Hillary Dawkins and Isar Nejadgholi. 2022. Region-dependent temperature scaling for certainty calibration and application to class-imbalanced token classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 538–544, Dublin, Ireland. Association for Computational Linguistics.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Vasco Grossmann, Lars Schmarje, and Reinhard Koch. 2022. Beyond hard labels: investigating data label distributions. *arXiv preprint arXiv:2207.06224*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. 2021. Calibration of neural networks using splines. In *International Conference on Learning Representations*.

Guande He, Jianfei Chen, and Jun Zhu. 2023. Preserving pre-trained features helps calibrate fine-tuned language models. *arXiv preprint arXiv:2305.19249*.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Haozhe Ji, Pei Ke, Zhipeng Hu, Rongsheng Zhang, and Minlie Huang. 2023. Tailoring language generation models under total variation distance. In *The Eleventh International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023a. Clam: Selective clarification for ambiguous questions with generative language models.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023b. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, pages 12295–12305.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2805–2814. PMLR.

Matthieu Labeau and Shay B. Cohen. 2019. Experimenting with power divergences for language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4104–4114, Hong Kong, China. Association for Computational Linguistics.

Benjamin LeBrun, Alessandro Sordoni, and Timothy J O'Donnell. 2022. Evaluating distributional distortion in neural language modeling. *arXiv preprint arXiv:2203.12788*.

Dongkyu Lee, Ka Chun Cheung, and Nevin L Zhang. 2022. Adaptive label smoothing with self-knowledge in natural language generation. *arXiv preprint arXiv:2210.13459*.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models infer and disagree like humans? *arXiv preprint arXiv:2305.13788*.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022. Calibration meets explanation: A simple and effective approach for model confidence estimates. *arXiv preprint arXiv:2211.03041*.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Radford M Neal. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

OpenAI. 2022. Introducing chatgpt. *Available at* `https://openai.com/blog/chatgpt`.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

M Ángeles Serrano, Alessandro Flammini, and Filippo Menczer. 2009. Modeling statistical properties of written text. *PloS one*, 4(4):e5372.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2019. Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

David Widmann, Fredrik Lindsten, and Dave Zachariah. 2019. Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*, 32.

Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(4):495–525.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv e-prints*, pages arXiv–2210.

Shiyue Zhang, Shijie Wu, Ozan Irsoy, Steven Lu, Mohit Bansal, Mark Dredze, and David Rosenberg. 2023. Mixce: Training autoregressive language models by mixing forward and reverse cross-entropies. *arXiv preprint arXiv:2305.16958*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

Appendix

A Method 2 - Biased Model Estimate

We attempted constructing another estimator of the model distribution. Unlike the MC estimator in the main text, this estimator is biased due to it overestimating the probability of words in the distribution support and underestimating ones not belonging to it. This estimator forces the model to assign non-zero probabilities to humans responses; in an attempt to see if the model will, in this case, be able to predict human variability better.

We construct the support of the distribution as words that are 'likely' under the model. These include words generated with unbiased and nucleus sampling, the greedy word, as well as the original corpus word and human-answered words. For the words requiring sampling from the model, we follow a procedure similar to the unbiased estimator for ensuring sampled words are complete.

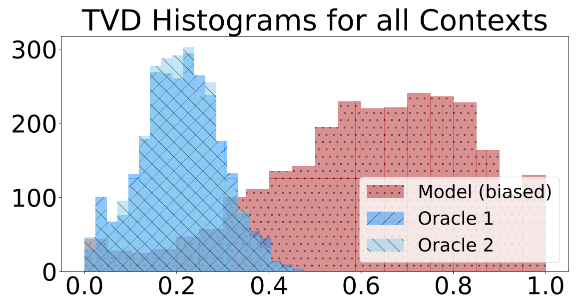


Figure 3: Histogram of TVD values for (biased) model and oracle distributions when compared to the full human distribution

The probability for each word is computed by renormalising the joint probabilities the model assigns for the corresponding token sequences:

$$\log q(w|c) = \log f(c,w) - \log f(c) - \text{logsumexp}_k[\log f(c,k) - \log f(c)]\,, \quad (1)$$

where $f(.)$ is the joint probability of the tokenised sequence, as assigned by the neural model.

We also evaluated the model's performance using such distributions. We use the same 1000 unbiased samples as before and an additional 100 nucleus samples for each of $p \in 0.7, 0.8, 0.9$. Results for ECE and TVD are shown in Table 2 and Figure 3 respectively. We observe similar results with the unbiased model in terms of both ECE and TVD.

Gold Label	ECE		
	Model	Oracle 1	Oracle 2
Corpus Word	0.068	0.116	0.185
Human Majority	0.138	0.563	0.458

Table 2: ECE results for the (Biased) Model and Oracle Distributions when considering the Gold-Label to be the corpus word or the human majority

B Predictors of TVD between model and oracle

We plot the target variable, TVD between the human and the model cpds against different predictors of interest (Figure 4 - 7). One particular predictor, the TVD between Oracles (Figure 4) is of interest, since it provides support for the claim made in Section 5; regarding GPT2's ability to predict variability well when the next word prediction task is less constrained. The results seem to support this theory - in the very low disagreement range between humans (TVD < 0.15), the model seems to predict variability well - or better, the lack of it. We also investigate context length as a predictor of the model's ability to predict human variability (Figure 5) - but surprisingly, we observe how the two seem to not be correlated. The plot with the human entropy and model entropy as the predictors, show a positive correlation (Figure 6 and 7 respectively). The results from the Bayesian regression with automatic feature determination are in Table 3, where each predictor and its coefficient are shown.

Predictor	Coefficient
Human Entropy	0.053
Model Entropy	0.095
TVD between Oracles	0.117
Context Length	0

Table 3: Bayesian Regression Predictors and Coefficients

C Larger Models

Due to the high computational inference costs of such large models, sampling 1000 ancestral generations for each context is infeasible. Hence, we opt for a lower number of samples - chosen on the basis of a subsampling experiment based on GPT-2. From the 1000 ancestral samples, we randomly selected subsamples of varying sizes (size = 10, 20, 40 and 100). For each of these, we re-computed the model distribution and computed the TVD values with an oracle. The Mean Squared Error between the TVD values of the subsampled distributions and the full-sampled distributions were computed and visualised through a histogram, as seen in Figure 8. We opted for a sample size of 40, since we considered it to be a good trade-off between computational costs and error.

C.1 ChatGPT prompting

Since ChatGPT is a conversational model - we prompt it to provide us with possible continuations to given contexts. We prompt it in two ways:

1. You are ChatGPT, a large language model trained by OpenAI. I want you to answer which word is a plausible continuation to the context <CONTEXT>. I have no specific intent, I just want your

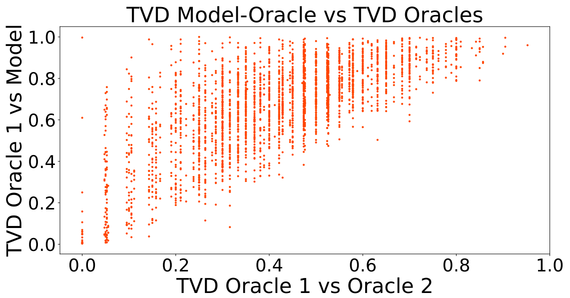


Figure 4: TVD values between oracles and TVD values between model and an oracle

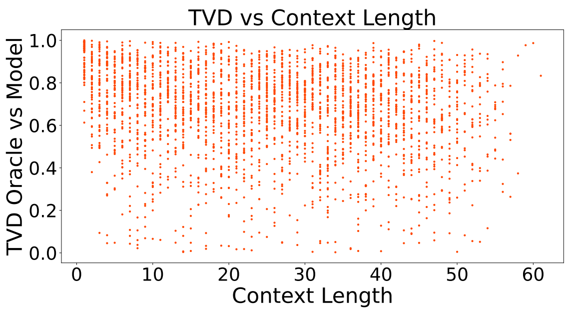


Figure 5: TVD values (between model and oracle) against Context Length

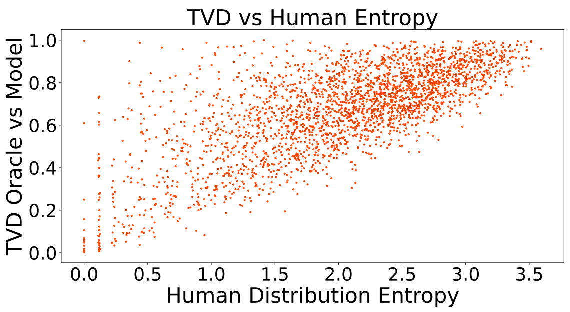


Figure 6: TVD values (between model and oracle) against Human Entropy

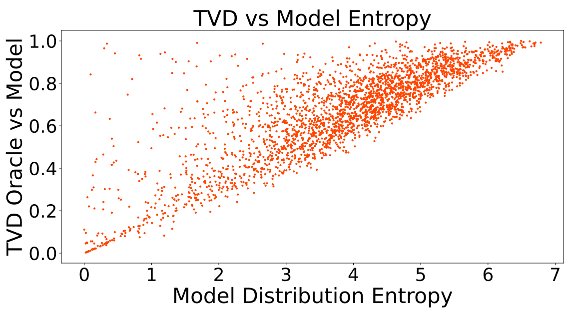


Figure 7: TVD values (between model and oracle) against Model Entropy

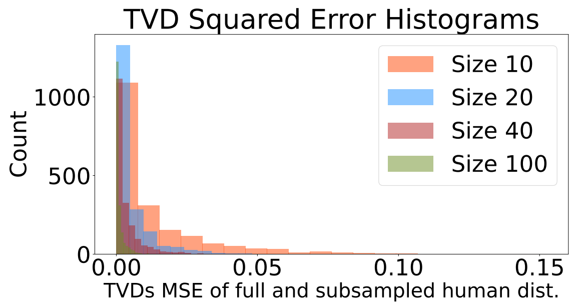


Figure 8: Histograms of MSE values between TVD values

guess. Return only the word and nothing else.

2. You are ChatGPT, a large language model trained by OpenAI. I want you to answer which 40 words are plausible continuations to the context <CONTEXT>. I have no specific intent, I just want your guess. Return only the words and nothing else.

For the former, we request 40 generations and for the latter only one (for both, temp = 1); both ways returning eventually of 40 continuations - which are ensured to be whole words. The first procedure imitates unbiased sampling more closely than the second - but due to the fact that minimal variability was observed, we implemented both methods.

For both BLOOM and ChatGPT generations we used the Hugging Face and OpenAI API subscriptions respectively, for two months. Regarding GPT2, we run generations using 1 NVIDIA A100 GPU, each passage needing approximately 2 hours to compute 1000 generations for all contexts in the passage.

C.2 TVD Differences

We additionally visualise the histograms of the difference in TVD values between the model and the human distribution minus the oracle and human distributions (Figure 9).

D Token-Level Experiment

One could claim that by estimating next-word distributions instead of next-token ones, we introduce some level of bias towards the model - since they are trained on BPE tokens rather than words. Despite finding this artificial, we repeat a subset of the experiments on a token level: instead of finding a method to sample sequences of tokens that form complete words from the model, we tokenize

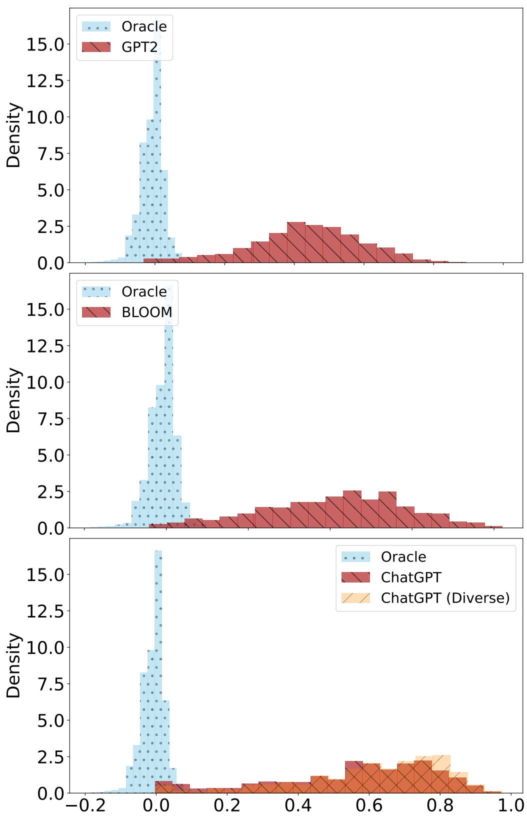


Figure 9: Histogram of TVD differences for model and oracle distributions when compared to the full human distribution. The vertical axis corresponds to density (normalizing counts so that the total histogram area equals 1).

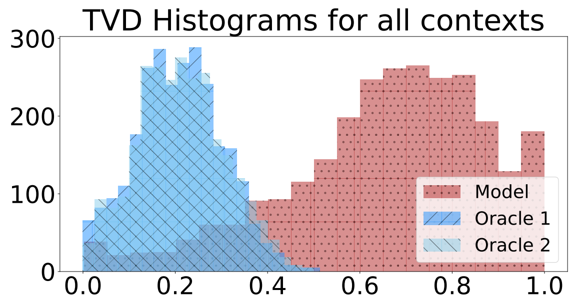


Figure 10: Histogram of TVD values for model and oracle distributions when compared to the full human distribution on a BPE-level analysis

human answers and create the target distribution of tokens. More specifically, we obtain from the model the distribution of next-tokens given a context. For the human distribution, we tokenize all human responses and take the first token of each one. We obtain the MLE of the human next-token distribution (and oracles) in a similar fashion to Section 3. Then, we perform a similar analysis for ECE and TVD values. Results are similar to the word-level analysis (Table 4 and Figure 10). We refrain from using token level analysis for calibration because it's not clear how to compare LMs with different tokenizers, whose vocabulary sizes differ.

Gold Label	ECE		
	Model	Oracle 1	Oracle 2
Human Majority	0.141	0.500	0.396

Table 4: ECE results for the Biased Model and Oracle Distributions

E Improving Model Experiments

We repeat the experiment where we artificially improve GPT2's performance (Section 5). This time, we create two types of disjoint oracles (by sampling from the human cpd without replacement) varying in size - a pair of size 20 and a pair of size 10. For each size, we sample 10 different pairs (using different seeds). For each pair, we compute the TVD value between them and the TVD value between an oracle and the model. As before, we randomly choose k% of model-oracle TVD instances to be replaced by the respective oracle-oracle instances. The aggregated results for the 10 seeds can be found in Figures 12 and 13 for the oracles of size 10 and 20 respectively. Results are very similar as before, showing that results are robust to

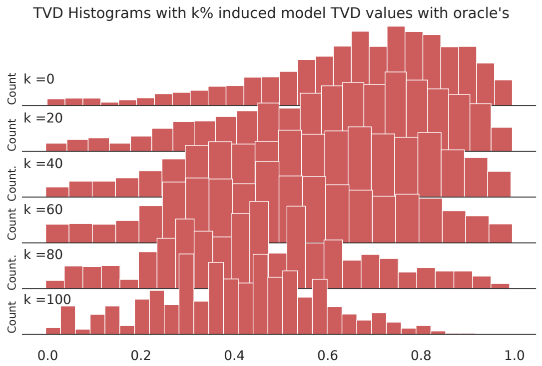


Figure 11: We artificially improve the Model-Oracle TVD histogram, by randomly replacing k% of the TVD values with the respective TVD values between oracles.

the oracle size and the sampled split itself.

F Out-Of-Distribution Effect Experiment

One could claim that we evaluate on a dataset, Provo Corpus, that does not necessarily originate from the distribution of the training dataset. To reinforce the validity of our results and establish that they are not just stemming from a domain mismatch of training and evaluation data, we complete experiments by fine-tuning on a subset of Provo Corpus. This way we, at least partly, remove the potential out-of-distribution effect - and re-evaluating calibration. Due to the Provo Corpus' limited size, the fine-tuning procedure has the following two aspects:

(1) A k-fold cross validation split (k=4), using the first 40 passages (Paragraphs 1-40) of Provo Corpus to create the 4 equal splits - each 10 passages long. We iteratively train on 3 of the splits and evaluate on the last 15 passages of Provo Corpus (Paragraphs 41-55). The paragraphs from the unused split are used for the evaluation of uncertainty. Overall, we end up with 4 different models, each used to create model distributions for 10 paragraphs - which, in turn, are used to measure TVD values for all their contexts.

(2) We do not fine-tune on the whole model - we freeze all parameters except those of the last two layers of GPT2-Small, since our training dataset is very small. We train using the cross-entropy loss, the AdamW optimizer (epsilon = 1e-8), for 10 epochs, with a 5e-4 learning rate, a batch size of 5, using 0 as the seed value.

The TVD results for the fine-tuned models', along with the respective perplexity curves during fine-tuning are in Figure 14 and 15 respectively.

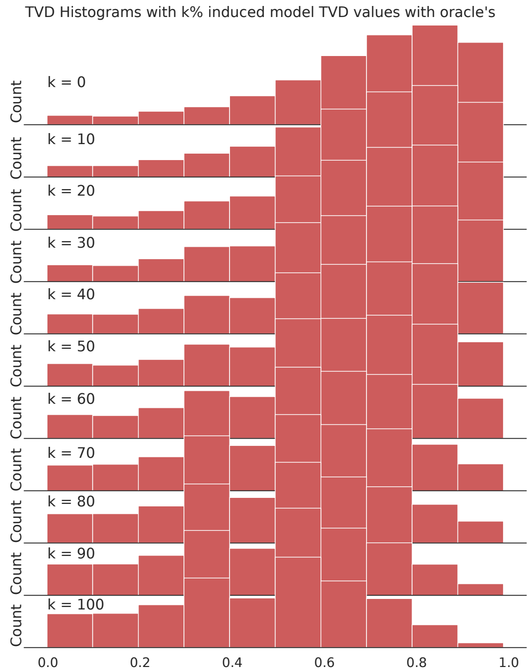


Figure 12: Improving the Model-Oracle TVD histogram, by randomly replacing k% of the TVD values with the respective TVD values between oracles, with an oracle size of 10, repeated on 10 seeds. k=0 corresponds to model performance and k = 100 to human performance.

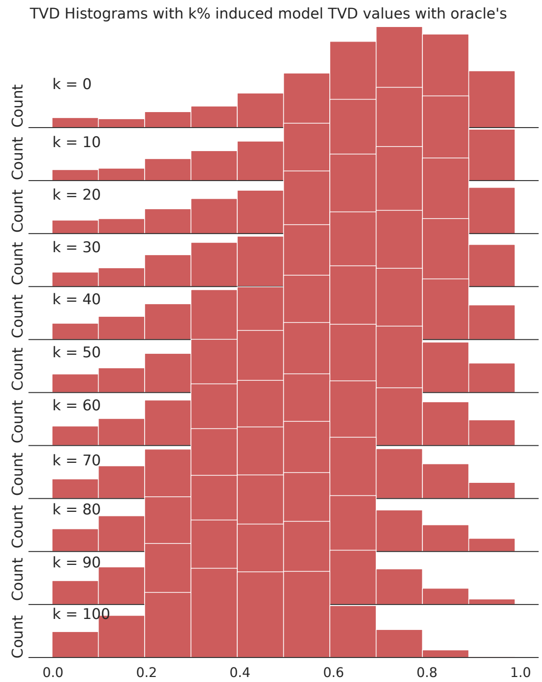


Figure 13: Improving the Model-Oracle TVD histogram, by randomly replacing k% of the TVD values with the respective TVD values between oracles, with an oracle size of 20, repeated on 10 seeds. k=0 corresponds to model performance and k = 100 to human performance.

G Visual Analysis of Distributions

We randomly choose one full passage (Paragraph 8) to illustrate further our conclusions. For all contexts, we provide the human and GPT2 distributions for the 15 most probable words of each cpd.

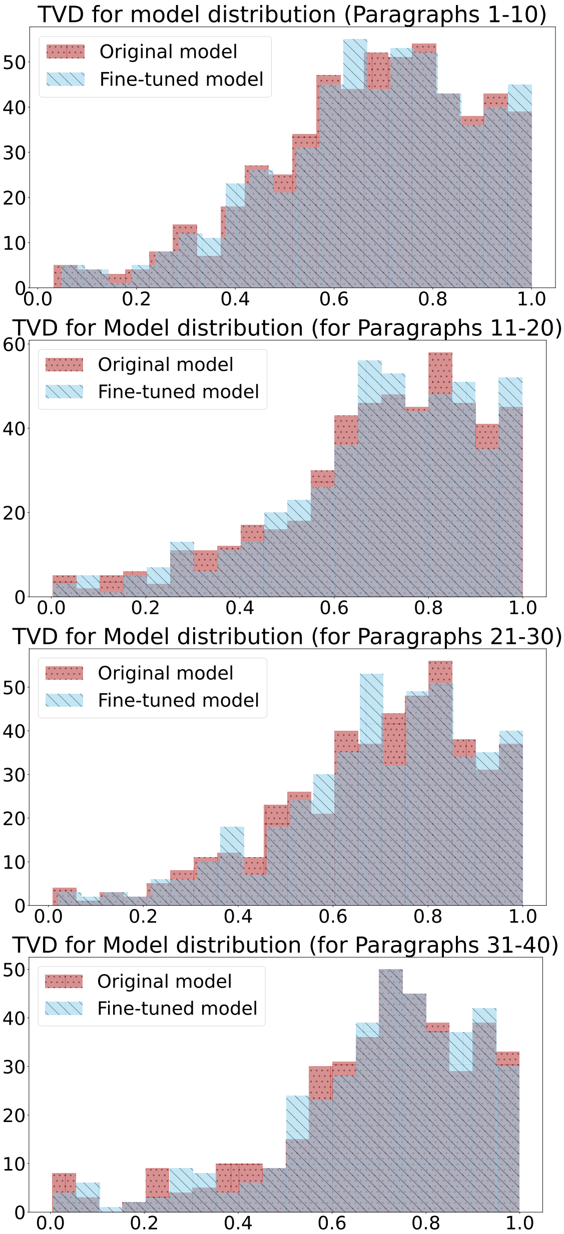


Figure 14: TVD histograms for all contexts between models (original and fine-tuned) and humans

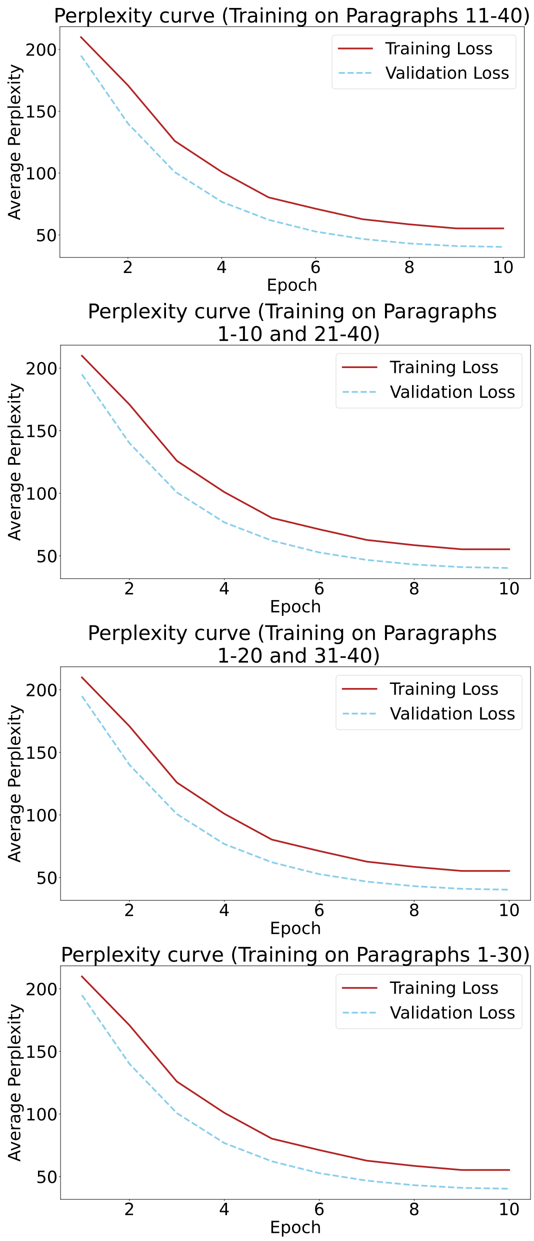
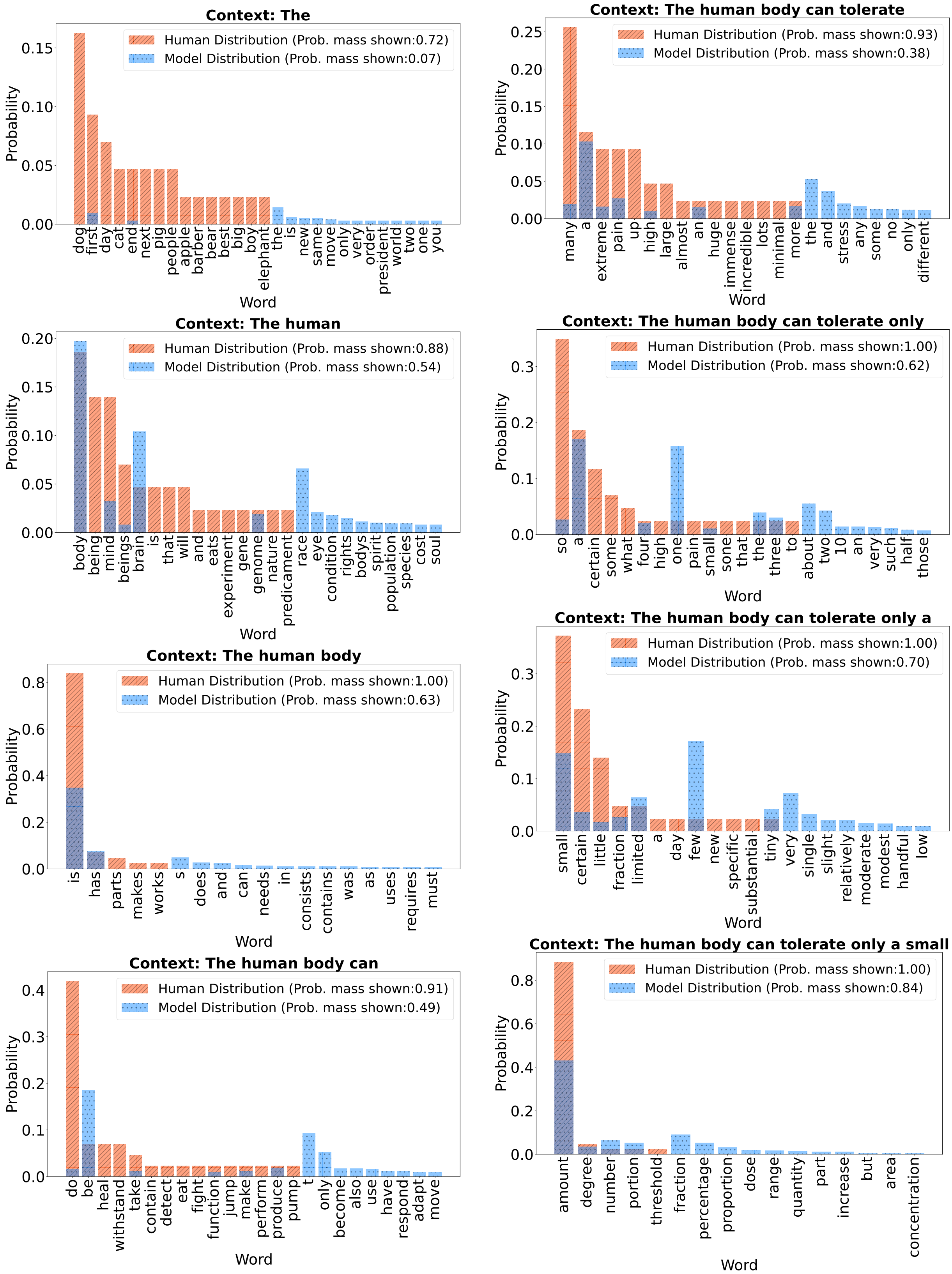
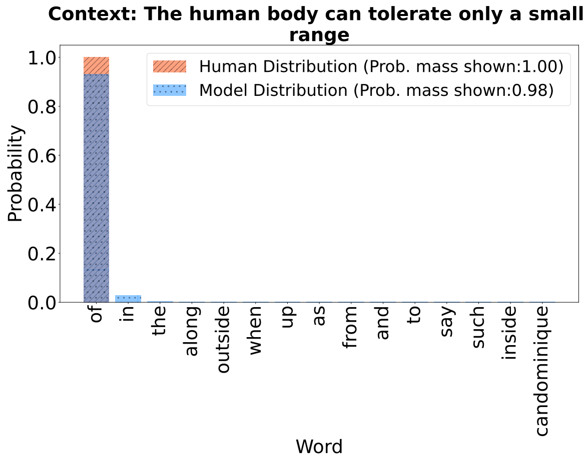
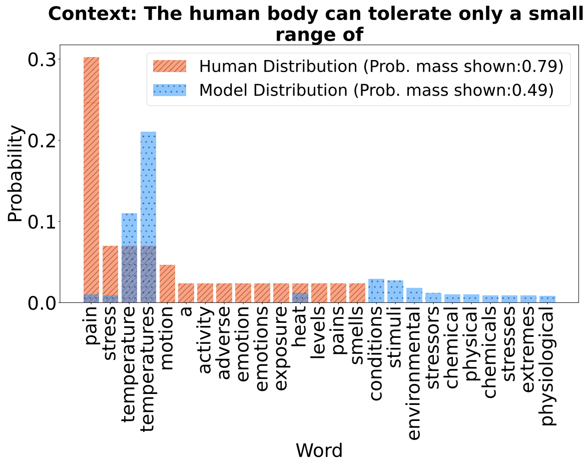
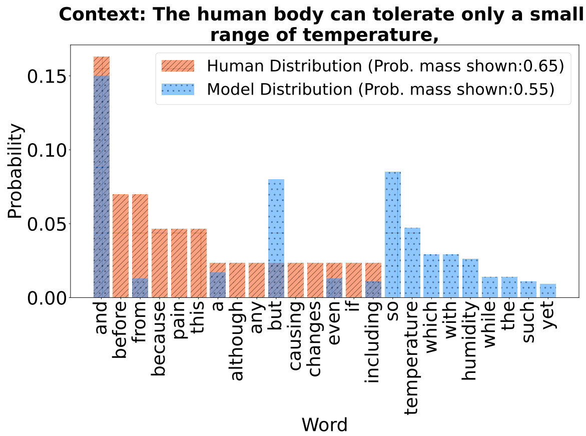


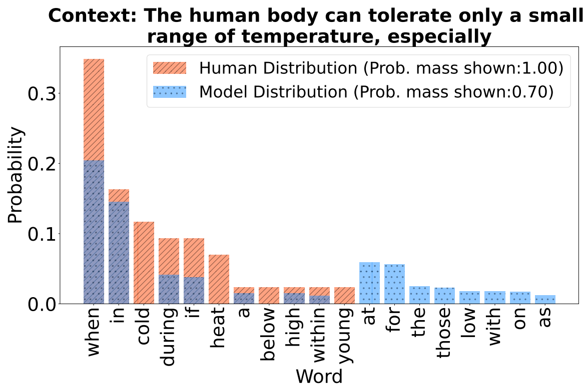
Figure 15: Training and Validation loss during the fine tuning of our model on a subset of Provo Corpus

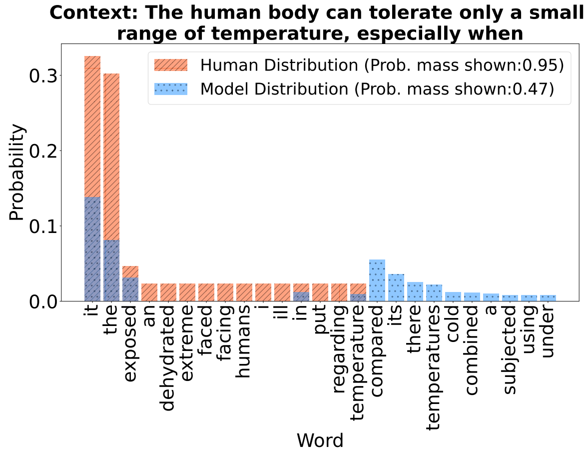

Context: The
Human Distribution (Prob. mass shown:0.72)
Model Distribution (Prob. mass shown:0.07)
Probability
Word
Context: The human body can tolerate
Human Distribution (Prob. mass shown:0.93)
Model Distribution (Prob. mass shown:0.38)
Context: The human
Human Distribution (Prob. mass shown:0.88)
Model Distribution (Prob. mass shown:0.54)
Context: The human body can tolerate only
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.62)
Context: The human body can tolerate only a
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.70)
Context: The human body
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.63)
Context: The human body can
Human Distribution (Prob. mass shown:0.91)
Model Distribution (Prob. mass shown:0.49)
Context: The human body can tolerate only a small
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.84)

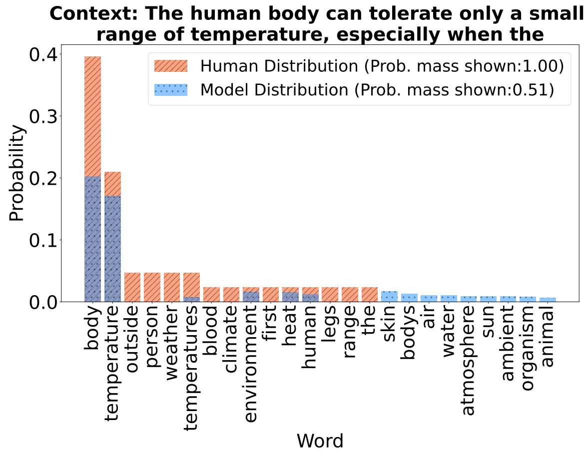

Context: The human body can tolerate only a small range
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.98)
Probability
1.0
0.8
0.6
0.4
0.2
0.0
of in the along outside when up as from and to say such inside candominique
Word

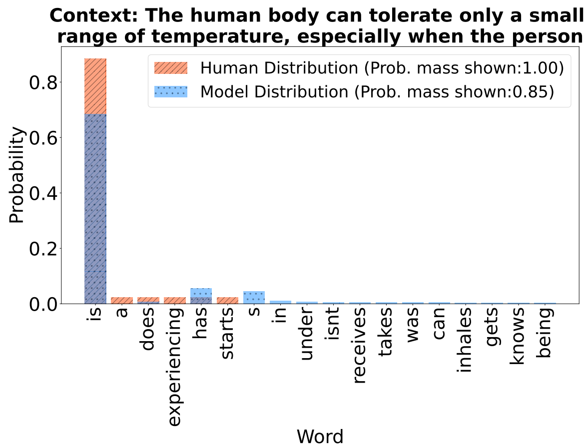

Context: The human body can tolerate only a small range of
Human Distribution (Prob. mass shown:0.79)
Model Distribution (Prob. mass shown:0.49)
Probability
0.3
0.2
0.1
0.0
pain stress temperature temperatures motion a activity adverse emotion emotions exposure heat levels pains smells conditions stimuli environmental stressors chemical physical chemicals stresses extremes physiological
Word

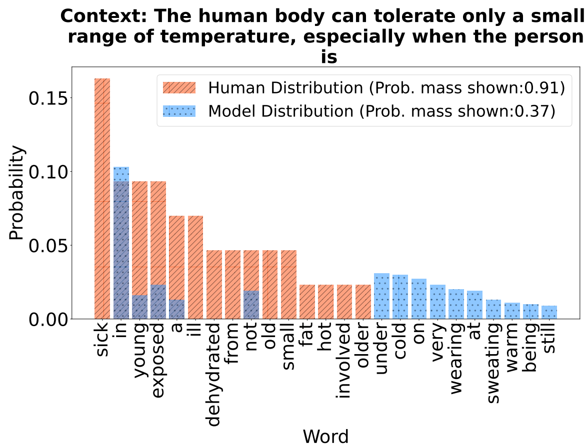

Context: The human body can tolerate only a small range of temperature,
Human Distribution (Prob. mass shown:0.65)
Model Distribution (Prob. mass shown:0.55)
Probability
0.15
0.10
0.05
0.00
and before from because pain this a although any but causing changes even if including so temperature which with humidity while the such yet
Word

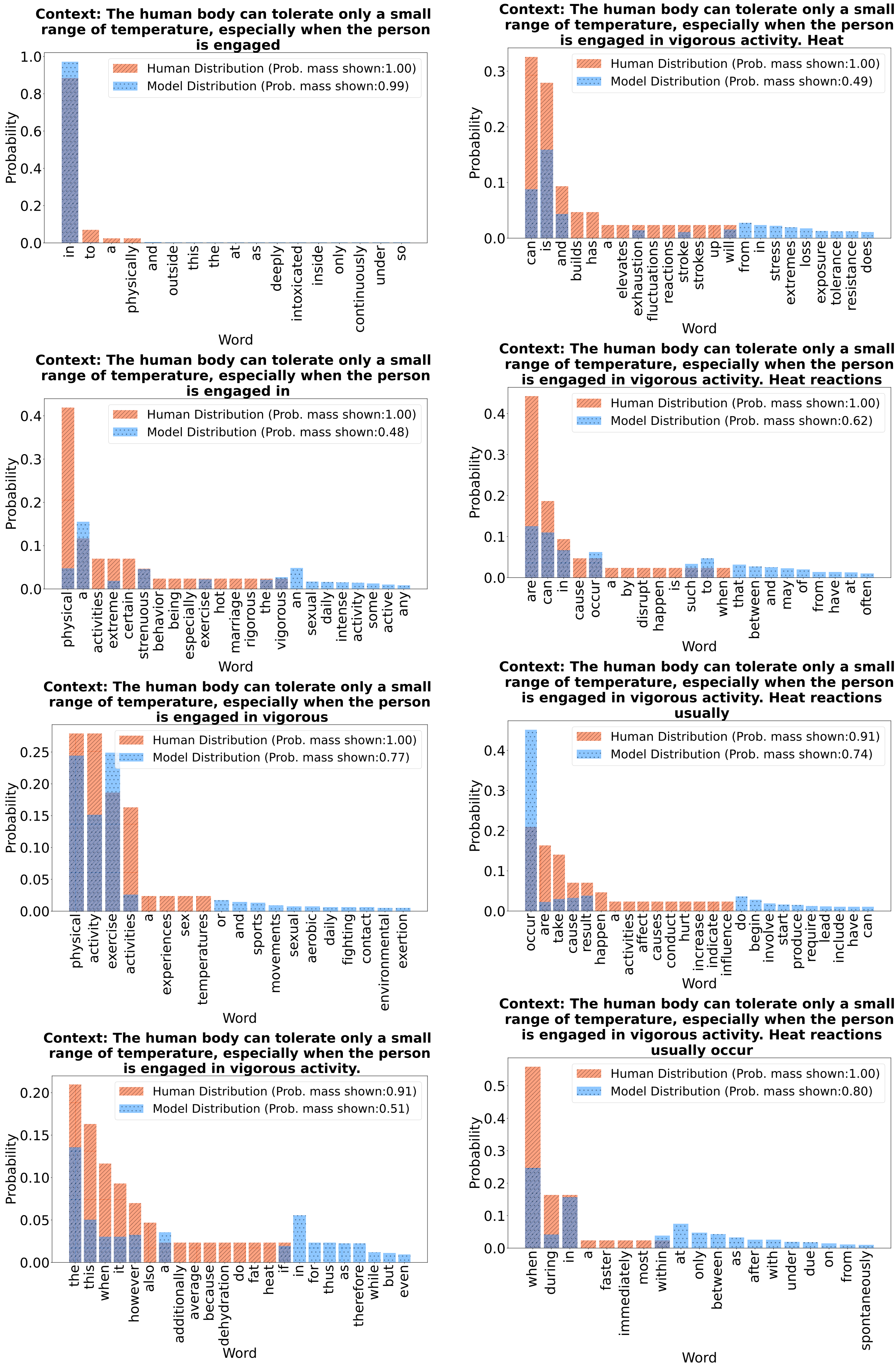

Context: The human body can tolerate only a small range of temperature, especially
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.70)
Probability
0.3
0.2
0.1
0.0
when in cold during if heat a below high within young at for the those low with on as
Word

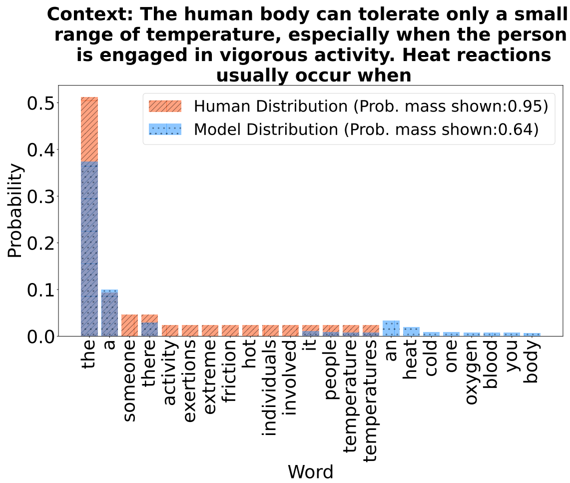

Context: The human body can tolerate only a small range of temperature, especially when
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.47)
Probability
0.3
0.2
0.1
0.0
it the exposed an dehydrated extreme faced facing humans i ill in put regarding temperature compared its there temperatures cold combined a subjected using under
Word

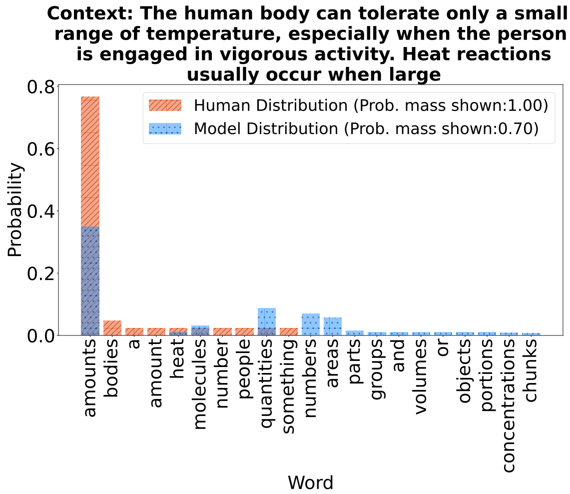

Context: The human body can tolerate only a small range of temperature, especially when the
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.51)
Probability
0.4
0.3
0.2
0.1
0.0
body temperature outside person weather temperatures blood climate environment first heat human legs range the skin bodys air water atmosphere sun ambient organism animal
Word

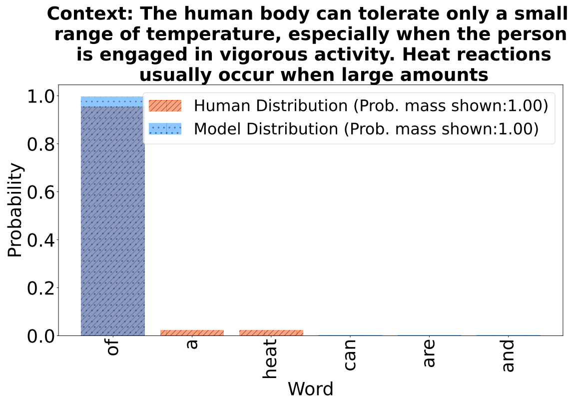

Context: The human body can tolerate only a small range of temperature, especially when the person
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.85)
Probability
0.8
0.6
0.4
0.2
0.0
is a does experiencing has starts s in under isnt receives takes was can inhales gets knows being
Word

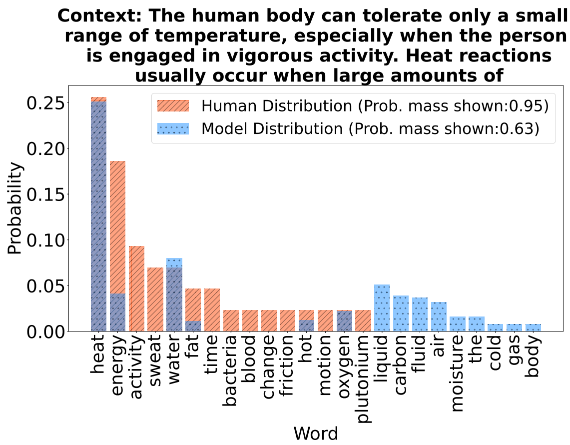

Context: The human body can tolerate only a small range of temperature, especially when the person is
Human Distribution (Prob. mass shown:0.91)
Model Distribution (Prob. mass shown:0.37)
Probability
0.15
0.10
0.05
0.00
sick in young exposed a ill dehydrated from not old small fat hot involved older under cold on very wearing at sweating warm being still
Word

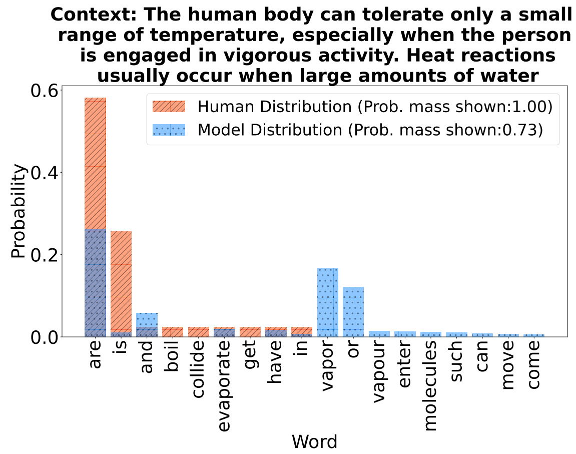

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.99)
Probability
Word
in to a physically and outside this the at as deeply intoxicated inside only continuously under so
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.48)
Probability
Word
physical a activities extreme certain strenuous behavior being especially exercise hot marriage rigorous the vigorous an sexual daily intense activity some active any
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.77)
Probability
Word
physical activity exercise activities a experiences sex temperatures or and sports movements sexual aerobic daily fighting contact environmental exertion
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity.
Human Distribution (Prob. mass shown:0.91)
Model Distribution (Prob. mass shown:0.51)
Probability
Word
the this when it however also a additionally average because dehydration do fat heat if in for thus as therefore while but even
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.49)
Probability
Word
can is and builds has a elevates exhaustion fluctuations reactions stroke strokes up will from in stress extremes loss exposure tolerance resistance does
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.62)
Probability
Word
are can in cause occur a by disrupt happen is such to when that between and may of from have at often
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually
Human Distribution (Prob. mass shown:0.91)
Model Distribution (Prob. mass shown:0.74)
Probability
Word
occur are take cause result happen a activities affect causes conduct hurt increase indicate influence do begin involve start produce require lead include have can
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.80)
Probability
Word
when during in a faster immediately most within at only between as after with under due on from spontaneously

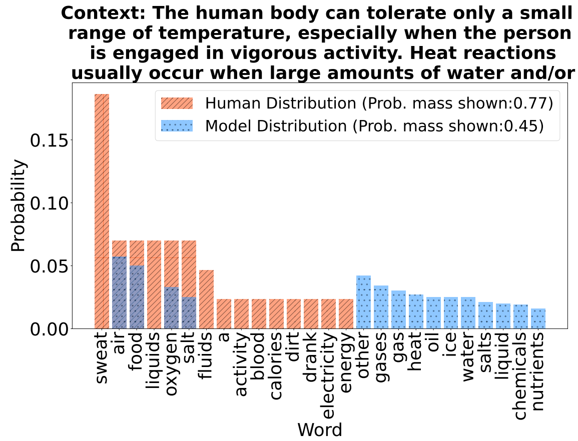

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.64)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
the
a
someone
there
activity
exertions
extreme
friction
hot
individuals
involved
it
people
temperature
temperatures
an
heat
cold
one
oxygen
blood
you
body
Word

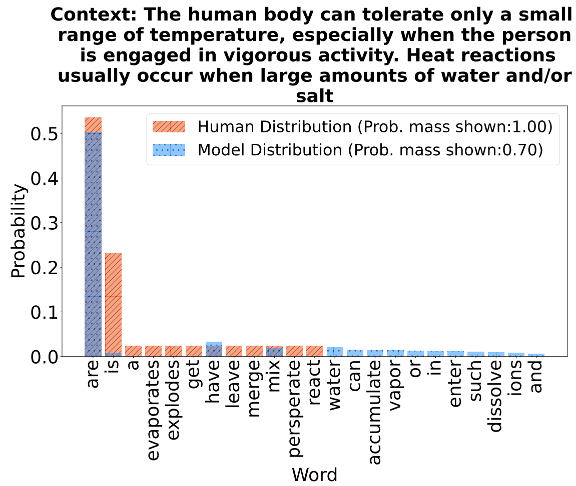

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.70)
Probability
0.0
0.2
0.4
0.6
0.8
amounts
bodies
a
amount
heat
molecules
number
people
quantities
something
numbers
areas
parts
groups
and
volumes
or
objects
portions
concentrations
chunks
Word

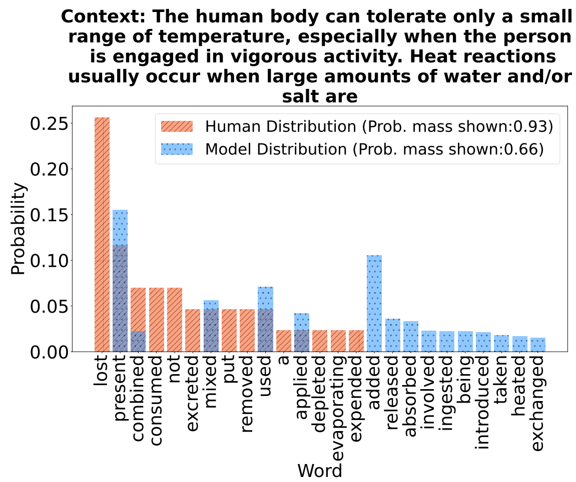

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:1.00)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
of
a
heat
can
are
and
Word

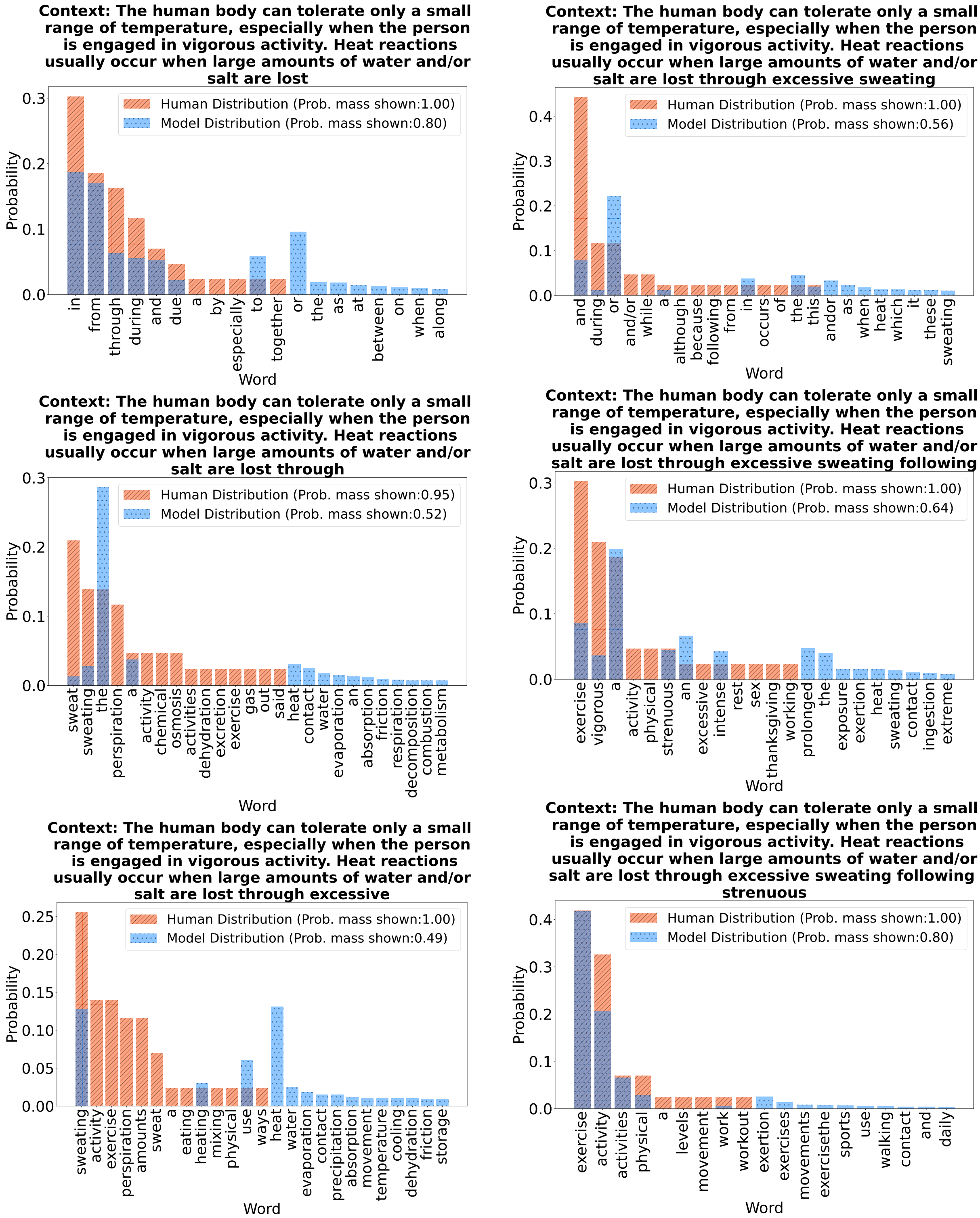

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.63)
Probability
0.00
0.05
0.10
0.15
0.20
0.25
heat
energy
activity
sweat
water
fat
time
bacteria
blood
change
friction
hot
motion
oxygen
plutonium
liquid
carbon
fluid
air
moisture
the
cold
gas
body
Word


Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.73)
Probability
0.0
0.2
0.4
0.6
are
is
and
boil
collide
evaporate
get
have
in
vapor
or
vapour
enter
molecules
such
can
move
come
Word

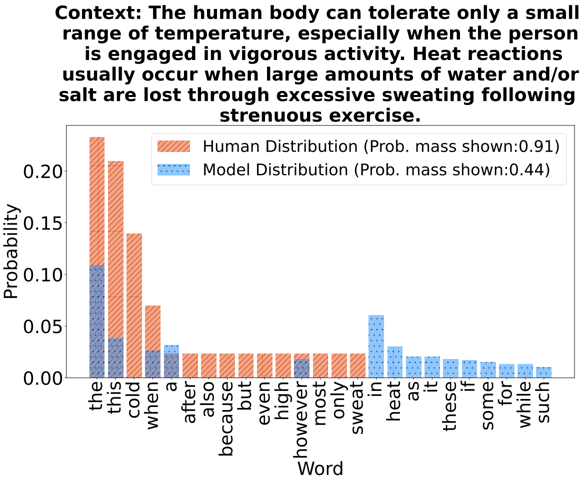

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or
Human Distribution (Prob. mass shown:0.77)
Model Distribution (Prob. mass shown:0.45)
Probability
0.00
0.05
0.10
0.15
sweat
air
food
liquids
oxygen
salt
fluids
a
activity
blood
calories
dirt
drank
electricity
energy
other
gases
gas
heat
oil
ice
water
salts
liquid
chemicals
nutrients
Word

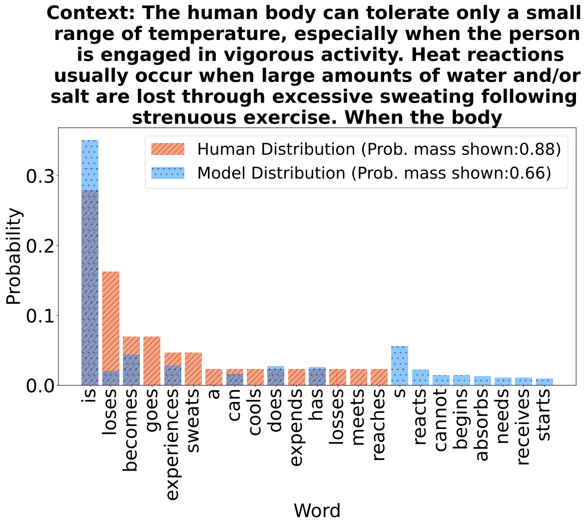

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.70)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
are
is
a
evaporates
explodes
get
have
leave
merge
mix
perspirate
react
water
can
accumulate
vapor
or
in
enter
such
dissolve
ions
and
Word

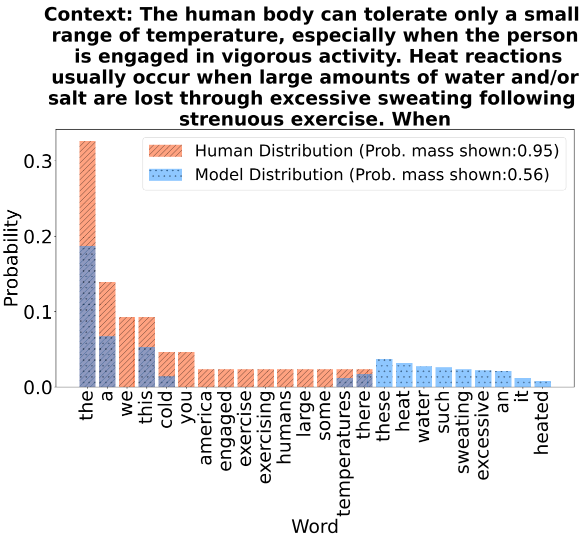

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are
Human Distribution (Prob. mass shown:0.93)
Model Distribution (Prob. mass shown:0.66)
Probability
0.00
0.05
0.10
0.15
0.20
0.25
lost
present
combined
consumed
not
excreted
mixed
put
removed
used
a
applied
depleted
evaporating
expended
added
released
absorbed
involved
ingested
being
introduced
taken
heated
exchanged
Word

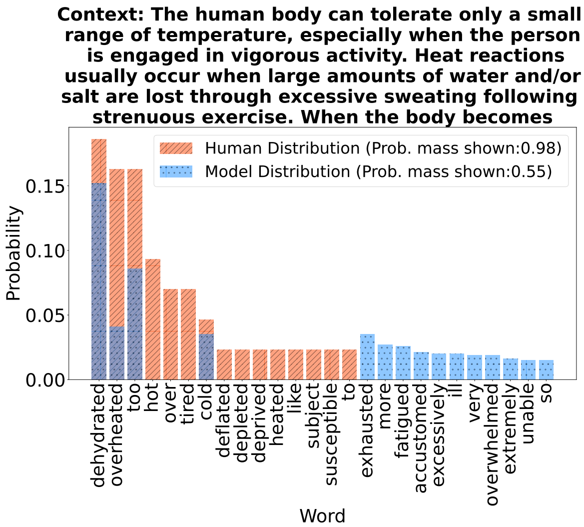


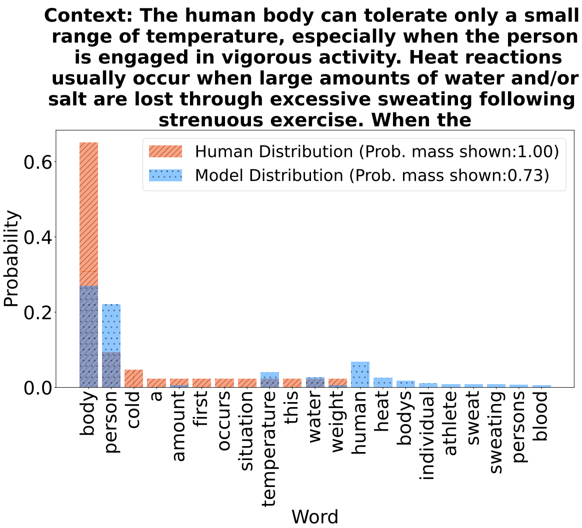
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.80)
Probability
Word
in from through during and due a by especially to together or the as at between on when along
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.56)
Probability
Word
and during or and/or while a although because following from in occurs of the this andor as when heat which it these sweating
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.52)
Probability
Word
sweat sweating the perspiration a activity chemical osmosis activities dehydration excretion exercise gas out said heat contact water evaporation an absorption friction respiration decomposition combustion metabolism
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.64)
Probability
Word
exercise vigorous a activity physical strenuous an excessive intense rest sex thanksgiving working prolonged the exposure exertion heat sweating contact ingestion extreme
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.49)
Probability
Word
sweating activity exercise perspiration amounts sweat a eating heating mixing physical use ways heat water evaporation contact precipitation absorption movement temperature cooling dehydration friction storage
Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.80)
Probability
Word
exercise activity activities physical a levels movement work workout exertion exercises movements exercisethe sports use walking contact and daily

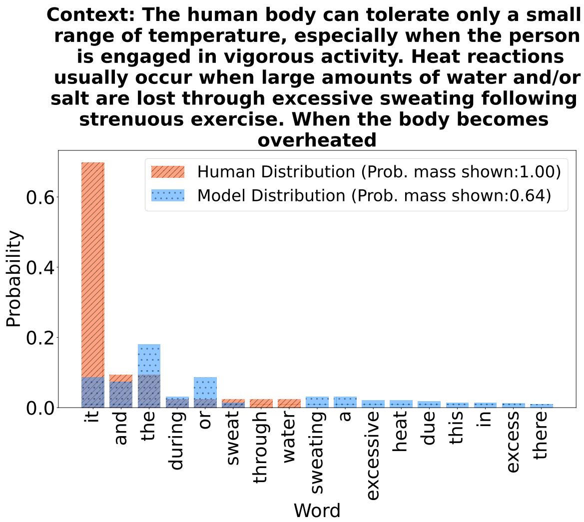

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise.
Human Distribution (Prob. mass shown:0.91)
Model Distribution (Prob. mass shown:0.44)
Probability
Word
the this cold when a after also because but even high however most only sweat in heat as it these if some for while such

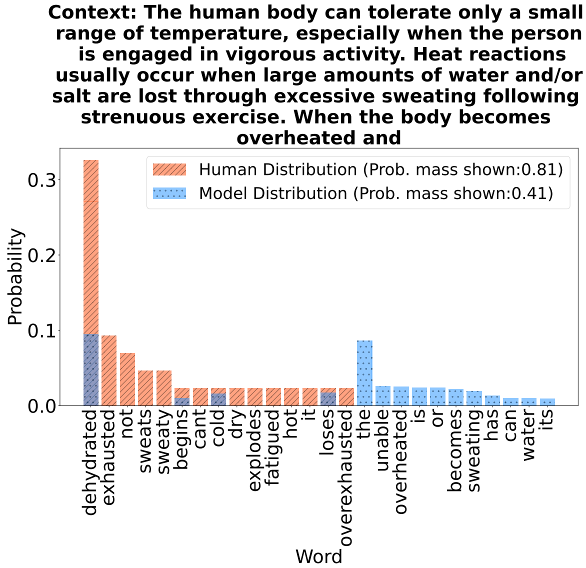

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body
Human Distribution (Prob. mass shown:0.88)
Model Distribution (Prob. mass shown:0.66)
Probability
Word
is loses becomes goes experiences sweats a can cools does expends has losses meets reaches s reacts cannot begins absorbs needs receives starts

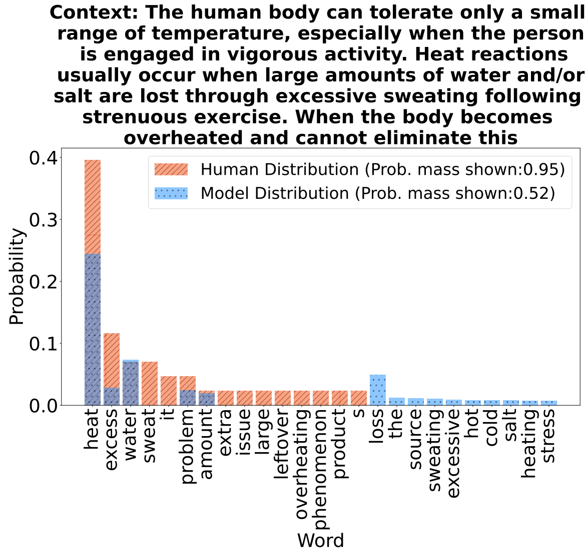

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.56)
Probability
Word
the a we this cold you america engaged exercise exercising humans large some temperatures there these heat water such sweating excessive an it heated

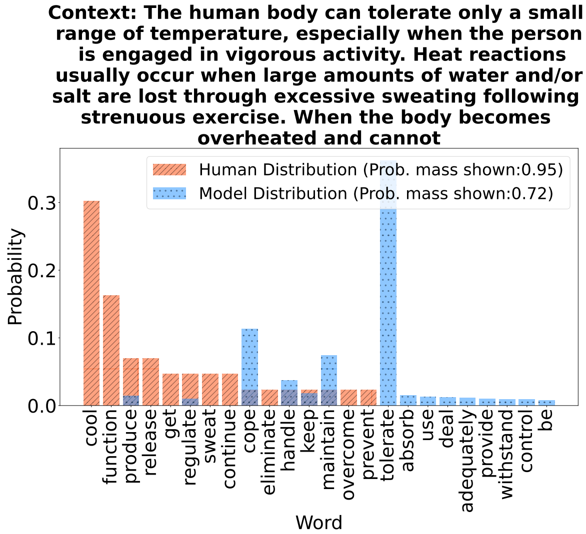

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes
Human Distribution (Prob. mass shown:0.98)
Model Distribution (Prob. mass shown:0.55)
Probability
Word
dehydrated overheated too hot over tired cold deflated depleted deprived heated like subject susceptible to exhausted more fatigued accustomed excessively ill very overwhelmed extremely unable so

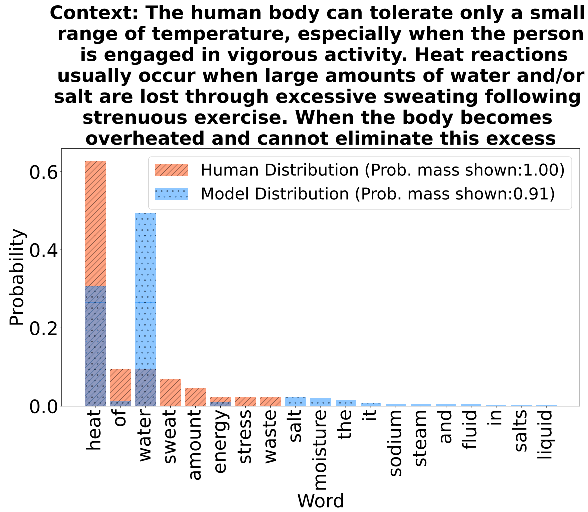

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.73)
Probability
Word
body person cold a amount first occurs situation temperature this water weight human heat bodys individual athlete sweat sweating persons blood

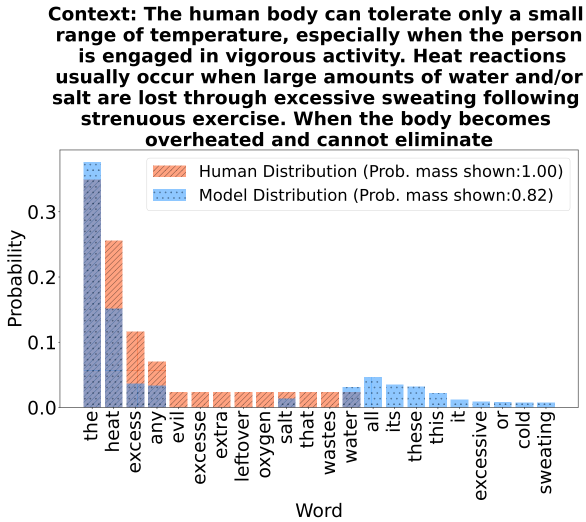

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.64)
Probability
Word
it and the during or sweat through water sweating a excessive heat due this in excess there

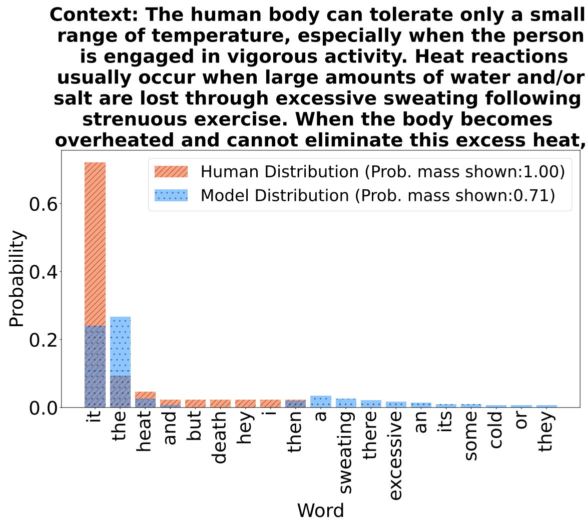

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and
Human Distribution (Prob. mass shown:0.81)
Model Distribution (Prob. mass shown:0.41)
Probability
0.0
0.1
0.2
0.3
Word
dehydrated exhausted not sweats sweaty begins cant cold dry explodes fatigued hot it loses overexhausted the unable overheated is or becomes sweating has can water its

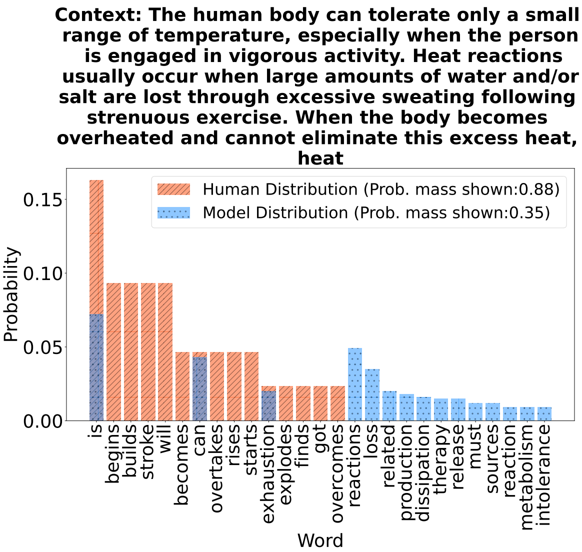

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.52)
Probability
0.0
0.1
0.2
0.3
0.4
Word
heat excess water sweat it problem amount extra issue large leftover overheating phenomenon product s loss the source sweating excessive hot cold salt heating stress

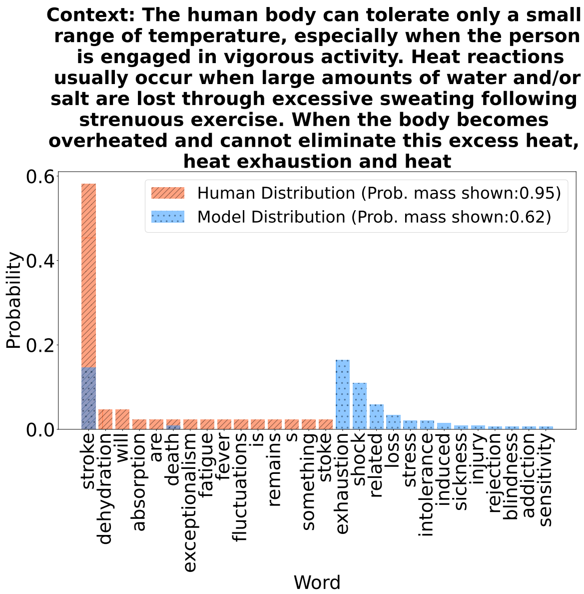

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.72)
Probability
0.0
0.1
0.2
0.3
Word
cool function produce release get regulate sweat continue cope eliminate handle keep maintain overcome prevent tolerate absorb use deal adequately provide withstand control be

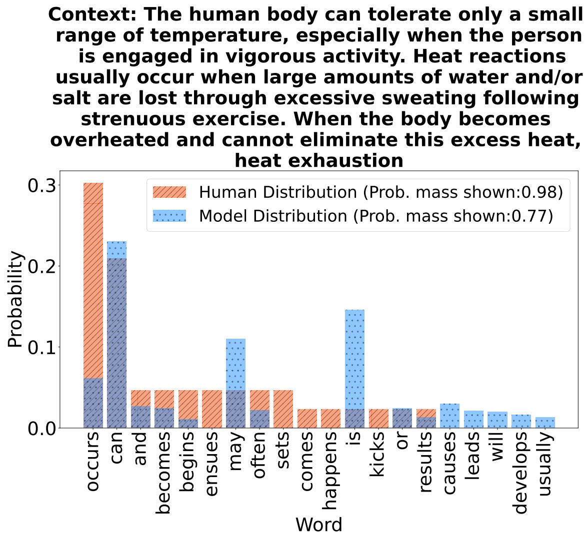

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.91)
Probability
0.0
0.2
0.4
0.6
Word
heat of water sweat amount energy stress waste salt moisture the it sodium steam and fluid in salts liquid

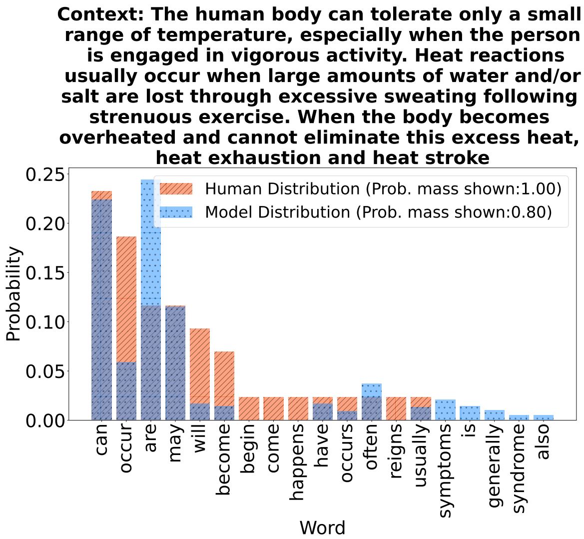

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.82)
Probability
0.0
0.1
0.2
0.3
Word
the heat excess any evil excesse extra leftover oxygen salt that wastes water all its these this it excessive or cold sweating

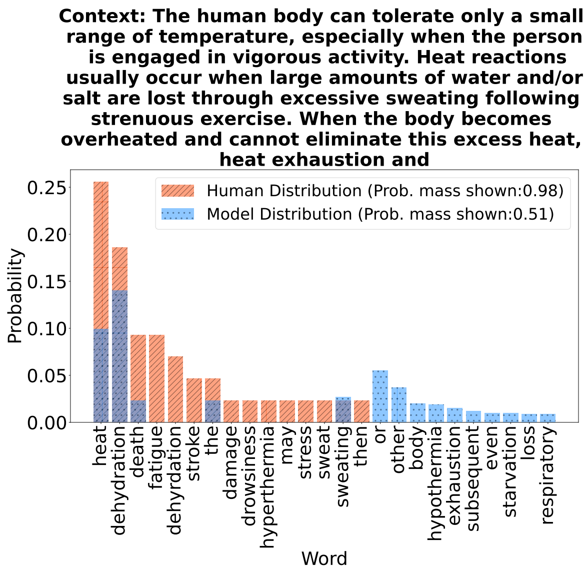

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat,
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.71)
Probability
0.0
0.2
0.4
0.6
Word
it the heat and but death hey i then a sweating there excessive an its some cold or they

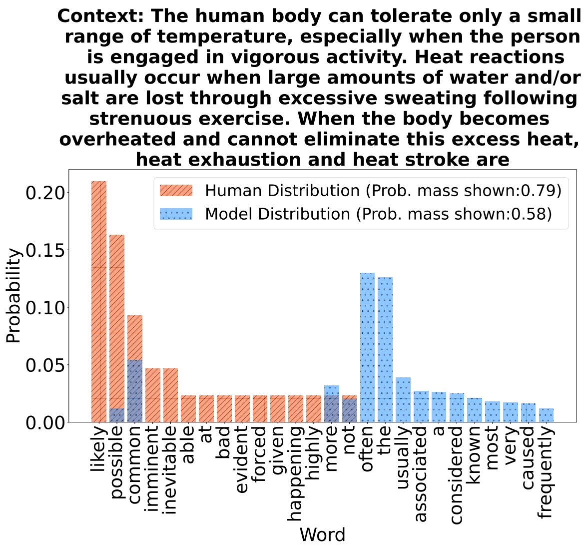

Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat
Human Distribution (Prob. mass shown:0.88)
Model Distribution (Prob. mass shown:0.35)
Probability
0.00
0.05
0.10
0.15
is begins builds stroke will becomes can overtakes rises starts exhaustion explodes finds got overcomes reactions loss related production dissipation therapy release must sources reaction metabolism intolerance
Word


Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat exhaustion and heat
Human Distribution (Prob. mass shown:0.95)
Model Distribution (Prob. mass shown:0.62)
Probability
0.0
0.2
0.4
0.6
stroke dehydration will absorption are death exceptionalism fatigue fever fluctuations is remains s something stoke exhaustion shock related loss stress intolerance induced sickness injury rejection blindness addiction sensitivity
Word


Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat exhaustion
Human Distribution (Prob. mass shown:0.98)
Model Distribution (Prob. mass shown:0.77)
Probability
0.0
0.1
0.2
0.3
occurs can and becomes begins ensues may often sets comes happens is kicks or results causes leads will develops usually
Word


Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat exhaustion and heat stroke
Human Distribution (Prob. mass shown:1.00)
Model Distribution (Prob. mass shown:0.80)
Probability
0.00
0.05
0.10
0.15
0.20
0.25
can occur are may will become begin come happens have occurs often reigns usually symptoms is generally syndrome also
Word


Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat exhaustion and
Human Distribution (Prob. mass shown:0.98)
Model Distribution (Prob. mass shown:0.51)
Probability
0.00
0.05
0.10
0.15
0.20
0.25
heat dehydration death fatigue dehyrdation stroke the damage drowsiness hyperthermia may stress sweat sweating then or other body hypothermia exhaustion subsequent even starvation loss respiratory
Word


Context: The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat exhaustion and heat stroke are
Human Distribution (Prob. mass shown:0.79)
Model Distribution (Prob. mass shown:0.58)
Probability
0.00
0.05
0.10
0.15
0.20
likely possible common imminent inevitable able at bad evident forced given happening highly more not often the usually associated a considered known most very caused frequently
Word