Operationalizing Specifications, In Addition to Test Sets for Evaluating Constrained Generative Models

Vikas Raunak **Matt Post** **Arul Menezes**
Microsoft Azure AI
{viraunak,mpost,arulm}@microsoft.com

Abstract

In this work, we present some recommendations on the evaluation of state-of-the-art generative models for constrained generation tasks. The progress on generative models has been rapid in recent years. These large-scale models have had three impacts: firstly, the fluency of generation in both language and vision modalities has rendered common *average-case* evaluation metrics much less useful in diagnosing system errors. Secondly, the same substrate models now form the basis of a number of applications, driven both by the utility of their representations as well as phenomena such as in-context learning, which raise the abstraction level of interacting with such models. Thirdly, the user expectations around these models and their feted public releases have made the technical challenge of out of domain generalization much less excusable in practice. Subsequently, our evaluation methodologies haven't adapted to these changes. More concretely, while the associated utility and methods of interacting with generative models have expanded, a similar expansion has not been observed in their evaluation practices. In this paper, we argue that the scale of generative models could be exploited to raise the abstraction level at which evaluation itself is conducted and provide recommendations for the same. Our recommendations are based on leveraging specifications as a powerful instrument to evaluate generation quality and are readily applicable to a variety of tasks.

1 Introduction

Recent advances in generative models across different modalities have enabled a myriad of applications, making the reliability failures of such models a frontier of research. Adjacent advances such as prompting [Brown et al., 2020], open-vocabulary classification [Radford et al., 2021], etc. have significantly enhanced the utility of such models, implicitly raising the abstraction level of human-model interactions. We believe that modern generative models have had three main impacts:

1. **Fluent Generations**: The fluency of generation in both language and vision modalities has rendered existing benchmarks and metrics much less useful in diagnosing system errors [Gehrmann et al., 2022, Srivastava et al., 2022], with previously widely used benchmarks and metrics on tasks such as machine translation [Raunak et al., 2022] or summarization [Goyal et al., 2022] becoming less useful in gauging system problems.
2. **Foundational Role**: The same substrate models now form the basis of a number of applications [Bommasani et al., 2021] and this is driven both by the utility of their representations as well as phenomena such as in-context learning [Brown et al., 2020], which raise the abstraction level of interacting with such models.
3. **Greater User Expectations**: The user expectations around these models have made the technical challenge of out of domain generalization much less excusable in practice.

Proceedings of The Workshop on Human Evaluation of Generative Models, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Subsequently, our evaluation methodologies haven't adapted to these changes and the evaluation of modern generative models remains a challenge, with many hitherto standard benchmarks and metrics becoming less and less useful with increasing model capabilities.

2 Specifications for Evaluating Generative Models

In this paper, we posit specifications as a powerful instrument for the evaluation of large-scale generative models. The evaluation analogy here is akin to the invention of fast transportation means, where platforms and guardrails had to be built for the society to safely and reliably use them. As these large-scale generative models enter more and more application domains, our primary concern here is about the observable behaviors of such models. And at this juncture, we think that the idea of specifications (as in engineering) is a more powerful way to think about the quality of these systems than the traditional evaluation practices in the machine learning community, primarily metrics and test cases. Besides empirical observations of generative models' failure modes, a strong basis for proposing a specifications-based evaluation framework stems from the fact that tails of the data distributions are notoriously hard to model and errors could arise as catastrophic failures on certain specific inputs despite the model's high average-case performance [Nair et al., 2022, Belinkov and Bisk, 2018]. By proposing a specifications-based evaluation framework, we also aim to naturally quantify model reliability as adherence to specifications.

To be precise, specifications are expressions of user intent on program behaviors [Gulwani et al., 2017]. In traditional machine learning evaluation, references are a common way to express such specifications. And fundamentally, metrics are measurements based on those references, which are designed to mimic (correlate highly with) human judgements. Thereby, using references is one concrete instantiation of leveraging specifications. But, we can actually jump up an abstraction level and use arbitrary specifications for evaluation. We think this is important for modern generative models, owing to the insufficiency of metrics for evaluating system quality as well as due to the limitations of a strictly test-case based evaluation in eliciting system errors comprehensively.

2.1 Insufficiency of Metrics

Even though the metrics we use are getting better [Zhang et al., 2020, Rei et al., 2020, Sellam et al., 2020], the assumptions behind metrics are problematic. These assumptions include: joint fluency and adequacy modeling through a single-dimension of system quality and no explicit sensitivity to salient errors at the instance-level. These assumptions behind system quality evaluation are not only detrimental to the downstream model user, but a lack of multidimensional view into model quality also makes the system developers unaware of different error types during development. We argue that this is much more problematic for large scale models where one can not iterate too fast.

2.2 Insufficiency of Test Cases

Methodologies such as CheckList [Ribeiro et al., 2020], which build test cases for behavioral testing of models don't generalize to generative tasks. There are multiple reasons as to why a curated test-case (ground-truth) based evaluation cannot scale for a comprehensive evaluation of large-scale generative models, namely: (a) there are multiple equally plausible outputs corresponding to the same input, (b) errors in state-of-the-art generative systems are rare and vast amounts of data are required to elicit long-tail errors, (c) the errors or undesirable behaviors produced by the models are highly contextual, and static or limited-diversity test cases cannnot capture this, (d) error distributions across system iterations (whether in data, tokenization, model or learning) change, making a cache of test cases (however adversarial, e.g. as in [Nie et al., 2020]) obsolete in the long-run.

Further, with a view towards addressing the limitations of present evaluation protocols, we believe that a specifications based framework for evaluation should address four concerns:

1. **Instance-Level Measurements**: The evaluation framework should provide targeted measurements of specification violations at the instance-level, unlike metrics, which typically work at the the level of corpora or sets.
2. **Scalabilily**: The framework should be scalable to address error rarity and consume only input data, i.e. it should have no requirement for references.

3. **Invariance to (unstable) model error distributions**: The framework should not rely on a cache of test cases and be amenable to evaluate models across data or modeling iterations.
4. **Trustworthy Measurements**: The framework should yield measurements of specification violations in a trustworthy manner, avoiding any Type I errors.

In addition to the above desired characteristics, an implicit feature of specifications based evaluation is that it raises the abstraction level at which we conduct evaluations for such models, providing much more flexibility for the inclusion of arbitrary measurements (which could be catered even towards specific evaluation datasets' attributes). However, this gain in flexibility comes at the cost of operationalizing such specifications for evaluation, which could be non-trivial. In the next section, we demonstrate a case study on Machine Translation (MT) where we operationalized arbitrary specifications successfully to elicit and reliably measure a number of previously invisible errors. Further, in section 4, we posit some reliability failures of generative models as avenues ripe for similar specifications based evaluation. In section 5, we present a few sketches for operationalizing specifications on different generative model applications.

Property	**Source-Translation Instance**
Physical Unit	Teacher's hallway song and dance reminds students to stay 6 feet apart. Lehrer Flur Lied und Tanz erinnert die Schüler zu bleiben 6 Meter auseinander.
Currency	Floorpops Medina Self Adhesive Floor Tiles, £14 from Dunelm - buy now Floorpops Medina selbstklebende Bodenfliesen, 15 € von Dunelm günstig kaufen

Table 1: Specification based evaluation for Google Machine Translation [Raunak et al., 2022]

3 Operationalizing Specifications for Evaluation: A Case Study on MT

In this section, we describe a case study on MT, where operationalizing a specification based framework for evaluation yielded measurements to quantify previously invisible errors in state-of-the-art systems. At a high-level, we start with specifications of correct behavior and then build *detectors*, which check for violations of such specifications. The behavior specification is expressed for arbitrary input attributes or output properties. Some of these input attributes as such as physical units are at the token level, while some properties such as coverage are at the sequence level. More formally, the operationalization of an arbitrary specification as a trustworthy measurement is done is through a detector, which is an algorithm (iteratively constructed with human-in-the-loop) that, given an input-output instance, returns a boolean value indicating the presence of a specification violation with very high precision. We build detectors for 7 different specifications in MT [Raunak et al., 2022].

We find that by just passing a large number of input-output instances through these detectors, we gained quantifiable visibility into problems that were previously eliding measurements, such as translation of salient content or catastrophic errors such as hallucinations. For example, we find that errors such as incorrect conversion of physical units or currencies could be discovered and quantified at-scale using a specification which checks if the units/currencies have been carried over without any semantic change (an example is presented in Table 1). Further, we find that even though such errors are rare, they are quite pervasive across state-of-the-art systems, posing a reliability threat for translation models. We believe that this approach to evaluation is well suited towards measuring reliability problems in other constrained generative models as well, which often fail in similar ways.

To summarize, our recommendation here is that by starting with an explicit role for specifications in evaluation, the effective evaluation set can be expanded dramatically by building error detectors which implement those specifications and generate trustworthy measurements. Such an evaluation effectively augments average-case performance metrics with specification based measurements, which could reliably quantify worst-case model behaviors during system development and beyond.

4 Reliability Challenges of Large Scale Generative Models

Throughout, this paper we posit *reliability* in terms of *adherence to certain specifications*. Under this view, we can interpret coarse-to-fine evaluation as going from looser specifications toward tighter

(a) Plausible Output

(b) Implausible Output

Figure 1: An example of specification violation with DALL-E. The two images are generated from the input "*cheetah is playing tabla on stage while the crowd is cheering on a rainy day*". Behaviors such as this could be enumerated by conducting reference-free specifications-based evaluation.

ones for characterizing a model's failure modes. This way of characterizing model errors seems like a departure from standard practice, where we are used to thinking about quality in terms of average-case performance metrics such as BLEU, ROUGE, etc. [Papineni et al., 2002, Lin, 2004]. However, the framing of generation quality in terms of specifications helps us unify several phenomena under a single pedagogy. For example, the simplest specification for an generative system is that given a valid input, it should not output something absolutely irrelevant, and hallucinations in any modality could be characterized as samples on which a constrained generation model breaks this basic specification.

Further, below, we enumerate some of the reasons as to why the nature of errors in state-of-the-art generative systems necessitates an evaluation protocol in which average-case performance measures are augmented with specifications-based measurements. Specifically, the reasons include:

1. **Memorization**: High-capacity neural models are known to memorize the long-tail and templatic/repeated data. This poses a reliability risk inherent in any generative application of large neural models [Raunak and Arul, 2022], including privacy risks [Carlini et al., 2021].
2. **Real-world Data Distributions**: In general, it is very hard to map model behavior under out-of-distribution data settings. Further challenges, such as neural models' vulnerability to noise [Belinkov and Bisk, 2018] or adversaries [Goodfellow et al., 2014, Elsayed et al., 2018] necessitate an evaluation protocol which could be scaled to arbitrary sets of inputs.
3. **Biases and Spurious Correlations**: Besides spurious correlations that threaten model reliability [Ilyas et al., 2019], biases in the models could also impact their utility and exacerbate harms [Buolamwini and Gebru, 2018, Bolukbasi et al., 2016].
4. **Lack of Abstractions for Model Editing**: Large or *gigantic* models with huge training costs run the risk of becoming software monoliths, due to a lack of abstractions in understanding, debugging and updating such models. This itself presents a risk towards their utility in sensitive domains where rapid interactivity vis-à-vis different stakeholders is required. Error visibility during model development becomes even more important in such cases.

Coupling these challenges with the fact that these models have significantly expanded the scope of consumption of generative technologies [Radford et al., 2021, Brown et al., 2020], it is implied that their evaluation protocol needs to be more comprehensive than traditional smaller scale models.

5 Sketches for Operationalizing Specifications

In this section, we present a few sketches for applying specifications based evaluation for the evaluation of generative models, under some well known applications:

1. **Text to Image Generation**: The design of specifications becomes considerably harder and compute-intensive, but still feasible when the transduction modalities change. Consider for example, Figure 1, which shows two images generated by DALL-E [Ramesh et al., 2021] for the same input text. Here, in case (b), the basic specification of the output having some support in the input is violated. For detecting such specification violations, an object

detector could flag this instance as an error by checking for the object label in the text. And we can build trustworthy measurements by making this algorithm high precision.

2. **Constrained Text Generation**: One of the harder cases to tackle for designing specifications is the case of constrained text generation for tasks which have no strong alignment between the input and output (unlike MT), such as data-to-text generation. However, even in such cases, many safety and reliability attributes are amenable to specifications based evaluation. Consider the safety violation involving gratuitous toxic text generation from the models; in this case, the evaluation set can be dramatically expanded through a specifications based approach which checks for toxic text attributes in generations from *arbitrary* inputs.

6 Leveraging Specifications for Human Evaluations

Earlier, we argued that in order to obtain a trustworthy and multi-dimensional view into system quality during model evaluation or development, the procedure for checking specification violation (detector) be made high-precision. This allows generation of trustworthy error statistics at an arbitrary scale, allowing evaluation at scale without the need for any references. This simple high-precision rule of thumb for building measurements from specifications could then be leveraged for system evaluations, system comparisons or targeted evaluation of particular data or modeling interventions. For example, the specification of the output having some support in the input in the case of image generations automatically becomes a measurement of hallucinations, if its implementation, a hallucination detector (which checks for the violation of the specification in an instance) is made high-precision.

In this section, we propose some ideas on how new (arbitrary) categories of measurements designed for targeted evaluation of specific properties could be directly leveraged to guide human evaluations at different granularities. The idea of using fine-grained specifications to structure large-scale human evaluation has been explored in some of the more mature applications such as MT. For example, by grounding human evaluation in categories developed through error analysis for MT, the MQM framework in Freitag et al. [2021] posits a hierarchy of translation errors which serve as annotation slots for human evaluation. Similarly, our key proposition for human evaluation is to leverage fine-grained specifications-based error analysis to explicitly guide the human annotation categories. Leveraging fine-grained specifications for human evaluation purposes could allow annotators to provide targeted signals which typically get lost in coarse grained categories. For example, the human evaluation in DALL-E was done on accuracy and realism axes. However, using categories grounded in specifications-based error analysis (or evaluation), new measurements such as omissions, additions, inconsistencies, etc. could be constructed to elicit and quantify more subtle effects in generation.

Further, even without the construction of an explicit error typology informed through the use of specifications, human evaluation can benefit by using specification violations as a sample selection step, i.e. the outputs flagged by using the detectors implemented to verify the specifications could be used by human evaluators as samples for system comparisons or targeted evaluations. This has the second-order effect of creating specialized test cases from inputs only, upon which further human evaluation could be conducted to elicit more characteristics of the models' failure modes.

7 Summary and Conclusions

To summarize, in this work, we presented some recommendations for the evaluation of generative models. We recommended evaluations based on explicitly formulated specifications, in order to raise the abstraction level of evaluation and allow it to scale arbitrarily. We believe such an evaluation framework could become increasingly important as large-scale constrained generative models enter real-world application domains, where safety and reliability problems are heavily penalized. To implement specifications, we proposed the use of high-precision as a rule of thumb for building trustworthy measurements. However, one limitation of this mode of operationalizing specifications is that not all errors could be readily enumerated, since the design itself necessitates a loss of recall. This could be partially mitigated by making the specifications very fine-grained. However, even specifications at coarser granularities (e.g., 'coverage' in the MT case study [Raunak et al., 2022]), will give visibility to error categories that are hard to elicit through curated test cases or through metric scores just through the advantage of *evaluation scale*. Further, we presented how such specifications could further be leveraged to guide traditional human evaluations. Finally, we note that the design of specifications itself is a component of human evaluation, only at a higher abstraction level.

References

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. URL https://papers.nips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3914–3924, Red Hook, NY, USA, 2018. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2018/hash/8562ae5e286544710b2e7ebe9858833b-Abstract.html.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. volume 9, pages 1460–1474, Cambridge, MA, 2021. MIT Press. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez-Beltrachini, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. Gemv2: Multilingual nlg benchmarking in a single line of code, 2022. URL https://arxiv.org/abs/2206.11249.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. URL https://arxiv.org/abs/1412.6572.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022. URL https://arxiv.org/abs/2209.12356.

Sumit Gulwani, Oleksandr Polozov, Rishabh Singh, et al. Program synthesis. *Foundations and Trends® in Programming Languages*, 4(1-2):1–119, 2017. URL https://www.nowpublishers.com/article/Details/PGL-010.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Jayakrishnan Nair, Adam Wierman, and Bert Zwart. *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2022. URL https://www.cambridge.org/core/books/fundamentals-of-heavy-tails/3B1A35A6E72551E50E4723A4785044EE.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P02-1040.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Vikas Raunak and Menezes Arul. Finding memo: Extractive memorization in constrained sequence generation tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Online, December 2022. Association for Computational Linguistics. URL https://arxiv.org/abs/2210.12929.

Vikas Raunak, Matt Post, and Arul Menezes. Salted: A framework for salient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Online, December 2022. Association for Computational Linguistics. URL https://arxiv.org/abs/2205.09988.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.213.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.442.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.704.

Aarohi Srivastava, Abhinav Rastogi, Abhishek B Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou,

Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Saleem Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J. Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O. Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu F Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yushi Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615, 2022. URL https://arxiv.org/abs/2206.04615.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.