First Tragedy, then Parse: History Repeats Itself in the New Era of Large Language Models

Naomi Saphra
Kempner Institute at Harvard University
nsaphra@fas.harvard.edu

Eve Fleisig
University of California - Berkeley
efleisig@berkeley.edu

Kyunghyun Cho
New York University & Genentech
kyunghyun.cho@nyu.edu

Adam Lopez
University of Edinburgh
alopez@inf.ed.ac.uk

Abstract

Many NLP researchers are experiencing an existential crisis triggered by the astonishing success of ChatGPT and other systems based on large language models (LLMs). After such a disruptive change to our understanding of the field, what is left to do? Taking a historical lens, we look for guidance from the first era of LLMs, which began in 2005 with large n-gram models for machine translation (MT). We identify durable lessons from the first era, and more importantly, we identify evergreen problems where NLP researchers can continue to make meaningful contributions in areas where LLMs are ascendant. We argue that disparities in scale are transient and researchers can work to reduce them; that data, rather than hardware, is still a bottleneck for many applications; that meaningful realistic evaluation is still an open problem; and that there is still room for speculative approaches.

1 Introduction

Picture this scene: A renowned NLP researcher at a hot seven-year-old startup steps onstage to deliver a keynote. The speaker describes an ambitious new system to the packed room, building up to the results slide: a bar chart in which the x-axis shows the number of training words, and the y-axis shows system accuracy. As each data point is revealed, performance rises relentlessly, culminating in a system trained on well over a trillion words using over a thousand processor cores. It smashes the state of the art by a margin previously thought impossible.

Attendees are visibly shaken as they realize, over the course of a minute, that years of research have just been rendered utterly inconsequential. Established academics panic, anticipating the wholesale rejection of already-submitted grant applications. PhD students despair, contemplating the irrelevance of their unfinished dissertations. Many ponder an exit to industry or a change of fields. They will speak of little else this week.

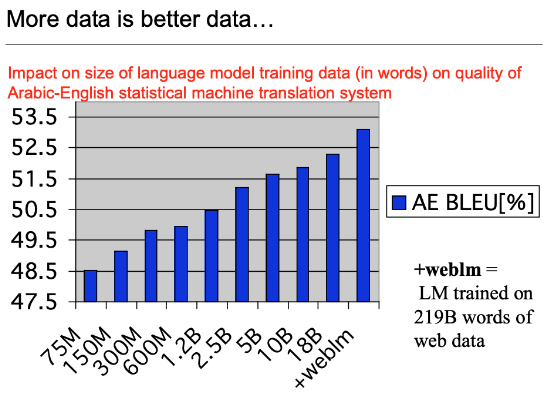

Figure 1: Results slide (reproduced from Och, 2005) of Franz Och's keynote talk at the 2005 ACL Workshop on Building and Using Parallel Texts, a predecessor to the Conference on Machine Translation.

Does this scene sound like one that might have happened in the past year? In fact, it happened 19 years ago, in 2005, launching the first era of Large Language Models (LLMs): the **Statistical Machine Translation (SMT) era**. The speaker, Franz Och, had co-invented key methods in SMT (Och and Ney, 2003; Koehn et al., 2003; Och, 2003), but had not published new work since joining Google in 2004, instead revealing it in an invited talk prior to the launch of a new Google Translate (Och, 2006).[1] The provocative results slide from that talk (Figure 1) shows how Google improved its SMT system simply by expanding the training corpus of a phrase-based language model (Brants et al., 2007).[2]

[1]The description of the talk and its aftermath is based on the vivid recollections of one of the authors, who was present.

[2]By **language model** (LM), we mean a generative probabilistic model $\Pr(\mathbf{x})$ of a string $\mathbf{x}$. MT requires a *conditional* LM $\Pr(\mathbf{y} \mid \mathbf{x})$ of target string $\mathbf{y}$ given source string $\mathbf{x}$. In SMT it was originally modeled using the *noisy channel* formulation as $\Pr(\mathbf{y} \mid \mathbf{x}) \propto \Pr(\mathbf{y}) \times \Pr(\mathbf{x} \mid \mathbf{y})$ (Brown et al., 1993). The *translation model* $\Pr(\mathbf{x} \mid \mathbf{y})$ must be trained on a corpus of example translations, but the LM $\Pr(\mathbf{y})$ can be trained on *any* data in the target language, making it amenable to scaling. Like modern LLMs, LMs of the SMT era were generative probabilistic models, albeit based on n-grams (Shannon,

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2310–2326
June 16-21, 2024 ©2024 Association for Computational Linguistics

The first era of LLMs initially provoked great anxiety among MT researchers about the state of their field, but MT research has continued to flourish in academia, industry, and government. Even in the modern era of deep learning, MT has been a locus of innovations that have fundamentally altered NLP and all of machine learning (Bahdanau et al., 2015; Vaswani et al., 2017; Sutskever et al., 2014).

We believe that this history offers lessons for the *current* era of LLMs, an era during which massive proprietary models have become a de facto baseline for many tasks (Rogers, 2023). The expense of state-of-the-art research has led many to question the role of smaller and publicly funded groups in AI (Lee et al., 2023), a phenomenon we will call the **scale crisis**.[3] Researchers without direct access to LLMs have publicly fretted over their research directions, with Togelius and Yannakakis (2023) suggesting pivots in research direction to sidestep scale, and Ignat et al. (2023) sketching research areas that are "not within the purview of LLMs." But what should researchers do if they care about problems that *are* within the purview of LLMs? To answer this question, we look to the first era of LLMs. What were the durable lessons of that time and evergreen research problems that still matter today? We arrive at several recurrent lessons:

Scale is supreme (Section 2). We argue that, for areas where data is plentiful, NLP researchers cannot escape the **Bitter Lesson** (Sutton, 2019) that general purpose methods exploiting scale will outperform methods that leverage informed priors. We recommend that researchers take advantage of improvements in hardware as they enable scale at affordable budgets (Section 2.1) and that they remember small-scale problems (Section 2.2).

Evaluation is a bottleneck (Section 3). The Bitter Lesson favors generic methods, which require evaluation metrics to optimize over. But improved models create an evaluation bottleneck, since error detection becomes harder when most remaining mistakes are subtle or associated with edge cases. At scale, automated metrics show their flaws. We recommend that researchers work on improving metrics (Section 3.1).

There is no gold standard (Section 4). When one can afford the annotation costs, it may be tempting to consider human feedback as the ideal solution to the evaluation problem. Unfortunately, history has repeatedly shown that naïve methods of soliciting human preferences result in poor feedback, prioritizing superficial properties of model outputs. This lesson may serve as a counterpoint to the impulse to collect massive quantities of low-quality data in response to the Bitter Lesson. Instead, we recommend grounding performance measurement in concrete downstream tasks (Section 4.1).

Progress is not continuous (Section 5). The ascent of neural MT abruptly ended SMT's decade of seemingly unbeatable growth. This change in directions was enabled by new hardware-based paradigms, so we recommend that researchers continue exploring new methods that might scale well on future hardware (Section 5.1).

We conclude with a simple message: **Do research** (Section 6). We remind the reader that engineering achievements do not render scientific achievements insignificant, and we encourage the NLP community to renew their commitment to foundational scientific research even in areas where scale is currently a dominant factor.

2 Scale is supreme.

The first lesson offered by the history of SMT is that data and compute scale are the dominant factors in system performance. In all eras of MT, improvements in BLEU are logarithmic in training data size (Brants et al., 2007; Koehn and Knowles, 2017). This is immediately obvious from Figure 1: in order to achieve each linear step of improvement in accuracy (y axis), training data size must double (x axis). Figure 2 (reproduced from Kaplan et al., 2020) shows a strikingly similar log-linear relationship between training data size and system performance for LLMs. Indeed, such relationships are observed across many application areas of machine learning, including vision (Mahajan et al., 2018) and speech (Moore, 2003).

In a research landscape centered on performance metrics, scale will dominate. Sutton (2019) named the resulting malaise the "Bitter Lesson": "General methods that leverage computation are ultimately the most effective, and by a large margin." Both

1948) rather than neural networks. Early LMs were widely used across applications, beginning in speech recognition (Jelinek et al., 1975), though, unlike contemporary LLMs, they were rarely end products themselves. Although the LM of Brants et al. (2007) had a very different architecture from contemporary LLMs, it was an LLM in an important sense: it was trained on 2 trillion tokens, which is comparable to the training data size of modern LLMs.

[3]We use the term *crisis* deliberately since others have done so. For example, following the announcement of GPT-4 (OpenAI, 2023), @andriy_mulyar (2023) posted on Twitter that his feed was "full of ph.d. students having an existential crisis."

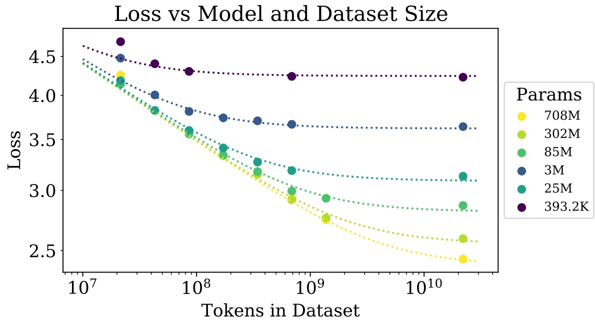

Figure 2: Figure from Kaplan et al. (2020) illustrating a power law relationship between dataset size and test loss for LLMs with varying numbers of parameters.

SMT and LLMs exemplify this lesson, and in fact Sutton explicitly references natural language processing. Many NLP researchers accordingly feel lost without access to large-scale systems. However, there are limits to scale, and as we will show by reviewing the history of the SMT era, its disparities are often transient. **The scale crisis is not a permanent state**.

2.1 Follow the hardware.

For several years following the release of Google Translate, large-scale commercial systems dominated the rankings of translation into English, where data was plentiful (NIST, 2008; Callison-Burch et al., 2009, 2010, 2011, 2012). Open source tools such as Moses (Koehn et al., 2007) and collaborations that pooled the resources of small labs narrowed the gap but did not close it until 2013, when translation into English was convincingly won by an academic group using modest hardware (Bojar et al., 2013). The decisive tool was KenLM, an efficient language modeling library (Heafield, 2011; Heafield et al., 2013) that demonstrated how, with the right software, contemporary hardware had made LLM training widely accessible. This end to the SMT scale crisis was the outcome of trends in hardware and software advancement.

The advent of LLMs in the SMT era and their later academic availability were both consequences of **Moore's law**, a six-decade trend in which computing power has doubled biannually: as Sutton (2019) observes, "over a slightly longer time than a typical research project, massively more computation inevitably becomes available." SMT-era LLMs arrived when researchers noticed that they had ignored Moore's law for too long. They rapidly closed the gap: Brants et al. (2007) ended a brief race to scale n-gram models to web-scale data (Zhang et al., 2006; Emami et al., 2007). But once the gap was closed, further incremental improvements—which required doubling the training data—necessarily required doubling the hardware cost or waiting for its capacity to double. Well-funded research sought more immediate gains elsewhere, while researchers with longer horizons rode Moore's law towards parity through collaboration and algorithmic advances. The new era of LLMs has already followed the first part of this pattern: computational requirements of LLMs have been doubling at a rate of less than a year or perhaps faster (Sevilla et al., 2022; Amodei and Hernandez, May 16, 2018), much faster than Moore's law.

The end of the SMT scale crisis was by no means inevitable or foreseeable at the beginning of the first era of LLMs in 2005: it resulted from the efforts of many researchers. We are encouraged to see a similar trajectory forming now. Already, startups advertise cheap large-scale training to the public (Portes et al., 2023). Like the groups that competed with commercial translation software in the SMT era, large cross-institutional collaborations are currently pooling resources to build public models (Scao et al., 2022). A community has developed around efficient ML, spawning new publication venues like MLSys and developing algorithms already employed in many LLMs (Hernandez and Brown, 2023). BERT (Devlin et al., 2019), regarded as inaccessible to many small academic labs at its release, now runs on a consumer-grade M1 MacBook laptop (Roesch and Mazenett, 2021).

In short, small labs do not need to abandon their entire research direction if they are interested in working with state-of-the-art models. Algorithmic efficiency guarantees usually hold across different resource scales, so a method developed on inexpensive hardware can be directly applied at industrial scale. Therefore, *all researchers can seek opportunities to collaborate and develop better algorithms.*

2.2 Remember small-scale problems.

While directly tackling scale is one strategy, we also recommend pursuing research on problems where data, not compute, is the bottleneck.

Small-scale settings provide a fertile ground for innovation in data-driven methods; in the previous era of LLMs, SMT researchers often used linguistic structure to improve performance when using smaller data. When such methods showed promise in these development settings, Google Translate inevitably tested them at industrial scale. Nonetheless, at the end of its lifespan, Google Translate's SMT system remained a phrase-based lookup table. The exploration of classical SMT researchers led to no lasting inventions based on syntax or semantics, and so the assumption that resulting improvements could transfer to large scale settings may have been based on unfounded optimism. Unlike their predecessors, modern NLP researchers preemptively recognize the futility of scaling up data-informed methods, because many of the improvements they offer are already provided by scale. But by leaning too far into the bitter lesson's pessimism now, we risk neglecting settings where, for practical or financial reasons, we *must* learn from limited data.

When Google Translate launched the first era of LLMs, it was only available between Arabic and English (Och, 2006). Data was the bottleneck that stood between SMT and its application to many meaningful problems, including, most obviously, the translation between many other language pairs. Solving this bottleneck required such diverse efforts as the collection of the Europarl corpus (Koehn, 2005), the OPUS corpus (Tiedemann and Thottingal, 2020), the JHU Bible Corpus (McCarthy et al., 2020) and the Nunavut Hansard (Martin et al., 2003); the rapid development of a Haitian Creole corpus in the aftermath of the Haiti earthquake (Lewis, 2010); the crowdsourcing of corpora for many Arabic dialects (Zbib et al., 2012); and the development of open-source web crawlers for parallel text (Smith et al., 2013). It is ongoing today in projects led by groups like Masakhane (Adelani et al., 2022; Nekoto et al., 2020; Emezue and Dossou, 2021) and No Language Left Behind (NLLB team et al., 2022). And yet, MT is still impossible for the vast majority of the world's estimated 7,000 languages.

Just as no one would have claimed in 2006 that Google had solved all translation problems, no one should claim now that LLMs have solved all NLP problems.[4] An identical bottleneck persists in the current era of LLMs, dominated by anglophone systems like ChatGPT. While these models can handle many languages to some degree due to the incidental multilinguality of any large training corpus (Blevins and Zettlemoyer, 2022), the training data is overwhelmingly English, and supervision data for learning from human feedback is overwhelmingly from English-speaking Kenyans (Perrigo, 2023). The hegemony of English has made it a presumed default, inciting the creation of the Bender Rule: "Always name the language(s) you're working on" (Bender, 2019).

Just as MT researchers have done since the SMT era, LM researchers today develop tools which rely less on scale (Alabi et al., 2022; Meyer et al., 2022; Park et al., 2021) for underserved languages. Researchers can leverage international collaborations with local linguists, incentives outside a profit model, and noncommercial resources to broaden the population that has access to technology in their own language. Furthermore, as compute costs continue to decline, even English corpora will become relatively "low resource" for future highly overparameterized models. To exploit these datasets more effectively, we turn to a perennial need across AI: quality evaluation metrics.

3 Evaluation is a bottleneck.

The next lesson offered by reflection on the SMT scale era is that the quality of evaluation methods makes a substantial difference in the effectiveness of training because a good evaluation can be used as a training signal. In SMT, this epiphany was delivered by minimum error rate training, which trained directly on target metrics like BLEU (Och, 2003). Likewise, train-time feedback metrics are often adapted for test-time evaluation: language modeling work may present validation loss, or equivalently perplexity, as the direct measurement of language modeling performance. It is therefore easy to use **symmetric** evaluations, applying the same metric for training feedback and test-time performance assessment. Like model evaluation, training can be based on comparison with a ground truth, as in conventional training; quality estimation based on output alone, as used often in Reinforcement Learning (RL) settings (Konda and Tsitsiklis, 1999; Silver et al., 2014; Bai et al., 2022b); or direct feedback, as provided by RL from Human Feedback (RLHF) and related methods of human assessment (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a). The evaluation metric then becomes a crucial lever to improve model quality.

[4]While we focus on data scarcity in underserved languages, some settings may provide limited data even in English. Such data scarcity may be due to practical hurdles to data collection (e.g., legally protected medical data) or an insufficient profit incentive (e.g. data for speakers of English from lower socioeconomic classes; Curry et al., 2024) While we cannot say the degree to which pure scale can solve problems like robustness or handling longer contexts, we can identify many problems where data is not collected at scale.

With large scientific projects in both industry (StabilityAI, 2021; Thoppilan et al., 2022) and non-profit sectors (Biderman et al., 2023; Scao et al., 2022) spending millions on training LLMs, it may be surprising to point away from scale to other bottlenecks in system performance. The focus on scale bottlenecks is reasonable, as for any given compute budget, there is an optimal quantity of training data that yields the most accurate model (Hoffmann et al., 2022). Either compute or data size, therefore, can become a limiting factor, and there are still financial barriers to compute (Lee et al., 2023) and limits on how much unstructured natural language data is available (Villalobos et al., 2022). However, training requires data and compute resources to be connected by evaluation feedback; this connection determines the effectiveness of training.

Unfortunately for efficient automatic approaches to evaluation, automatic metrics often fail to predict human evaluation, as shown for language modeling loss (Liu et al., 2023) and BLEU (Reiter, 2018). Even a metric that mirrors the training objective slightly can artificially inflate model performance; for example, it was observed early on that a phrase-based evaluation metric like BLEU might favor phrase-based SMT (Riezler and Maxwell III, 2005). Identifying good metrics is challenging and becomes more difficult with each improvement, as the remaining errors become increasingly subtle or complex. This challenge was also recognized early in SMT, with calls for a "BLEU++" (Och, 2005).

Evaluation, therefore, has become a crucial goal for modern LLM research. While training often relies on cross-entropy loss or other simple comparisons between each token in a sequence, benchmarking a trained LLM typically uses different evaluation criteria, such as checking the final answer in a word problem or the accuracy of a prompt-based classifier (Laskar et al., 2023). However, these tests are plagued by data contamination: benchmark exposure during training has created illusory gains in tasks ranging from code generation (Khan et al., 2023) to theory-of-mind puzzles (Ullman, 2023). Clearly language models are improving, but we cannot say precisely how, or by how much.

3.1 Improve the metrics.

Because evaluation is a bottleneck, we recommend greater focus on improving metrics, a goal which can yield rewards even with limited access to scale. Straightforward increases in computational infrastructure and raw data collection yield predictably diminishing returns on investment. Evaluation, by contrast, provides a conceptual space that can reward innovation and careful work with new insights and unknown improvements in system capabilities. What are the fundamental problems in this space, and how might we approach them?

One argument for why automatic metrics and static benchmarks are poor methods of evaluation is that they fail at measurement modeling; that is, these metrics do not actually measure what they purport to measure. This concern is reflected in objections to benchmarks and metrics that fail to reflect human evaluation (Liu et al., 2016; Novikova et al., 2017) or improvement on natural language understanding more broadly (Raji et al., 2021). These discussions parallel the unease in the MT community when studies found that automated metrics such as BLEU did not always correlate with human judgments (Callison-Burch et al., 2006). In response, SMT saw a flurry of incrementally improved bitext-based metrics (Stanojević and Sima'an, 2014; Popović, 2015; Mutton et al., 2007). In a scale crisis, improving evaluation metrics that leverage naturally available data like bitext can be a worthwhile focus.

Unfortunately, the naturally available data used in evaluation can still contaminate training corpora. Furthermore, automated metrics that rely on a static ground truth cannot reflect general quality (Raji et al., 2021), model conditions under interactive deployment, or provide on-policy reward feedback for reinforcement learning. These issues motivate automated metrics that do not require ground truth, although proposals based on AI supervision are themselves difficult to evaluate due to the same issues of dataset bias and contamination.

Despite a research community strongly motivated to improve them, even the best automated metrics are far from perfect. As automated metrics and static benchmarks fail, researchers with resources are increasingly hiring humans to assess model outputs. Facing similar challenges in SMT, researchers also called for human evaluation to be prioritized, both for benchmarking (Callison-Burch et al., 2006, 2007) and for training (Hopkins and May, 2011). However, human evaluation does not intrinsically solve problems with measurement modeling, and raises challenges of its own.

4 There is no gold standard.

Language evolved to be interpreted by humans. This fact leads us to a tempting myth: that we can easily evaluate synthetic natural language outputs by simply asking a human for their opinion. To the contrary, the next lesson we discuss from the SMT era is that human annotation cannot provide a universal "gold standard" of quality feedback. When MT competitions proudly turned to human evaluation as the highest-quality and most reliable option for choosing a winner, critics pointed out that these evaluations failed the basic expectations of consistency needed for a fair ranking (Bojar et al., 2011; Lopez, 2012). Even soliciting useful, let alone perfect, evaluations from humans turned out to require careful thought and trade-off decisions.

Practitioners often rely on naive methods of soliciting human feedback on LLM outputs, such as single rating scales or ranking model outputs by quality, which do not distinguish why annotators prefer a particular model output and thus offers limited guidance. For example, OpenAI's ChatGPT annotator interface asks the user to rank outputs from "best" to "worst" (Ouyang et al., 2022). Anthropic adds extra dimensions but with limited guidance, as annotators evaluate the extent to which generated text is "helpful" and "harmful," claiming that the vagueness of these guidelines permits versatile human preferences (Bai et al., 2022a). Recent work goes even further by using freeform text feedback, rather than predefined numerical axes (Shuster et al., 2022; Andreas et al., 2022; Scheurer et al., 2022, 2023), although how best to incorporate these explanations remains an unsolved problem.

The NLP community, however, is rediscovering that eliciting human preferences without clear guidance produces data that is not only noisy, but introduces systematic errors in models trained on the data. When many dimensions of quality are collapsed into a single preference scale, outputs that are worse along some dimensions may have higher ratings because they perform well along others. In these cases, annotators prioritize fluency over other aspects of the text, such as factuality or consistency (Clark et al., 2021). LLMs consequently prioritize fluency of large language model outputs over factuality (Ji et al., 2023), mirroring concerns from the SMT era that models prioritized fluency over the faithfulness of translations (Dorr et al., 2011)—concerns that have since been empirically confirmed (Martindale and Carpuat, 2018). Belz and Hastie (2014) and van der Lee et al. (2021) note that overall quality of generated text is often "too abstract" to be measured and both recommend the use of separate criteria for different dimensions of the text to distinguish what specific issues are present in a model output. Gehrmann et al. (2023) and van der Lee et al. (2021) warn that vague annotation guidelines can exacerbate annotator confusion, underscoring the importance of clearly defining the different dimensions on which to rate text quality.

Even after these refinements to the evaluation process, human evaluation for MT has encountered issues that remain unsolved, and current research suggests that evaluation of current models will increasingly encounter similar issues, including the following challenges.

Specifying evaluation criteria is hard. Even when evaluation criteria are separated into several axes, these scores are correlated, suggesting that human evaluators have difficulty in separating out criteria such as adequacy and fluency (Novikova et al., 2018). In addition, many studies fail to define their axes (van der Lee et al., 2021), permitting evaluators to differ even more in their interpretation of the task and thus increasing variation among annotators. That is, separating out axes of evaluation is necessary but not sufficient to identify multiple desirable traits of model output, a recurring problem in the history of MT evaluation (Chatzikoumi, 2019). Even when evaluation criteria can be defined clearly, crowdsourced annotators often lack the necessary expertise to follow them. Crowdworkers therefore align poorly with expert annotators, even underperforming against automatic evaluation metrics (Freitag et al., 2021). Some problems with objective specification can be resolved by defining multiple objectives and consulting expert annotators rather than crowdworkers.

Individual preferences are inconsistent. Classic SMT results reveal another fundamental problem in human evaluation: pairwise human rankings often fail to produce a consistent order (Bojar et al., 2011; Lopez, 2012). Any approach based on comparing outputs therefore reflects an unrealistic expectation of consistency in human preferences. The signal provided by ranking is noisy.

Disagreement isn't just noise. When human evaluators disagree on the quality of text, this does not necessarily reflect "noise" or "random variation" but rather genuine differences in opinion among evaluators (Larimore et al., 2021; Patton et al., 2019; Prabhakaran et al., 2021; Pavlick and Kwiatkowski, 2019; Basile et al., 2021; Plank, 2022), a problem that has long plagued MT evaluation (Lommel et al., 2014). The management of diverse annotator preferences is only exacerbated when benchmarking and training on freeform text from varied sources (Giulianelli et al., 2023). Furthermore, aggregation of annotator judgments obscures the opinions of underrepresented groups (Prabhakaran et al., 2021; Fleisig et al., 2023), and use of inter-annotator agreement as a quality metric causes additional erasure of perspectives by denying that these priorities are contested (Blodgett, 2021).

These issues collectively prevent human evaluation from providing clear feedback on model outputs. The fact that they have remained major concerns in MT despite decades of research suggests that current researchers would do well not to underestimate the challenges posed by these issues. Furthermore, issues of both task specification and disagreement may be even more central to the evaluation of current models that can handle more varied tasks. Whereas fluency and faithfulness to a source text might cover major concerns in MT, there is a broader range of criteria that generated text must fulfill, such as informativeness and coherence (van der Lee et al., 2021). These requirements, along with increased freedom to produce text on topics where there is real-world disagreement, including social, ethical, and political concerns (Abid et al., 2021; Blodgett et al., 2020; Liu et al., 2021; Zhao et al., 2021), mean that human evaluation issues will be pressing problems for the NLP community to solve.

4.1 Focus on concrete tasks.

Due to the inherent flaws of evaluation based on human assessment, we recommend measuring concrete tasks under deployment conditions. Extrinsic evaluations (Belz and Reiter, 2006), wherein model output quality is evaluated based on utility for specific downstream applications, are still uncommon in evaluation of text generation (van der Lee et al., 2021). However, they may be more useful for evaluating the quality of content or meaning (Reiter and Belz, 2009; Reiter, 2023) because human assessment often fails to predict performance on downstream applications (Kunz et al., 2022).

In MT, concrete downstream objectives have long been used in evaluation. Snover et al. (2006) examined how many manual edits human translators had to make to model output, reflecting the desiderata of human-AI collaboration settings. Other metrics rely on the user's ability to accomplish specific tasks using model output, such as answering reading comprehension questions based on translations (Jones et al., 2005; Callison-Burch, 2009; Scarton and Specia, 2016) or summaries (Wang et al., 2020). A recent and growing body of research attempts to measure the effectiveness of MT in second language education (Lee, 2023).

In general, the best evaluations are likely to rely on realistic assessment of what LLMs enable humans to do. In modern LLMs, work on the challenges of evaluation is likely to draw on insights from human-computer interaction.[5] Good user trials require careful study design and consideration of human variety, as well as an understanding of individual psychology.

5 Progress is not continuous.

Our final lesson is that new paradigms can unlock new orders of scale and even new scaling coefficients, leading to abrupt improvements in performance. The SMT era, fueled by large n-gram models, lasted for over a decade, with scale providing increasing improvements over time. But Moore's law was threatened during this era due to the breakdown of Dennard scaling, the observation that smaller transistors require commensurately less power, meaning that they can be miniaturized while keeping power consumption constant. To drive continued improvement, hardware manufacturers turned to parallelization. Graphical processing units (GPUs), which favor high parallelization of code with minimal branching—and thus simpler and smaller processors—were soon being repurposed to train neural networks (Hooker, 2020).

Neural networks had been investigated in SMT for years. Indeed their earliest use in SMT was as n-gram language models, when Schwenk et al. (2006) built an SMT decoder using the neural n-gram model of Bengio et al. (2003)—an idea that only began to gain traction almost a decade later

[5] Interdisciplinarity between machine learning applications and HCI is a perennial concern. HCI researchers are periodically invited to speak at ML and ML applications conferences, e.g., NeurIPS hosted HCI-centered keynotes from Deborah Estrin in 2013 and Juho Kim in 2023.

when revisited by Devlin et al. (2014), whose ACL best paper award marked a shift in NLP establishment attitudes towards neural networks. The increasing power of GPUs and their use in training neural networks fueled new research in end-to-end neural MT (NMT), enabling Kalchbrenner and Blunsom (2013) to revive the even older idea of an encoder-decoder architecture (Ñeco and Forcada, 1997). Advances such as attention (Bahdanau et al., 2015), seq2seq (Sutskever et al., 2014), and transformers (Vaswani et al., 2017) followed in a flurry of activity. Within two years, NMT swept the annual shared tasks (Jean et al., 2015; Chung et al., 2016; Bojar et al., 2016), and in 2016, Google Translate announced that it had switched to NMT (Wu et al., 2016; Turovsky, 2016). Research on SMT quickly faded.

GPUs effectively introduced a new dominant **paradigm** by creating conditions that favored deep learning. Kuhn (1962) described scientific advancement as a cycle of scientific revolutions in which paradigms such as phrase-based SMT or deep learning emerge, followed by periods of **normal science** when researchers aim to apply, articulate, and expand the fact base of the paradigm. Often, scientific revolutions result from the availability of new tools. Hooker (2020) drew on this framework of scientific revolution to analyze the landscape of AI research, identifying the **Hardware Lottery** as a situation in which hardware dictates methods.[6] Under the Hardware Lottery, GPUs offered a winning ticket for deep learning to reshape MT. What research objectives are recommended by the resulting NMT revolution?

5.1 Shape the hardware.

The Hardware Lottery tells us that hardware guides the direction of research, but researchers can also direct the design of hardware. While these new tools may enable scientific revolutions, Kuhn (1962) pointed out that the development of new tools is itself shaped by the reigning paradigm and by the normal scientific process. Hardware design itself is an example, having been driven for many years by incremental improvements to a paradigm of miniaturization and parallelization of transistors. But hardware manufacturers are approaching the physical limits of miniaturization, and the path forward is again uncertain (Lundstrom and Alam, 2022), as it was at the end of Dennard scaling . Therefore, we recommend that researchers focus not only on developing and using new hardware, but on anticipating potential hardware developments and developing algorithms for platforms before they are widely available.

By creating software tools and algorithms that can take advantage of hardware designed for sparsity (Krashinsky et al., 2020) or new sources of parallelism (Launay et al., 2020), researchers can develop techniques preemptively for future technologies. At the same time, they also create a market to motivate the development of new hardware that can enable the next revolutionary development. Researchers may even co-design hardware and software jointly, a strategy likely to drive future computing advances (Leiserson et al., 2020; Lundstrom and Alam, 2022). It is the possibility of reshaping tools for the future that makes alternative paradigms worth exploring under a scale crisis.

6 Conclusion: Do research.

As pure engineering efforts and institutional wealth outstrip novel scientific work, some in the AI community are pessimistic about the prospects of foundational research. Our position, articulated over the course of this paper, is that there is much exciting, timely work yet to be done.

These lessons are not particular to LLMs, but apply to any field subject to the Bitter Lesson. For example, issues in human evaluation plague many disciplines in machine learning. In computer vision, annotator idiosyncrasies account for many of the remaining inaccuracies of modern ImageNet models (Shankar et al., 2020). If a constrained labeling task such as image classification is subject to varied human judgment (Parrish et al., 2023), how much harder is it to annotate free text generation?

Beyond our specific recommendations for researchers interested in improving the capabilities of language models, we would also point to scientific opportunities across related fields. From interpretability to empirical training analysis to public policy, many research areas only become more relevant and complex as models rapidly improve. Furthermore, while novel modeling work suffers in a scale crisis, we can focus on new architectures and algorithms that take advantage of existing hardware and even anticipate future tools.

[6] Gururaja et al. (2023), whose oral history of NLP—including comments on the current scale crisis and the cyclic nature of what they call exploit-explore incentives—complements our work, also point to similar paradigm shifts in NLP emerging from a *software* lottery.

Some speculative alternatives to phrase-based SMT, such as explicitly modeling syntax (Galley et al., 2004; Collins et al., 2005; Chiang, 2007) and semantics (Jones et al., 2012), were obviated by the Bitter Lesson and its expression in NMT. However, other proposals formed the basis of the NMT era. Many enduring careers in NLP research were forged in areas that are now forgotten, and without risky exploration of unproven directions, the field could not have achieved many breakthroughs. Our anxieties should not discourage us from seizing the opportunities presented by a new era of LLMs.

Limitations

The positions taken in this paper are based on both experience and reading of historical trends in natural language processing. While we believe that the lessons we identify in this paper are durable, history does not always repeat, and our oracular powers are otherwise limited. Even after considering our position, researchers should use their own best judgement on directions to pursue.

Ethical Considerations

The authors received permission from Andriy Mulyar to feature his Twitter post as an example of the March 2023 scale crisis discourse.

Acknowledgements

We thank Kenneth Heafield, Nikolay Bogoychev, and Steven Kolawole for helpful discussion; and Arya McCarthy, Kevin Yang, Sanjay Subramanian, and Nicholas Tomlin for comments on previous drafts. We thank our reviewers for their constructive feedback.

This work was supported by Hyundai Motor Company (under the project Uncertainty in Neural Sequence Modeling) and the Samsung Advanced Institute of Technology (under the project Next Generation Deep Learning: From Pattern Recognition to AI). This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. During the majority of work on this paper, Naomi Saphra was employed by New York University.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pretrained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dario Amodei and Danny Hernandez. May 16, 2018. AI and compute. *OpenAI Blog*.

Jacob Andreas, Karthik Narasimhan, and Aida Nematzadeh, editors. 2022. *Proceedings of the First Workshop on Learning with Natural Language Supervision*. Association for Computational Linguistics, Dublin, Ireland.

@andriy_mulyar. 2023. my Twitter feed is full of ph.d. students having an existential crisis. *Twitter*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson 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

Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

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 Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi 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. 2022b. Constitutional AI: Harmlessness from AI feedback.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 313–320.

Anja Belz and Helen Hastie. 2014. *Comparative evaluation and shared tasks for NLG in interactive systems*, page 302–350. Cambridge University Press.

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Terra Blevins and Luke Zettlemoyer. 2022. Language Contamination Helps Explains the Cross-lingual Capabilities of English Pretrained Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *WMT@ACL*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Chris Callison-Burch, Cameron S. Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *WMT@ACL*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark A. Przybocki, and Omar Zaidan.

2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *WMT@ACL*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *WMT@NAACL-HLT*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *WMT@EACL*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *WMT@EMNLP*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Eirini Chatzikoumi. 2019. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26:137 – 161.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33:201–228.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. NYU-MILA neural machine translation systems for WMT'16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 268–271, Berlin, Germany. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. Classist tools: Social class correlates with performance in nlp.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Bonnie Dorr, Joseph Olive, John McCary, and Caitlin Christianson. 2011. *Machine Translation Evaluation and Optimization*, pages 745–843. Springer New York, New York, NY.

Ahmad Emami, Kishore Papineni, and Jeffrey Scott Sorensen. 2007. Large-scale distributed language modeling. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4:IV–37–IV–40.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of HLT-NAACL*, pages 273–280.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Gray Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Danny Hernandez and Tom Brown. 2023. AI and efficiency. *OpenAI Blog*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and L. Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Neural Information Processing Systems*.

Sara Hooker. 2020. The hardware lottery. *Communications of the ACM*, 64:58 – 65.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, et al. 2023. A PhD student's perspective on research in NLP in the era of very large language models. *arXiv preprint arXiv:2305.12544*.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

F. Jelinek, L. Bahl, and R. Mercer. 1975. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256.

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

Bevan K. Jones, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight. 2012. Semantics-based machine translation with hyperedge replacement grammars. In *International Conference on Computational Linguistics*.

Douglas Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005. Measuring translation quality by testing English speakers with a new defense language proficiency test for Arabic. In *Proceedings of the 2005 International Conference on Intelligence Analysis*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Conference on Empirical Methods in Natural Language Processing*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*. ArXiv: 2001.08361.

Mohammad Abdullah Matin Khan, M. Saiful Bari, Xuan Long Do, Weishi Wang, Md Rizwan Parvez, and Shafiq Joty. 2023. xCodeEval: A Large Scale Multilingual Multitask Benchmark for Code Understanding, Generation, Translation and Retrieval. ArXiv:2303.03004 [cs].

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*.

Philipp Koehn, Hieu T. Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Vijay Konda and John Tsitsiklis. 1999. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Ronny Krashinsky, Olivier Giroux, Stephen Jones, Nick Stam, and Sridhar Ramaswamy. 2020. NVIDIA Ampere Architecture In-Depth.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets.

Julien Launay, Iacopo Poli, Kilian Müller, Gustave Pariente, Igor Carron, Laurent Daudet, Florent Krzakala, and Sylvain Gigan. 2020. Hardware Beyond Backpropagation: a Photonic Co-Processor for Direct Feedback Alignment. ArXiv:2012.06373 [cs, stat].

Ji-Ung Lee, Haritz Puerto, Betty van Aken, Yuki Arase, Jessica Zosa Forde, Leon Derczynski, Andreas Rücklé, Iryna Gurevych, Roy Schwartz, Emma Strubell, and Jesse Dodge. 2023. Surveying (dis)parities and concerns of compute hungry nlp research.

Sangmin-Michelle Lee. 2023. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1-2):103–125.

Charles E. Leiserson, Neil C. Thompson, Joel S. Emer, Bradley C. Kuszmaul, Butler W. Lampson, Daniel S. Sanchez, and Tao B. Schardl. 2020. There's plenty of room at the top: What will drive computer performance after moore's law? *Science*, 368.

William D. Lewis. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *European Association for Machine Translation Conferences/Workshops*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023. Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14857–14866.

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, pages 31–37. Language Resources and Evaluation Conference Reykjavik.

Adam Lopez. 2012. Putting human assessments of machine translation systems in order. In *WMT@NAACL-HLT*.

Mark S. Lundstrom and Muhammad Ashraful Alam. 2022. Moore's law: The journey ahead. *Science*, 378:722 – 723.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, pages 185–201.

Joel D. Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an english-inuktitut parallel corpus. In *ParallelTexts@NAACL-HLT*.

Marianna J. Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect mt. In *Conference of the Association for Machine Translation in the Americas*.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus.

Roger K. Moore. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2581–2584.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic Evaluation of Sentence-Level Fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Ramón P. Ñeco and Mikel L. Forcada. 1997. Asynchronous translations with recurrent neural nets. *Proceedings of International Conference on Neural Networks (ICNN'97)*, 4:2535–2540 vol.4.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

NIST. 2008. NIST Open MACHINE TRANSLATION 2008 OFFICIAL Results.

NLLB team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Franz Och. 2006. Statistical machine translation live.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och. 2005. Statistical machine translation: Foundations and recent advances. In *Proceedings of Machine Translation Summit X: Tutorial notes*, Phuket, Thailand.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

OpenAI. 2023. Gpt-4 technical report.

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. 2022. Training language models to follow instructions with human feedback.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation.

Alicia Parrish, Sarah Laszlo, and Lora Aroyo. 2023. "Is a picture of a bird a bird": Policy recommendations for dealing with ambiguity in machine vision models. ArXiv:2306.15777 [cs].

Desmond Upton Patton, Philipp Blandfort, William R. Frey, Michael B. Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Hawaii International Conference on System Sciences*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Billy Perrigo. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Portes, Alex Trott, Daniel King, and Sam Havens. 2023. MosaicBERT: Pretraining BERT from Scratch for $20.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter. 2023. Future of NLG evaluation: LLMs and high quality human eval?

Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558.

Stefan Riezler and John T Maxwell III. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64.

Jared Roesch and Phil Mazenett. 2021. OctoML's BERT Model Acceleration on Apple M1 Pro and Max Chips.

Anna Rogers. 2023. Closed AI Models Make Bad Baselines.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla A. Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian

Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim T Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, A. Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, R. Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yun chao Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia. European Language Resources Association (ELRA).

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale.

Holger Schwenk, Daniel Déchelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating Machine Accuracy on ImageNet. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8634–8644. PMLR. ISSN: 2640-3498.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y.-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. ArXiv:2208.03188 [cs].

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 387–395, Bejing, China. PMLR.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association* *for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

StabilityAI. 2021. Stability AI Launches the First of its StableLM Suite of Language Models.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215.

Rich Sutton. 2019. The Bitter Lesson.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. ArXiv:2201.08239 [cs].

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Julian Togelius and Georgios N Yannakakis. 2023. Choose your weapon: Survival strategies for depressed AI academics. *arXiv preprint arXiv:2304.06035*.

Barak Turovsky. 2016. Found in translation: More accurate, fluent sentences in Google Translate.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Y. Zhang, Almut Silja Hildebrand, and Stephan Vogel. 2006. Distributed language modeling for n-best list re-ranking. In *Conference on Empirical Methods in Natural Language Processing*.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021. Ethical-advice taker: Do language models understand natural language interventions?

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.