REWARD MODELS ARE METRICS IN A TRENCH COAT

Anonymous authors
Paper under double-blind review

ABSTRACT

The emergence of reinforcement learning in post-training of large language models has sparked significant interest in reward models. Reward models assess the quality of sampled model outputs to generate training signals. This task is also performed by evaluation metrics that monitor the performance of an AI model. We find that the two research areas are mostly separate, leading to redundant terminology and repeated pitfalls. Common challenges include susceptibility to spurious correlations, impact on downstream reward hacking, methods to improve data quality, and approaches to meta-evaluation. Our position paper argues that a closer collaboration between the fields can help overcome these issues. To that end, we show how metrics outperform reward models on specific tasks and provide an extensive survey of the two areas. Grounded in this survey, we point to multiple research topics in which closer alignment can improve reward models and metrics in areas such as preference elicitation methods, avoidance of spurious correlations and reward hacking, and calibration-aware meta-evaluation.

1 INTRODUCTION

Reinforcement learning (RL) plays a major role in post-training, aligning, and adapting language models (LLMs) to a broad range of tasks (OpenAI, 2025; Comanici et al., 2025; xAI, 2025; Kimi et al., 2025; Guo et al., 2025). Scaling laws apply to reinforcement learning from human feedback (RLHF, Christiano et al., 2017) similarly as to the rest of the training stack (Bai et al., 2022a). As such, scalable alternatives to human feedback have become popular, either in the form of verifiable rewards (Lambert et al., 2024) or in the form of models that assess the quality of model outputs (Li et al., 2018). Developing robust and reliable *reward models* is crucial, as the downstream RL models can experience reward hacking (Amodei et al., 2016), optimizing for spurious correlations in the reward model rather than learning the intended behavior. To overcome these issues, reward models have experienced significant research interest.

In parallel to research on reward models, *model-based evaluation* of generated text has similarly seen a surge in interest, enabling a switch from "traditional" lexical metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to learned models (Ma et al., 2018; 2019) and prompt-based approaches referred to as LLM-as-a-judge (Zheng et al., 2023). The two fields present two sides of the same coin: while reward models assess output quality to directly improve models, evaluation metrics assess output quality to identify potential areas of improvement. Both fields seek to develop classifiers that consume generated content as input and assign a goodness score as output. Both fields strongly benefit from rigor, consideration of the sociotechnical context in which a system is deployed, and improved correlation between model-based judgments and expert human raters. The key difference between the two is that while metrics tend to be more specialized, reward models tend to assess broad capabilities spanning many tasks. Due to their similarities, one might expect the fields learn from and inform each other, and that breakthroughs transfer quickly between them.

Our position paper argues that while this should be the case, it is not. Instead, the academic literature in these fields only infrequently informs each other and the fields are actively developing and using different terminology for the same methods. While metrics are commonly used to generate training data for reward models (Malik et al., 2025), and are thus instrumental to reward model performance, little attention is being paid to which metrics generate that data. We demonstrate this phenomenon by analyzing the citation graphs of papers in each sub-field, showing that inter-field citations account for fewer than 10% of total cited papers. We further support this claim by presenting results from two small experiments: one in which we apply a metric to a reward modeling benchmark and one

where we apply reward modeling techniques to a factuality evaluation benchmark. The results show that reward modeling approaches lag behind dedicated metrics for these specialized tasks, providing opportunities for improvements and motivating cross-testing on their respective benchmarks.

Motivated by these findings, we conduct an extensive survey of the two fields and their intersection. We lay out scenarios in which we can use all the tools at our disposal and showcase how it could lead to better reward models and evaluation metrics. Specifically, we argue that a closer collaboration could lead to major progress in overcoming reward hacking, in preference elicitation, and in meta-evaluation. We also discuss areas in which the fields differ and should not interact, and how this relates to Goodhart's law, which states that a measure ceases to be a good measure when it becomes a target. Grounded in these discussions, we make specific recommendations to researchers working on reward models and evaluation metrics on how the separation can be overcome.

2 How did modern Reward Models come about?

As part of the rising popularity of deep learning, RL started to be explored for tasks like structured prediction (Daumé et al., 2009; Ross et al., 2011), image recognition (Mnih et al., 2014; Ba et al., 2015) and for agents like the Neural Turing Machine (Zaremba & Sutskever, 2015). Successfully training a model via RL hinges on being able to generate reward signals. This includes being able to derive the value of intermediate states. As Sutton & Barto (2018) argue, "the most important component of almost all reinforcement learning algorithms we consider is a method for efficiently estimating values." Commenting on this issue, Yann LeCun famously criticized RL for having much sparser rewards than self-supervised learning during his talk "Predictive Learning" (LeCun, 2016).

This issue applies to generated language: generation has a combinatorially large state space with its sequential token choices from a large vocabulary, and no single objective number can represents the value of an output (Gehrmann et al., 2023). For that reason, generation models are typically trained via teacher-forcing, a supervised approach that shows the model a ground-truth token at each prediction step. This happens only during training, not at test-time. Moreover, while models are trained with a cross-entropy objective, they are evaluated via different metrics. Ranzato et al. (2016) coined the term *exposure bias* for this mismatch between training and test time.

If there was a way to directly optimize for the metric(s) we care about, the exposure bias could be overcome. Evaluation metrics are designed to act as a proxy for human judgments and are thus well-suited to serve as a reward function. While some inference-time methods optimize metrics (Wiseman & Rush, 2016; Freitag et al., 2021b), reinforcement learning is a natural fit to optimize for these metrics during training. REINFORCE (Williams, 1992) and minimum risk training (Duda & Hart, 1974) generate metric-based reward signals using sampled token sequences, and actor-critic approaches estimate partial rewards for predicted tokens (Bahdanau et al., 2017). Various instantiations of these approaches were used for machine translation (Ranzato et al., 2016; Shen et al., 2016), image captioning (Rennie et al., 2017), video captioning (Pasunuru & Bansal, 2017), and summarization (Paulus et al., 2018).

At that time, the reward models were measuring lexical overlap between a generated sequence and a ground truth (e.g., Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015). These metrics have well-understood drawbacks (e.g., Reiter, 2018; Freitag et al., 2020), especially for RL (Choshen et al., 2020). Among others, they lead to *reward hacking* where models generate non-fluent language that maximizes reward scores (Amodei et al., 2016). Researchers worked to overcome these issues, for example by regularizing the training process by combining cross-entropy losses with RL or by hand-crafting additional reward functions (Pasunuru & Bansal, 2018; Kryściński et al., 2018; Wu et al., 2018a). The advent of metrics measuring semantic rather than lexical similarity led to significantly reduced reward-hacking since the new models avoided over-optimizing for the generation of relevant words without fluent context (Li et al., 2018; Yasui et al., 2019; Scialom et al., 2019).[1] These models led to a clear path whereby new metrics could be validated and then used as reward models. For example, the metric BLEURT was introduced (Sellam et al., 2020), evaluated as part of the WMT Metrics shared task (Mathur et al., 2020), and then assessed as reward model (Shu et al., 2021).

[1] This is also related to Generative Adversarial Networks (Goodfellow et al., 2014) for generation (e.g., Yu et al., 2017; Wu et al., 2018b) where a discriminator differentiates generated text from the ground truth, thus similarly generating a model-based signal for human-likeness.

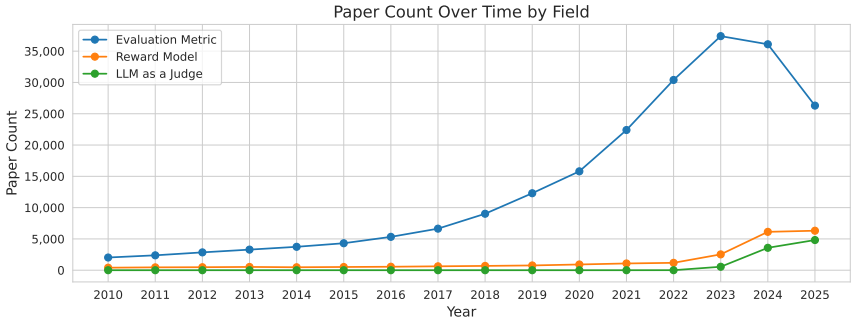


Figure 1: The figure shows the number of publications per year in the three subfields according to a keyword search on Google Scholar. Publications on evaluation metrics have slowed, even though research on reward modeling and LLM-as-a-judge is quickly rising in popularity. If the fields were actively learning from one another, one could assume that mentions of "evaluation metrics" should be growing alongside these newly emerging fields, but they are not.

In parallel to the work above, the concept of *reinforcement learning from human feedback (RLHF)* was introduced for game playing and robotics (Christiano et al., 2017). In an essay titled "Scalable agent alignment via reward modeling: a research direction", Leike et al. (2018) propose capturing human preferences via dedicated *reward models*. This research culminated in the work on RLHF for summarization (Stiennon et al., 2020) which popularized Proximal Policy Optimization (PPO, Schulman et al., 2017) as RL approach for text generation. While Stiennon et al. (2020) analyzed correlations between ROUGE and their human preference data, they did not use widely accepted alignment metrics, existing human preference corpora, or the semantic similarity evaluation metrics discussed above. In the followup work on InstructGPT (Ouyang et al., 2022), there are no references to the generation RL literature and no evaluation of the reward model. Subsequent work introduced the notion of AI Feedback as reward models (Bai et al., 2022b; Lee et al., 2024a) and argued that language model probabilities can be directly used to model rewards in a Bradley-Terry model (Rafailov et al., 2023; Bradley & Terry, 1952). Neither draws the connection to the role that perplexity and model probabilities played in existing evaluations (e.g., Lewis et al., 2020; Min et al., 2023).

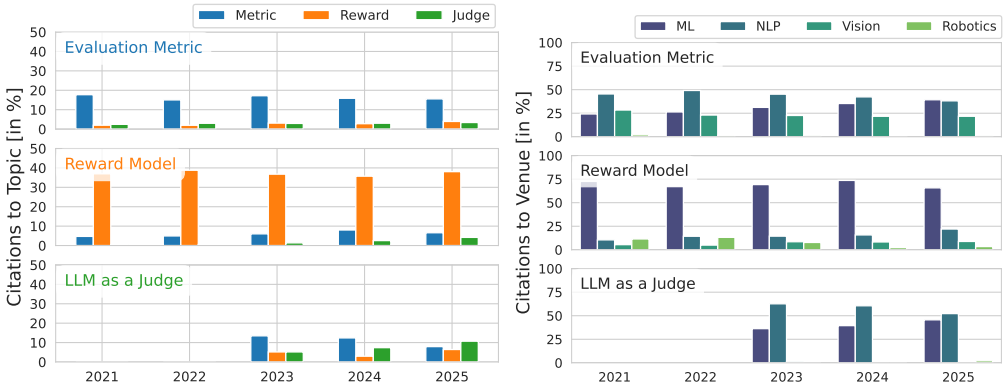
This separation culminated in benchmarks for reward models (e.g., Frick et al., 2025; Lambert et al., 2025; Liu et al., 2025c; Zhou et al., 2025) and for metrics (e.g., Honovich et al., 2022; Clark et al., 2023; Freitag et al., 2024) that exist in parallel without meaningful interaction. This raises the question of whether this disconnect is part of a broader trend. And if the two fields were integrated tighter, would we be in a better state today? And what should one learn from the other?

3 Quantifying the Research Field Separation

Figure 1 establishes the need for this investigation by showing the number of papers found on Google Scholar per year that contain the exact strings "Evaluation Metric", "Reward Model", and "LLM-as-a-judge". We include "LLM-as-a-judge" as an emerging field that has similarly experienced rapid growth and which also uses language models to estimate the quality of generated output. Notably, despite the exponential growth of the two emerging topics, the number of papers mentioning evaluation metrics decreased in 2024, with the trend continuing into 2025.

If the terminology was merely changing, one would expect the new literature to still build on the older one. For that reason, we empirically study the cause of this phenomenon by conducting a citation analysis. We select up to 300 papers per field per year (2021–2025) via the Semantic Scholar Graph API, with sensitivity checks at 100/200 yielding similar trends (Kinney et al., 2023).[2] For each paper, we additionally retrieve its citations, yielding approximately 10,000 citations per year for each field to analyze. As a proxy to identify whether a paper in field A cites a paper in field B, we

[2]Documented at `https://api.semanticscholar.org/api-docs/graph`. The search results in the maximum 300 papers for the first two fields and 8, 25, and 43 papers respectively for LLM-as-a-judge over the past three years. More details on this analysis in Appendix A.



(a) The fraction of citations from one field to another, based on keywords in cited papers.

(b) The fraction of citations to papers in venues associated with a research area.

Figure 2: In our analysis of citation dynamics across the three fields, we find that evaluation papers tend to cite other evaluation papers across research fields, while reward model papers mostly cite each other and are highly focused on machine learning venues. LLM-as-a-judge work mostly cites ML and NLP venues, but has less clear citation dynamics.

select *signaling terms* for each field: (1) "metric(s)", (2) "reward", "reinforcement", "policy", and (3) "judge". If any of those terms appears in the title or abstract of a cited paper, we count this as an inter-field or intra-field citation. The results in Figure 2a show that evaluation metrics and reward models are distinct fields, with only few inter-field citations but many intra-field citations. This is especially pronounced for reward models where almost 40% intra-field citations. The numbers for evaluation metrics trend lower at around 15–20% which we attribute to the heterogeneity of the field; for example, papers on metrics for summarization cite summarization papers rather than only evaluation papers. LLM-as-a-judge is an outlier, with too few papers to draw definitive conclusions.

We find more evidence for the field separation when we analyze the venues of the cited papers. For this, we categorize venues into fields (e.g., ICLR as ML venue) and calculate the percentage of citations to papers published in the various fields. The results in Figure 2b reveal that reward model research predominantly cites research in machine learning venues and not NLP and Computer Vision. In contrast, evaluation metric work is evenly distributed and LLM-as-a-judge work focuses on ML and NLP.[3] Since all observed trends are stable across years, we conclude that the three research fields are largely separate with limited interaction.

4 What can one learn from the other?

A rebuttal to our proposition that the two fields should learn from each other is that maybe there is little to learn. We thus highlight two scenarios in which a closer cross-field interaction could have changed conclusions or yielded additional insights.

4.1 Metrics can perform well on Reward Model Benchmarks

The recently introduced RewardBench-M (Gureja et al., 2025) uses a subset of the MAPLE dataset (Zhu et al., 2024) to assess reward models on translation evaluation. The task requires identifying which of two translation outputs was rated higher by human evaluators. The data is split into an easy and difficult subset based on the difference of human scores of the two provided translations. While all their tested models perform nearly perfectly on the easy subset, Gureja et al. (2025) remark that "*models that perform well on easy tasks can struggle to maintain the same level of performance on harder translations, indicating the need for more sophisticated mechanisms to handle [...] challenging scenarios*". However, no machine translation evaluation metric was assessed as a

[3]We omit fields with $\leq 5\%$ of citations in all years, including Speech, IR, and HCI. Ambiguous venues like preprint servers or broad venues like AAAI are excluded from this analysis. Detailed list in Appendix C.

	de→en	en→de	zh→en	en→zh
GPT-4o	**71.0**	61.0	**77.0**	80.0
Aya Expanse 32B	62.0	**69.0**	76.0	79.0
COMETKIWI-DA (2022)	59.0	**68.0**	59.0	**86.0**

Table 1: Results on the hard machine translation evaluation subset of RewardBench-M. For non-English evaluations, a 3 year old model with 550M parameters outperforms much larger LLMs.

baseline. Thus, to test this hypothesis, we evaluate the three-year-old metric CometKiwi (Rei et al., 2022) which is based on InfoXLM (Chi et al., 2021) and has only 550M parameters.

The results of CometKiwi alongside the two best-performing models on the challenging translation test set of RewardBench-M (Dang et al., 2024; Hurst et al., 2024) are shown in Table 1. Despite its age and being significantly smaller, CometKiwi performs similarly on German and outperforms the other models on Chinese, with the overall best evaluation performance for the non-English generated text, demonstrating that the "sophisticated mechanisms" needed in current reward models already exist. Building on this observation, the MetaMetrics approach (Anugraha et al., 2024) has been evaluated on the latest MT metrics shared task (Freitag et al., 2024) and on the RewardBench leaderboard (Lambert et al., 2025), scoring highly in both, although not with the same model.

4.2 REWARD MODELS CAN UNDERPERFORM ON METRICS BENCHMARKS

Another area in which reward model and metrics benchmarks are aligned in their goals is the assessment of how well models can assess factuality and attribution. There exist benchmarks for metrics (Honovich et al., 2022), model performance (Jacovi et al., 2025), and reward models (Malik et al., 2025) that assess factuality. Recent work demonstrates the effectiveness of LLM judge (and reward) models (e.g., Calderon et al., 2025; Hashemi et al., 2024), some even finding that dedicated finetuned evaluation models underperform LLM judges (Huang et al., 2025).

Among these benchmarks, the metrics benchmark SEAHORSE (Clark et al., 2023) is the largest with over 100,000 human judgments of summarization quality aspects across multiple languages. For this experiment, we prompt various LLMs with the same instructions provided to human annotators in SEAHORSE to give a binary judgment whether a summary is attributable to an article.[4] Due to a lack of data availability, we exclude the WikiLingua (Ladhak et al., 2020) subset of SEAHORSE and focus only on XLSum (Hasan et al., 2021) and MLSum (Scialom et al., 2020), retaining 7,793 of the 18,330 test examples. We report Pearson correlation (ρ) and accuracy.

The results (Table 2) show that LLMs underperform the dedicated model trained on in-domain data. This remains true even if we assess judge models with a high reasoning budget like Gemini 2.5 Pro and GPT-5. In fact, the two reasoning models have an 89% agreement rate, higher than the inter-rater agreement of 73% reported in the paper, indicating that the models look for similar input and output-features to make their prediction. The results are fairly consistent across languages. Interestingly, all models score lowest on English among the evaluated languages.

Overall, our findings disagree with Huang et al. (2025) who show that LLM judges can outperform dedicated metrics in similar setups, while agreeing with Bavaresco et al. (2025) who show low LLM judge correlations for summarization evaluation. Multiple explanations exist for the results presented here, including annotation artifacts that cause a lower performance of the LLM judge setup. However, we can conclude that for evaluating attribution for summarization, it remains unclear whether LLM judges have caught up to dedicated models, a question that requires further rigorous study. This conclusion mirrors the argument by Chehbouni et al. (2025) that the "rapid and widespread adoption [of LLM judges] may have occurred prematurely". Moreover, as shown in Table 2, the LLM judge setup outperforms a strong Natural Language Inference model (NLI) baseline (Conneau et al., 2018). As such, this setup could still be useful in cases where dedicated training data for a reward model or evaluation metric is unavailable.

[4]We optimized performance on the validation set to minimize the effect of the prompt format.

	ρ	Acc. %
ROUGE-L	0.13	
mT5_{XNLI}	0.43	
$\text{mT5}_{\text{SEAHORSE}}$	0.59	
GPT-4o	0.47	73.3
Gemini 2.0 Flash	0.42	72.1
Gemini 2.5 Flash	0.48	73.8
Claude Sonnet 4	0.45	70.1
+ Reasoning		
Gemini 2.5 Pro	0.50	73.9
GPT-5	0.47	70.2

Table 2: Pearson ρ coefficient and binary prediction accuracy on SEAHORSE for identifying whether a summary is **attributable** to a source article. The baselines are finetuned mT5_{XXL} models by Clark et al. (2023). The LLM-as-a-judge approach is outperformed by a dedicated trained metric.

5 Metrics and reward models are (not) the same

Metrics and reward models both judge quality aspects of generated content with the goal of being aligned with human preferences. Yet, they are not the same: they can differ in their design, application, training, and testing. To explore these aspects, we provide a survey of the two fields and discuss themes where a closer interaction could lead to mutually helpful insights.

5.1 Designing Reward Models and Evaluation Metrics

Sociotechnical Context matters Evaluation metrics tend to be narrowly focused on specific quality aspects. These quality aspects should follow clear and standardized definitions such that the metrics are transferable and produce scores that are understandable across organizations. A lack of transparency in how metrics are designed and the subsequent lack of reproducibility has been subject of much past criticism (Rankel et al., 2013; Post, 2018; Gehrmann et al., 2023).

In contrast, if a reward model is the provider of training signals during reinforcement learning, it must therefore be able to score a myriad of tasks and output types. As such, modeling human preferences encompasses many aspects of preference beyond output quality, including whether a model correctly refuses undesired requests or avoids producing toxic language (Bai et al., 2022b). These judgments depend on the specific application the model is used for and the policies governing this application. Reward models that measure these application-specific aspects are inherently less transferrable and tied to the specific organizations that develop them (Gehrmann et al., 2025). Following this reasoning, these reward models are inherently not comparable to another, which calls into question the utility of non-specific reward-modeling benchmarks.

Aspect-aligned reward models Fine-grained assessments of (partial) generations are areas with extensive recent work in RL that more closely align with work on evaluation metrics (Gunjal et al., 2025; Lightman et al., 2024). It is of particular interest, since a diverse set of reward signals can mitigate issues that arise from single-objective optimization (Freitag et al., 2021b; Zhang et al., 2024; Fisch et al., 2024). A popular approach for this is to use reward models that score rubrics instead of providing generic preferences (Gunjal et al., 2025). Rubrics are fine-grained evaluation criteria (Arora et al., 2025; Hashemi et al., 2024), similar to those traditionally assessed by dedicated metrics. Rubric-based prompted scoring, alongside learned reward models, is mentioned as an instrumental ingredient for post-training of models like Gemini 2.5 (Comanici et al., 2025).

Successfully assessing rubrics requires clear definitions of the evaluation categories (Howcroft et al., 2020). Yet, even for popular concepts like "hallucination rate", definitions can vary widely (Maynez et al., 2020; Rashkin et al., 2023; Ji et al., 2023). Increasing the consistency of these definitions will be crucial as reward models become more specific, and thus are designed more similar to evaluation

metrics where the topic of fine-grained assessments is well-studied (e.g., Eyal et al., 2019; Wang et al., 2020; Fabbri et al., 2021; Scialom et al., 2021; Lee et al., 2024b; Wei et al., 2025).

5.2 Training Reward Models and Evaluation Metrics

Data Collection The data collection methodology for any model must be aligned with its design goals. For reward models, this means reflecting the preferences of the intended audience of the downstream model, which can be extremely broad. Since many aspects of generated text cannot be objectively assessed, this necessitates collecting feedback from diverse sources (Casper et al., 2023; Metz et al., 2025). The culture and lived experience of raters can lead to drastically different subjective preference judgments (e.g., Aroyo et al., 2023; Rastogi et al., 2024; 2025).

Another critical question to consider is whether the selected raters have sufficient expertise, as changes in annotation quality can lead to drastically different insights (Freitag et al., 2021a; Wei et al., 2024). In many cases, existing metrics and reward models already outperform non-expert raters, and only the highest quality annotations can further improve the models (Cui et al., 2023; Liu et al., 2024; Wen et al., 2025b). Moreover, Wen et al. (2025a) find that RL may produce errors that are increasingly difficult for humans to detect. However, hiring raters with expertise to judge long-form generation is notoriously challenging (Zhang et al., 2023).

Optimization targets Design aspects such as access to a ground truth and the output format (pairwise preferences, categorical labels, or continuous scores) influence how models are developed. These choices depend on the downstream use case, and can have significant impact on model efficacy. For example, reference-less evaluation has improved significantly in recent years but still underperforms reference-based metrics (Ma et al., 2019; Freitag et al., 2024). Similar results were found for LLM-as-a-judge setups (Krumdick et al., 2025). These choices are reflected in benchmarking practices; many reward models produce pairwise comparisons, and their benchmarks consequently focus on this binary setup (Frick et al., 2025). In contrast, metrics typically generate continuous outputs, allowing for more flexible evaluation. By focusing primarily on pairwise judgments, reward model development may be ignoring the potential benefits of continuous scoring.

Another shared goal is the development of lightweight models that can run efficiently alongside larger models during inference or training. Advances in distillation, quantization, parallelization, and pruning are therefore highly relevant to both fields. Consequently, approaches to model compression that seek to train student models to outperform their teachers can equally benefit the development of both reward models and evaluation metrics (Kim et al., 2024; Sun et al., 2023).

5.3 Testing Reward Models and Evaluation Metrics

Identifying and Debugging Reward Hacking A lack of correlation between reported reward model and downstream RL model performance has been attributed to limitations of the reward model (Ivison et al., 2024; Kim et al., 2025; Wen et al., 2025c). When the reward model does not robustly generalize, or focuses on spurious correlations, it can lead to *reward hacking*. Amodei et al. (2016) describe reward hacking as the process of "gaming" flaws in the reward model to maximize the rewards without learning the intended behavior. This phenomenon was empirically observed for text (Pasunuru & Bansal, 2017; Kryściński et al., 2018; Wu et al., 2018a) and non-text RL (Amodei & Clark, 2016; Krakovna et al., 2020; Nagarajan et al., 2021).

It is not specific to reward models, as most classification models suffer from spurious correlations (Ribeiro et al., 2016; McCoy et al., 2019) and spurious correlations were found in reward models (Liu et al., 2025b) and metrics (Sun et al., 2019). Among the effects, reward models may prefer more confident-sounding answers (Leng et al., 2025), exhibit a verbosity bias (Saito et al., 2023), focus more on style than content (Feuer et al., 2025), and results may be confounded by the order in which outputs are shown (Wang et al., 2024). Relatedly, the problem of *sycophancy* has been characterized as models learning to match user beliefs over generating truthful responses (Sharma et al., 2024). Murugadoss et al. (2025) and Hu et al. (2024) further show that the detail of LLM-as-a-judge prompts have little influence on its performance, implying that models rely too much on their implicitly learned quality criteria definitions. These issues motivate work on diagnostic datasets (Gabriel et al., 2021), distractor generation (Qiu et al., 2020; Dhole et al., 2023), and model interpretability (Jacovi et al., 2023), to become aware of and overcome spurious correlations.

Meta-Evaluation Frameworks The field of meta-evaluation is concerned with the question of how we evaluate evaluators. Callison-Burch et al. (2007) popularized this practice in NLP through a shared task series that performs a yearly assessment of MT metrics. Meta-evaluation measures two aspects: *segment-level* and *system-level* performance. A high system-level performance means that system rankings in a leaderboard are trustworthy, while segment-level assessments look at whether individual pairs of system outputs are ranked correctly. These two measures are not always correlated (Wei & Jia, 2021), motivating an approach that matches how a model is used.

Algorithms like DPO (Rafailov et al., 2023) use the reward score difference between a chosen and rejected model output as training signal. This directly matches the segment-level meta-evaluation. However, a known issue is that evaluation metrics are often not well-calibrated (Kocmi et al., 2024), which may cause issues if they are applied as reward models. Moreover, reward model benchmarks like RewardBench 2 (Lambert et al., 2025; Gureja et al., 2025) do not consider score calibration, instead reporting overall accuracy on the task of identifying the highest rated system output, which more closely matches a system-level assessment. As a result, calibration issues may be overlooked if one focuses only on reward model benchmark performance. This oversight of segment-level assessments could further contribute to the lack of correlation between reward model and downstream model performance. Thus, future work on reward model benchmarking could benefit from reporting segment-level rather than system-level performance, including assessments of score calibration.

Meta-Evaluation Targets A complicating factor for the meta-evaluation of reward models is the breadth of tasks for which they need to assess output quality. Their meta-evaluations thus need to strike a balance of breadth, validity, and relevance. Ivison et al. (2024) suggest that existing reward model benchmarks are too narrow, especially considering their performance variance across tasks (Bavaresco et al., 2025). Benchmarks like RewardBench 2 already average multiple categories, but the question of how to aggregate sub-category scores into a single ranking becomes important. To that end, Frick et al. (2025) find that pessimistic reward model evaluations instead of average performance are more indicative of downstream model performance, motivating alternative leaderboard designs that focus on finding shortcomings, rather than averaging performance numbers. These issues are further exacerbated when the systems that are being evaluated by evaluation metrics and reward models improve. As these models become harder to distinguish, biases in the evaluation setup become more noticeable (Wei & Jia, 2021) and tie handling procedures need to be introduced (Thompson et al., 2024; Sun et al., 2025).

5.4 Recommendations

While developing an unhackable reward model is likely impossible (Skalse et al., 2022), metrics, and more directly reward models, share a symbiotic relationship with the downstream models where improvements in one translate to improvements in the other (Gehrmann et al., 2021). This means, we should strive to produce the most accurate estimate of human preferences. To that end, both fields benefit from having high-quality training and meta-evaluation data. This data needs to be grounded in clear definitions in the sociotechnical context that the to-be-assessed models are deployed in. While it is unavoidable to introduce spurious correlations, in both fields it is crucial to identify and measure them and to mitigate their impact on downstream uses.

Modeling choices and optimization targets similarly align between the two fields, whether that is applying LLM-as-a-judge, or training classification models on human-curated data. Due to this overlap, newly introduced methods for modeling human preferences should be evaluated on metrics and reward model benchmarks alike to paint a more accurate and complete picture of how these methods perform. More generally, meta-evaluation and the development of leaderboards is an area in which the fields have significantly diverged. They should come together to address the poor correlation between reward model benchmark scores and downstream model performance. Shared best practices on tie handling, segment-level correlation measures, conducting model calibration assessments, and collecting test datasets will benefit both fields.

However, as reward modeling matures as a field, it will be important to avoid falling into traps like developing default models and benchmarks that, despite being outdated, continue to be broadly used (Bommasani & Cardie, 2020). Instead, evaluation research could adopt the practice from reward modeling of moving to new and better performing model as they become available.

While reward models and evaluation metrics should be developed using the same best practices, one cannot be used to replace the other. As Goodhart's law states, *when a measure becomes a target, it ceases to be a good measure* (Goodhart, 1984). Applying to both fields, models may perform well as a proxy for the distribution over human preferences but will diverge in the tail (i.e., rarely seen model inputs) and may over-generalize and focus on spurious patterns (Manheim & Garrabrant, 2018; Gao et al., 2023). Similar arguments apply to the utility of shared tasks and leaderboards which similarly are a frequent target of criticism (e.g., Scott & Moore, 2006; Ethayarajh & Jurafsky, 2020; Thomas & Uminsky, 2020; Raji et al., 2021; Bowman & Dahl, 2021). Our practical recommendation is thus for the fields to share insights into methodologies, but not to collapse into one field.

A full collapse would risk creating a monoculture where only a few benchmarks dictate model optimization, rather than having many different targets (Koch & Peterson, 2024). As Singh et al. (2025) state, an "over-reliance on a single leaderboard creates a risk that providers may overfit to the aspects of leaderboard performance, without genuinely advancing the technology in meaningful ways". Issues with broadly adopted evaluation setups lead to overspecialization and lack of generalization beyond what a specific leaderboard measures (Liu et al., 2025a; Zouhar et al., 2024). Being too rigid in how meta-evaluations are conducted could also exclude new methods from being investigated fairly (Perrella et al., 2024).

Furthermore, while methods for reward modeling may be informed by insights from evaluation metric development, the specific reward models may not perform well on the same benchmarks. As discussed above, reward models are often specific to a sociotechnical context, and would thus not perform well on public reward modeling benchmarks. This may cause a rift between industry and academic research where the best reward modeling approaches are not publicly disclosed because they are too entangled within this context. Yet, especially for models that measure human preferences for whether a model output is considered offensive or undesirable, it is critical to develop public standards and be transparent about the underlying policies a model is trying to enact.

6 Conclusions

In this work, we have argued that evaluation metrics and reward models share many similarities. Their developers need to make the same choices about their inputs and outputs, their collection of training and validation data, and the resulting models suffer from the same drawbacks. While the application areas and the specific choices made during development may differ, at their core, both seek to model human preferences of model output quality. This supports our thesis that the fields should look at and learn from each other's advances, rather than continuing to exist in parallel.

We grounded this discussion in a citation analysis that demonstrated that the research fields are developing mostly in isolation from each other. This separation of fields can lead to missed opportunities, the rediscovery of established findings, and potentially flawed conclusions. We quantified the separation through two experiments that show that reward models may be lacking when assessed on domains for which evaluation metrics are already available. We provided an extensive survey and discussed several areas in which future work on metrics and reward models, meta-evaluations, and benchmark creation could incorporate insights from both fields. While we recommend against the development of a monoculture with too few relevant benchmarks, we encourage researchers to consider work from both fields and work on unifying both methodologies and terminologies.

Beyond the scope of this work, we acknowledge efforts in reinforcement learning to solve tasks with verifiable rewards, for example math problems (Ke et al., 2025), for which reward models play a less central role. In this domain, model-based reward modeling approaches largely perform string matching between the verified and generated answer, and thus do not require as complex approaches as those discussed here. Training models for these verifiable domains can induce reasoning capabilities and has led to broader generalization (Guo et al., 2025; Comanici et al., 2025). While this finding does not make reward models for non-verifiable domains obsolete, it presents a possible alternative or parallel path in which reward models do not play such a central role. Moreover, we note that improvements in reward models may not always translate into downstream model improvements. The *superficial alignment hypothesis* by Zhou et al. (2023) poses that the reinforcement learning stage primarily changes 'how' a model responds, rather than contributing new world knowledge. Thus, even a perfect reward model cannot overcome fundamental knowledge gaps from pre-training.

REFERENCES

Dario Amodei and Jack Clark. Faulty reward functions in the wild. https://openai.com/index/faulty-reward-functions/, 2016.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 459–469, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.32. URL https://aclanthology.org/2024.wmt-1.32/.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *CoRR*, abs/2505.08775, 2025. doi: 10.48550/ARXIV.2505.08775. URL https://doi.org/10.48550/arXiv.2505.08775.

Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. The reasonable effectiveness of diverse evaluation data. *arXiv preprint arXiv:2301.09406*, 2023.

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.7755.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJDaqqveg.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022a. URL https://api.semanticscholar.org/CorpusID:248118878.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022b. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce

Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Rishi Bommasani and Claire Cardie. Intrinsic evaluation of summarization datasets. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8075–8096, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.649. URL https://aclanthology.org/2020.emnlp-main.649/.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.782. URL https://aclanthology.org/2025.acl-long.782/.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz (eds.), *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0718/.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. Neither valid nor reliable? investigating the use of llms as judges. *arXiv preprint arXiv:2508.18076*, 2025.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280/.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. On the weaknesses of reinforcement learning for neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eCw3EKvH.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9397–9413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.584. URL https://aclanthology.org/2023.emnlp-main.584/.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://api.semanticscholar.org/CorpusID:52271711.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Hal Daumé, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine Learning*, 75(3):297–325, 2009. doi: 10.1007/S10994-009-5106-X. URL https://doi.org/10.1007/s10994-009-5106-x.

Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahadiran, Simon Mille, Ashish Shrivastava, Samson Tan, Tongshang Wu, Jascha Sohl-Dickstein, Jinho Choi, Eduard Hovy, Ondřej Dušek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Tanya Goyal, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honoré, Ishan Jindal, Przemysław Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxine Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicholas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos Samus, Ananya Sai, Robin Schmidt, Thomas Scialom,

Tshephisho Sefara, Saqib Shamsi, Xudong Shen, Yiwen Shi, Haoyue Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, Aditya Srivatsa, Tony Sun, Mukund Varma, A Tabassum, Fiona Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Zijie Wang, Gloria Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyu Wu, Witold Wydmanski, Tianbao Xie, Usama Yaseen, Michael Yee, Jing Zhang, and Yue Zhang. NL-augmenter: A framework for task-sensitive natural language augmentation. *Northern European Journal of Language Technology*, 9, 2023. doi: 10.3384/nejlt.2000-1533.2023.4725. URL https://aclanthology.org/2023.nejlt-1.5/.

Richard O. Duda and Peter E. Hart. Pattern classification and scene analysis. In *A Wiley-Interscience publication*, 1974. URL https://api.semanticscholar.org/CorpusID:12946615.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393/.

Matan Eyal, Tal Baumel, and Michael Elhadad. Question answering as an automatic evaluation metric for news article summarization. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3938–3948, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1395. URL https://aclanthology.org/N19-1395/.

Alexander R. Fabbri, Chien Sheng Wu, Wenhao Liu, and Caiming Xiong. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *North American Chapter of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:245218667.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P Dickerson. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MzHNftnAM1.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*, 2024.

Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 61–71, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.5. URL https://aclanthology.org/2020.emnlp-main.5/.

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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021a. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2021b. URL https://api.semanticscholar.org/CorpusID:248392447.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 47–81, Miami,

Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.2. URL https://aclanthology.org/2024.wmt-1.2/.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cbttLtO94Q.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. GO FIGURE: A meta evaluation of factuality in summarization. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 478–487, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.42. URL https://aclanthology.org/2021.findings-acl.42/.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In Antoine Bosselut, Esin Durmus, Varun Prashant Gangal, Sebastian Gehrmann, Yacine Jernite, Laura Perez-Beltrachini, Samira Shaikh, and Wei Xu (eds.), *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 96–120, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10/.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal Artificial Intelligence Research*, 77:103–166, 2023. doi: 10.1613/JAIR.1.13715. URL https://doi.org/10.1613/jair.1.13715.

Sebastian Gehrmann, Claire Huang, Xian Teng, Sergei Yurovski, Arjun Bhorkar, Naveen Thomas, John Doucette, David Rosenberg, Mark Dredze, and David Rabinowitz. Understanding and mitigating risks of generative ai in financial services. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025. URL https://api.semanticscholar.org/CorpusID:278170940.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pp. 91–121. Springer, 1984.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *CoRR*, abs/2507.17746, 2025. doi: 10.48550/ARXIV.2507.17746. URL https://doi.org/10.48550/arXiv.2507.17746.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-RewardBench: Evaluating reward models in multilingual settings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43–58, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.3. URL https://aclanthology.org/2025.acl-long.3/.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413/.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745. URL https://aclanthology.org/2024.acl-long.745/.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada (eds.), *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.inlg-1.23. URL https://aclanthology.org/2020.inlg-1.23/.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. Are LLM-based evaluators confusing NLG quality criteria? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9530–9570, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.516. URL https://aclanthology.org/2024.acl-long.516/.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.306. URL https://aclanthology.org/2025.findings-acl.306/.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633, 2024.

Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. Diagnosing ai explanation methods with folk concepts of behavior. *Journal of Artificial Intelligence Research*, 78:459–489, 2023.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *CoRR*, abs/2501.03200, 2025. doi: 10.48550/ARXIV.2501.03200. URL https://doi.org/10.48550/arXiv.2501.03200.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13252–13280, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.649. URL https://aclanthology.org/2025.acl-long.649/.

Kimi, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023. URL https://api.semanticscholar.org/CorpusID:256194545.

Bernard J. Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution. *CoRR*, abs/2404.06647, 2024. doi: 10.48550/ARXIV.2404.06647. URL https://doi.org/10.48550/arXiv.2404.06647.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1999–2014, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.110. URL https://aclanthology.org/2024.acl-long.110/.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/, 2020.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*, 2025.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. Improving abstraction in text summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1808–1817, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1207. URL https://aclanthology.org/D18-1207/.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4034–4048, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.360. URL https://aclanthology.org/2020.findings-emnlp.360/.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Yann LeCun. Predictive learning. https://www.youtube.com/watch?v=Ount2Y4qxQo&t=1072s, 2016.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=uydQ2W41KO.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. *arXiv preprint arXiv:2403.18771*, 2024b.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *ArXiv*, abs/1811.07871, 2018. URL https://api.semanticscholar.org/CorpusID:53745764.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Piji Li, Lidong Bing, and Wai Lam. Actor-critic based training framework for abstractive summarization. *CoRR*, abs/1803.11070, 2018. URL http://arxiv.org/abs/1803.11070.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

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

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

Siqi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. Re-evaluating open-ended evaluation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=kbOAIXKWgx.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, Daniel Sohn, Anastasia Makarova, Jeremiah Zhe Liu, Yuan Liu, Bilal Piot, Abe Ittycheriah, Aviral Kumar, and Mohammad Saleh. RRM: Robust reward model training mitigates reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=88AS5MQnmC.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Rm-bench: Benchmarking reward models of language models with subtlety and style. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025c. URL https://openreview.net/forum?id=QEHrmQPBdd.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 671–688, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6450. URL https://aclanthology.org/W18-6450/.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 62–90, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302. URL https://aclanthology.org/W19-5302/.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *CoRR*, abs/2506.01937, 2025. doi: 10.48550/ARXIV.2506.01937. URL https://doi.org/10.48550/arXiv.2506.01937.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. Results of the WMT20 metrics shared task. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 688–725, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.77/.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Yannick Metz, Andras Geiszl, Raphaël Baur, and Mennatallah El-Assady. Reward learning from multiple feedback types. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9Ieq8jQNAl.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2204–2212, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html.

Bhuvanashree Murugadoss, Christian Pölitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. Evaluating the evaluator: Measuring llms' adherence to task evaluation instructions. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 19589–19597. AAAI Press, 2025. doi: 10.1609/AAAI.V39I18.34157. URL https://doi.org/10.1609/aaai.v39i18.34157.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fSTD6NFIW_b.

OpenAI. Gpt-5 system card, Aug 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 979–985, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1103. URL https://aclanthology.org/D17-1103/.

Ramakanth Pasunuru and Mohit Bansal. Multi-reward reinforced summarization with saliency and entailment. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 646–653, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2102. URL https://aclanthology.org/N18-2102/.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkACIQgA-.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16216–16244, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.856. URL https://aclanthology.org/2024.acl-long.856/.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Zhaopeng Qiu, Xian Wu, and Wei Fan. Automatic distractor generation for multiple choice questions in standard tests. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2096–2106, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.189. URL https://aclanthology.org/2020.coling-main.189/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In Hinrich Schuetze,

Pascale Fung, and Massimo Poesio (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 131–136, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2024/.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, December 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Zoe Ashwood, Aida Mostafazadeh Davani, Mark Diaz, Michela Paganini, Alicia Parrish, Ding Wang, Vinodkumar Prabhakaran, Lora Aroyo, and Verena Rieser. Insights on disagreement patterns in multimodal safety perception across diverse rater groups. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=8TI0lUrJBP.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Díaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, et al. Whose view of safety? a deep dive dataset for pluralistic alignment of text-to-image models. *arXiv preprint arXiv:2507.13383*, 2025.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60/.

Ehud Reiter. A structured review of the validity of BLEU. *Comput. Linguistics*, 44(3), 2018. doi: 10.1162/COLI_A_00322. URL https://doi.org/10.1162/coli_a_00322.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1179–1195. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.131. URL https://doi.org/10.1109/CVPR.2017.131.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *CoRR*, abs/2310.10076, 2023. doi: 10.48550/ARXIV.2310.10076. URL https://doi.org/10.48550/arXiv.2310.10076.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Answers unite! unsupervised metrics for reinforced summarization models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3246–3256, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1320. URL https://aclanthology.org/D19-1320/.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. MLSUM: The multilingual summarization corpus. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8051–8067, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.647. URL https://aclanthology.org/2020.emnlp-main.647/.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.529. URL https://aclanthology.org/2021.emnlp-main.529/.

Donia Scott and Johanna Moore. An nlg evaluation competition? eight reasons to be cautious. *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, 2006.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1683–1692, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1159. URL https://aclanthology.org/P16-1159/.

Raphael Shu, Kang Min Yoo, and Jung-Woo Ha. Reward optimization for neural machine translation with learned metrics. *CoRR*, abs/2104.07541, 2021. URL https://arxiv.org/abs/2104.07541.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Joar Max Viktor Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yb3HOXO3lX2.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking reward modeling in preference-based large language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rfdblE10qm.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. Dialect-robust evaluation of generated text. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6010–6028, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.331. URL https://aclanthology.org/2023.acl-long.331/.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf (eds.), *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 21–29, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2303. URL https://aclanthology.org/W19-2303/.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction, 2nd Edition*. MIT Press, 2018. URL http://www.incompleteideas.net/book/the-book-2nd.html.

Rachel Thomas and David Uminsky. The problem with metrics is a fundamental problem for ai. *arXiv preprint arXiv:2002.08512*, 2020.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1222–1234, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.118. URL https://aclanthology.org/2024.wmt-1.118/.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15,*

2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Johnny Wei and Robin Jia. The statistical advantage of automatic NLG metrics at the system level. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 6840–6854, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.533. URL https://aclanthology.org/2021.acl-long.533/.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. Rocketeval: Efficient automated LLM evaluation via grading checklist. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zJjzNj6QUe.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=xJljiPE6dg.

Xueru Wen, Jie Lou, Zichao Li, Yaojie Lu, XingYu XingYu, Yuqiu Ji, Guohai Xu, Hongyu Lin, Ben He, Xianpei Han, Le Sun, and Debing Zhang. Cheems: A practical guidance for building and evaluating Chinese reward models from scratch. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15187–15211, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.737. URL https://aclanthology.org/2025.acl-long.737/.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, XingYu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=Cnwz9jONi5.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Sam Wiseman and Alexander M. Rush. Sequence-to-sequence learning as beam-search optimization. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1296–1306, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1137. URL https://aclanthology.org/D16-1137/.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. A study of reinforcement learning for neural machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3612–3621, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1397. URL https://aclanthology.org/D18-1397/.

Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial neural machine translation. In Jun Zhu and Ichiro Takeuchi (eds.), *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pp. 534–549. PMLR, 14–16 Nov 2018b. URL https://proceedings.mlr.press/v95/wu18a.html.

xAI. Grok 4 model card, Aug 2025. URL https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

Go Yasui, Yoshimasa Tsuruoka, and Masaaki Nagata. Using semantic similarity as reward for reinforcement learning in sentence generation. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics: Student Research Workshop, pp. 400–406, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2056. URL https://aclanthology.org/P19-2056/.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In Satinder Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 2852–2858. AAAI Press, 2017. doi: 10.1609/AAAI.V31I1.10804. URL https://doi.org/10.1609/aaai.v31i1.10804.

Wojciech Zaremba and Ilya Sutskever. Reinforcement learning neural turing machines. *CoRR*, abs/1505.00521, 2015. URL http://arxiv.org/abs/1505.00521.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14944–14982, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.835. URL https://aclanthology.org/2023.acl-long.835/.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:267095304.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KBMOKmX2he.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. RMB: comprehensively benchmarking reward models in LLM alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=kmgrlG9TR0.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. A preference-driven paradigm for enhanced translation with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3385–3403, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.186. URL https://aclanthology.org/2024.naacl-long.186/.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. Fine-tuned machine translation metrics struggle in unseen domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 488–500, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.45. URL https://aclanthology.org/2024.acl-short.45/.

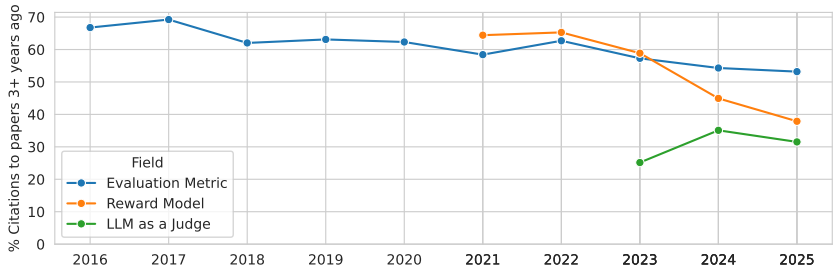


Figure 3: We show the percentage of citations to papers that were published more than three years ago. Reward model literature exhibits outlier behavior in which this percentage is decreasing drastically every year.

A Detailed Methods for Citation Analysis

The results in the main text present a high-level overview of the key findings of our citation analysis. We chose to present results for the last 5 years since reward models that resemble those discussed in this paper were only popularized in this time frame. While the Semantic Scholar API yields results for the years prior, they are mostly irrelevant to the discussion at hand. We extended the analysis for evaluation metrics back to 2016 but find no significant difference in results. Therefore, we omit them for readability.

Similarly, we choose to present results of an analysis of up to 300 papers per year as a result of qualitative assessment of the data. A qualitative assessment found that, beyond the first 300 papers, search results became too noisy, with irrelevant papers being retrieved. We repeated the analysis with only the top 100 and 200 papers with no significant differences in results.

The specific keywords for the analysis in Figure 2a were selected based on repeated trials to maximize precision at the cost of recall. For example, including "evaluation" as a proxy keyword for evaluation metrics would have yielded any paper that discusses their evaluation results, not necessarily discusses how to build evaluation metrics. Similarly, we included generic RL-related terms like "policy" for reward models since the terminology was evolving and papers only fairly recently converged on this term. To not miss citations to relevant papers prior to 2020, we included them at the risk of overestimating the true citation count.

As a result, the specific numerical results are a side effect of this keyword-based identification and should be interpreted with caution. While an LLM-based identification process may yield more accurate results, it would require processing a significant number of tokens. Since we were only interested in aggregate trend information, we found the results from keyword-based searches sufficient and stable across many variants.

B Recency Bias in Citations

We quantify recency bias in citations across the three fields of study. Citations only to recent papers would provide an additional piece of evidence that insights from work before LLMs became popular are not being considered. Indeed, we find that while the average age of a cited paper for evaluation metrics published in 2025 is 5.0 years, cited papers by reward modeling and LLM-as-a-judge papers are only 3.6 and 3.8 years old. 62.2% of citations in reward modeling papers are to papers published less than 2 years ago (68.5% for LLM-as-a-Judge), in contrast to 46.8% for evaluation metrics. Critically, Figure 3 shows how citations to older papers have been decreasing, especially in literature on reward models. This result is an indicator that reward modeling research is evolving quickly and that benchmarks are quickly made irrelevant by new results.

C ASSIGNMENTS OF CONFERENCE TO SUBFIELD

For our analysis of citations to research areas in Section 3, we assign academic venues to an area if the venue is clearly affiliated with it. For example, AAAI's scope is all of AI and we therefore do not include it in this analysis. We include a venue in this analysis if papers published there received at least 50 citations among all the 50,000+ papers included in our analysis. We account for various misspellings, capitalization differences, and abbreviations, but only list each venue once in the following list of assignments:

Machine Learning COLT, ICLR, ICML, JMLR, NeurIPS / NIPS, TMLR, TNNLS

Natural Language Processing ACL (including Findings of ACL), CONLL, EACL, EAMT, EMNLP, INLG, LREC, NAACL, SemEval, WMT

Robotics CoRL, ICRA, IROS, IJRR, RSS

Vision CVPR, ECCV, ICCV, IJCV, MICCAI, TIP, TOG, WACV

D USE OF LLMS FOR WRITING

After manually writing the entire draft, we used ChatGPT and Google Gemini to improve its writing. Specifically, we prompted them to find spelling and grammatical errors and suggest improvement to sentence structures. Additionally, we prompted them to provide feedback on whether the argument structure was clear throughout all paragraphs, asking for specific improvements can can be made. All suggestions were manually reviewed.