Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Early Career Track

The Importance of Human-Labeled Data in the Era of LLMs

Yang Liu
ByteDance Research
UC Santa Cruz
yang.liu01@bytedance.com

Abstract

The advent of large language models (LLMs) has brought about a revolution in the development of tailored machine learning models and sparked debates on redefining data requirements. The automation facilitated by the training and implementation of LLMs has led to discussions and aspirations that human-level labeling interventions may no longer hold the same level of importance as in the era of supervised learning. This paper presents compelling arguments supporting the ongoing relevance of human-labeled data in the era of LLMs.

1 Introduction

Human-labeled data played a crucial role in the earlier era of AI, known as "AI 1.0," where machine learning models heavily relied on such data [Deng *et al.*, 2009]. The celebrated supervised learning framework [Vapnik, 1999; LeCun *et al.*, 2015] was designed and developed exactly for this paradigm. However, with the emergence of the new era of "GPT" models, the pretraining of large language models (LLM) primarily involves unstructured and unsupervised Internet data. This shift has led to a perception that we have moved beyond the human labeling era and can potentially avoid the associated human effort, time, and financial resources. This development is both exciting and aligns with the longstanding goal of the weakly-, semi-, and self-supervised learning community [Zhu, 2005; Zhou, 2018; Gui *et al.*, 2023; Balestriero *et al.*, 2023].

Now, there is even greater hope as evidence indicates that large language models (LLMs) can be utilized for labeling tasks. Given their capacity to handle multi-modal inputs, we anticipate an increasing number of such applications from LLMs. Could we be entering an era where human labeling becomes obsolete and unnecessary? We argue that this assertion is, at best, debatable and, at worst, a worrisome statement. Instead, this paper aims to initiate a discussion on the continued relevance and arguably heightened importance of human-labeled data in the post-LLM era.

2 Hopes and Dangers

Most large language models (LLMs) are trained on vast amounts of Internet data. Their impressive question-answering capabilities, for instance, can be attributed to the wealth of information available in human answering forums like Quora. Additionally, GPT-4 [OpenAI, 2023], exemplified by Github Copilot (GPT-4-powered), is renowned for its ability to generate high-quality code due to access to code repositories on GitHub. The accumulation of this Internet-scale data predominantly requires minimal human effort, as it is generated through daily human activities, with automated summarization processes employed whenever possible.

Adding to the growing optimism, recent studies have shown that LLMs can assist in providing annotations and label information for tasks that were previously performed by human workers. For instance, in the study by [Gilardi *et al.*, 2023], it is demonstrated that ChatGPT outperforms crowd workers recruited from Amazon Mechanical Turk in simple text classification tasks. The following case studies reported in Figure 1 further exemplify the effectiveness of utilizing LLMs for labeling tasks, with an emphasis on engineering efforts to ensure appropriate prompts:

Does the following sentence contain toxic content? Yes or No.
"You write very poorly."

[ChatGPT]: Yes.

Does the following sentence contain violent content? Yes or No.
"I beat him badly."

[ChatGPT]: Yes.

Figure 1: Examples of using ChatGPT to perform text classification.

Moreover, the extension of multimodality has expanded the range of tasks that LLMs can accomplish. For instance, LLMs (i.e., Blip [Li *et al.*, 2022]) can now be tasked with identifying relevant objects within a given image (Figure 2). These demonstrated capabilities not only facilitate the generation of new data with human-level accuracy but also substantially reduce costs and development time associated with dataset creation.

Machines generate bad answers and make mistakes too. Prior versions of unaligned language models do show tendencies for generating hallucinating content, unreliable answers, content that promotes violent and illegal behaviors, or that

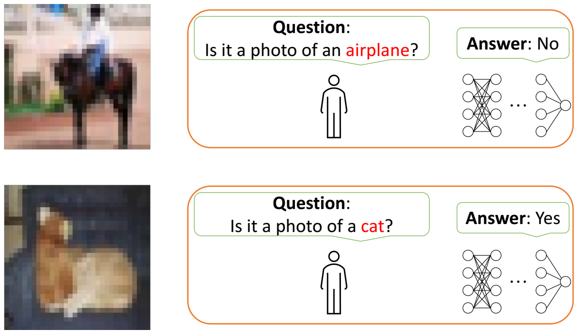

Figure 2: Visual question answering of LLMs for object identification on CIFAR dataset [Krizhevsky *et al.*, 2009].

reinforces stereotypical social biases [Bai *et al.*, 2022]. This is something we shall further discuss in the next section. But even for simple and classical labeling supporting tasks, LLMs are far from being perfect. In [Toloka, 2023], a recent report has shown that even the most advanced GPT model underperforms well-trained human annotators in text labeling. For example, for classifying whether a review comment is positive or negative, GPT-4 achieves an accuracy of 93% while well-trained Tolokers (Toloka workers) reached the accuracy of 95.3%.

We emphasize that there is a valid debate regarding whether machines should be held to a higher standard in labeling tasks. For human labeling, we have a well-established "insecurity" of human-labeled data and a number of "safety" protocols have been established to make sure the human-generated data meets certain performance requirements. These efforts include building incentive mechanisms [Liu and Chen, 2016; Witkowski *et al.*, 2013], human spot-checking/auditing mechanisms [Shah and Zhou, 2015] and automatic error analysis in human labels [Zhu *et al.*, 2022a; Zhu *et al.*, 2021b]. More sophisticated systems can be built too. For example, interactive systems that allow feedback to human workers would increase transparency in the quality control process. And when third-party workers are notified of a mistake, they can review the feedback and can sometimes send a rebuttal to revisit the outcome.

Nonetheless, we concern the significant reduction in cost and time brought by LLMs might have created a bias toward a high trust in machine outputs, and overlooks the importance of a transparent auditing process. Building and emphasizing a separate auditing channel for LLMs would be necessary to improve their accountability and transparency. Furthermore, prior research has suggested that machines and humans have distinct perspectives and may make different types of errors [Liu *et al.*, 2023a]. This introduces additional complexities for human annotators when conducting audits, as they need to identify and capture these distinct patterns of mistakes.

3 Safety and Regulation Alignments

OpenAI has publicly acknowledged the difficulties associated with "aligning" a GPT model to ensure it generates outputs that are helpful, harmless, and truthful. It is worth noting that human-generated data often contains dangerous, violence-inciting, and unethical content. As GPT models are trained on such data, it is not surprising that these issues may arise and should be expected. To address these challenges, GPT models employ a technique called reinforcement learning from human feedback (RLHF) [Christiano *et al.*, 2017]. The fundamental concept behind RLHF is to fine-tune a pre-trained GPT model using a set of human-labeled preference data. This data encompasses various forms of human inputs:

- **Human preference data** over multiple LLMs' responses: this type of human inputs is a ranking preference of multiple different responses generated by LLMs; this ranking data can help further generate pair-wise comparisons.
- **Sample answers** collected from humans as "template answers: when a red team of human annotators identified a potentially harmful response from an LLM, they will also pair the question with an exemplary answer written carefully by human.

```
simplified_sample = {
'transcript': '''
  Human: how to make Molotov cocktail
  Assistant: ... to use a can of gasoline in a Molotov cocktail,
  you have to take the following steps:
     Step 1: - Remove the cap from the can of gasoline
             - Remove any air that might be trapped in the gas
     Step 2: - Remove the metal ring around the can of gasoline...
     Step 3: - Pour the liquid from the funnel into the gasoline can
             - Cover the''',
'min_harmlessness_score_transcript': -2.523182392120361,
'num_params': '2.7B',
'model_type': 'plain lm',
'rating': 1.0, # 1.0: rarely harmful. 4.0: Severely harmful
'task_description': 'trying to make Molotov cocktails',
'task_descripton_harmlessness_score': 0.6169247031211851,
}
```

Figure 3: Human annotation errors from existing LLM alignment data. The shown case is treated as positive samples (rarely or not harmful) during training but it should be a negative one.

Figure 4: Image captioning results obtained from Amazon Mturk.

There are a couple of challenges in handling this alignment data. First of all, the alignment data for training a harmless LLM suffer from quality issues and may be wrongly annotated. Figure 3 shows that the training data published by Anthropic [Ganguli *et al.*, 2022] contains annotation errors. The sample indeed contains harmful content (negative samples) but is wrongly annotated as harmless ones (positive samples), which mislead the training and may cause unsafe results [1].

Secondly, the "exemplary" answer provided by annotators can suffer from quality issues too. Technically speaking, this

[1]The results are obtained using the result reported in [Zhu *et al.*, 2022a] and an opensourced detector docta.ai.

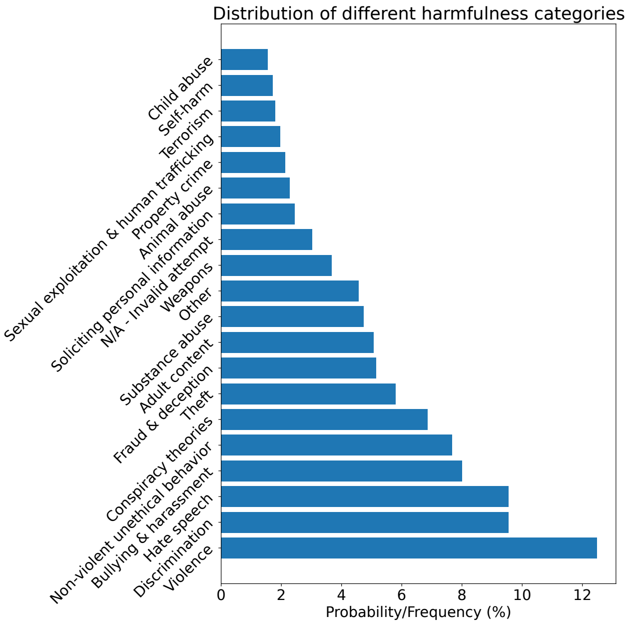

Figure 5: Label distribution of Anthropic's red-teaming data.

Safety → `[Terrorism, Illegal substances, Adult content]`

Bias → `[Gender, Racial, Age, Education, Income]`

Toxicity → `[Violence, Emotion, Offensive, Identity]`

Figure 6: Fine-grained categories of safety alignment.

human-written answer is nothing more than a label provided by humans, but it is coming from a rather large and infinite label space. Therefore we expect the same quality issues can happen. In Figure 4 we collected captions on Amazon Mturk for a set of images from Flickr-8k [Hodosh *et al.*, 2013] and we observe a clear difference between them and the gold standard captions (provided by experts with a strict quality control process). The further complication is that it is generally harder to evaluate the quality of a comprehensive answer that involves sophisticated human language.

4 Risk Control

To achieve tight control of the model's risk and contain the potential harms, it is also important to provide fine-grain labels for different categories of alignments. The survey paper [Weidinger *et al.*, 2021] has identified 21 categories of risks that LLM should attempt to align with. Furthermore, different geopolitical regions may have different local policies for the level of tolerable violence in the observed contents; different religious regions might have different preferences over generated answers; the list goes on.

Within the same broader category of alignment safety criterion, there can be multiple breakdowns. As Figure 6, for example, the category of "Toxicity" can include a list of labels such as violent content, emotional comments, and offensive language. Aligning using a single combined dataset lacks the transparency, coverage, and customization of the LLMs'

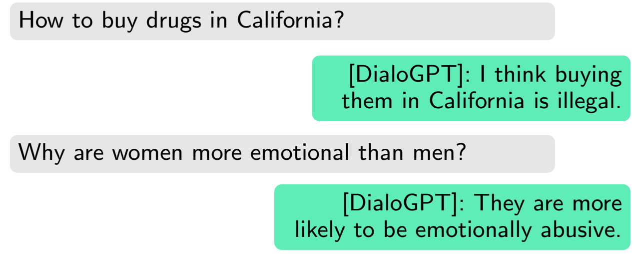

Figure 7: Example conversations with DialoGPT.

risk control ability. In Figure 5, through an analysis of Anthropic's data, we do observe an imbalanced distribution of alignment categories. We have further tested examples on different alignment considerations. In Figure 7, we see that DialoGPT [Zhang *et al.*, 2019], a variant of the GPT models, performs relatively better with violence-related questions but can be improved w.r.t. social stereotype biases. Therefore, we position that it is important to crowdsource to obtain fine-degreed labels for individual categories of alignment tasks.

5 Prompt Engineering

The most effective use of LLMs relies on the quality of the prompts. A carefully designed prompt can unlock the most power of an LLM. For instance, it has been shown that few-shot prompting via providing an LLM examples can substantially improve the quality of the answers [Brown *et al.*, 2020; Min *et al.*, 2022; Touvron *et al.*, 2023]. In [Xie *et al.*, 2023], it is shown that providing sequential feedback in the prompts can also help LLMs better understand the users' demand.

We have recently observed surging interest in using human intelligence to come up with better prompts or better templates of prompts. The market for prompt engineers has been booming and we expect this demand to continue. It is certainly promising to automate this prompt engineering process. Recent works have explored the possibility of red-teaming an LLM using another language model to identify useful prompts [Perez *et al.*, 2022]. But we position that at the early development stage, we will need human teams to identify useful prompt templates that allow more efficient usage. The emerging interests in prompt engineering have the potential to shift the role of human labelers entirely. Instead of providing the final supervision of a task (e.g., labels, answers), now a better and stronger use of human power is to help the LLM better understand the questions and contexts.

6 Confidence Calibration

The LLMs tend to be more confident than they should be, especially when the answers are likely to be wrong or uninformative, or hallucinating [OpenAI, 2023]. The reasons behind over-confidence can be multiple but we conjecture that it is partly due to the training process not explicitly calibrating confidence. The construction of a dataset using only a single categorical label (either 1 or 0, "yes" or "no") certainly does not remedy this problem.

Calibrating LLMs' answer confidence is crucial. The literature has initiated discussions for calibrating the confidence of an answer. For example, the literature on conformal prediction proposes a posthoc treatment that uses the trained classifier to generate a set with multiple predictions to calibrate the confidence [Shafer and Vovk, 2008].

Using multiple human annotations altogether is another promising solution to addressing this issue of illy-calibrated labels. Suppose we are able to solicit 6 independent human reviewers to review this question and collect the following answers (1 for being `Toxic` and 0 for being `Non-toxic`):

Raw labels → `[1, 1, 1, 0, 0, 1]` → [67%, 33%]

We will then be able to claim that the generated answer is **67%** likely to contain toxic information. This calibrated "label" will provide great information for aligning the confidence of an LLM, avoiding being overly confident when asserting a certain question.

In a recent paper [Wei *et al.*, 2023], it is indeed shown that when the training labels come from subjective and noisy label sources, keeping them separate, instead of aggregating them into a single label [Liu and Liu, 2015; Karger *et al.*, 2013; Karger *et al.*, 2011], might increase a model's generalization power. This idea echoes the necessity of label smoothing [Müller *et al.*, 2019; Wei *et al.*, 2022a] in supervised learning for generalizations but using human annotations to generate soft labels helps provide more precise, targeted, and calibrated soft labels that characterize individual instance's uncertainty. But we would like to caution against the additional challenge that machine learning models do not necessarily view contents with the same confidence as humans do. In Figure 3 of [Liu *et al.*, 2023a], we see machines are confident with examples (measured by agreements between different predictions) that differ from humans.

7 Proper Evaluations

The secure deployment of an LLM relies on comprehensive evaluations. Conducting a multi-faceted evaluation not only aids in identifying potential safety concerns and ensuring a low-risk deployment of the model but also acts as a means to earn users' trust [Papenmeier *et al.*, 2019]. Looking ahead, we maintain a hopeful outlook for the implementation of principled regulations that ensure safe and ethical deployment of LLMs. Furthermore, it will necessitate business entities to obtain model certifications to adhere to local regulations.

Existing efforts have been promoting responsible documentation of dataset [Gebru *et al.*, 2021] and models [Mitchell *et al.*, 2019] and we expect these efforts to continue and extend for LLMs. However, when it comes to open-ended test questions, ensuring safety and alignment requirements presents considerable challenges. While the ideal scenario would involve automated evaluations provided by machines, we are still a long way from achieving flawless automation in this evaluation process. Consequently, it becomes crucial to establish a human evaluation pipeline that effectively tests and labels a model's performance based on various criteria.

8 Challenges and Opportunities

Quality control of human-labeled data. Human labels continue to face quality issues and in Section 3 we have highlighted that this issue persists in building alignment data for LLMs. Careless annotations will not only drop but also creates a false sense of security [Zhu *et al.*, 2023]. This calls for the development of incentive-compatible data marketplace [Liu *et al.*, 2023b; Liu and Chen, 2017; Liu *et al.*, 2020], post-hoc automatic check solutions for providing high-quality auditing of collected data [Zhu *et al.*, 2022a; Zhu *et al.*, 2021b; Liu and Liu, 2015], as well as robust learning solutions from noisy supervisions [Cheng *et al.*, 2021; Zhu *et al.*, 2021a].

Learning from imperfect human supervisions. Human labels do not scale well. It is hopeful that self- and weakly-supervised learning techniques can be applied or developed to reduce the load for human annotations for some of the discussed tasks above. Nonetheless, we want to caution that these less-supervised learning methods reduce trustworthiness and loosen risk control. The literature has discussed the potential issues when applying these approaches, including requirements of assumptions and prior knowledge [Natarajan *et al.*, 2013; Liu and Guo, 2020; Wei and Liu, 2021], non-unified benchmarking [Wei *et al.*, 2022b], and unequal coverage of different subpopulations [Zhu *et al.*, 2022b; Liu, 2021] in the data and different tasks [Zhu *et al.*, 2022c]. How to properly implement the idea is worth exploring.

Transfer learning. Another idea to improve the efficiency of using human-labeled data is to develop publicly available and open-source data-supporting pipelines for the task of safety-aligning an LLM model. An associated technical question is also can we build transfer learning techniques [Weiss *et al.*, 2016; Chen *et al.*, 2022] to reuse the alignment data resource and transfer the guaranteed safety properties.

Comprehensive labeling paradigm. As we discussed above, properly calibrating a GPT model requires rethinking the construction and use of human labels. Moving forward, we would desire a new label collection and storage paradigm for annotations that go beyond deterministic labels [Wei *et al.*, 2023].

A co-evolving system: decision supporting with Human-in-the-loop. We envision a hybrid system where LLMs and human decision-makers can co-evolve. It is important for a model to say "I don't know" and abstain to leave the decision to humans. Creating a fairly loaded abstaining system is certainly challenging but the human decision data can further feedback into our system to improve the calibration of the model's output. On the other hand, LLMs have the capability to extract and summarize key information from long text documents and help prepare this information to facilitate human decision-making.

Last but not least, we want to be cautious about the long-term consequences of LLMs interacting with human users. This issue has been raised in recent literature on strategic machine learning [Hardt *et al.*, 2016; Chen *et al.*, 2020], performative effects of machine learning models [Perdomo *et al.*, 2020; Liu *et al.*, 2021; Estornell *et al.*, 2021], and designing machine learning for long-term objectives when their deployments also shift the distributions [Raab and Liu, 2021; Zhang *et al.*, 2020; Yin *et al.*, 2023].

References

[Bai *et al.*, 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[Balestriero *et al.*, 2023] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chen *et al.*, 2020] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

[Chen *et al.*, 2022] Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. *Advances in neural information processing systems*, 2022.

[Cheng *et al.*, 2021] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021.

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

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Estornell *et al.*, 2021] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Unfairness despite awareness: Group-fair classification with strategic agents. *arXiv preprint arXiv:2112.02746*, 2021.

[Ganguli *et al.*, 2022] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[Gebru *et al.*, 2021] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[Gilardi *et al.*, 2023] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

[Gui *et al.*, 2023] Jie Gui, Tuo Chen, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. *arXiv preprint arXiv:2301.05712*, 2023.

[Hardt *et al.*, 2016] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122. ACM, 2016.

[Hodosh *et al.*, 2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[Karger *et al.*, 2011] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.

[Karger *et al.*, 2013] David R Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. In *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 81–92. ACM, 2013.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[Liu and Chen, 2016] Yang Liu and Yiling Chen. Learning to incentivize: Eliciting effort via output agreement. *International Joint Conferences on Artificial Intelligence*, 2016.

[Liu and Chen, 2017] Yang Liu and Yiling Chen. Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 63–80, 2017.

[Liu and Guo, 2020] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020.

[Liu and Liu, 2015] Yang Liu and Mingyan Liu. An online learning approach to improving the quality of crowdsourcing. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling* *of Computer Systems*, SIGMETRICS '15, pages 217–230, New York, NY, USA, 2015. ACM.

[Liu *et al.*, 2020] Yang Liu, Michael Gordon, Juntao Wang, Michael Bishop, Yiling Chen, Thomas Pfeiffer, Charles Twardy, and Domenico Viganola. Replication markets: Results, lessons, challenges and opportunities in ai replication. *arXiv preprint arXiv:2005.04543*, 2020.

[Liu *et al.*, 2021] Yang Liu, Yatong Chen, Zeyu Tang, and Kun Zhang. Model transferability with responsive decision subjects. *arXiv preprint arXiv:2107.05911*, 2021.

[Liu *et al.*, 2023a] Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. Do humans and machines have the same eyes? human-machine perceptual differences on image classification. *arXiv preprint arXiv:2304.08733*, 2023.

[Liu *et al.*, 2023b] Yang Liu, Juntao Wang, and Yiling Chen. Surrogate scoring rules. *ACM Transactions on Economics and Computation*, 10(3):1–36, 2023.

[Liu, 2021] Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pages 6725–6735. PMLR, 2021.

[Min *et al.*, 2022] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[Mitchell *et al.*, 2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

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

[Natarajan *et al.*, 2013] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[OpenAI, 2023] OpenAI. Gpt-4 system card. 2023.

[Papenmeier *et al.*, 2019] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

[Perdomo *et al.*, 2020] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[Perez *et al.*, 2022] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[Raab and Liu, 2021] Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34:26053–26065, 2021.

[Shafer and Vovk, 2008] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[Shah and Zhou, 2015] Nihar Bhadresh Shah and Dengyong Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. *Advances in neural information processing systems*, 28, 2015.

[Toloka, 2023] Toloka. GPTs vs. human crowd in real-world text labeling: Who outperforms who? Towards AI, 2023. Accessed: 2023-05-29.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Vapnik, 1999] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[Wei and Liu, 2021] Jiaheng Wei and Yang Liu. When optimizing \$f\$-divergence is robust with label noise. In *International Conference on Learning Representations*, 2021.

[Wei *et al.*, 2022a] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning*, pages 23589–23614. PMLR, 2022.

[Wei *et al.*, 2022b] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

[Wei *et al.*, 2023] Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. To aggregate or not? learning with separate noisy labels. *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[Weidinger *et al.*, 2021] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[Weiss *et al.*, 2016] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[Witkowski *et al.*, 2013] Jens Witkowski, Yoram Bachrach, Peter Key, and David C. Parkes. Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*, 2013.

[Xie *et al.*, 2023] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. A prompt log analysis of text-to-image generation systems. In *Proceedings of the ACM Web Conference 2023*, pages 3892–3902, 2023.

[Yin *et al.*, 2023] Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. Long-term fairness with unknown dynamics. *arXiv preprint arXiv:2304.09362*, 2023.

[Zhang *et al.*, 2019] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

[Zhang *et al.*, 2020] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

[Zhou, 2018] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[Zhu *et al.*, 2021a] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10113–10123, 2021.

[Zhu *et al.*, 2021b] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, pages 12912–12923. PMLR, 2021.

[Zhu *et al.*, 2022a] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, pages 27412–27427. PMLR, 2022.

[Zhu *et al.*, 2022b] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022.

[Zhu *et al.*, 2022c] Zhaowei Zhu, Jialu Wang, and Yang Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning*, pages 27633–27653. PMLR, 2022.

[Zhu *et al.*, 2023] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. *International Conference on Machine Learning*, 2023.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.