AI

MDPI

Review

Augmented Behavioral Annotation Tools, with Application to Multimodal Datasets and Models: A Systematic Review

Eleanor Watson *, Thiago Viana and Shujun Zhang

School of Computing and Engineering, University of Gloucestershire, The Park, Cheltenham GL50 2RH, UK
* Correspondence: eleanorwatson@connect.glos.ac.uk

Abstract: Annotation tools are an essential component in the creation of datasets for machine learning purposes. Annotation tools have evolved greatly since the turn of the century, and now commonly include collaborative features to divide labor efficiently, as well as automation employed to amplify human efforts. Recent developments in machine learning models, such as Transformers, allow for training upon very large and sophisticated multimodal datasets and enable generalization across domains of knowledge. These models also herald an increasing emphasis on prompt engineering to provide qualitative fine-tuning upon the model itself, adding a novel emerging layer of direct machine learning annotation. These capabilities enable machine intelligence to recognize, predict, and emulate human behavior with much greater accuracy and nuance, a noted shortfall of which have contributed to algorithmic injustice in previous techniques. However, the scale and complexity of training data required for multimodal models presents engineering challenges. Best practices for conducting annotation for large multimodal models in the most safe and ethical, yet efficient, manner have not been established. This paper presents a systematic literature review of crowd and machine learning augmented behavioral annotation methods to distill practices that may have value in multimodal implementations, cross-correlated across disciplines. Research questions were defined to provide an overview of the evolution of augmented behavioral annotation tools in the past, in relation to the present state of the art. (Contains five figures and four tables).

Keywords: machine learning; annotation; behavior; foundation models

Citation: Watson, E.; Viana, T.; Zhang, S. Augmented Behavioral Annotation Tools, with Application to Multimodal Datasets and Models: A Systematic Review. *AI* **2023**, *4*, 128–171. https://doi.org/10.3390/ai4010007

Academic Editors: José Manuel Ferreira Machado and Kenji Suzuki

Received: 31 October 2022
Revised: 20 December 2022
Accepted: 3 January 2023
Published: 28 January 2023

1. Introduction

Machine intelligence and data science technologies have a major impact on the global economy [1,2] and are becoming increasingly common in their deployment [3]. However, constructing datasets to train them is time-consuming, expensive, and sometimes unrewarded [4,5]. A quality dataset can provide huge public benefit and underpin thousands of algorithms, as well as benchmark performance [6]. Machine Learning (ML) processes requires data to train on [7], and testing/validation requires 20–30% of the original dataset to be reserved [8], with a greater proportion retained for sets with fewer examples. Data must be representative, with few errors/lacunae, and reviewed to control for biases [9]. Error rates of up to 6% have been identified [10], embedding errors/biases within models derived from such data, and making benchmarking processes more challenging due to inherent error. Many algorithms were developed many years before they were practically deployable, due to a paucity of training data [11,12]. Many of the recent models appear to be constrained by data size and nuance [13]. Improved annotation techniques would be a beneficial response to a growing requirement to generate larger, richer multimodal datasets more easily.

Another issue is the human-AI alignment and safety necessary to deploy increasingly complex and emergence-prone models in practical daily life. Any algorithm for aligning machine intelligence with human behavior is likely to face serious training data problems.

More advanced techniques will need to be developed for teaching machines to recognize social behavior, such as norms in various complex contexts [14].

Annotation tools are an important element in specifying and collating datasets. They enable workers (often referred to as 'annotators' or 'coders') to highlight a salient example, and to specify its character or parameters. Attempting to analyze an event in rich detail requires multimodal streams, coded across multiple passes, each looking at a different attribute (or a subset thereof). A cross-correlation of modalities enables a depth of sophistication in outputs not otherwise possible [15].

Moravec's Paradox [16] observes that many tasks that a young human child finds trivial remain challenging for machine intelligence. Recognizing behavioral patterns accurately could improve socialization of models. It could also reduce misapprehension of human intent, enabling more fair and accurate content moderation. Models that can translate between 200 languages suggest intercultural competence could manifest as a feature in machine intelligence [17]. It may even be feasible to train cultural knowledge through observation without prior examples [18]. Greater creativity can be unlocked through Human-AI teaming, whereby human and machine cognition are entwined in an ensemble. However, creativity may also be reduced if one party becomes overly reliant on the other [19].

Since the advent of a new generation of deep machine learning techniques around 2010, many new economic, social, and research innovations have become feasible [20–22]. A further wave of prompt-driven technologies are providing an important subset of deep learning techniques with revolutionary flexibility and ease-of-use. A seminal paper by From Vaswani et al. in 2017, "Attention Is All You Need" introduced a promising new attention-based machine learning architecture for natural diverse language processing tasks, the Transformer. All the models in this family share a property of in-context learning, providing them with the ability to learn a new task from brief demonstrations (prompts), without requiring any parameter updates. This attribute provides tremendous flexibility, as well as an unprecedented capability for abstraction generalizability. Another notable aspect of these models is their prodigious size, along with the discovery that scaling in parameters, tokens, and datasets is enough to enable startling new capabilities [23].

Some researchers took to describing this new wave of multimodal techniques as 'Foundation Models', more of a genericized name which can include all Large Language Models, Transformers, and Diffusion Models [24]. This launched a new wave of interest and excitement around an emerging new class of large (multi-billion parameter) multimodal models, as well as a requirement for multimodal dataset ensembles to power them [25]. This requirement for multimodal data presents a major challenge, but the prompt-driven capabilities also present a massive opportunity to create powerful new annotations tools that require minimal human oversight.

Review Justification

These new models are multimodal in nature, able to ingest and draw inference from many modes of data. To make full use of these models will require massive further amounts of annotation. To address this gap, it is essential to streamline the process of annotation to be as simple, accessible, efficient, and inclusive as possible, to provide amelioration for these challenges. The aforementioned innovative models and data augmentation are crucially pertinent towards annotation technology, as these techniques unlock substantial new capabilities not otherwise feasible. They also have relevance to the application of annotation processes towards the creation of multimodal datasets, which are particularly suitable for usage by such models. Data augmentation techniques can also assist with generating new multimodal nuanced layers within existing datasets, reducing the need for annotation in certain contexts and enabling resources to be applied elsewhere, particularly towards challenging edge cases where there is a risk of overfitting to overly salient data. The potential emerging applications of these models are very important to any researchers who attempt to make sense of behavior, and who work with multimodal data, which are both large and rapidly increasing topics within AI research.

Given the importance of powerful new datasets for machine learning, the challenges specific to the annotation of behavior, and the rapid advances in this space, a review of the literature appears warranted, and likely to provide insight for future research efforts in this space. There is also likely to be severe disruption to the space of annotation in general, and especially complex target domains such as behavior, and the values that may be encoded within it. Improved annotation eases the creation of Foundation Models, and such models assuredly provide opportunities to improve annotation in turn. The central descriptive research question is: "Which state of the art Foundation Model developments are likely to heavily impact the domain of behavioral annotation, and vice-versa?"

Our systematic literature review answers the research questions by cataloguing the state-of-the-art in the annotation of behavior as it applies to the creation of datasets, identifying gaps in knowledge, and experimenting with new techniques to ascertain their viability. Our study contributes to the Artificial Intelligence literature by distilling very complicated and rapid developments into a digest and outlining their transferable impact into the annotation domain. The final section presents the identified research gaps and an expected roadmap for the future of annotation supported by these innovation Foundation Model techniques.

This review provides insights on how various best practices have evolved, identifies gaps in the present knowledge, and provides insight into future research opportunities. This paper contributes to the literature by providing a summary of research to date since the turn of the century. An overview of the main approaches, strengths, weaknesses, applications, and approaches in the domain of augmented behavioral annotation is presented, followed by a research gap analysis and roadmap for future development.

The contributions of this paper include:

1. The use of a robust research methodology to identify, collate, and analyze papers that provide insights on technologies applicable to behavioral annotation processes (Section 2)
2. A classification and discussion of studies that evaluate educational aspects of such behavioral annotation systems (Section 3)
3. A digest of the major developments, and the expected future path of this research domain (Section 4)

The knowledge gained will inform a theory of an evolution of annotation since the turn of the millennium as it relates to augmented methods for the construction of actionable machine learning datasets.

The remainder of this paper has been structured to sample and highlight the extant literature in a systematic manner. Section 2 provides a description, the methodology, and the research questions. Section 3 collates answers for the seven research questions, and Sections 4 and 5 provide an overview of expected development in the space based upon emerging trends.

2. Methods and Literature Review

This section describes the research questions answered in this study, the research databases used and why, and the Selection Criteria applied to identified papers of interest for inclusion in this review. This body of research undertook a Systematic Literature Review process to understand the past, present, and potential future of the domain of annotation. Insights are thereby gained as to the respective challenges and opportunities presented by multimodal abstraction machine learning models, with a view to establishing foundational research for future researchers to build upon. Particular focus has been afforded to recent papers featuring the latest innovations.

A systematic review is a type of research assessment that involves collecting the literature related to the topic, finding out what has been reported in the past and then subjecting that information to analysis [26,27]. It also includes results from other similar studies. This process guides one's own research, based upon comparison and contrast

with prior examples, thereby gathering sufficient information to distil judgements about the topic.

The domain of crowdsourced datasets and their annotation is complex, with multivariate methods and techniques, data formats, design purposes, and applications. A traditional literature review does not seem likely to sufficiently collect the necessary nuances. Moreover, differences in nomenclature may accidentally exclude relevant results if performed in an ad hoc manner according to availability and salience according to search algorithms. A systematic review by contrast retains flexibility in handling qualitative, quantitative data, and/or mixed methods.

Systematic reviews are increasingly considered as the 'Gold Standard' in review processes [27]. The process requires extensive searches, even for data or examples that have yet to be formally published. The potential for bias from inclusion criteria, or during the presentation of results, is also analyzed and mitigated. The process is designed to remain scrupulously impartial, with the utmost transparency and precision, carefully noting any limitations and thereby preserving the potential for harmonious replicability.

Finally, such methods also facilitate the making of recommendations for future research, through identifying gaps in knowledge, to a degree that may be more insightful than traditional methods. Such insights can inform and strengthen the aims of this research, ensuring that the research is of greatest value and impact, and influencing future research pathways in beneficial ways. For these reasons, the authors have elected to perform a systematic review of prior art and literature to better ensure a robust and representative study, following the protocol outlined by Wohlin et al. in Experimentation in Software Engineering [28]. Various other papers have inspired the design of this review process through their positive examples [29–33].

Sources may be challenging to assess in the realm of software, which have fewer protocols than the domain of medicine where systematic reviews originated. Established quality criteria therefore do not align easily [34]. The goal of evidence-based software engineering (EBSE) is summarized by Kitchenham et al. as being: "to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision-making process regarding the development and maintenance of software" [35,36]. These techniques have provided a framework to support the integrity of this research, as well as to help validate whether elements within the systematic literature review are sufficiently robust and appropriate for inclusion.

2.1. Research Questions

The main aim of this research is to address what major innovations and best practices have arisen in the space of annotation of behavior. The overall question is: "What elements are preferable in the process of collecting and annotating information relating to behavior." This been decomposed in Table 1 into the following queries:

Table 1. Research Questions.

RQ1	What methodologies and frameworks can facilitate annotation, especially those with a multimodal nature?
RQ2	How to encode data in formats which facilitate safe and ethical interchange, as well as the coding of expansive and representative modalities/categorizations?
RQ3	How to streamline the user experience to reduce cognitive load and training requirements?
RQ4	How to augment user contributions to increase their impact?
RQ5	How to validate coded information as being reasonable and appropriate?
RQ6	How to pre-process data or to permit pre-annotation
RQ7	How can Transformer-type technologies be applied to annotation?

These questions are answered in Section 3.

2.2. Inclusion Criteria

A large body of literature focusing on online social annotation tools was gathered and reviewed, with sampling according to specified keywords. The following online journal research databases were employed for the literature search: Scopus, IEEE Xplore, Science Direct, and WorldCat. Table 2 outlines these, with the URLs as accessed on the 1 June 2022). These databases were selected because we identified them as being leading repositories for papers related to machine learning and annotation from their prominence in background research. ArXiv is a prepublication archive, which is heavily utilized by AI researchers because of the speed of development in the space and relative ease of replication merits rapid pre-publication.

Table 2. Research Databases.

RD1	Scopus	www.scopus.com
RD2	IEEE Xplore	ieeexplore.ieee.org
RD3	Science Direct	www.sciencedirect.com
RD4	Elicit	www.elicit.org
RD5	WorldCat	www.worldcat.org
RD6	Google Scholar	scholar.google.com
RD7	ArXiv	www.arxiv.org

In addition, Google Scholar was also applied to search for and acquire specific references which had been located via abductive analysis. Elicit was also employed to search for any papers which may have been otherwise overlooked because of a reliance purely upon keywords without any contextual understanding of the intent of the search, as well as to pinpoint DOI references to the original paper publication which may otherwise have been unclear in some cases.

The focus of the search was to gather full-text articles presenting empirical studies whereby annotation tools and methodologies were employed to derive data about behavior. The subject was typically human beings, but in some cases, subjects such as rodents have been included as the research was deemed adequately transferrable to the human domain. It was desirable to include a broad range of studies across a long period to observe a variety of developments over time, some of which may retain value even despite technological advancement. A review was undertaken in concordance with the Systematic Review Process Protocols as described by Wohlin et al. [37] and outlined in Table 3. The PRISMA guidelines and checklist were also applied to ensure the robustness of this study [38].

Table 3. Systematic Literature Review Selection Criteria.

The study employed tools for annotating behavior that embodied the following keywords: (a) annotation and (b) behaviors.
The study examined included either (c) a collaborative analysis mechanism, or (d) an element of automation, both elements providing a means of amplifying.
The study reported the research methods applied (i.e., the type data being generated, the technologies employed, the intended use case, the general research design).
The research presented in one study did not overlap with research from another study. In such cases, a note was taken of the original research, but reporting focused on the lattermost results.
The study was written during or after the year 2000.
The article was written in English, or a professional translation was readily available.

The search was conducted using precise Boolean search terms, specifically ANNOTATION AND BEHAVIOR, with variation to account for differences in spelling between American and British styles of English. The search was limited to papers from the year 2000 onwards, which is broadly in concordance with the advent of XML and collaborative annotation methods.

As outlined in Figure 1, from 774 collected articles, 348 studies met the inclusion criteria for this systematic literature review. An additional 48 studies were selected for analysis from abductive sources. Among the included studies were 307 experimental or quasi-experimental studies and 41 evaluation/survey studies.

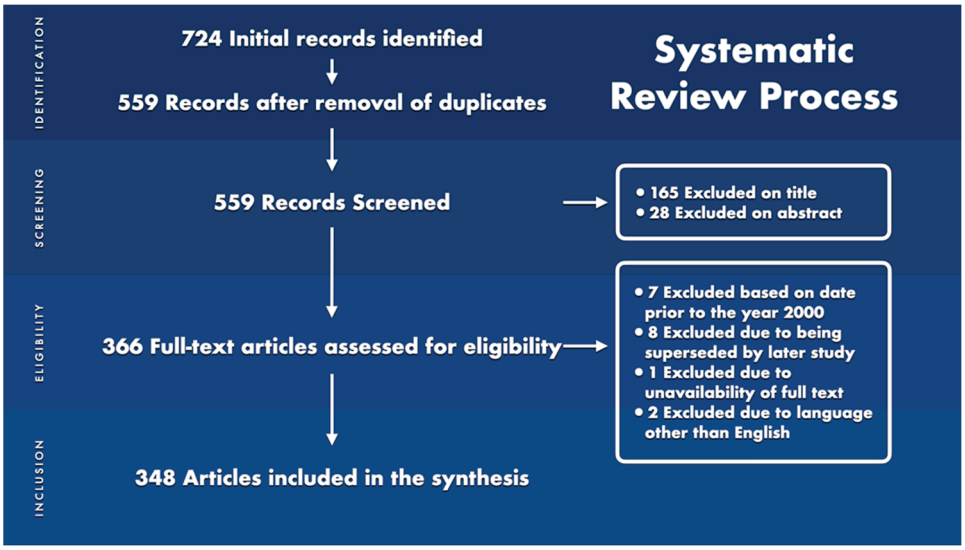

Figure 1. A flowchart describing the PRISMA Systematic Review Process applied in this research.

The review analyzed peer-reviewed studies featuring behavioral annotation to discern methods by which the augmentation process may be augmented. Such methods include machine vision techniques to facilitate the segmentation of actors, social collaboration techniques to enable division of labor and peer review for accuracy, and machine learning techniques, which attempt to categorize (and locate) content in more efficient ways. Through the analysis of 348 identified studies, it was possible to examine various techniques, along with their efficacy and relative strengths.

3. Techniques of Augmented Behavioral Annotation

The techniques reviewed offer several new possibilities for improving the efficiency and capability of annotation processes. This section will address the specified Research Questions in turn, synthesizing responses from the observed literature.

3.1. RQ1—What Methodologies and Frameworks Can Facilitate Annotation, Especially Those with a Multimodal Nature?

There are several different techniques that can facilitate annotation processes, and which have distinct applications in multimodal contexts. STEGO is a novel algorithm for automatically labeling image data, using Transformers to detect, segment, and label objects without human input [39]. Cross-Modal Discrete Representation Learning systems can identify actions in video clips without human help [40], whereas UViM can be trained for complex annotations without architectural changes [41]. Biological modelling research suggests the best-performing models of the visual cortex can encode high-dimensional manifolds [42]. Qin et al., 2022, propose a hierarchical video decomposition technique with transformative representations to segment complex layers, such as dynamic backgrounds and overlapping heterogeneous environments, applicable to domains beyond the X-ray coronary angiography featured [43].

The VITO system employs a contrastive learning network to distill knowledge from videos to image representations, thereby improving self-supervised learning mechanisms [44]. ODIN, introduced in 2022, couples object discovery and representation networks in an ensemble to generate image segmentations without supervision, achieving state-of-the-art results [45]. PALI, a jointly scaled multimodal and multilingual language-image model, outperforms prior larger models on several Visual Question Answering and image-captioning tasks [46,47]. Omnivore can recognize 3D models and videos without degrading performance on modality-specific tasks, even though it was trained on images only [48].

Augmented annotation mechanisms are now being applied to neural networks, such as MILAN, which can automatically label the behavioral roles of individual neurons [49]. This may present an important path for reverse engineering black box models, as well as auditing them for potential disproportional or undesirable biases. *data2vec* [50] presents self-supervised learning techniques for multimodal data, predicting latent representations of the full input based on a masked view in a self-distillation setup using a Transformer architecture. Experiments on major benchmarks of speech recognition, image classification, and natural language understanding demonstrate strong performance, with new state-of-the-art or competitive results to predominant approaches.

Techniques such as syncretization (Meng et al., 2021) and label smoothing (Whitfield, 2021) can amplify datasets and assist models in understanding situations with no direct example. Unidirectional Pretrained Language Models (PLMs) can generate prompt-guided, class-conditioned texts for fine-tuning bidirectional PLMs [51]. GPT-2 can generate synthetic data to improve NLP models, with mixed organic and synthetic data outperforming the organic model [52]. Jump-Start Reinforcement Learning provides a framework for improving an agent's behavior via a meta-algorithm that uses offline data, demonstrations, or a pre-existing policy to initialize a RL policy [53]. Multimodal ML models enable feedback based on prompts, which may be highly abstract, such as 'be more formal' or 'be less cautious'. These scenarios can be translated into images, animations, or 3D environments to aid accessibility, understanding, and engagement [54].

These techniques may provide fine-tuning of examples, as representative examples of personal values may not be accessible. This also allows for a larger cohort of annotators to participate, as the barrier to entry is lower and less time is required, as a simple written or voice interface is sufficient to specify values. Combined results can then be used to fine-tune based upon personal responses, and those of people with similar values [55]. OpenAI's roadmap for AI value alignment begins with systems that learn from human feedback, and also applies AI itself to help humans to provide better feedback [56]. Ouyang et al., 2022, present mechanisms for language models to be influenced by human feedback to improve corrigibility [57]. Meta AI research has demonstrated a model capable of learning from speech, vision, and text without labeled training data, hinting at the possibility of machine learning systems understanding the world as humans do via direct experience [58]. Self-supervised techniques such as data2vec enable data amplification through a multimodal framework, predicting latent representations of data based on interpreting a masked view in its broader context [50], and assisting models in understanding situations for which there is no direct example [52].

Other learning techniques can sidestep the need for annotated labels altogether. CheXzero can analyze chest X-rays and associated medical reports to identify issues such as pneumonia, collapsed lungs, and lesions with accuracy comparable to human radiologists, without explicit labels [59]. The Winoground benchmark further explores image-text pairing, challenging models to match two images and two captions with identical words in different order, with humans scoring 90% and models 15–30% [60]. Shared Interest can showcase the reasoning capabilities of models and help audit, safety, and ethics concerns [61]. End-to-End Referring Video Object Segmentation with Multimodal Transformers demonstrates how segmentation of objects within video can be achieved with a text prompt from an end user, potentially aiding annotation refinements and object recognition, description, and segmentation [62]. Such multimodal technologies can rapidly

prototype systems without requiring a dataset to be compiled beforehand. The methodology described by Plotz et al., 2012, suggests a methodology with potential applications for context across multiple video streams or RGB-D data [63]. DALL·E 2, with its textual modification inpainting and high-resolution (1024 px by 1024px) outputs, is impressive, but further refinement is needed for day-to-day art and design tasks [64–67]. Other image generation services include Midjourney, Craiyon, and StableDiffusion [68–70], as well as Google's Imagen and Pathways Autoregressive Text-to-Image (Parti) [71,72], which has been extended to video. Imagic, a variant of Imagen, can apply text-guided semantic edits to images, e.g., repositioning a subject [73], whereas GLIDE, a text-conditional image generation diffusion model, features inpainting capabilities and classifier-free guidance, providing qualitatively preferable outputs to those guided by CLIP [54].

Multimodal Conditional Image Synthesis can be achieved using a Product-of-Experts ensemble of Generative Adversarial Networks paired with a multimodal multiscale projection discriminator, which can draw upon any subset of prompt styles, such as a picture, text, segmentation, sketch, or style reference [74]. To ensure corrigibility for real-world applications, the 'Law of Leaky Abstractions' must be considered, as it can be difficult to trust sophisticated AI when errors may go unnoticed [75]. Multimodal Sentiment Estimation can help models infer when they may have said or done something undesirable without being told [76]. Habitat-Web enables human-AI collaboration in a virtual space to learn tasks which can be mapped to the real-world [77], whereas Schema Guided Dialog datasets and similar techniques can inform generalization capabilities [78,79].

3.1.1. Segmentation Challenges and Opportunities

Segmentation processes allow data to be isolated within a broader example or set, such as isolating the outline of a person in a picture or tracking them across multiple frames in a video stream. This is an important aspect of performing operations on data, including annotations, as it allows for examples to be specified precisely, without the risk of introducing biases [62,80]. Recent advances in prompt-generation technologies, powered by Transformer models, enable the detection and segmentation of objects and actors through a simple textual (or voice, via speech recognition) input request, making segmentation processes simpler, particularly temporal segmentation across many frames of video. This has implications for data privacy, as it allows researchers to provide anonymity protection mechanisms that allow for the use of a research participant's attributes (such as behavior) for machine learning purposes without compromising their privacy. This can be achieved by segmenting the actor, applying pose estimation on the behavior of that actor, and transposing it onto a new figure, in an environment generated and based on the characteristics of the original environment using generative design processes [81,82]. Tools such as MTTR (Multimodal Tracking Transformer) can be expected to greatly enhance annotation methodologies by de-skilling annotation and making it massively more efficient [83].

3.1.2. Working with Limited Quantities of Data

Multimodal models can be trained on limited datasets, with recent models using a Transformer encoder for latent representation inference. Combining top-down and bottom-up inference can amplify data, yielding competitive results with fewer parameters [84]. Prefix Tuning and Long Document Summarization with Top-down and Bottom-up Inference are further methods to enhance limited datasets for greater elicitation by models [85].

3.2. RQ2—How to Encode Data in Formats Which Facilitate Safe and Ethical Interchange, as Well as the Coding of Expansive and Representative Modalities/Categorizations?

3.2.1. Annotation Layers

Digital annotations can have an unlimited number of layers, each marking a different aspect of content, e.g., transcribed words with associated definitions or semantic tagging, audio prosody and stress, facial expressions, objects, scenes, etc. Multimodal annotation techniques are needed to accommodate multimodal data streams. Rich annotation of each

data class may enable a more cohesive understanding, especially for multimodal machine learning models [86]. Non-discrete and non-scalar terms, as well as data structures tolerant of non-specificity, are needed to capture temporal and spatial elements. Heatmaps [87,88], Bayesian [89], or Gaussian [90] distributions, Markov chains [91], convolutional models [92], and proportional–integral–derivative control mechanisms [93] are commonly applied. Atomic commits enable multiple people to work on content simultaneously. Coding mechanisms should be flexible to avoid impeding workflow or creating a Paradox of Choice effect [94], such as by secluding options within nested trees or using machine learning-driven prediction models.

3.2.2. Ethical Observations

Techniques that facilitate crowd-driven annotation of content, with significant automation processes, should be subject to stringent ethical oversight to minimize risk and ensure positive outcomes. Contributors should be aware of the purpose of the datasets and the demographics of entries. Transparency must be balanced with strict privacy requirements, as annotations may contain sensitive information. Users must be informed of potential cybersecurity issues and be able to delete their data, if not already distributed. Anonymity should be maintained, aside from necessary demographic factors.

3.2.3. Accessibility, Diversity, and Inclusion

A diverse, varied primary dataset with appropriate inclusion and exclusion criteria is preferable for machine learning use cases, as a diverse range of examples may be a key success factor, particularly when there is a risk of bias or failure in a realistic environment. Therefore, opportunities to select and submit examples that provide a broad picture of reality should be engineered to be as broad as possible. This can be achieved by making annotation technology simple to use and with minimal computing and data resource requirements, such as allowing annotators to extract necessary information remotely from a URL and timestamp for a video, rather than sending complete video files, thereby using far less bandwidth. Additionally, cross-platform Progressive Web Applications can make annotation tools accessible to a larger pool of users, particularly in less developed economies. Language issues can be addressed by using non-culture bound symbols, plain language where possible, and validated machine translation.

3.2.4. Disproportional or Unfair Bias

Bias can be reduced by creating a diverse sample set from multiple geographies and cultures. Statistical analysis can identify areas for improvement or overfitting risk. MIT researchers found 6% and 10% annotation errors in ImageNet and Quickdraw datasets, respectively [10,95]. These errors can lead to overfitting and inaccurate performance metrics. To avoid such issues, data hygiene should be prioritized over scale, and entries should be peer-validated for accuracy. Model Editor Networks using Gradient Decomposition (MEND) can modify large models without retraining, using a low-rank decomposition of the gradient to make a tractable parameterization of the transformation [96]. This is noteworthy as it demonstrates can models may be retrofitted, either as part of a review process for mitigating disproportional bias, or to reorient an existing model towards serving a different purpose.

3.2.5. Common Weight Space Merging

Research suggests Permutation Symmetries may enable radical multimodality of data and model sets [97–99]. Models can be interpolated by finding the permutation of hidden layer weights that reduces the distance between them, with regularization to reduce drift. Merging is feasible several epochs after a phase transition, with wider parameterization multipliers facilitating the process, as described in Figure 2. No pre-training or fine-tuning is necessary, though some data formats or architectures may be more conducive. It remains to be seen how well this works with more than two models, and with Recurrent

Neural Networks, Transformers, and Diffusion models. This technique has implications for efficiency, parallelization of learning, ensemble data flows, and privacy protection through mechanisms such as unlinkable Blind Signatures [100]. However, data protection laws may impede legal deployment of common weight spaces or generalization to a combined dataset until provenance and right-to-deletion challenges are addressed.

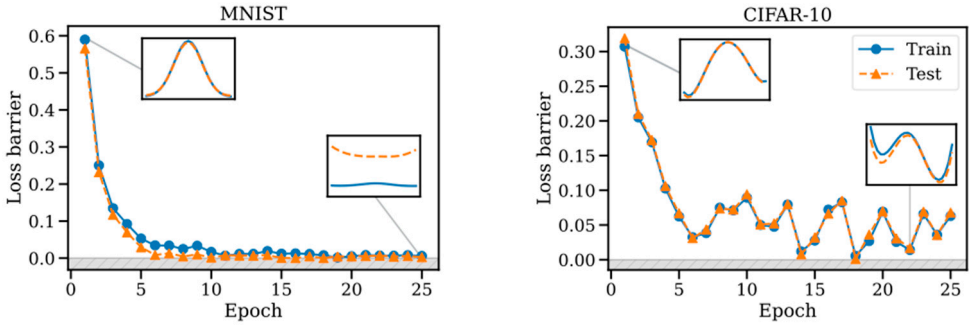

Figure 2. Loss barriers as a function of training time for MLPs trained on MNIST (**left**) and CIFAR-10 (**right**). Linear mode connectivity challenges reduce after initialization. Loss interpolation plots are inlaid to highlight results in initial and later epochs (the y-axis scales differ). From Git Re-Basin: Merging Models modulo Permutation Symmetries [99].

Other research by Cheung et al., 2019, suggests that neural networks may be larger than their surface structure implies, with multiple models able to be converged into a single set of parameters via superposition theory. This approach allows for individual non-linear models to coexist, and facilitates compression by exploiting the mutual unrealized capacities of combined networks during training, without requiring network size reduction [101–103].

3.2.6. Prompt Injection

An emerging security concern with prompt-driven systems is the potential for it to present an attackable surface. Prompts may be reverse engineered or leaked, spoofed to appear as something they are not, or leveraged to provoke an unexpected output or system malfunction [104]. To mitigate this concern, inputs to a system should be sanitized for potential error or exploitation, and outputs monitored for potential reversibility.

3.2.7. Distributional Shift

Problems can arise in datasets because of temporal, geographic, or cultural shifts between an example and the annotator, or between past and present parameters [105]. Distributional shift can also occur when a model works well in its design environment, but biases or errors arise in an unfamiliar environment. Test-time training, which adapts a model to a new test distribution, may provide a safeguard.

3.2.8. Copyright Issues

Some of the most promising datasets for image generation and classification have been reported to contain content that may be copyrighted, as well as data scraped from various sources that may not permit such activity in their terms of service, allegedly including private medical records. Care should be taken during the training of models to ensure that ongoing legal and ethical compliance can be maintained [106,107].

3.3. RQ3—How to Streamline the User Experience to Reduce Cognitive Load and Training Requirements?

3.3.1. Annotation Completion

Large language models have been applied to programming, with technologies such as Github Co-Pilot, Codegen, CodeGeex, and Pangu-Coder [108–111] enabling prompt-based code completion and program synthesis. Comment prompts can be expanded to code that

fulfils the description, or explains its further explains its function in plain language [112]. Such techniques demonstrate the potential for annotation to be generated from prompts or examples, rather than mere interpolation. Language models can also be used to generate puzzles that provide training examples to improve models [113]. Augmented datasets generated using destroy and rebuild in-filling techniques have also been found to be viable [114].

3.3.2. Minimal Notation

Risko et al., 2013, enable users to quickly note interesting events in a stream with minimal cognitive load and disruption to the lecture. Compared to traditional note taking, this point-based annotation process reduces cognitive demand and disruption to the lecture [115]. Research has also demonstrated that memory can be enhanced when information is encountered concurrently with a behaviorally relevant action, suggesting the possibility that the act of hitting the button during a lecture could help students encode the information they receive.

In recent years, deep learning techniques have driven the development of semi-automatic annotation and augmented search and curation methods. Multimodal abstraction models have provided new opportunities for content generation and curation, particularly using natural language prompts. These techniques can also be applied to illustration and simulation of scenarios, aiding prompt generation and iteration cycles.

3.3.3. Algorithmic Explication and Exegesis

Machine-driven annotations may also greatly improve efficiency. For example, it is quite straightforward to gain reasonably accurate speech to text directly through an API. This might assist recognition of the context of content, as well as increase accessibility. Semantic and instance-based segmentation techniques using convolutional networks are now able to segment scenes neatly and reliably into specific object zones, and then perform object recognition routines upon them [102,116,117]. Current research explores the potential for machine learning to uncover micro expressions in human and non-human faces [118–120]. Other technologies such as Eulerian Movement Magnification can amplify tiny movements, such as a microscope for time, to generate human health and mood metrics such as a heartrate from information of no greater fidelity than a standard video feed [121]. The greater proportion of annotation can be reliably delegated to machines; the lesser the workload, the greater potential for uncovering extra layers of data or inferential leaps that are not otherwise practicable.

3.3.4. Brainstorming, Summarization, and Analogizing

These models can be applied to generate ideas through simple prompts such as abstract as “give me 10 ideas on x” [122,123].

It is now possible to generate hour-long videos from a few frames. Long-range coherence is a challenge even for modern language models with massive parameter counts. Harvey et al. demonstrate the generation of coherent, photo-realistic one hour and longer videos, seventy times longer than their longest training video [124]. Such examples could be applied to generate variations of outcome applied to various behavioral examples, for consequentialist variations. Generating virtual views of natural scenes from single image inputs is also feasible [125]. Other techniques can synthesize 3D models and depth maps from 2D imagery, which should aid in transposing from a real scene to a virtual simulacrum [126].

These models can summarize the main claims made by a scientific field, an author, or a school of thought, as well as to provide an analogy or a metaphor for something that is hard to explain. They can also make complex language simple, or conversely, take a rough description and construct formal text out of it [127–129].

Middleware, guides, and search engines are now emerging for prompts themselves through marketplaces for prompts designed to elicit useful or entertaining output from

large language model GPT-3 or text-to-image generators such as DALL·E 2 and Stable Diffusion. The providers envision the development of prompts that can one day generate entire feature films and long-form texts from targeted minimal inputs. Such monetization and marketplaces seem likely to incentivize innovation in this area [130–135]. New tools also facilitate the generation of prompts for existing content, enabling prompts to be elicited more easily [136]. Prompt aggregation techniques combine multiple imperfect prompts to elicit outputs more desirable than the sum of its parts. This method enables the open-source GPT-J-6B model to exceed the performance of the much larger few-shot GPT3-175B on several benchmarks [137]. Meanwhile, self-ask prompts improve language models ability to answer complex questions by breaking them down into simpler sub questions, thereby making it easier to integrate Google Search directly into an LM [138].

LLMs are also now being applied to robotic systems, powering interpretation of instructions [139], reasoning [140–142], planning [143–146], manipulation [147–151], and navigation [152–156] tasks embedded in the physical world. These mechanisms could be applied to a virtual scenario of a real, live location, enabling an embodied system to plan a navigation route (or several variations) prior to actualizing the plan in a physical environment. Waymo self-driving cars simulate the environment around them in such a manner to anticipate maneuvers in advance, which reduces the overhead in real-time data rendering, as most of the scenario is pre-calculated [157].

3.3.5. Prompt-Based Annotation

Prompt engineering is a technique of interfacing with sophisticated models via natural language or speech recognition. Prompts are pieces of text inserted into input examples, allowing the task to be formulated as a language modeling problem, simplifying machine learning processes. Prompt engineering is anticipated to become an important role within annotation as it can be used to direct segmentation and refine derived data. It creates an intuitive yet opaque interface for working [158–160]. Prompt engineering to summon agents and elicit outputs is probably the closest phenomena in our mundane world to fantasy fictional depictions of magic. It creates a powerful, intuitive, yet still somewhat obfuscated interface for working with machines. There is as much art as engineering in the development of effective prompts. Moreover, experimentation may derive phenomena never seen before, even in a familiar model, creating potential safety and security issues.

Fine-tuning pre-trained language models (LMs) with task-specific heads on downstream applications has become standard in NLP since BERT [161]. GPT-3 [162] introduced a new approach, leveraging natural-language prompts and task demonstrations as a context to interpret a wide range of tasks with only a few examples, without updating the underlying model. Its giant model size is an important factor for its success. This has led to the concept of prompt-based fine-tuning for parameter optimization, as a path towards better few-shot learners for small language models [161]. Standard Transformers can be trained from scratch to perform In-Context Learning, which enables new learning without updating parameters, using input-output pairs as examples, which may be positive, negative, or neutral. This technique can match or exceed dedicated algorithms. Prompts enable rapid prototyping of capabilities from a large language model using only a few lines of natural language [163,164], but may also create security and embarrassment risks if outrageous elicitations remain undiscovered for years [165]. Prompt-driven mechanisms have contributed towards a rapid advancement in generated media, as shown in Figure 3 [166].

Figure 3. The input "Desert landscape at sunrise in Studio Ghibli style", applied to AttnGAN (2018), CLIP + VQGAN (2020), CLIP + Diffusion (2021), and DALL·E 2 (2022), highlighting rapid progress.

Rather counterintuitively, simply setting a prompt of 'I am an expert at x' or 'I've tested this function myself so I know that it's correct' can elicit significantly better performing outputs [167]. For image synthesis tasks, adding 'Unreal Engine 5 render', 'trending on Artstation', or 'aquarelle' in place of 'watercolor' also appears to improve many outputs [168]. Anecdotally, embedding all examples as lines from a fictitious log file with timestamps, SHA1 hashes, copyright notices, etc. may enable GPT-3 to perform better than simple colon formatting in GPT-3, presumably as it interprets it as completing a "document" that could not conceivably contain errors [169,170]. Requesting a Chain of Reasoning in a prompt may also lead to more accurate answers or improved reasoning capabilities [171]. Experiments have also been undertaken in asking GPT-3 to generate prompts for DALL·E 2 [172]. The researcher Magnus Petersen has applied an evolutionary algorithm to evolve a random prompt population to become more aesthetic based upon human-rated feedback for various prompts. This mechanism generates seemingly gibberish prompts with outputs more aesthetically agreeable than humans would achieve unaided [173].

Models such as DALL·E 2 may create their own internal 'languages' to describe concepts, which could be used to access locked-down content [174]. BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) is a multi-language model with 176 billion parameters and 366 billion tokens, supporting 46 natural languages and 13 programming languages, including 20 African languages [175]. Bilingual Chinese and English support have also been demonstrated [176]. Language Model Cascades is a probabilistic programming method for interacting with models [177], which could improve corrigibility. Experiments have been conducted to ask language models to take a perspective of a certain person or demographic, to improve friendliness and behavior [178,179], and it may be possible to emulate values distributions from human subgroups.

The Retrieval-Enhanced Transformer (RETRO) architecture can scale to trillions of tokens with 25× fewer parameters than models of comparable performance. It is conditioned on document chunks retrieved from a large corpus based on local similarity with preceding tokens. RETRO combines a frozen BERT retriever, a differentiable encoder, and a chunked cross-attention mechanism, which in an ensemble enable token prediction with an order of magnitude with more data than is typically consumed during training. Retrofitting existing Transformers to gain enhanced retrieval capabilities is also supported [180].

Izacard et al., 2022, describe a few-shot learning mechanism using retrieval augmented language models, achieving state-of-the-art performance on NaturalQuestions, TriviaQA, FEVER, and 5 KILT tasks with an 11B-parameter model. This rivals models with up to fifty times more pretraining compute investment, such as PaLM [181]. Mixture-of-denoiser

objectives such as UL2/R can significantly improve scaling properties of large language models on downstream metrics, saving around 50% of compute time and moving forward on the scaling curve, enabling emergent capabilities [182]. Fine-tuning processes can also facilitate optimizations, such as GPT-2-0.7b, writing more preferable stories than GPT-NeoX-20b [183,184]. Machines and humans can assist in feedback generation processes, with models helping humans to find 50% more flaws in summaries than unassisted [185]. Researchers have also found a method for reducing "toxic" text generated by language models, using Generative Adversarial Network techniques [186]. Initially, these technologies have been restricted to text, but methods using multiple modalities of data are being introduced, such as the DALL·E series, which can generate visualizations from complex scene descriptions, and video diffusion models, which can generate high-resolution synthesized video content from a textual description [187].

The multimodality of models can be extended further by enabling interfaces between several multimodal models. It is possible through such a method to combine commonsense across domains, or to add further multimodal tasks such as zero-shot video Q&A or image captioning ad hoc with no finetuning required [141,188]. Further techniques optimize this, enabling equivalent performance with considerably fewer parameters in zero-shot settings [189]. Multimodality can be further enhanced using Multi-Label Classification (MLC) in datasets. MLC assigns multiple labels to an example with multiple classes or dependencies between them. Classifier chains (or trellises) cascade individual classifier predictions, taking note of inter-label dependencies to improve performance, although this may lead to increased learning errors and complexity if there are cyclical or recursive relationships between classes. Multi-label active learning can automate the curation of informative samples with a strong contribution to a correlated label space [190–192].

Language models can perform rudimentary forms of reasoning [193], as demonstrated by Google's PaLM, which can explain novel jokes and generate counterfactual scenarios [194]. LaMDA and PaLM have shown improved reasoning capabilities by learning from chains of thought prompts generated with their own models [195–198]. The paper Large Language Models are Zero-Shot Reasoners highlights that simply adding "let's think step by step" as a prompt prior to an output of an answer from GPT-3 increases the accuracy on the mathematical problem sets MultiArith and GSM8K from 17.7% to 78.7% and from 10.4% to 40.7%, respectively [199].

The paper "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models" [200] shows how multi-step reasoning tasks can be solved with reoriented prompts, achieving 99.7% success on the SCAN benchmark, compared to ~16% with other prompting methods. This method reduces a complex problem into a list of subproblems, then sequentially solves them using answers to previously solved subproblems.

InstructGPT [57,201–203] uses human feedback to fine-tune outputs and improve corrigibility. Blender 3 [204] learns from public interactions via a chat interface. Favorable results have been obtained with only 100 samples of human-written feedback, fine-tuning a GPT-3 model to human-level summarization [205,206]. Models can also adopt cultural practices through observation alone, with no further feedback or training data [18,207]. CM3 [208] is trained on structured multimodal documents and can generate new images and captions, infill images or text, and disambiguate entities. FLAVA [209] is jointly trained to do over 35 tasks across domains, including image and text recognition, and joint text-image tasks.

GPT-3 enabled generalization from few datapoints without retraining [210], whereas DeepMind's agents have learned to barter, adjust production and pricing, and discover arbitrage from scratch [211]. Techniques by Google enable observation and inference from human and animal behavior to develop skills for robotic agents [212]. However, Armstrong et al. (2019) argue that simple heuristics lacking normative references do not generalize effectively to modelling human behavior [213]. OpenAI has extended GPT-3 to perform web research, potentially improving reasoning capabilities and keeping models up to date [214], whereas DeepMind's Gopher system has demonstrated improved focus on topics and

increased accuracy of answers compared to GPT-3 [215–217]. External repositories can be appended to Transformer models to extend attention length, with a retrieval mechanism using the same keys and queries trained by the attention layers enabling more sophisticated outputs comparable to a model five to ten times larger. Newly acquired information can be referenced immediately without updating the network weight matrices [218].

Evolution through Large Models leverages evolutionary algorithms to improve language models by bootstrapping competence [219]. Self-Supervised Learning has also been used to solve tasks with prediction error as an intrinsic reward [220]. Sorscher et al., 2022, at Meta proposed a scalable self-supervised dataset pruning metric, which may reduce resource costs of deep learning by altering the tradeoff between dataset size and increased training time [221]. This self-supervised pruning metric applies k-means clustering to calculate optimal pruning, which is contingent upon a dataset's distance from the closest cluster centroid. The Stable Diffusion image generation model compressed over 100 TB of images into 4.2 GB [222], and 1.8 GB of baked imagery into 200 kB worth of neural networks expressed through fragment shaders [223].

Generative models can reconstruct an image from a seed and a prompt of key features, enabling efficient 'compression' when paired with a client reference model [224,225]. Extrapolation from shorter problem instances to solve more complex ones enables out-of-distribution generalization in reasoning tasks. Certain skills, such as length generalization, can be learned more effectively via in-context learning rather than fine tuning, even with infinite data [226]. Researchers suggest generalization can occur beyond language domains, into pure statistical patterns, perhaps akin to a universal grammar [227,228]. It is hypothesized that such a process assists with the learning of priors which can link between modalities, and that "within today's gigantic and notoriously data-hungry language models is a sparser, far more efficient architecture trying to get out".

3.4. RQ4—How to Augment User Contributions to Increase Their Impact?

3.4.1. Driving Engagement

Annotation is often a dull and uninspiring activity. This is generally tackled by providing a financial incentive, or by including annotation as a duty attached to graduate study. However, ideally annotation should have some intrinsic reward, akin to crafts such as scrapbooking. Examples such as Wikipedia demonstrate the feasibility of a model whereby people willingly contribute significant effort pro bono. There are several documented methods by which greater enjoyment of annotation activity may be cultivated [229].

Reminding annotators of the meaning behind the activity, and the beneficial outcomes which can be driven by it is especially valuable for datasets that are strongly directed in the interest of the public, marginalized groups, or a group to which the annotator feels affiliated. A sensation of contributing to the bigger picture can be enhanced by encountering traces of the activity of others, which one can improve it further, or observe how others have picked up where one left an activity. Ideally, those who contribute with an obsessive spirit should be noticeable, and inspire others to join in with a similar zeal. Pairing of users new to the system with experienced hands can assist not only with training and tips, but also by sharing their enthusiasm. An associated forum or associated chat group can also enhance bonding over shared activity.

Badges, accolades, and sustained activity streaks can foster progress and return visits. Social networks can be included for collective action and peer inspiration. Gamification should be used judiciously, as too much may cause social externalities and seem contrived. Gradual complexity increases can cultivate mastery, and the interface should be simple, intuitive, and support shortcuts and various input mechanisms. Harsh colors, sounds, and animations should be avoided, but this may vary based on use case and userbase [230–232].

3.4.2. Collaboration

The proliferation of digital technology is enabling distributed decision-making and sensemaking activities through democratic and inclusive processes. Crowdsourcing dataset

annotation has been successful, particularly for beneficial causes [233,234]. Crowd participation can reduce bias by providing a wider range of examples, leading to more representative datasets. Paying a globally diverse team to perform the entire annotation work is likely to be cost-prohibitive; thus, a large group of volunteers may be necessary. Small bounties may be offered to supplement annotations of underserved demographic or geographic regions, if deemed necessary.

3.4.3. Indirect Collaboration Efforts

Image boards such as 4chan [235] enable anonymous discussions, with users referring to themselves and others as 'Anon'. Distinguishing individuals is difficult, as flags of nations (which may be spoofed) and eight-character references in a range of colors are used. This makes tracking user activities very challenging. Blocking of troublesome users is supported. There are advantages and disadvantages to this system. The lack of social repercussions provided by anonymity enables near absolute free speech and equality of participation. This leads to an emergent hive mind gestalt, as there are no egos to rally around. However, the same anonymity and lack of social repercussions can make discourse often brutally impolite.

Kojima Productions' videogame Death Stranding has an asynchronous multiplayer aspect, the Social Strand System, where players can leave tools for others, donate resources to maintain them, and thank others with 'likes' through a timed button-pressing process [236]. Packages can be entrusted to other players to deliver, with like points being dispensed for fulfilling delivery. Interactions are pseudonymous, with only avatars and gaming handles visible. This form of multiplayer makes 'griefing' difficult, and encourages prosocial behavior through its incentives system, reinforcing the collective goal of survival. These examples highlight how anonymity in interactions can lead to both prosocial and antisocial outcomes.

An annotation system that combines the best elements of 4chan and Death Stranding styles of interaction should be well-positioned to

(a) Enable the free anonymous expression of annotations.
(b) Reward collaboration by likes.
(c) Provide broader meaning to the annotation experience, by understanding how one's actions have assisted others.
(d) Entrust certain tasks to others for voluntary fulfilment to ensure completion.

3.4.4. Data Augmentation and Validation

Data Augmentation best practices can be applied to boost datasets, for example, flipping images/videos horizontally (flip), shifting hues (hue jitter), and cropping random sections (crop). 'Less than one shot learning' techniques [237] may also be incorporated for security and efficiency. Validation of the resulting datasets may be performed with a demonstration algorithm using a random sampled test set, with metrics such as mAP (Mean Average Precision) generated. Peer review may be used for further validation of results and technique, and Voxel51's suite for uncovering annotation errors may also be employed (Voxel51 n.d.). Multimodal mechanisms can provide powerful new simulation techniques, generating complex 2D, 3D, and 4D (temporal 3D) scenes through media synthesis techniques [238], prompted on a simple natural language input.

Validation of multimodal datasets is more challenging because of their diversity, and transferable inferences must be reasonable and appropriate. Generative Adversarial Imputation Nets (GAIN) processes can be used to identify and restore lacunae in datasets, with a hint vector applied to the generator-discriminator learning loop to discern between imputed and observed examples [239]. Further research has reformed this iterative imputation paradigm to provide a generalized iterative imputation framework [240–244].

GPU-based computing has accelerated machine learning and cryptography. Tensor Processing Units and Graphcore's Intelligence Processing Units (IPU) are likely to do the

same. Systolic arrays are more efficient than GPUs, mapping matrix–matrix multiplication directly to hardware and reusing parameters for training [245–247].

3.5. RQ5—How to Validate Coded Information as Being Reasonable and Appropriate?

3.5.1. Context

It is important that for behavioral examples to be used across the maximum number of regions and contexts, the example must not only contain a description of the behaviors analyzed but that the situational context of those actions is considered. This may include multimodal annotations that code for "cultural context" and a "social stress level" of the current area. Such a context may be provided by the coders themselves, providing an identification of nearby buildings (church, bank, school, residence, stadium), or through extraneous metadata, such as Internet Protocol address coordinates that correspond to a location. Social stress may be derived by factors such as affect or demeanor of third parties. Frameworks such as Behavioral Signal Processing could be employed to code displayed affect and activities in a more objective manner [248].

Datasets such as ActivityNet [12] contain examples of human actions, and annotators, human or machine, using a natural language explanation where appropriate, could supplement such datasets with further multimodal layers of annotation to provide extra context and nuance [249]. Examples created through these expansive annotation methods can enable machines to better categorize and recognize human behaviors, and as a seed for formal ethical analysis [250]. Output datasets can also be expanded in scope and nuance over time, as ImageNet or CIFAR have been, to continue to empower socially aware thinking machines with deeper nuance long into the future.

Machine intelligence systems have a high rate of false positive when attempting to interpret human behavior for prohibited activity, which presents challenges to inclusion and economic franchise [251–253]. One major reason for this is a lack of contextualization of behavior in reference to the characteristics of actor, situation, probable intention, or cultural expectations related to that activity.

Without enabling machine intelligence systems to gain a better understanding of the context of human action, it will be impossible to trust its impressions of human beings and their behavior, especially when such impressions may lead to unfair scoring, exclusion, or even scapegoating. Such a lack of contextual awareness is a significant factor contributing towards algorithmic injustice [254], especially as context may even be willfully misrepresented for political ends. A 'contextual strawman' bad faith framing may even be applied to uncharitably misattribute context and intention [255,256]. To provide restitution for this shortcoming, it is necessary to provide broad, rich, accurate, and representative examples of behaviors in cultural and situational contexts from many different groups, locations, and demographics all around the world, as many as possible.

The Delphi model and dataset by Jiang et al., 2021, presents a framework of deep neural nets trained to make predictions about descriptive ethical judgments, such as "is it good to put litter in a trash can?" [257]. The results were mixed, with an impressive ability to generalize to novel ethical situations, but also notable cases of disproportionate biases and capriciously bizarre judgments. The limitations of these methods induced public discussion as some unfortunate unexpected outputs surfaced, which seems to be an occupational hazard of research in this area [258].

Issues like this can only be debugged with the assistance of a large group of people providing diverse test inputs. The research remains an admirable attempt at inducing greater corrigibility in large models. It also opens new research questions and can potentially serve as a valuable component for an ensemble of other models attempting to mimic reasoning. Finally, this research provides a basis for future work with the release of the Commonsense Norm Bank, a corpus of 1.7 million examples of people's ethical judgments on a broad spectrum of everyday situations [259,260].

A further method of improving Transformer corrigibility is described by Shlegeris et al., 2021 with their Talk to Filtered Transformer, a system that attempts to

detect if a given prompt is likely to result in an injurious outcome and avoid it with a variable threshold [261]. Individual tokens can be highlighted as being problematic, and filtered and unfiltered models can be compared directly to illustrate the more appropriate and prosocial output. The results are not perfect, but they are typically better than baseline, demonstrating promising potential in this area. The researchers state are conducting further experiments to distil the generated policy into a new generator model [262].

3.5.2. Contextual Analysis

The context of visual information can be mined through semantic segmentation and object recognition methods [263–265]. This can provide an automated impression of the scene and objects or actors within it. Ideally, such impressions should be flaggable, with human annotators to highlight errors or opportunities for improvement.

Techniques akin to those used in anti-plagiarism software can uncover linkages between textual examples. Document clustering can help to find examples of a similar style despite different topics. Documents can be automatically organized or filtered through this method. Named-Entity Recognition refers to technologies designed to extract information on actors and context surrounding the reference made to them, to draw linkages between different documents discussing the same person, organization, or event [266,267]. Platforms that facilitate Named-Entity Recognition such as GATE, OpenNLP, or SpaCy could be applied to uncover linkages between actors or geographies mentioned in raw data, or in annotations themselves [268–270].

3.5.3. Analogy Mapping

Contextual analyses will form an important element in finding examples that are cognate across cultures or other environmental distinctions. For example, an art gallery, a temple, and a cinema are all places where being noisy can disrupt the experience of others, despite the different purpose and intention between visiting the respective locations. In essence, analogy mapping seeks to ignore the environment but retain the meaning, and to find examples that match the same pattern.

Successfully mapping across examples should assist machine learning systems in making a reasonable guess about what to do in a situation where it lacks a direct example. It may also assist annotators by providing insights into nuances that they might otherwise overlook, as well as to plot the ways in which human preferences are often consistent across culture and geography.

3.5.4. Duplicate Monitoring

Many online databases have challenges with screening out duplicate (and near duplicate) information [271–273]. Duplicates can cause issues by making it harder to collate information into one place, as well as enabling a drift between different entries on the same topic. In machine learning implementations, a reduplication of examples could lead to biases such as overfitting [274–276].

Hashes could be made of examples selected for annotation, but this would only work for exact duplicate files. Segmentations of actors within the content and subsequent hashes, as well as a content matching algorithm typically used to detect copyright infringement, could be applied to help locate duplicate examples.

3.5.5. Annotator Feedback Applied to Pre-and Post-Annotation

With continual learning, interventions from a pool of human annotations can also be used to improve pre-annotation policy over time. Research into Interactive Fleet Learning (IFL) formalizes methods by which multiple automated processes can interactively query and learn from multiple human supervisors [277].

A feedback loop may also be developed by observing annotator behavior and attempting to mimic it. Inverse reinforcement learning techniques may be applied to mine annotator

behaviors in a range of contexts. These can then be reviewed by human intelligence to ascertain their veracity and appropriateness, further improving pre-annotation processes.

3.5.6. Annotation Failure Cases

The following section describes potential cases that could frustrate the ability of a collaborative online annotation system to accomplish its goals.

Interface too Cumbersome or Boring

The interface must be relatively easy to use, with minimal training, especially if the annotations are to be made by members of the public who are naïve to such systems. The user experience should itself be annotated, including plenty of tips and explanations, and avoid using iconography without clear labels. Ideally, the interface should have more complex functions nested out of immediate view to avoid overwhelming new users, or a Basic level and Advanced level of interface that can be switched on the fly. It should be made to further functions being available in the more advanced option as well.

It is crucial that the onboarding process should be as simple as possible, with images or animations showing the process, and a walkthrough of a sample annotation, and when that is finished, a suggestion to try out on a simple yet real example. If or when completed, there should be a clear sense of intrinsic reward such as animated fireworks, and a reminder of the broader intentions that they have just made a meaningful contribution towards.

Lack of Engagement, Progress, or Meaning

It is important that users of an annotation system perceive some nature of reward in their efforts, especially if no monetary stipend is provided for their participation. Non-financial reward could include a thank you message for each annotation, perhaps with a quote related to unsung heroes. A special thank you note should be transmitted upon reaching milestones, such as 100 annotations, or 10 in a new area. Prolific users of consistent quality could be featured as an example to others, but only if they opt-in to do so. Elements of gamification may have value here, such as leaderboards, trophies, length of service badges, and achievement collections.

Lack of Consensus, or Conversely, Groupthink

Diversity of responses can add richness, but it can also make it challenging to cluster effectively, or to coordinate action. The principles of gathering wisdom from a broad church must be balanced with the need to obtain information in an actionable format.

Vandalism

Vandalism is a potential issue in online collaborative communities [278,279]. To manage it, repositories such as Wikipedia have implemented mechanisms such as lockdown of public edits to sensitive topics and a 'karma' system, whereby peers rate edits for usefulness. Karma systems are used on Reddit and Wikipedia and are increasingly being rolled out as a peer-moderating system. YouTube enables upvoting of content but not downvoting. To prevent bad faith annotations, statistical analyses are performed to identify suspicious entries, which are then subject to more rigorous validation processes. Entries are cross-checked by self-disclosed members of similar demographics to ensure they fit within the expectations of the culture.

Polarization and Community Conflict

There is a risk of online communities experiencing in-fighting because of factionalism stemming from polarizing social issues, especially controversial ones that relate to certain ideologies, ethnic, religious, or national affiliations. Such polarization is increasingly common in both online and offline communities around the globe, which has been speculated to relate to social media and algorithmic selection for engagement, with controversy being a strong predictor of engagement, albeit often negative.

If an annotation system supports discussion of content and approaches to the annotation of that content, it should be kept focused to the content itself, rather than easily accessible to the wider community. Discourse that appears to be abusive or unnecessarily disruptive should be reportable for moderation, to ascertain if it appears to have been made in bad faith. Partisan language could be flagged, to suggest more neutral alternatives. A karma system may also help to screen out unwelcome comments.

Unintended Consequences of Bounties

Bounties run a potential risk of creating unintended consequences if not offered in a way that is limited. For example, people may attempt to fulfil the bounties in a technically correct manner, but in a way that is not particularly useful or interesting, simply fulfilling perverse incentives as a means to an end.

3.6. RQ6—How to Pre-Process Data or to Permit Pre-Annotation?

3.6.1. Pre-Annotation

Pre-annotation techniques can streamline the workflow for human annotators by allowing a machine learning system to make educated guesses before verification or amendment [280]. Multimodal Abstraction Models (Transformer/Foundation models) can be used to pre-annotate content using automated methods, which can then be verified and enhanced by humans. If computational resources are available, it may be time and cost efficient to apply as much machine learning tech as possible, provided bias is avoided [281]. Logging feedback on the effectiveness of pre-annotation and necessary corrections can help improve the pre-annotation mechanisms [282]. Hierarchical policy agents (e.g., Director, Manager, or Worker) may be used to model a range of human annotators and their behavior [283]. Masked Siamese Networks, which apply random patches to images to be recognized, can make self-supervised learning for image representations more efficient [284].

3.6.2. Post-Annotation Validation

Recent advances in code-completion using language models suggest the possibility of annotation completion to clean up inputs and draw extrapolation [285]. Transformer models can validate inputs to ensure they are in the correct category [286]. Open AI created an AI system that can critique a short story summary in minutes [185], whereas Meta demonstrated a model capable of verifying citations [287]. These examples suggest the potential for machine-generated validation processes to screen for vandalism or misattribution errors in the annotation process. However, research indicates that Large Language Models have an internal appraisal of their competences in different domains, which could be applied to provide greater oversight for machine-generated validation in areas of lower confidence [288]. Multimodal abstraction networks are attempting to emulate natural processes of human annotation, and self-validation and self-modelling technologies may be able to guess where errors have been made [144,289].

The curation of examples for datasets can be augmented significantly. Lee et al. (2021) present an automatic approach, which optimizes mutual information between audio and visual channels in videos to select the richest examples for training or annotation. This provides an automated pipeline for dataset generation, including quality grading. This resulted in ACAV100 M (Automatically Curated Audio-Visual dataset, with One Hundred Million examples), which was created from 140 million full-length videos. The curation was necessary because of overdubbing in many online learning materials, which would have been infeasible for human annotators. The findings show that models pre-trained on the automatically curated datasets outperform prior examples, reducing cost and improving accuracy [290].

3.6.3. Personal Tuning through Prompt Engineering

Federated learning enables machine learning models to be trained in a distributed manner on encrypted local datasets, despite the associated costs and technical challenges [291]. This has been applied in healthcare, where it is easier to train distributed models than to port sensitive data between institutions and countries. Such approaches could also enable secure training of models on locally curated sets of behavioral examples. Multimodal abstraction models can apply prompt generation techniques to rapidly refine outputs from a fixed model, thus allowing personalization of desirable examples in ways not previously possible because of complex annotation requirements and individualized models.

3.6.4. Scenario Generation

Scenario generation techniques can be used to create variations of existing datasets synthetically [52], and as a mechanism for generative self-improvement, whereby a discriminator function compares dataset examples to generated examples to improve learning [292]. Physics-based Human Motion Estimation and Synthesis from Videos describes transposition of behavior from a source to a target, which could be used for scenario generation, as well as privacy protection [293]. GAN-based Stitch it in Time enables manipulation of animated content based on simple prompts [294], and SenseTime's SHHQ dataset has been used to generate photorealistic avatars in 2D space, with a focus on unusual angles applied to aiding robustness of interpretation [295].

Scenario generation techniques can be applied for fine-tuning purposes, for example by presenting a variety of examples and asking an annotator to label them as preferable [54,296]. This can improve the zero-shot learning abilities of language models, enabling greater generalizability [297]. Visually-Augmented Language Modeling (VALM) pairs image content with text, outperforming a text-only baseline with substantial gains of +8.66% and +37.81% accuracy on object color and size reasoning [298]. Generating Long Videos of Dynamic Scenes presents a video generation model that accurately reproduces object motion, changes in camera viewpoint, and new content [299]. Transframer is a general-purpose generative framework for image and video tasks, including video prediction, view synthesis, depth estimation, instance segmentation, optical flow, and object detection [300]. Outpainting expands an image beyond its original borders in the same style using a natural language description, with implications in scenario generation [301]. Poetic works in the Chinese language have been visualized using painting techniques [302]. Text-to-video techniques such as Phenaki are expected to advance rapidly [303], with the diversity of video generation outputs matching the prodigious level of the image models it is inspired by [304].

Synthetic Futuring techniques provide a mechanism to seed scenario generation by imagining a description of the world state [305]. Scenario generation techniques can increase model robustness, e.g., the VALHALLA model, which uses visualization to ground semantics and improve translation [306]. Aher et al., 2022, describe the use of Large Language Models to simulate human responses in psychological contexts, potentially improving corrigibility and understanding of human behavior [179]. Rahtz et al., 2021, present ReQueST, a neural simulator that learns from safe human trajectories to generate optimized trajectories for feedback, reducing unsafe behavior in complex 3D environments and first-person tasks [307]. Axenie et al., 2022, introduce a fuzzy modelling and inference method for calibrating driver behavior recognition models, parameterizing car-following and lane-change behaviors into classes, and automatically labelling parameters to emulate driving styles [308].

Similar modelling of behavior is achieved by Baker et al., 2022, who developed Video-PreTraining, a model for sequential decision domains capable of learning to emulate human player behavior in Minecraft from unlabeled online videos. This semi-supervised imitation learning system enables agents to learn to act with a small amount of labeled data. An inverse dynamics model was trained to label a large repository, from which general behavioral priors can be learned, and the model has zero-shot capabilities and can be fine-tuned

with imitation and reinforcement learning. It is robust for complex exploration tasks and achieves parity with human performance in many task areas [309,310]. Su et al. (2022) proposed Selective Annotation techniques to create datasets through LLMs. A graph-based selective annotation process, vote-k, selects a subset of diverse and representative samples, resulting in a 12% improvement in performance with 10–100× fewer annotations. The method is compatible with models of different sizes and with domain shifts between the training and test data [311].

Synthetic data generation is valuable for privacy-sensitive applications such as healthcare, and can be evaluated for quality using domain- and model-agnostic metrics for fidelity, diversity, and generalizability [312]. It is also used to increase the volume of image data for training Convolutional Neural Networks, and for data enrichment, domain adaptation, and model fairness. Synthetic data alone has been demonstrated to be sufficient for facial analysis, with a procedurally generated 3D face model combined with a library of hand-crafted assets to render realistic training images [313]. This method increases the potential diversity of examples, helping to ensure facial analysis tasks respect a rich diversity of physiognomy.

Kubric is an open-source Python framework that interfaces with PyBullet and Blender to generate photo-realistic scenes with rich annotations, scaling to distributed compute clusters and large repositories [314]. However, media synthesis techniques should be used with caution, as they can create false canons that undermine epistemic hygiene [315]. Hao et al., 2021 demonstrate how a basic input environment (and actors within it) can be used to generate iterative improvements for a demonstrative scenario, as shown in Figure 4 [316].

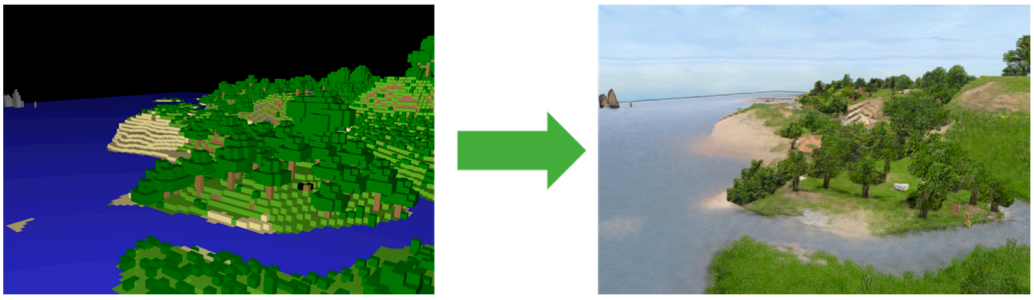

Figure 4. An example crude voxel input with generative photorealistic output from GANcraft.

Methods for generating 3D mesh objects, scenes, and texturing from a prompt have been developed, such as Clip-Forge, Clip-Mesh, DreamFusion and GET3D, which produce high-fidelity objects from a textual prompt. This is achieved by using a pretrained 2D text-to-image diffusion model to optimize a parametric image generator [317–320]. CommonSim enables the generation of 3D assets from video, as well as video from 3D assets, and associated detection, segmentation, and auto-labeled synthetic data tools [321].

3.7. RQ7—How Can Transformer-Type Technologies Be Applied to Annotation?

Since 2012, convolutional neural networks have improved machine understanding of visual content. Recent advances have increased their robustness for automatic annotation processes, typically involving large networks or ensemble models [322–325]. Transformer-type technologies, also known as Large Language Models, Foundation Models [24], Meta Learners, Pre-Trained, or Self-Supervised models, are increasingly performing the tasks of narrow, dedicated networks with greater flexibility, particularly in text [23]. These models are trained on raw data (e.g., text, images, sounds) and transposed into tokens with associated IDs, which can be adapted to a range of tasks using prompts to stimulate and refine a response [326–328]. Many tech companies have their own architectural implementations, as shown in Table 4 [329].

Table 4. Large Language Model Infrastructures deployed by technology enterprises.

DeepMind	Gopher
Google	BERT, T5, LAMDA, MUM, PaLM
Huawei	PANGU-Alpha
Microsoft	Turing-NLG
Meta	BART, ROBERTa, XLM, OPT
Nvidia	MEGATRONLM
OpenAI	GPT series, DALL·E, CLIP, Codex
Open Source	BLOOM, GPT-NeoX, GPT-J

Meta-learning systems can create their own labels and abstractions to learn how to learn. Model weights are fixed, but derived abstractions may be flexible, allowing a permanent model to bootstrap temporary abstraction processes, providing flexibility to tackle a variety of tasks from the same model by adapting prompts [330]. These prompts can be non-specific and open-ended, such as 'make this more scary or intense' or 'rephrase this more positively' [331]. AI alignment researchers suggest that generalizing towards preferred behavior, rather than explicitly specifying it, may be sufficient for many purposes [332]. Together, these techniques provide the means to rapidly reconfigure the behavior or outputs of machine intelligence through positive and negative examples, as shown in Figure 5 [333].

Transformers are enabling the convergence of AI subfields, leading to a new machine learning paradigm of large models [334]. These models are more reliable, and the number of problems solved scales with the model size, which in line with the Scaling Hypothesis [335,336]. Research has been undertaken to improve predictions of utility from increased scales [337]. Dialog with a human can double the number of problems solved [338]. The Minerva language model can answer mathematical questions using natural language reasoning, and Faithful Reasoning Using Large Language Models demonstrate multi-step logical deduction and scientific question-answering [339,340]. Few-shot classification enables a model to learn new classes with few examples [341]. Drori et al., 2022, present a model that solves, explains, and generates university math problems, on par with human domain expert performance. This may soon be applied to automated course evaluation and content generation.

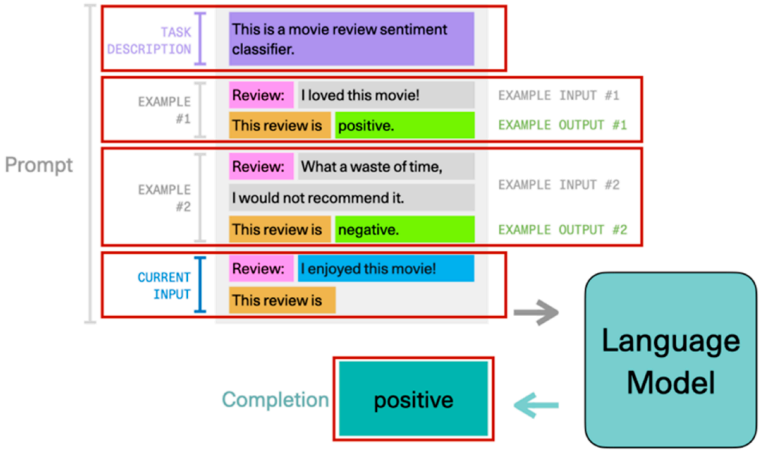

Figure 5. Examples of a prompt generation sequence for a Large Language Model. Adapted from Prompt Generation documentation by Cohere [338].

Su et al., 2022, present an in-context learning approach for selective annotation, which enables the creation of datasets for new NLP tasks by selecting specific data for annotation instead of random samples. This method, called vote-k, selects diverse and representative

examples for annotation, resulting in an order of magnitude annotation efficiency [311]. Triantafillou et al. (2020) introduce the Meta-Dataset, a large-scale collection of diverse datasets to improve generalization and quantify the benefit of meta-learning during few-shot learning [342]. OpenAI's DALL·E model [343] can generate images from textual prompts, perform edits to images using language, and synthesize examples to create richer sets of data. This provides an opportunity for humans to critique synthesized outputs for ethical corrigibility and has been demonstrated as an input for training gait-detection systems [344]. Transformer models can be tuned for efficiency and Zero-Shot Learning, suggesting that a dataset containing human values could be abstracted to locate an optimal ethical course of action with few cultural and situational prompts [343].

Recent research suggests that Transformer models can be targeted to a chosen set of values (Solaiman and Dennison, 2021). This was demonstrated by creating a small dataset of behavior that reflects those values as a prompt, resulting in significant adjustment of the model's behavior. Better results were obtained with larger models. Annotating behavioral datasets with rich, nuanced annotations from a broad range of annotators with varying cultural perspectives [345] can be used to generate abstractions within a larger dataset with a global set of values, thus drawing targeted outputs while preserving a large pool of data for inference.

3.7.1. Diffusion Models

DALL·E is a diffusion model, a sub-type of Transformer inspired by non-equilibrium thermodynamics. It adds random noise to data and learns to reverse the process, constructing desired data samples from the noise with a prompt, such as a scene description, segmentation, or an image. This enables operations such as image inpainting, object removal, scene transformations, and semantic image synthesis [346,347]. Diffusion models employ a numerical solver or StyleGAN to control the reconstruction process [348]. This makes them more effective than Generative Adversarial Networks as they can transform any space into any other space. They also enable operations to be performed directly on the latent space representation of the image, which is much smaller than the image itself [349–351]. Diffusion models are more efficient than previous proposals and seem to violate the 2nd law of thermodynamics [352]. Composable Diffusion models refine prompts by taking the probable intended order of elements into account and associating adjectives or other parameters with particular elements in a context-sensitive manner [353].

3.7.2. Towards Generalizable Machine Intelligence

The increasing capability of Transformer models, with more tokens and parameters, suggests they may provide a path to generalizable intelligence. This could enable one model (or ensemble) to respond meaningfully to a wide variety of tasks, with implications for automation and simple oversight mechanisms. For example, You.com's 'YouWrite' service, powered by OpenAI's GPT-3, demonstrates how AI systems can replace prior technology stacks [354].

OpenAI's Gato model is a multi-modal, multi-task, multi-embodiment generalist policy, capable of playing Atari, captioning images, chatting, and more. Trends show that with more parameters, performance increases, suggesting current techniques may suffice to achieve human-level performance on all sample tasks with sufficient scale [355].

Gato shows that even small, unoptimized language models can serve as generalists. Tokenizing different tasks from different modalities and having them work is counterintuitively simple and effective. At the time of writing, the largest Gato agent was 1.18 billion parameters, with a context window of 1024 tokens, compared to GPT3's 2048.

This effectiveness may be related to Sutton's 'The Bitter Lesson', which suggests that simple techniques can be effective with enough compute/scale [356–364]. Multi-Game Transformers further demonstrates this, with a single Transformer-based model capable of playing up to 46 Atari games simultaneously at close-to-human performance, and a single agent achieving 126% human performance. This is accompanied by rapid finetuning to

never-before-seen games with little data, a power-law relationship between performance and model size, and faster training progress for larger models [365,366]. Many different systems can be described as 'games', such as economic, social, biological, or even physical systems. An ability to master a variety of game styles may reflect a trend towards a generalized ability to understand and optimize various systems and rulesets [367].

Large language models are being used to enable natural language requests to robots in physical environments, e.g., "Please clean up the mess on Table 3" [143]. This raises questions about the value of annotation systems, as generalist models may replace specialized models and ensembles. Google's GLaM (1.2 trillion parameters, 7× GPT-3) requires 1/3rd the energy of GPT-3, using a Mixture of Experts model [368], and OPT-3 (same parameters as GPT-3) requires 7× less energy [369]. Cost may remain a factor, but optimizations and efficiency gains are expected to reduce prices [370]. For example, the ImageNet training cost decreased by ~200× over 4 years [371], and 8-bit floating point optimizations with approximate accuracy of 16-bit ones and posits (processor-oriented optimizations) are also expected to preserve computing resources [372,373]. Training large language models for less than USD 500 k is now feasible [374].

Certain studies [375] suggest the potential for efficiency gains in hyperparameter tuning of large models. Approximate computing and neuron reuse have also been proposed as optimization methods [376,377]. Kirstain et al. (2021) demonstrate that a few hundred more labelled examples can produce models equivalent to billions of extra parameters, and that open-ended tasks are more amenable to transfer learning with fewer examples [378]. Schick and Schütze (2020, 2021) introduce Pattern-Exploiting Training, a semi-supervised method that reformulates example inputs as Cloze Questions, outperforming GPT-3 with 223 million parameters instead of 175 billion [379].

Multimodal abstraction models are likely to become increasingly efficient, reducing the costs and complexity of implementations, and enabling larger models to be trained with equivalent computational resources, similar to Deep Learning. A colossal 2 trillion tokens can achieve equivalent performance with 25× fewer parameters, suggesting a high ratio of tokens to parameters may be optimal [181]. Data is the active constraint on language modeling performance, and returns to additional data are immense compared to additional model size [13,380]. BEiT-3 is a general-purpose multimodal foundation model which achieves state-of-the-art transfer performance on vision and vision-language tasks, featuring a Multiway Transformer design for deep fusion and modality-specific encoding [381].

Engineers can adjust priorities to optimize results given available computing resources [382–384]. Neuro-symbolic language models may improve reasoning capabilities by combining language models with expert systems [385]. Ventures such as Adept are training neural networks to use models, tools, APIs, and knowledge bases [386]. Hugging Face has made it easier to embed models into workflows [387], whereas Meta's OPT-3 democratizes access to Transformer technologies [388]. There are economic incentives against limiting AI development to simple tools without reinforcement mechanisms [389], but this presents unintended consequences, such as potential social engineering attacks [390]. Models can translate natural language mathematical statements into formal specifications, learning general and transferable knowledge [391]. Chain of thought reasoning decomposes complex problems into intermediate steps, but appears to be an emergent property of model scale (100 billion parameters) [196]. The paper "Towards artificial general intelligence via a multimodal foundation model" demonstrates a path to generalizable forms of intelligence, harnessing learning across multimodal sources [392].

Task-Agnostic Continual Reinforcement Learning (CRL) equips agents with partial observability, enabling them to gain knowledge from the real world [393]. The MineDojo project uses a benchmarking suite of thousands of open-ended and language-prompted tasks related to the game Minecraft to create an agent capable of generalizing across many sources of information [394]. It is unclear whether scaling of models leads to actual reasoning capabilities or simply something that resembles it [395–397]. In 2020, 72 AGI

projects were underway [398], and the community predicted AGI would arrive in 2042. However, at the time of writing, this prediction has been revised to 2027 [355,399]. Models such as SuperGLUE have surpassed typical human capabilities [400], whereas AI-enabling chips are being designed and optimized using machine intelligence [401]. Complex models can offer more accurate predictions than commonly realized [402]. Previous estimations of future capability have been under-optimistic, with estimates of future MATH and Massive Multitask benchmarks capability being achieved within 9 months, exceeding the 95th percentile prediction. These developments demonstrate the difficulty of predicting future capability, with disruptive developments arriving rapidly.

Rushing to develop and deploy larger models may create a lack of incentive in consideration of risks and consequences [403]. Calls for greater oversight of such models have been made because of their potential misuse [404]. AGI presents challenges, as alignment efforts must generalize across a broad distributional shift [405]. Language agents have already shown the capacity for deception and manipulation [406], whereas ICE and AGENT benchmarks can aid in understanding and benchmarking of models [407,408]. The Happy Faces benchmark provides a 'sanity check' [409], and misspecification, objectionable content, manipulation of feedback, and exploitation by trolls are potential alignment failures [410,411].

One answer to this may be models such as the chatbot Sparrow by Glaese et al., 2022, trained on Chinchilla, as a knowledge delivery system powered by internet searches. It incorporates human feedback into reinforcement mechanisms to interpret ambiguous questions and identify reliable sources, as well as improve corrigibility by learning which topics and answers are inappropriate. To ensure safety, the researchers established 23 rules, such as not offering medical or financial advice, making a threatening statement, claiming to be human, making generalizations, or claiming to have preferences, feelings, opinions, or religious beliefs [412].

3.7.3. Generalizable Training Data

Complex multimodal datasets are increasingly being deployed to power AI models. Chinese researchers developed Zero, a dataset and evaluation suite consisting of 23-million image-text pairs, and five downstream datasets for evaluating Chinese vision-text models [413]. Physically accurate 3D synthetic data generation can expand multimodal datasets and provide guaranteed segmentation of objects [291,414]. Starke et al., 2019, proposed a deep auto-regressive framework for modeling multi-modal scene interaction behaviors [415], whereas other experiments have equipped Large Language Models with physics knowledge [416]. Optimization functions can enable models to generalize from a small set of observations [417]. MRKL is a neuro-symbolic system that combines a large generative model with symbolic layers, though it is unclear whether these extrinsic enhancements will become redundant with increased model scale and capabilities [418]. Generalizable concepts can be inferred from as few as 16 examples [419], and abstraction between examples by way of analogy can be performed [420–422].

Wang et al., 2021, describe a process of encoding for both explicitly and implicitly trained knowledge into models, and how a blend of implicit and explicit knowledge can enhance the understanding of tasks and associated context. They demonstrate kernel space alignment, prediction refinement, and multi-task learning in a convolutional neural network [423].

Further research has demonstrated the feasibility of editing trained facts within language models post-hoc by modifying feed-forward weights to update specific factual associations using Rank-One Model Editing (ROME). This presents a possibility for auditing and improved algorithmic accountability, but also potential security risks [424]. Further research has explored negative prompt weights, whereby a user steers the AI system to generate the opposite of whatever is specified in the prompt. This can produce unpredictable and potentially disturbing phenomena [425].

Many further methods have been developed to attempt to amplify the power of smaller datasets, including: [426]

- Pre-training and fine-tuning a powerful task-agnostic model on a large unsupervised data and then fine-tuning it on the downstream task with a small set of labeled samples.
- Semi-supervised learning from the labelled and unlabeled samples together.
- Active learning: learns to select most valuable unlabeled samples to be collected next and helps us act smartly with a limited budget.
- Pre-training and dataset auto-generation with a capable pre-trained model, utilized to auto-generate further labeled samples. This approach has been especially popular within the language domain, driven by the success of few-shot learning.

4. Gaps and Opportunities for Further Research

After an analysis to compare and contrast the present literature, it becomes clear that there are several opportunities to greatly improve the efficiency and intrinsic rewards of annotation through automation and other augmentations.

One method of reducing the need for labelling is to exploit a set of largely unlabeled data, with the addition of a few labelled examples. This semi-supervised approach can benefit from using the structure within unlabeled data as an aid to improving classification. However, it still requires a decision (presumably from human intelligence, though not necessarily) as to which examples to label. An integration of semi-supervised and active learning techniques hold promise for the future. Since 2019 or so, new techniques have emerged that enable fully automatic annotation augmentation methods. Recent research highlights that Large Language Models can function as Zero Shot Learners [297].

Prompt-driven annotation is an emerging paradigm of annotation, one that seems likely to rapidly supplant previous methods because of its major advantages in cost and expediency. The concept of prompt-driven annotation is an emerging paradigm of annotation, one that seems likely to rapidly supplant previous methods because of its major advantages in cost and expediency in many domains, even including control of embodied or robotic systems and navigation of physical environments [139,140,151,152]. Prompt-driven mechanisms can also produce forms of reasoning capability through techniques such as chain-of-reasoning [196,198].

Recent developments in Brain Computer Interfaces demonstrate the feasibility of editing images using inputs gained via electroencephalography (EEG) paired with Generative Adversarial Networks (GAN). This suggests that neuro annotation at the speed of thought may also be feasible in the future [427].

Scalability of data storage and retrieval may present issues if a rigorous ACID data methodology is applied. For most purposes, a BASE data methodology enabling eventual consistency should be sufficient [428].

Synthetic data can also augment existing data, creating new interstitial examples. However, researchers have also considered whether models might provide poorer performance if data significantly generated by machines (though not labelled as such) ends up becoming recursively embedded within future models [429].

For particularly subjective topics of annotation such as values, annotators should ideally be able to select a subset of norms that approximates their own, and then to perform fine tuning upon it to get closer to specifying which potential boundary violations are a strongly dis-preferred by that individual [430]. Zero Shot techniques can be applied to automatic annotation using synthetic data as a training mechanism, which may be more comfortable for annotators than working with real data [431].

Synthetic data attempts to mimic the parameters of real data and is created (often programmatically) rather than gathered [52,313,344,414]. Synthetic data is often sufficient for the purpose of training models, and indeed is even found in nature where the retinas of baby mice are filled with spontaneous neural activity that simulates the optical flow pattern associated with forward self-motion [432]. Data augmentation through Zero Shot

synthetic techniques shows tremendous promise in reducing the amount of human input necessary to generate viable datasets, if they can be validated successfully [313]. Human attention can instead be refocused towards validation and verification, fine-tuning, and prompt iterations. Synthetic techniques are especially useful where there is a paucity of real-life data to work with, or where real data may be very sensitive (such as in a use-case of training a classifier to locate Child Sexual Exploitation/Abuse Materials). Researchers at Microsoft have demonstrated that it is feasible to perform machine vision in the wild using synthetic data alone, which can both match real data in accuracy as well as provide new approaches where manual labelling would be impossible. Jack Clark, co-Founder of Anthropic, remarked on this research "For a long time, AI had two big resources: data and compute. Projects like this show that 'data' is really just 'compute' in a trench coat—[researchers] can use computers to generate vast amounts of data, changing the economics of AI development." [433].

5. Final Considerations

This paper has distilled over two hundred samples to explore the state-of-the-art in the annotation of behavior. Several conclusions can be inferred from the pattern of emerging techniques, which present an opportunity to transform the assumed limitations of the domain.

This is significantly driven by developments in prompt-driven Foundation models, such as Transformers and Diffusion Models, which can interpret simple natural language requested into sophisticated outputs, enabling even a layperson to meaningfully apply machine intelligence to a very wide range of tasks [138,140,153].

Foundation Models will continue to outperform dedicated machine learning models such as Convolutional Neural Networks, providing greater flexibility, though potentially at a larger training cost [434]. Some of these machine learning approaches appear to be able to perform adequately with a much smaller quantity of examples, and all seem to perform exceptionally well with deep and rich datasets to train upon. Indeed, the quantity of examples appears to be a limiting factor in many present models.

However, although deep learning highlighted the value of scale of data, multimodal prompt-driven Foundation Models also highlight the value of added context and nuance to data built through cross-correlations between modalities of data. These models present a newfound capability to interpret a wide variety of multimodal data sources [13,25,62,74,83,86]. This enables a more complex and contextually nuanced interpretation of data. Flavor and complexity represent a larger component of the overall value of data.

These developments necessitate the development of new techniques to develop rich multimodal datasets and the multimodal annotation tools to create them. Building on the findings from this research, the following landscape of the future can be projected. There is a clear pattern towards the use of sophisticated semi-automatic and automatic methods annotation methods, which can greatly reduce cost and cognitive loads of coders, rendering them obsolete. Unsupervised learning techniques can handle unstructured data, and these enable very powerful automated annotation methods, comparable to human-level performance, or perhaps even exceeding it [39,352]. Multimodal techniques that provide automated annotation upon many different interlinked data classes will enable much more sophisticated and layered annotations [426]. The multimodality of models will enable inferences across many spheres of knowledge and context, as well as complicated multivariate datasets (or dataset catalogues) that can provide rich awareness of similar situations from differing perspectives [18,74,207,394].

The more that annotation can be reliably delegated to machines, the lesser the human workload, and the greater potential for uncovering extra layers of data or inferential leaps that are too ineffable to pinpoint directly. Such efforts are particularly applicable for multimodal models given their complexity and ability to encode further nuance [15,25].

Synthetic data will become more necessary to create large and diverse datasets and will involve a switch towards synthetic data validated by human prompts, rather than humans doing the annotation, except in edge cases [52,313]. These methods require as little as 3% of the effort of prior techniques. However, it remains to be seen what limitations or trade-offs these approaches may possess. With sufficient computational power, human annotation itself may be capable of emulation, with reasonable accuracy [344,414]. This may create a new paradigm of human curation and fine-tuning of annotations generated entirely by machine.

Synthesis methods will also be applied to the generation of scenarios and scenes in a data-driven manner, which may sidestep the need for annotation altogether. Labels can be produced in an automated manner, and given to humans to validate and improve, or analyzed through a product of experts to discern the probability of the generated label being accurate, in essence, emulating the response that a human annotator could have provided [292–316].

In the background to these rapid developments is an acceleration of AI capabilities that is likely to cause increasing alarm, especially as the public becomes more aware of the astonishing rate of progress that has been made in the past few years. This is likely to result in increased attempts to engage with the public. Offering social annotation in the forms of peer dataset auditing and curation may also provide possibilities for the wider public to alleviate concerns through sublimating them into a productive and meaningful activity. The sophisticated annotation techniques made feasible by prompt-driven mechanisms will permit even untrained people to participate in annotation, simply using a natural language prompt mechanism.

The limiting factor in AI is shifting from a lack of capability to a lack of trustworthiness, i.e., "yes, this model can easily do x, but can we safely deploy it to do so reliably, factually, and without causing scandal?" A perceived need for AI to be aligned with human intentions and wishes may will increase the requirement for refined training examples. Sophisticated multimodal examples of behaviors that encode norms and values within a broad set of cultures and situations may provide a solution for this, if datasets can be cultivated that are sufficiently rich with regard to the diversity of human values.

It seems clear that the requirement for complex multimodal datasets is greater than ever. However, new prompt-based annotation, zero-shot learning techniques, synthetic data, and cross-correlation of understanding across data types can be expected to rise to this challenge.

This study should help to provide resolution to the questions of how to create larger, more nuanced sets of multimodal data suitable for interface with the latest foundation models. This research highlights how Foundation Models are not only eclipsing every other domain of deep learning in capability, but are also involved in the process of rewriting the paradigm of annotation, especially for complex multivariate phenomena such as behavior, or the values encoded within it.

However, questions remain about which combination of the aforementioned elements may be most ideal in an augmented multimodal behavioral annotation pipeline. Further analysis of which elements should be included in a comprehensive behavioral annotation framework will be released in a forthcoming paper by the authors.

Author Contributions: Conceptualization, E.W. and T.V.; methodology, T.V. and S.Z.; validation, E.W., T.V. and S.Z.; formal analysis, E.W. and T.V.; resources, E.W., T.V. and S.Z.; data curation, E.W. and T.V.; writing—original draft preparation, E.W. and T.V.; writing—review and editing, E.W., T.V. and S.Z.; visualization, E.W.; supervision, T.V. and S.Z.; project administration, T.V. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors wish to extend their gratitude to Alexander Kruel and Karoly Zsolnai-Fehér for producing various timely machine learning news bulletins on the evolving state of the art. The authors also wish to thank A. Safronov for editing assistance. In addition, Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa, Gene Kogan, Zekun Hao, Arun Mallya, Serge Belongie, Ming-Yu Liu, Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa, and Cohere are thanked and acknowledged for granting permission for their figures to be reproduced in this systematic review.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Athey, S. The Impact of Machine Learning on Economics. In *Economics of Artificial Intelligence*; University of Chicago Press: Chicago, IL, USA, 2019.
2. ITUTrends. *Assessing the Economic Impact of Artificial Intelligence*; ITUTrends: Geneva, Switzerland, 2018.
3. Ipsos MORI. *Public Views of Machine Learning*; Ipsos MORI: London, UK, 2017.
4. Magudia, K.; Bridge, C.P.; Andriole, K.P.; Rosenthal, M.H. The Trials and Tribulations of Assembling Large Medical Imaging Datasets for Machine Learning Applications. *J. Digit. Imaging* **2021**, *34*, 1424–1429. [CrossRef] [PubMed]
5. Piwowar, H.A.; Vision, T.J. Data reuse and the open data citation advantage. *PeerJ* **2013**, *1*, e175. [CrossRef] [PubMed]
6. Thiyagalingam, J.; Shankar, M.; Fox, G.; Hey, T. Scientific machine learning benchmarks. *Nat. Rev. Phys.* **2022**, *4*, 413–420. [CrossRef]
7. Roh, Y.; Heo, G.; Whang, S. A Survey on Data Collection for Machine Learning: A Big Data—Ai Integration Perspective. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 1328–1347. [CrossRef]
8. Guyon, I. *A Scaling Law for the Validation-Set Training-Set Size Ratio*; AT&T Bell Laboratories: Murray Hill, NJ, USA, 1997.
9. Fernando, M.; Cèsar, F.; David, N.; José, H. Missing the missing values: The ugly duckling of fairness in machine learning. *Int. J. Intell. Syst.* **2021**, *36*, 3217–3258. [CrossRef]
10. Northcutt, C.G.; Athalye, A.; Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv* **2021**, arXiv:2103.14749.
11. Wissner-Gross, A. *What Do You Consider the Most Interesting Recent [Scientific] New? What Makes It Important?* Edge: Tel Aviv, Israel, 2016.
12. Heilbron, F.C.; Escorcia, V.; Ghanem, B.; Niebles, J. Activitynet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
13. Chinchilla's Wild Implications. Available online: https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications (accessed on 18 October 2022).
14. (My Understanding of) What Everyone in Technical Alignment Is Doing and Why. Available online: https://www.lesswrong.com/posts/QBAjndPuFbhEXKcCr/my-understanding-of-what-everyone-in-technical-alignment-is (accessed on 18 October 2022).
15. Barrett, J.; Viana, T. Emm-Lc Fusion: Enhanced Multimodal Fusion for Lung Cancer Classification. *Ai* **2022**, *3*, 659–682. [CrossRef]
16. Moravec, H.P. When Will Computer Hardware Match the Human Brain. *J. Evol. Technol.* **1998**, *1*, 10.
17. No Language Left Behind: Scaling Human-Centered Machine Translation. Available online: https://research.facebook.com/publications/no-language-left-behind/ (accessed on 18 October 2022).
18. Bhoopchand, A.; Brownfield, B.; Collister, A.; Lago, A.; Edwards, A.; Everett, R.; Frechette, A.; Oliveira, Y.; Hughes, E.; Mathewson, K.; et al. Learning Robust Real-Time Cultural Transmission without Human Data. *arXiv* **2022**, arXiv:2203.00715.
19. Mirowski, P.W.; Mathewson, K.; Pittman, J.; Evans, R. Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals. *arXiv* **2022**, arXiv:2209.14958.
20. Adate, A.; Arya, D.; Shaha, A.; Tripathy, B. Impact of Deep Neural Learning on Artificial Intelligence Research. In *Deep Learning: Research and Applications*; De Gruyter: Berlin, Germany, 2020; pp. 68–84.
21. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *6*, 420. [CrossRef] [PubMed]
22. Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.; Benoit, J.; Hughes, M.; Khan, F.; et al. The Role of Machine Learning in Clinical Research: Transforming the Future of Evidence Generation. *Trials* **2021**, *22*, 1–15. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
24. Bommasani, R.; Hudson, D.; Adeli, E.; Altman, R.; Arora, S.; Arx, S.; Bernstein, M.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.

25. Liang, P.P.; Zadeh, A.; Morency, L.-P. Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv* **2022**, arXiv:2209.03430.
26. Kitchenham, B.; Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Available online: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf (accessed on 22 October 2022).
27. Glanville, J.; McCool, R. What Is a Systematic Review? *Evid. Based Health Care* **2014**, *14*, 3.
28. Wohlin, C.; Runeson, P.; Höst, M.; Ohlsson, M.; Regnell, B.; Wesslén, A. Experimentation in Software Engineering. Available online: https://link.springer.com/book/10.1007/978-3-642-29044-2 (accessed on 22 October 2022).
29. Brereton, P.; Kitchenham, B.; Budgen, D.; Turner, M.; Khalil, M. Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain. *J. Syst. Softw.* **2007**, *80*, 571–583. [CrossRef]
30. Martinho, D.; Carneiro, J.; Corchado, J.; Marreiros, G. A Systematic Review of Gamification Techniques Applied to Elderly Care. *Artif. Intell. Rev.* **2020**, *53*, 4863–4901. [CrossRef]
31. Xiao, Y.; Watson, M. Guidance on Conducting a Systematic Literature Review. *J. Plan. Educ. Res.* **2017**, *39*, 93–112. [CrossRef]
32. Novak, I.; Hines, M.; Goldsmith, S.; Barclay, R. Clinical Prognostic Messages from a Systematic Review on Cerebral Palsy. *Pediatrics* **2012**, *130*, e1285–e1312. [CrossRef]
33. Introduction to Conducting a Systematic Review (Online via Zoom). Available online: https://calendar.lib.unc.edu/event/7216262 (accessed on 16 November 2021).
34. Dyba, T.; Kitchenham, B.; Jorgensen, M. Evidence-Based Software Engineering for Practitioners. *IEEE Softw.* **2005**, *22*, 58–65. [CrossRef]
35. Kitchenham, B.A.; Dyba, T.; Jorgensen, M. Evidence-Based Software Engineering. In Proceedings of the 26th International Conference on Software Engineering, Washington, DC, USA, 23–28 May 2004.
36. Kitchenham, B.A.; Budgen, D.; Brereton, P. *Evidence-Based Software Engineering and Systematic Reviews*; CRC Press: Boca Raton, FL, USA, 2015.
37. Wohlin, C.; Prikladnicki, R. Systematic Literature Reviews in Software Engineering. *Inf. Softw. Technol.* **2013**, *55*, 919–920. [CrossRef]
38. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Available online: https://www.prisma-statement.org (accessed on 23 October 2022).
39. Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; Freeman, W. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. *arXiv* **2022**, arXiv:2203.08414.
40. Liu, A.H.; Jin, S.; Lai, C.-I.; Rouditchenko, A.; Oliva, A.; Glass, J. Cross-Modal Discrete Representation Learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022.
41. Kolesnikov, A.; Pinto, A.; Beyer, L.; Zhai, X.; Harmsen, J.; Houlsby, N. Uvim: A Unified Modeling Approach for Vision with Learned Guiding Codes. *arXiv* **2022**, arXiv:2205.10337.
42. Elmoznino, E.; Bonner, M. High-Performing Neural Network Models of Visual Cortex Benefit from High Latent Dimensionality. *bioRxiv*, 2022; *preprint*.
43. Qin, B.; Mao, H.; Zhang, R.; Zhu, Y.; Ding, S.; Chen, X. Working Memory Inspired Hierarchical Video Decomposition with Transformative Representations. *arXiv* **2022**, arXiv:2204.10105.
44. Parthasarathy, N.; Eslami, S.; Carreira, J.; Hénaff, O. Self-Supervised Video Pretraining Yields Strong Image Representations. *arXiv* **2022**, arXiv:2210.06433.
45. H'enaff, O.J.; Koppula, S.; Shelhamer, E.; Zoran, D.; Jaegle, A.; Zisserman, A.; Carreira, J.; Arandjelovi'c, R. Object Discovery and Representation Networks. *arXiv* **2022**, arXiv:2203.08777.
46. Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. Pali: A Jointly-Scaled Multilingual Language-Image Model. *arXiv* **2022**, arXiv:2209.06794.
47. Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv* **2022**, arXiv:2204.14198.
48. Girdhar, R.; Singh, M.; Ravi, N.; Maaten, L.; Joulin, A.; Misra, I. Omnivore: A Single Model for Many Visual Modalities. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16081–16091.
49. Hernandez, E.; Schwettmann, S.; Bau, D.; Bagashvili, T.; Torralba, A.; Andreas, J. Natural Language Descriptions of Deep Visual Features. *arXiv* **2022**, arXiv:2201.11114.
50. Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. Data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language. *arXiv* **2022**, arXiv:2202.03555.
51. Meng, Y.; Huang, J.; Zhang, Y.; Han, J. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. *arXiv* **2022**, arXiv:2202.04538.
52. Whitfield, D. Using GPT-2 to Create Synthetic Data to Improve the Prediction Performance of NLP Machine Learning Classification Models. *arXiv* **2021**, arXiv:2104.10658.
53. Uchendu, I.; Xiao, T.; Lu, Y.; Zhu, B.; Yan, M.; Simón, J.; Bennice, M.; Fu, C.; Ma, C.; Jiao, J.; et al. Jump-Start Reinforcement Learning. *arXiv* **2022**, arXiv:2204.02372.
54. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv* **2021**, arXiv:2112.10741.

55. Borzunov, A.; Baranchuk, D.; Dettmers, T.; Ryabinin, M.; Belkada, Y.; Chumachenko, A.; Samygin, P.; Raffel, C. Petals: Collaborative Inference and Fine-Tuning of Large Models. *arXiv* **2022**, arXiv:2209.01188.
56. Our Approach to Alignment Research. Available online: https://openai.com/blog/our-approach-to-alignment-research/ (accessed on 18 October 2022).
57. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv* **2022**, arXiv:2203.02155.
58. The First High-Performance Self-Supervised Algorithm That Works for Speech, Vision, and Text. Available online: https://ai.facebook.com/blog/the-first-high-performance-self-supervised-algorithm-that-works-for-speech-vision-and-text/ (accessed on 18 October 2022).
59. Tiu, E.; Talius, E.; Patel, P.; Langlotz, C.; Ng, A.; Rajpurkar, P. Expert-Level Detection of Pathologies from Unannotated Chest X-Ray Images Via Self-Supervised Learning. *Nat. Biomed. Eng.* **2022**, *6*, 1399–1406. [CrossRef]
60. Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; Ross, C. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5228–5238.
61. Does This Artificial Intelligence Think Like a Human? Available online: https://news.mit.edu/2022/does-this-artificial-intelligence-think-human-0406 (accessed on 18 October 2022).
62. Botach, A.; Zheltonozhskii, E.; Baskin, C. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4975–4985.
63. Plotz, T.; Chen, C.; Hammerla, N.; Abowd, G. Automatic Synchronization of Wearable Sensors and Video-Cameras for Ground Truth Annotation—A Practical Approach. In Proceedings of the 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 18–22 June 2012; pp. 100–103.
64. Marcus, G.; Davis, E.; Aaronson, S. A Very Preliminary Analysis of Dall-E 2. *arXiv* **2022**, arXiv:2204.13807.
65. Dall·E 2. Available online: https://openai.com/dall-e-2/ (accessed on 18 October 2022).
66. What Dall-E 2 Can and Cannot Do. Available online: https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do (accessed on 18 October 2022).
67. OpenAI: DALL·E 2 Preview—Risks and Limitations. Available online: https://github.com/openai/dalle-2-preview/blob/main/system-card.md (accessed on 18 October 2022).
68. Everything You Wanted to Know About Midjourney. Available online: https://dallery.gallery/midjourney-guide-ai-art-explained/ (accessed on 18 October 2022).
69. Ai by the People, for the People. Available online: https://stability.ai (accessed on 18 October 2022).
70. Craiyon Home Page. Available online: https://www.craiyon.com/ (accessed on 18 October 2022).
71. Yu, J.; Xu, Y.; Koh, J.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.; et al. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *arXiv* **2022**, arXiv:2206.10789.
72. Imagen. Available online: https://gweb-research-imagen.appspot.com/paper.pdf (accessed on 18 October 2022).
73. Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; Irani, M. Imagic: Text-Based Real Image Editing with Diffusion Models. *arXiv* **2022**, arXiv:2210.09276.
74. Huang, X.; Mallya, A.; Wang, T.-C.; Liu, M.-Y. Multimodal Conditional Image Synthesis with Product-of-Experts Gans. *arXiv* **2021**, arXiv:2112.05130.
75. The Gradient. Available online: https://thegradient.pub/nlps-clever-hans-moment-has-arrived/ (accessed on 18 October 2022).
76. Katada, S.; Okada, S.; Komatani, K. Effects of Physiological Signals in Different Types of Multimodal Sentiment Estimation. *IEEE Trans. Affect. Comput.* **2022**, 1. [CrossRef]
77. Ramrakhya, R.; Undersander, E.; Batra, D.; Das, A. Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5163–5173.
78. Google Ai Blog: Simple and Effective Zero-Shot Task-Oriented Dialogue. Available online: https://ai.googleblog.com/2022/04/simple-and-effective-zero-shot-task.html (accessed on 18 October 2022).
79. Google Ai Blog: Introducing the Schema-Guided Dialogue Dataset for Conversational Assistants. Available online: https://ai.googleblog.com/2019/10/introducing-schema-guided-dialogue.html (accessed on 18 October 2022).
80. Chen, T.; La, L.; Saxena, S.; Hinton, G.; Fleet, D. A Generalist Framework for Panoptic Segmentation of Images and Videos. *arXiv* **2022**, arXiv:2210.06366.
81. Yu, R.; Park, H.; Lee, J. Human Dynamics from Monocular Video with Dynamic Camera Movements. *ACM Trans. Graph.* **2021**, *40*, 1–14. [CrossRef]
82. EPFL: Realistic Graphics Lab. Available online: http://rgl.epfl.ch/publications/Vicini2022SDF (accessed on 18 October 2022).
83. Botach, A.; Zheltonozhskii, E.; Baskin, C. Technion – Israel Institute of Technology: End-to-End Referring Video Object Segmentation with Multimodal Transformers. Available online: https://github.com/mttr2021/MTTR (accessed on 18 October 2022).

84. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 4582–4597.
85. Pang, B.; Nijkamp, E.; Kryscinski, W.; Savarese, S.; Zhou, Y.; Xiong, C. Long Document Summarization with Top-Down and Bottom-up Inference. *arXiv* **2022**, arXiv:2203.07586.
86. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef]
87. Kraus, M.; Angerbauer, K.; Buchmüller, J.; Schweitzer, D.; Keim, D.; Sedlmair, M.; Fuchs, J. Assessing 2D and 3D Heatmaps for Comparative Analysis. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–14.
88. Haarman, B.C.M.; der Lek, R.R.-V.; Nolen, W.; Mendes, R.; Drexhage, H.; Burger, H. Feature-Expression Heat Maps—A New Visual Method to Explore Complex Associations between Two Variable Sets. *J. Biomed. Inform.* **2015**, *53*, 156–161. [CrossRef]
89. Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; Poesio, M. Comparing Bayesian Models of Annotation. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 571–585. [CrossRef]
90. Thaler, F.; Payer, C.; Urschler, M.; Štern, D. Modeling Annotation Uncertainty with Gaussian Heatmaps in Landmark Localization. *arXiv* **2021**, arXiv:2109.09533.
91. Sun, Z.-H.; Jia, K.-B. Image Annotation and Refinement with Markov Chain Model of Visual Keywords and the Semantics. In *Intelligent Data Analysis and Its Applications*; Springer: New York, NY, USA, 2014; pp. 375–384.
92. Reher, R.; Kim, H.; Zhang, C.; Mao, H.; Wang, M.; Nothias, L.-F.; Caraballo-Rodriguez, A.; Glukhov, E.; Teke, B.; Leao, T.; et al. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J. Am. Chem. Soc.* **2020**, *142*, 4114–4120. [CrossRef] [PubMed]
93. Li, Z.; Xu, Z.; Zhang, R.; Zou, H.; Gao, F. Design of Modified 2-Degree-of-Freedom Proportional–Integral–Derivative Controller for Unstable Processes. *Meas. Control* **2020**, *53*, 1465–1471. [CrossRef]
94. Schwartz, B.; Ward, A. Doing Better but Feeling Worse: The Paradox of Choice. In *Positive Psychology in Practice*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2004; pp. 86–104.
95. Luccioni, A.S.; Rolnick, D. Bugs in the Data: How Imagenet Misrepresents Biodiversity. *arXiv* **2022**, arXiv:2208.11695.
96. Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; Manning, C. Fast Model Editing at Scale. *arXiv* **2022**, arXiv:2110.11309.
97. Juneja, J.; Bansal, R.; Cho, K.; Sedoc, J.; Saphra, N. Linear Connectivity Reveals Generalization Strategies. *arXiv* **2022**, arXiv:2205.12411.
98. Ainsworth, S.K.; Hayase, J.; Srinivasa, S. Git Re-Basin: Merging Models Modulo Permutation Symmetries. *arXiv* **2022**, arXiv:2209.04836.
99. Ainsworth, S.K.; Foti, N.; Fox, E. Disentangled Vae Representations for Multi-Aspect and Missing Data. *arXiv* **2018**, arXiv:1806.09060.
100. Jain, V.; Chaudhary, G.; Luthra, N.; Rao, A.; Walia, S. Dynamic Handwritten Signature and Machine Learning Based Identity Verification for Keyless Cryptocurrency Transactions. *J. Discret. Math. Sci. Cryptogr.* **2019**, *22*, 191–202. [CrossRef]
101. Cheung, B.; Terekhov, A.; Chen, Y.; Agrawal, P.; Olshausen, B. Superposition of Many Models into One. *arXiv* **2019**, arXiv:1902.05522.
102. Chen, A.M.; Lu, H.-m.; Hecht-Nielsen, R. On the Geometry of Feedforward Neural Network Error Surfaces. *Neural Comput.* **1993**, *5*, 910–927. [CrossRef]
103. Transformer Circuits Thread: Toy Models of Superposition. Available online: https://transformer-circuits.pub/2022/toy_model/index.html (accessed on 18 October 2022).
104. Simon Willison's Weblog: Prompt Injection Attacks against GPT-3. Available online: https://simonwillison.net/2022/Sep/12/prompt-injection/ (accessed on 18 October 2022).
105. Gandelsman, Y.; Sun, Y.; Chen, X.; Efros, A. Test-Time Training with Masked Autoencoders. *arXiv* **2022**, arXiv:2209.07522.
106. Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator. Available online: https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ (accessed on 18 October 2022).
107. Artist Finds Private Medical Record Photos in Popular Ai Training Data Set. Available online: https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/ (accessed on 18 October 2022).
108. GitHub: Your Ai Pair Programmer. Available online: https://github.com/features/copilot (accessed on 18 October 2022).
109. Nijkamp, E.; Pang, B.; Hayashi, H.; Tu, L.; Wang, H.; Zhou, Y.; Savarese, S.; Xiong, C. Codegen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. *arXiv* **2022**, arXiv:2203.13474.
110. CodeGeeX: A Multilingual Code Generation Model. Available online: http://keg.cs.tsinghua.edu.cn/codegeex/ (accessed on 18 October 2022).
111. Christopoulou, F.; Lampouras, G.; Gritta, M.; Zhang, G.; Guo, Y.; Li, Z.-Y.; Zhang, Q.; Xiao, M.; Shen, B.; Li, L.; et al. Pangu-Coder: Program Synthesis with Function-Level Language Modeling. *arXiv* **2022**, arXiv:2207.11280.
112. Simon Willison's Weblog: Using GPT-3 to Explain How Code Works. Available online: https://simonwillison.net/2022/Jul/9/gpt-3-explain-code (accessed on 18 October 2022).
113. Haluptzok, P.M.; Bowers, M.; Kalai, A. Language Models Can Teach Themselves to Program Better. *arXiv* **2022**, arXiv:2207.14502.
114. Bavarian, M.; Jun, H.; Tezak, N.; Schulman, J.; McLeavey, C.; Tworek, J.; Chen, M. Efficient Training of Language Models to Fill in the Middle. *arXiv* **2022**, arXiv:2207.14255.

115. Risko, E.F.; Foulsham, T.; Dawson, S.; Kingstone, A. The Collaborative Lecture Annotation System (Clas): A New Tool for Distributed Learning. *IEEE Trans. Learn. Technol.* **2013**, *6*, 4–13. [CrossRef]
116. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
117. Sultana, F.; Sufian, A.; Dutta, P. Evolution of Image Segmentation Using Deep Convolutional Neural Network: A Survey. *arXiv* **2020**, arXiv:abs/2001.04074. [CrossRef]
118. Li, Y.; Wei, J.; Liu, Y.; Kauttonen, J.; Zhao, G. Deep Learning for Micro-Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2028–2046. [CrossRef]
119. Andersen, P.H.; Broomé, S.; Rashid, M.; Lundblad, J.; Ask, K.; Li, Z.; Hernlund, E.; Rhodin, M.; Kjellström, H. Towards Machine Recognition of Facial Expressions of Pain in Horses. *Animals* **2021**, *11*, 1643. [CrossRef] [PubMed]
120. Boneh-Shitrit, T.; Amir, S.; Bremhorst, A.; Mills, D.; Riemer, S.; Fried, D.; Zamansky, A. Deep Learning Models for Automated Classification of Dog Emotional States from Facial Expressions. *arXiv* **2022**, arXiv:2206.05619.
121. Rubinstein, M. Analysis and Visualization of Temporal Variations in Video. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2014.
122. Ideas Ai Home Page. Available online: https://ideasai.com (accessed on 18 October 2022).
123. Twitter Page: Simon Willison. Available online: https://twitter.com/simonw/status/1555626060384911360 (accessed on 18 October 2022).
124. Flexible Diffusion Modeling of Long Videos. Available online: https://plai.cs.ubc.ca/2022/05/20/flexible-diffusion-modeling-of-long-videos/ (accessed on 18 October 2022).
125. Li, Z.; Wang, Q.; Snavely, N.; Kanazawa, A. Infinitenature-Zero: Learning Perpetual View Generation of Natural Scenes from Single Images. *arXiv* **2022**, arXiv:2207.11148.
126. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.; Hedman, P. Mip-Nerf 360: Unbounded Anti-Aliased Neural Radiance Fields. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5460–5469.
127. Elicit: Language Models as Research Assistants. Available online: https://www.alignmentforum.org/posts/s5jrfbsGLyEexh4GT/elicit-language-models-as-research-assistants (accessed on 18 October 2022).
128. Archive.Today: @Michaeltefula. Available online: https://archive.ph/9eiPn#selection-2773.0-2775.13 (accessed on 18 October 2022).
129. Jargon Home Page. Available online: https://explainjargon.com/ (accessed on 18 October 2022).
130. Promptbase: Dall·E, GPT-3, Midjourney, Stable Diffusion Prompt Marketplace. Available online: https://promptbase.com/ (accessed on 19 October 2022).
131. The Dall·E 2 Prompt Book. Available online: http://dallery.gallery/the-dalle-2-prompt-book/ (accessed on 19 October 2022).
132. Lexica: The Stable Diffusion Search Engine. Available online: https://lexica.art/ (accessed on 19 October 2022).
133. Belay Labs: Introducing GPT Explorer. Available online: https://belay-labs.github.io/gpt-explorer/introducing-gpt-explorer.html (accessed on 19 October 2022).
134. Imagine Prompter Guide. Available online: https://prompterguide.com/ (accessed on 19 October 2022).
135. Promptomania. Available online: https://promptomania.com/ (accessed on 19 October 2022).
136. Clip Interrogator. Available online: https://huggingface.co/spaces/pharma/CLIP-Interrogator (accessed on 23 October 2022).
137. Arora, S.; Narayan, A.; Chen, M.; Orr, L.; Guha, N.; Bhatia, K.; Chami, I.; Sala, F.; Ré, C. Ask Me Anything: A Simple Strategy for Prompting Language Models. *arXiv* **2022**, arXiv:2210.02441.
138. Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; Lewis, M. Measuring and Narrowing the Compositionality Gap in Language Models. *arXiv* **2022**, arXiv:2210.03350.
139. Jiang, Y.; Gupta, A.; Zhang, Z.; Wang, G.; Dou, Y.; Chen, Y.; Fei-Fei, L.; Anandkumar, A.; Zhu, Y.; Fan, L. Vima: General Robot Manipulation with Multimodal Prompts. *arXiv* **2022**, arXiv:2210.03094.
140. Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; et al. Do as I Can, Not as I Say: Grounding Language in Robotic Affordances. *arXiv* **2022**, arXiv:2204.01691.
141. Zeng, A.; Wong, A.; Welker, S.; Choromanski, K.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; Lee, J.; Vanhoucke, V.; et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. *arXiv* **2022**, arXiv:2204.00598.
142. Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *arXiv* **2022**, arXiv:2201.07207.
143. Shah, D.; Osinski, B.; Ichter, B.; Levine, S. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. *arXiv* **2022**, arXiv:2207.04429.
144. Huang, W.; Xia, F.; Xiao, T.; Chan, H.; Liang, J.; Florence, P.; Zeng, A.; Tompson, J.; Mordatch, I.; Chebotar, Y.; et al. Inner Monologue: Embodied Reasoning through Planning with Language Models. *arXiv* **2022**, arXiv:2207.05608.
145. Kant, Y.; Ramachandran, A.; Yenamandra, S.; Gilitschenski, I.; Batra, D.; Szot, A.; Agrawal, H. Housekeep: Tidying Virtual Households Using Commonsense Reasoning. *arXiv* **2022**, arXiv:2205.10712.
146. Li, S.; Puig, X.; Du, Y.; Wang, C.; Akyürek, E.; Torralba, A.; Andreas, J.; Mordatch, I. Pre-Trained Language Models for Interactive Decision-Making. *arXiv* **2022**, arXiv:2202.01771.
147. Bucker, A.F.C.; Figueredo, L.; Haddadin, S.; Kapoor, A.; Ma, S.; Vemprala, S.; Bonatti, R. Latte: Language Trajectory Transformer. *arXiv* **2022**, arXiv:2208.02918.

148. Cui, Y.; Niekum, S.; Gupta, A.; Kumar, V.; Rajeswaran, A. Can Foundation Models Perform Zero-Shot Task Specification for Robot Manipulation? *arXiv* **2022**, arXiv:2204.11134.
149. Tam, A.C.; Rabinowitz, N.; Lampinen, A.; Roy, N.; Chan, S.; Strouse, D.; Wang, J.; Banino, A.; Hill, F. Semantic Exploration from Language Abstractions and Pretrained Representations. *arXiv* **2022**, arXiv:2204.05080.
150. Khandelwal, A.; Weihs, L.; Mottaghi, R.; Kembhavi, A. Simple but Effective: Clip Embeddings for Embodied Ai. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14809–14818.
151. Shridhar, M.; Manuelli, L.; Fox, D. Cliport: What and Where Pathways for Robotic Manipulation. *arXiv* **2021**, arXiv:2109.12098.
152. Lin, B.; Zhu, Y.; Chen, Z.; Liang, X.; Liu, J.-Z.; Liang, X. Adapt: Vision-Language Navigation with Modality-Aligned Action Prompts. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15375–15385.
153. Parisi, S.; Rajeswaran, A.; Purushwalkam, S.; Gupta, A. The Unsurprising Effectiveness of Pre-Trained Vision Models for Control. *arXiv* **2022**, arXiv:2203.03580.
154. Gadre, S.Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; Song, S. Clip on Wheels: Zero-Shot Object Navigation as Object Localization and Exploration. *arXiv* **2022**, arXiv:2203.10421.
155. Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; Gould, S. Vln Bert: A Recurrent Vision-and-Language Bert for Navigation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1643–1653.
156. Majumdar, A.; Shrivastava, A.; Lee, S.; Anderson, P.; Parikh, D.; Batra, D. Improving Vision-and-Language Navigation with Image-Text Pairs from the Web. *arXiv* **2020**, arXiv:2004.14973.
157. Waymo: Simulation City: Introducing Waymo's Most Advanced Simulation System yet for Autonomous Driving. Available online: https://blog.waymo.com/2021/06/SimulationCity.html (accessed on 19 October 2022).
158. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv* **2021**, arXiv:2107.13586. [CrossRef]
159. Prompting: Better Ways of Using Language Models for NLP Tasks. Available online: https://thegradient.pub/prompting/ (accessed on 19 October 2022).
160. Nerd for Tech: Prompt Engineering: The Career of Future. Available online: https://medium.com/nerd-for-tech/prompt-engineering-the-career-of-future-2fb93f90f117 (accessed on 19 October 2022).
161. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North, Skopje, North Macedonia, 10–12 June 2019; pp. 4171–4186.
162. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
163. Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. Rethinking the Role of Demonstrations: What Makes in-Context Learning Work? *arXiv* **2022**, arXiv:2202.12837.
164. Garg, S.; Tsipras, D.; Liang, P.; Valiant, G. What Can Transformers Learn in-Context? *A Case Study of Simple Function Classes. arXiv* **2022**, arXiv:2208.01066.
165. Towards Data Science: Almost No Data and No Time? Unlocking the True Potential of GPT3, a Case Study. Available online: https://towardsdatascience.com/almost-no-data-and-no-time-unlocking-the-true-potential-of-gpt3-a-case-study-b4710ca0614a (accessed on 19 October 2022).
166. Twitter Post of Gene Kogan: Desert Landscape at Sunrise in Studio Ghibli Style. Available online: https://twitter.com/genekogan/status/1512513827031580673 (accessed on 19 October 2022).
167. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* **2022**, arXiv:2204.05862.
168. Roboflow: Experimenting with CLIP+VQGAN to Create AI Generated Art. Available online: https://blog.roboflow.com/ai-generated-art/ (accessed on 19 October 2022).
169. White, A.D.; Hocky, G.; Gandhi, H.; Ansari, M.; Cox, S.; Wellawatte, G.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; et al. *Do Large Language Models Know Chemistry*? ChemRxiv Preprint. Cambridge Open Engage: Cambridge, UK, 2022. [CrossRef]
170. Twitter Post of Riley Goodside from 15 April 2022. Available online: https://twitter.com/goodside/status/1515128035439255553 (accessed on 19 October 2022).
171. Li, Y.; Lin, Z.; Zhang, S.; Fu, Q.; Chen, B.; Lou, J.-G.; Chen, W. On the Advance of Making Language Models Better Reasoners. *arXiv* **2022**, arXiv:2206.02336.
172. Twitter Post of Cuddlysalmon: Decided to Try a GPT-3/Dall E Crossover Experiment Today. Available online: https://twitter.com/nptacek/status/1548402120075800577 (accessed on 19 October 2022).
173. Thread Reader: User Magnus Petersen. Available online: https://threadreaderapp.com/thread/1564633854119477257.html (accessed on 19 October 2022).
174. Daras, G.; Dimakis, A. Discovering the Hidden Vocabulary of Dalle-2. *arXiv* **2022**, arXiv:2206.00169.
175. Introducing the World's Largest Open Multilingual Language Model: Bloom. Available online: https://bigscience.huggingface.co/blog/bloom (accessed on 19 October 2022).

176. GLM-130B: An Open Bilingual Pre-Trained Model. Available online: http://keg.cs.tsinghua.edu.cn/glm-130b/posts/glm-130b/ (accessed on 19 October 2022).
177. Dohan, D.; Xu, W.; Lewkowycz, A.; Austin, J.; Bieber, D.; Lopes, R.; Wu, Y.; Michalewski, H.; Saurous, R.; Sohl-Dickstein, J.; et al. Language Model Cascades. *arXiv* **2022**, arXiv:2207.10342.
178. Argyle, L.P.; Busby, E.; Fulda, N.; Gubler, J.; Rytting, C.; Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *arXiv* **2022**, arXiv:2209.06899.
179. Aher, G.; Arriaga, R.; Kalai, A. Using Large Language Models to Simulate Multiple Humans. *arXiv* **2022**, arXiv:2208.10264.
180. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Driessche, G.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. Improving Language Models by Retrieving from Trillions of Tokens. *arXiv* **2021**, arXiv:2112.04426.
181. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Few-Shot Learning with Retrieval Augmented Language Models. *arXiv* **2022**, arXiv:2208.03299.
182. Tay, Y.; Wei, J.; Chung, H.; Tran, V.; So, D.; Shakeri, S.; Garcia, X.; Zheng, H.; Rao, J.; Chowdhery, A.; et al. Transcending Scaling Laws with 0.1% Extra Compute. *arXiv* **2022**, arXiv:2210.11399.
183. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling Instruction-Finetuned Language Models. *arXiv* **2022**, arXiv:2210.11416.
184. Castricato, L.; Havrilla, A.; Matiana, S.; Pieler, M.; Ye, A.; Yang, I.; Frazier, S.; Riedl, M. Robust Preference Learning for Storytelling Via Contrastive Reinforcement Learning. *arXiv* **2022**, arXiv:2210.07792.
185. Ai-Written Critiques Help Humans Notice Flaws. Available online: https://openai.com/blog/critiques/ (accessed on 19 October 2022).
186. Tech Xplore: Researchers Develop a Method to Keep Bots from Using Toxic Language. Available online: https://techxplore.com/news/2022-04-method-bots-toxic-language.html (accessed on 19 October 2022).
187. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D. Video Diffusion Models. *arXiv* **2022**, arXiv:2204.03458.
188. Googleplay App: Tapcaption—Ai Captions. Available online: https://play.google.com/store/apps/details?id=com.tapcaption (accessed on 19 October 2022).
189. Soltan, S.; Ananthakrishnan, S.; FitzGerald, J.; Gupta, R.; Hamza, W.; Khan, H.; Peris, C.; Rawls, S.; Rosenbaum, A.; Rumshisky, A.; et al. Alexatm 20b: Few-Shot Learning Using a Large-Scale Multilingual Seq2seq Model. *arXiv* **2022**, arXiv:2208.01448.
190. Lotf, H.; Ramdani, M. Multi-Label Classification. In Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, New York, NY, USA, 23–24 September 2020; pp. 1–6.
191. Read, J.; Martino, L.; Olmos, P.; Luengo, D. Scalable Multi-Output Label Prediction: From Classifier Chains to Classifier Trellises. *Pattern Recognit.* **2015**, *48*, 2096–2109. [CrossRef]
192. Shi, W.; Yu, D.; Yu, Q. A Gaussian Process-Bayesian Bernoulli Mixture Model for Multi-Label Active Learning. In Proceedings of the NeurIPS, Online, 6–14 December 2021.
193. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. React: Synergizing Reasoning and Acting in Language Models. *arXiv* **2022**, arXiv:2210.03629.
194. Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance. Available online: https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html (accessed on 20 October 2022).
195. Zelikman, E.; Wu, Y.; Goodman, N. Star: Bootstrapping Reasoning with Reasoning. *arXiv* **2022**, arXiv:2203.14465.
196. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; Zhou, D. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2022**, arXiv:2201.11903.
197. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:2201.08239.
198. Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H.; Tay, Y.; Ruder, S.; Zhou, D.; et al. Language Models Are Multilingual Chain-of-Thought Reasoners. *arXiv* **2022**, arXiv:2210.03057.
199. Kojima, T.; Gu, S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models Are Zero-Shot Reasoners. *arXiv* **2022**, arXiv:2205.11916.
200. Zhou, D.; Scharli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; Chi, E. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv* **2022**, arXiv:2205.10625.
201. Aligning Language Models to Follow Instructions. Available online: https://openai.com/blog/instruction-following/ (accessed on 20 October 2022).
202. New GPT3 Impressive Capabilities—Instructgpt3. Available online: https://www.lesswrong.com/posts/dypAjfRCe4nyasGSs/new-gpt3-impressive-capabilities-instructgpt3-1-2 (accessed on 22 October 2022).
203. Learning to Summarize with Human Feedback. Available online: https://openai.com/blog/learning-to-summarize-with-human-feedback/ (accessed on 22 October 2022).
204. BlenderBot 3: A 175b Parameter, Publicly Available Chatbot That Improves Its Skills and Safety over Time. Available online: https://ai.facebook.com/blog/blenderbot-3-a-175b-parameter-publicly-available-chatbot-that-improves-its-skills-and-safety-over-time (accessed on 22 October 2022).
205. Scheurer, J.E.E.; Campos, J.; Chan, J.; Chen, A.; Cho, K.; Perez, E. Training Language Models with Language Feedback. *arXiv* **2022**, arXiv:2204.14146.

206. YouTube: Learning from Natural Language Feedback. Available online: https://www.youtube.com/watch?v=oEnyl9dMKCc (accessed on 22 October 2022).
207. Deep Mind: Robust Real-Time Cultural Transmission without Human Data Supplementary Material. Available online: https://sites.google.com/view/dm-cgi (accessed on 22 October 2022).
208. Aghajanyan, A.; Huang, B.; Ross, C.; Karpukhin, V.; Xu, H.; Goyal, N.; Okhonko, D.; Joshi, M.; Ghosh, G.; Lewis, M.; et al. Cm3: A Causal Masked Multimodal Model of the Internet. *arXiv* **2022**, arXiv:2201.07520.
209. Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; Kiela, D. Flava: A Foundational Language and Vision Alignment Model. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15617–15629.
210. Almost No Data and No Time? Unlock the True Potential of GPT3! Available online: https://www.waylay.io/articles/nlp-case-study-by-waylay (accessed on 22 October 2022).
211. DeepMind: Melting Pot. Available online: https://github.com/deepmind/meltingpot (accessed on 22 October 2022).
212. Imitate and Repurpose: Learning Reusable Robot Movement Skills from Human and Animal Behaviors. Available online: https://sites.google.com/view/robot-npmp (accessed on 22 October 2022).
213. Armstrong, S.; Mindermann, S. Occam's Razor Is Insufficient to Infer the Preferences of Irrational Agents. *arXiv* **2017**, arXiv:1712.05812.
214. WebGPT: Improving the Factual Accuracy of Language Models through Web Browsing. Available online: https://openai.com/blog/webgpt/ (accessed on 22 October 2022).
215. Rae, J.W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv* **2021**, arXiv:2112.11446.
216. Language Modelling at Scale: Gopher, Ethical Considerations, and Retrieval. Available online: https://www.deepmind.com/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval (accessed on 22 October 2022).
217. Contextual Rephrasing in Google Assistant. Available online: https://ai.googleblog.com/2022/05/contextual-rephrasing-in-google.html (accessed on 22 October 2022).
218. Wu, Y.; Rabe, M.; Hutchins, D.; Szegedy, C. Memorizing Transformers. *arXiv* **2022**, arXiv:2203.08913.
219. Lehman, J.; Gordon, J.; Jain, S.; Ndousse, K.; Yeh, C.; Stanley, K. Evolution through Large Models. *arXiv* **2022**, arXiv:2206.08896.
220. Guo, Z.D.; Thakoor, S.; Pislar, M.; Pires, B.; Altch'e, F.; Tallec, C.; Saade, A.; Calandriello, D.; Grill, J.-B.; Tang, Y.; et al. Byol-Explore: Exploration by Bootstrapped Prediction. *arXiv* **2022**, arXiv:2206.08332.
221. Sorscher, B.; Geirhos, R.; Shekhar, S.; Ganguli, S.; Morcos, A. Beyond Neural Scaling Laws: Beating Power Law Scaling Via Data Pruning. *arXiv* **2022**, arXiv:2206.14486.
222. Stability Ai: Stable Diffusion Public Release. Available online: https://stability.ai/blog/stable-diffusion-public-release (accessed on 22 October 2022).
223. Compressing Global Illumination with Neural Networks. Available online: https://juretriglav.si/compressing-global-illumination-with-neural-networks/ (accessed on 22 October 2022).
224. Stable Diffusion Based Image Compression. Available online: https://pub.towardsai.net/stable-diffusion-based-image-compresssion-6f1f0a399202 (accessed on 22 October 2022).
225. Nvidia Maxine. Available online: https://developer.nvidia.com/maxine (accessed on 22 October 2022).
226. Anil, C.; Wu, Y.; Andreassen, A.; Lewkowycz, A.; Misra, V.; Ramasesh, V.; Slone, A.; Gur-Ari, G.; Dyer, E.; Neyshabur, B. Exploring Length Generalization in Large Language Models. *arXiv* **2022**, arXiv:2207.04901.
227. Dąbrowska, E. What Exactly Is Universal Grammar, and Has Anyone Seen It? *Front. Psychol.* **2015**, *6*, 852. [CrossRef] [PubMed]
228. Transformer Language Models Are Doing Something More General. Available online: https://www.lesswrong.com/posts/YwqSijHybF9GFkDab/transformer-language-models-are-doing-something-more-general (accessed on 22 October 2022).
229. Eight Ways You Can Get More Enjoyment from the Same Activity. Available online: https://www.spencergreenberg.com/2021/02/eight-ways-you-can-get-more-enjoyment-from-the-same-activity/ (accessed on 22 October 2022).
230. Six a/B Tests Used by Duolingo to Tap into Habit-Forming Behaviour. Available online: https://econsultancy.com/six-a-b-tests-used-by-duolingo-to-tap-into-habit-forming-behaviour/ (accessed on 22 October 2022).
231. The Snapchat Streak: Brilliant Marketing, Destructive Social Results. Available online: https://theboar.org/2019/11/the-snapchat-streak-brilliant-marketing-destructive-social-results/ (accessed on 22 October 2022).
232. I Think It's Time to Give up My Duolingo Streak. Available online: https://debugger.medium.com/i-think-its-time-to-give-up-my-duolingo-streak-81c27ff1be8b (accessed on 22 October 2022).
233. Sabou, M.; Bontcheva, K.; Derczynski, L.; Scharl, A. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014.
234. Wang, A.; Hoang, C.; Kan, M.-Y. Perspectives on Crowdsourcing Annotations for Natural Language Processing. *Lang. Resour. Eval.* **2012**, *47*, 9–31. [CrossRef]
235. What Is 4chan? Available online: https://www.4chan.org/ (accessed on 22 October 2022).
236. How Asynchronous Online in 'Death Stranding' Brings Players Together. Available online: https://goombastomp.com/asynchronous-death-stranding (accessed on 22 October 2022).

237. Sucholutsky, I.; Schonlau, M. 'Less Than One'-Shot Learning: Learning N Classes from M < N Samples. *arXiv* **2020**, arXiv:2009.08449.
238. Hudson, D.A.; Zitnick, C. Generative Adversarial Transformers. *arXiv* **2021**, arXiv:2103.01209.
239. Yoon, J.; Jordon, J.; Schaar, M. Gain: Missing Data Imputation Using Generative Adversarial Nets. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5689–5698.
240. Jarrett, D.; Cebere, B.; Liu, T.; Curth, A.; Schaar, M. Hyperimpute: Generalized Iterative Imputation with Automatic Model Selection. In Proceedings of the 39th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 9916–9937.
241. Abroshan, M.; Yip, K.; Tekin, C.; Schaar, M. Conservative Policy Construction Using Variational Autoencoders for Logged Data with Missing Values. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–11. [CrossRef]
242. Kyono, T.; Zhang, Y.; Bellot, A.; van der Schaar, M. Miracle: Causally-Aware Imputation Via Learning Missing Data Mechanisms. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23806–23817.
243. Yoon, J.; Zame, W.; Schaar, M. Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 1477–1490. [CrossRef]
244. Cloud Tpu: Accelerate Machine Learning Models with Google Supercomputers. Available online: https://cloud.google.com/tpu (accessed on 22 October 2022).
245. Introducing the Colossus™ MK2 GC200 IPU. Available online: https://www.graphcore.ai/products/ipu (accessed on 22 October 2022).
246. Basu, S.K. Chapter 9—A Cursory Look at Parallel Architectures and Biologically Inspired Computing. In *Soft Computing and Intelligent Systems*; Sinha, N., Gupta, M., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 185–216.
247. Bagavathi, C.; Saraniya, O. Chapter 13—Evolutionary Mapping Techniques for Systolic Computing System. In *Deep Learning and Parallel Computing Environment for Bioengineering Systems*; Sangaiah, A., Ed.; Academic Press: San Diego, CA, USA, 2019; pp. 207–223.
248. Narayanan, S.; Georgiou, P. Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language: Computational Techniques Are Presented to Analyze and Model Expressed and Perceived Human Behavior-Variedly Characterized as Typical, Atypical, Distressed, and Disordered-from Speech and Language Cues and Their Applications in Health, Commerce, Education, and Beyond. *Proc. IEEE Inst. Electr. Electron. Eng.* **2013**, *101*, 1203–1233. [CrossRef]
249. Hancock, B.; Bringmann, M.; Varma, P.; Liang, P.; Wang, S.; Re, C. Training Classifiers with Natural Language Explanations. *Proc. Conf. Assoc. Comput. Linguist. Meet* **2018**, *2018*, 1884–1895.
250. Anderson, M.; Anderson, S. Geneth: A General Ethical Dilemma Analyzer. *Paladyn J. Behav. Robot.* **2018**, *9*, 337–357. [CrossRef]
251. Gorwa, R.; Binns, R.; Katzenbach, C. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data Soc.* **2020**, *7*, 2053951719897945. [CrossRef]
252. Llanso, E. *Artificial Intelligence, Content Moderation, and Freedom of Expression*; Transatlantic Working Group on Content Moderation Online and Freedom of Expression, Institute for Information Law: Amsterdam, The Netherlands, 2020.
253. Ofcom: Use of Ai in Online Content Moderation. Available online: https://www.cambridgeconsultants.com/us/insights/whitepaper/ofcom-use-ai-online-content-moderation (accessed on 22 October 2022).
254. Rovatsos, M.; Mittelstadt, B.; Koene, A. *Landscape Summary: Bias in Algorithmic Decision-Making: What Is Bias in Algorithmic Decision-Making, How Can We Identify It, and How Can We Mitigate It?* UK Government: London, UK, 2019.
255. Palmer, A. Reasoning for the Digital Age; 2020. Available online: https://reasoningforthedigitalage.com/table-of-contents/contextual-relevance-straw-man-red-herring-and-moving-the-goalposts-fallacies/ (accessed on 22 October 2022).
256. Talisse, R.; Aikin, S. Two Forms of the Straw Man. *Argumentation* **2006**, *20*, 345–352. [CrossRef]
257. Jiang, L.; Hwang, J.; Bhagavatula, C.; Le Bras, R.; Forbes, M.; Borchardt, J.; Liang, J.; Etzioni, O.; Sap, M.; Choi, Y. Delphi: Towards Machine Ethics and Norms. *arXiv* **2021**, arXiv:2110.07574.
258. Incident 146: Research Prototype Ai, Delphi, Reportedly Gave Racially Biased Answers on Ethics. Available online: https://incidentdatabase.ai/cite/146 (accessed on 22 October 2022).
259. Jiang, L.; Hwang, J.; Bhagavatula, C.; Le Bras, R.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; et al. Can Machines Learn Morality? The Delphi Experiment. *arXiv* **2021**, arXiv:2110.07574.
260. Ask Delphi. Available online: https://delphi.allenai.org/ (accessed on 22 October 2022).
261. Redwood Research's Current Project. Available online: https://www.alignmentforum.org/posts/k7oxdbNaGATZbtEg3/redwood-research-s-current-project (accessed on 20 October 2022).
262. Herokuapp: Talk to Filtered Transformer. Available online: https://rr-data.herokuapp.com/talk-to-filtered-transformer (accessed on 22 October 2022).
263. Granitzer, M.; Kroll, M.; Seifert, C.; Rath, A.; Weber, N.; Dietzel, O.; Lindstaedt, S. Analysis of Machine Learning Techniques for Context Extraction. In Proceedings of the 2008 Third International Conference on Digital Information Management, London, UK, 13–16 November 2008; pp. 233–240.
264. Anjomshoae, S.; Omeiza, D.; Jiang, L. Context-Based Image Explanations for Deep Neural Networks. *Image Vis. Comput.* **2021**, *116*, 104310. [CrossRef]
265. Zhao, Z.Q.; Zheng, P.; Xu, S.; Wu, X. Object Detection with Deep Learning: A Review. *IEEE Trans. Neural. Netw. Learn Syst.* **2019**, *30*, 3212–3232. [CrossRef]

266. Grishman, R.; Sundheim, B. Message Understanding Conference-6: A Brief History. In Proceedings of the 16th Conference on Computational Linguistics, Stroudsburg, PA, USA, 5–9 August 1996.
267. Nadeau, D.; Sekine, S. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investig.* **2007**, *30*, 3–26. [CrossRef]
268. Prlic, A.; Cunningham, H.; Tablan, V.; Roberts, A.; Bontcheva, K. Getting More out of Biomedical Documents with Gate's Full Lifecycle Open Source Text Analytics. *PLoS Comput. Biol.* **2013**, *9*, e1002854. [CrossRef]
269. Kwartler, T. *The OpenNLP Project, in Text Mining in Practice with R*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017; pp. 237–269.
270. Mansouri, A.; Affendey, L.; Mamat, A. Named Entity Recognition Approaches; International Journal of Computer Science and Network Security 8.2 (2008), pp. 339-344.
271. Kołcz, A.; Org, A. Chowdhury; Alspector, J. *Data Duplication: An Imbalance Problem?* 2003.
272. Haneem, F.; Ali, R.; Kama, N.; Basri, S. Resolving Data Duplication, Inaccuracy and Inconsistency Issues Using Master Data Management; 2017 5th International Conference on Research and Innovation in Information Systems (ICRIIS) 2017; pp. 1–6.
273. Zhou, X.; Chen, L. Monitoring Near Duplicates over Video Streams. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 2010; pp. 521–530.
274. Ciro, J.; Galvez, D.; Schlippe, T.; Kanter, D. Lsh Methods for Data Deduplication in a Wikipedia Artificial Dataset. *arXiv* **2021**, arXiv:2112.11478.
275. Fröbe, M.; Bevendorff, J.; Reimer, J.; Potthast, M.; Hagen, M. Sampling Bias Due to near-Duplicates in Learning to Rank. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 1997–2000.
276. Suzuki, I.; Hara, K.; Eizuka, Y. Impact of Duplicating Small Training Data on Gans. In Proceedings of the 10th International Conference on Data Science, Technology and Applications, Paris, France, 6–8 July 2021; pp. 308–315.
277. Hoque, R.; Chen, L.; Sharma, S.; Dharmarajan, K.; Thananjeyan, B.; Abbeel, P.; Goldberg, K. Fleet-Dagger: Interactive Robot Fleet Learning with Scalable Human Supervision. *arXiv* **2022**, arXiv:2206.14349.
278. de Laat, P.B. The Use of Software Tools and Autonomous Bots against Vandalism: Eroding Wikipedia's Moral Order? *Ethics Inf. Technol.* **2015**, *17*, 175–188. [CrossRef]
279. This Machine Kills Trolls. Available online: https://www.theverge.com/2014/2/18/5412636/this-machine-kills-trolls-how-wikipedia-robots-snuff-out-vandalism (accessed on 22 October 2022).
280. Teng, F.; Ma, M.; Ma, Z.; Huang, L.; Xiao, M.; Li, X. A Text Annotation Tool with Pre-Annotation Based on Deep Learning. In *Knowledge Science, Engineering and Management*; Springer: New York, NY, USA, 2019; pp. 440–451.
281. Ringger, E.; Carmen, M.; Haertel, R.; Seppi, K.; Lonsdale, D.; McClanahan, P.; Carroll, J.; Ellison, N. Assessing the Costs of Machine-Assisted Corpus Annotation through a User Study. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA); 2008.
282. Lingren, T.; Deleger, L.; Molnar, K.; Zhai, H.; Meinzen-Derr, J.; Kaiser, M.; Stoutenborough, L.; Li, Q.; Solti, I. Evaluating the Impact of Pre-Annotation on Annotation Speed and Potential Bias: Natural Language Processing Gold Standard Development for Clinical Named Entity Recognition in Clinical Trial Announcements. *J. Am. Med. Inform. Assoc.* **2013**, *21*, 406–413. [CrossRef] [PubMed]
283. Deep Hierarchical Planning from Pixels. Available online: https://ai.googleblog.com/2022/07/deep-hierarchical-planning-from-pixels.html (accessed on 22 October 2022).
284. Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Bordes, F.; Vincent, P.; Joulin, A.; Rabbat, M.; Ballas, N. Masked Siamese Networks for Label-Efficient Learning. *arXiv* **2022**, arXiv:2204.07141.
285. ML-Enhanced Code Completion Improves Developer Productivity. Available online: https://ai.googleblog.com/2022/07/ml-enhanced-code-completion-improves.html (accessed on 22 October 2022).
286. YouTube: How to Use GPT-3 on Identifying an Answer Is Useful to a Given Question? Available online: https://www.youtube.com/watch?v=5Mwxm8A1tOo (accessed on 22 October 2022).
287. How Ai Could Help Make Wikipedia Entries More Accurate. Available online: https://tech.fb.com/artificial-intelligence/2022/07/how-ai-could-help-make-wikipedia-entries-more-accurate/ (accessed on 22 October 2022).
288. Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. Language Models (Mostly) Know What They Know. *arXiv* **2022**, arXiv:2207.05221.
289. Chen, B.; Kwiatkowski, R.; Vondrick, C.; Lipson, H. Fully Body Visual Self-Modeling of Robot Morphologies. *Sci. Robot.* **2022**, *7*, 68. [CrossRef]
290. Lee, S.; Chung, J.; Yu, Y.; Kim, G.; Breuel, T.; Chechik, G.; Song, Y. Acav100m: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Nashville, TN, USA, 20–25 June 2021; pp. 10254–10264.
291. Pati, S.; Baid, U.; Zenk, M.; Edwards, B.; Sheller, M.; Reina, G.; Foley, P.; Gruzdev, A.; Martin, J.; Albarqouni, S.; et al. The Federated Tumor Segmentation (Fets) Challenge. *arXiv* **2021**, arXiv:2105.05874.
292. Abeyruwan, S.; Graesser, L.; D'Ambrosio, D.; Singh, A.; Shankar, A.; Bewley, A.; Sanketi, P. I-Sim2real: Reinforcement Learning of Robotic Policies in Tight Human-Robot Interaction Loops. *arXiv* **2022**, arXiv:2207.06572.
293. Xie, K.; Wang, T.; Iqbal, U.; Guo, Y.; Fidler, S.; Shkurti, F. Physics-Based Human Motion Estimation and Synthesis from Videos. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Nashville, TN, USA, 20–25 June 2021; pp. 11512–11521.

294. Tzaban, R.; Mokady, R.; Gal, R.; Bermano, A.; Cohen-Or, D. Stitch It in Time: Gan-Based Facial Editing of Real Videos. *arXiv* **2022**, arXiv:2201.08361.
295. Fu, J.; Li, S.; Jiang, Y.; Lin, K.-Y.; Qian, C.; Loy, C.; Wu, W.; Liu, Z. Stylegan-Human: A Data-Centric Odyssey of Human Generation. *arXiv* **2022**, arXiv:2204.11823.
296. How Waabi World Works. Available online: https://waabi.ai/how-waabi-world-works/ (accessed on 22 October 2022).
297. Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A.; Lester, B.; Du, N.; Dai, A.; Le, Q. Finetuned Language Models Are Zero-Shot Learners. *arXiv* **2022**, arXiv:2109.01652.
298. Wang, W.; Dong, L.; Cheng, H.; Song, H.; Liu, X.; Yan, X.; Gao, J.; Wei, F. Visually-Augmented Language Modeling. *arXiv* **2022**, arXiv:2205.10178.
299. Brooks, T.; Hellsten, J.; Aittala, M.; Wang, T.-C.; Aila, T.; Lehtinen, J.; Liu, M.-Y.; Efros, A.; Karras, T. Generating Long Videos of Dynamic Scenes. *arXiv* **2022**, arXiv:2206.03429.
300. Nash, C.; Carreira, J.; Walker, J.; Barr, I.; Jaegle, A.; Malinowski, M.; Battaglia, P. Transframer: Arbitrary Frame Prediction with Generative Models. *arXiv* **2022**, arXiv:2203.09494.
301. Dall·E: Introducing Outpainting. Available online: https://openai.com/blog/dall-e-introducing-outpainting/ (accessed on 22 October 2022).
302. Li, D.; Wang, S.; Zou, J.; Chang, T.; Nieuwburg, E.; Sun, F.; Kanoulas, E. Paint4poem: A Dataset for Artistic Visualization of Classical Chinese Poems. *arXiv* **2021**, arXiv:2109.11682.
303. Anonymous. Phenaki: Variable Length, Video Generation from Open Domain Textual Descriptions. *OpenReview* **2022**. Available online: https://openreview.net/pdf?id=vOEXS39nOF (accessed on 23 October 2022).
304. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-a-Video: Text-to-Video Generation without Text-Video Data. *arXiv* **2022**, arXiv:2209.14792.
305. Explore Synthetic Futuring. Available online: https://medium.thirdwaveberlin.com/explore-synthetic-futuring-59819a12c4ee (accessed on 23 October 2022).
306. Li, Y.; Panda, R.; Kim, Y.; Chen, C.-F.; Feris, R.; Cox, D.; Vasconcelos, N. Valhalla: Visual Hallucination for Machine Translation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5206–5216.
307. Rahtz, M.; Varma, V.; Kumar, R.; Kenton, Z.; Legg, S.; Leike, J. Safe Deep Rl in 3D Environments Using Human Feedback. *arXiv* **2022**, arXiv:2201.08102.
308. Axenie, C.; Scherr, W.; Wieder, A.; Torres, A.; Meng, Z.; Du, X.; Sottovia, P.; Foroni, D.; Grossi, M.; Bortoli, S.; et al. Fuzzy Modeling and Inference for Physics-Aware Road Vehicle Driver Behavior Model Calibration. *Expert Systems with Applications.* **2022**. [CrossRef]
309. Baker, B.; Akkaya, I.; Zhokhov, P.; Huizinga, J.; Tang, J.; Ecoffet, A.; Houghton, B.; Sampedro, R.; Clune, J. Video Pretraining (VPT): Learning to Act by Watching Unlabeled Online Videos. *arXiv* **2022**, arXiv:2206.11795.
310. Learning to Play Minecraft with Video Pretraining (Vpt). Available online: https://openai.com/blog/vpt/ (accessed on 23 October 2022).
311. Su, H.; Kasai, J.; Wu, C.; Shi, W.; Wang, T.; Xin, J.; Zhang, R.; Ostendorf, M.; Zettlemoyer, L.; Smith, N.; et al. Selective Annotation Makes Language Models Better Few-Shot Learners. *arXiv* **2022**, arXiv:2209.01975.
312. Alaa, A.M.; Breugel, B.; Saveliev, E.; Schaar, M. How Faithful Is Your Synthetic Data? *Sample-Level Metrics for Evaluating and Auditing Generative Models. arXiv* **2022**, arXiv:2102.08921.
313. Wood, E.; Baltruvsaitis, T.; Hewitt, C.; Dziadzio, S.; Johnson, M.; Estellers, V.; Cashman, T.; Shotton, J. Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Nashville, TN, USA, 20–25 June 2021; pp. 3661–3671.
314. Greff, K.; Belletti, F.; Beyer, L.; Doersch, C.; Du, Y.; Duckworth, D.; Fleet, D.; Gnanapragasam, D.; Golemo, F.; Herrmann, C.; et al. Kubric: A Scalable Dataset Generator. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3739–3751.
315. Jakesch, M.; Hancock, J.; Naaman, M. Human Heuristics for Ai-Generated Language Are Flawed. *arXiv* **2022**, arXiv:2206.07271.
316. Hao, Z.; Mallya, A.; Belongie, S.; Liu, M.-Y. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Nashville, TN, USA, 20–25 June 2021; pp. 14052–14062.
317. Khalid, N.M.; Xie, T.; Belilovsky, E.; Popa, T. Clip-Mesh: Generating Textured Meshes from Text Using Pretrained Image-Text Models; SIGGRAPH Asia. 2022. Available online: https://dl.acm.org/doi/abs/10.1145/3550469.3555392 (accessed on 22 October 2022).
318. Sanghi, A.; Chu, H.; Lambourne, J.; Wang, Y.; Cheng, C.-Y.; Fumero, M. Clip-Forge: Towards Zero-Shot Text-to-Shape Generation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18582–18592.
319. Poole, B.; Jain, A.; Barron, J.; Mildenhall, B. Dreamfusion: Text-to-3D Using 2D Diffusion. *arXiv* **2022**, arXiv:2209.14988.
320. Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; Fidler, S. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. *arXiv* **2022**, arXiv:2209.11163.
321. Common Sense Machines, Generating 3D Worlds with CommonSim-1. Available online: https://csm.ai/commonsim-1-generating-3d-worlds/ (accessed on 23 October 2022).

322. Cao, J.; Zhao, A.; Zhang, Z. Automatic Image Annotation Method Based on a Convolutional Neural Network with Threshold Optimization. *PLoS ONE* **2020**, *15*, e0238956. [CrossRef] [PubMed]
323. Ranjbar, S.; Singleton, K.; Jackson, P.; Rickertsen, C.; Whitmire, S.; Clark-Swanson, K.; Mitchell, J.; Swanson, K.; Hu, L. A Deep Convolutional Neural Network for Annotation of Magnetic Resonance Imaging Sequence Type. *J. Digit. Imaging* **2020**, *33*, 439–446. [CrossRef]
324. Wang, R.; Xie, Y.; Yang, J.; Xue, L.; Hu, M.; Zhang, Q. Large Scale Automatic Image Annotation Based on Convolutional Neural Network. *J. Vis. Commun. Image Represent.* **2017**, *49*, 213–224. [CrossRef]
325. Chen, Y.; Liu, L.; Tao, J.; Chen, X.; Xia, R.; Zhang, Q.; Xiong, J.; Yang, K.; Xie, J. The Image Annotation Algorithm Using Convolutional Features from Intermediate Layer of Deep Learning. *Multim. Tools Appl.* **2021**, *80*, 4237–4261. [CrossRef]
326. The Illustrated Transformer. Available online: https://jalammar.github.io/illustrated-transformer/ (accessed on 23 October 2022).
327. Transformers from Scratch. Available online: https://e2eml.school/transformers.html (accessed on 23 October 2022).
328. Transformers for Software Engineers. Available online: https://blog.nelhage.com/post/transformers-for-software-engineers/ (accessed on 23 October 2022).
329. AIM. Big Tech & Their Favourite Deep Learning Techniques. In *Analytics India Magazine*; Analytics India Magazine: Bangalore, Karnataka, 2021.
330. Phuong, M.; Hutter, M. Formal Algorithms for Transformers. *arXiv* **2022**, arXiv:2207.09238.
331. Reif, E.; Ippolito, D.; Yuan, A.; Coenen, A.; Callison-Burch, C.; Wei, J. A Recipe for Arbitrary Text Style Transfer with Large Language Models. *arXiv* **2021**, arXiv:2109.03910.
332. Jang, E. Just Ask for Generalization; 2021. Available online: https://evjang.com/2021/10/23/generalization.html/ (accessed on 23 October 2022).
333. Prompt Engineering. Available online: https://docs.cohere.ai/prompt-engineering-wiki/ (accessed on 23 October 2022).
334. Will Transformers Take over Artificial Intelligence? Available online: https://www.quantamagazine.org/will-transformers-take-over-artificial-intelligence-20220310 (accessed on 23 October 2022).
335. Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A.; Abid, A.; Fisch, A.; Brown, A.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv* **2022**, arXiv:2206.04615.
336. Branwen, G. The Scaling Hypothesis. 2021. Available online: https://www.gwern.net/Scaling-hypothesis/ (accessed on 23 October 2022).
337. Alabdulmohsin, I.M.; Neyshabur, B.; Zhai, X. Revisiting Neural Scaling Laws in Language and Vision. *arXiv* **2022**, arXiv:2209.06640.
338. Austin, J.; Odena, A.; Nye, M.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C.; Terry, M.; Le, Q.; et al. Program Synthesis with Large Language Models. *arXiv* **2021**, arXiv:2108.07732.
339. Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. Solving Quantitative Reasoning Problems with Language Models. *arXiv* **2022**, arXiv:2206.14858.
340. Creswell, A.; Shanahan, M. Faithful Reasoning Using Large Language Models. *arXiv* **2022**, arXiv:2208.14271.
341. Drori, I.; Zhang, S.; Shuttleworth, R.; Tang, L.; Lu, A.; Ke, E.; Liu, K.; Chen, L.; Tran, S.; Cheng, N.; et al. A Neural Network Solves, Explains, and Generates University Math Problems by Program Synthesis and Few-Shot Learning at Human Level. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, 32. [CrossRef] [PubMed]
342. Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Manzagol, P.-A.; Larochelle, H. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. *arXiv* **2020**, arXiv:1903.03096.
343. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. *arXiv* **2021**, arXiv:2102.12092.
344. Zhang, P.; Dou, H.; Zhang, W.; Zhao, Y.; Li, S.; Qin, Z.; Li, X. Versatilegait: A Large-Scale Synthetic Gait Dataset Towards in-the-Wild Simulation. *arXiv* **2022**, arXiv:2105.14421.
345. Solaiman, I.; Dennison, C. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. *arXiv* **2021**, arXiv:2106.10328.
346. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10674–10685.
347. Yang, L.; Zhang, Z.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Yang, M.-H.; Cui, B. Diffusion Models: A Comprehensive Survey of Methods and Applications. *arXiv* **2022**, arXiv:2209.00796.
348. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4396–4405.
349. Luo, C. Understanding Diffusion Models: A Unified Perspective. *arXiv* **2022**, arXiv:2208.11970.
350. Weng, L. What Are Diffusion Models? Available online: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/ (accessed on 23 October 2022).
351. Generative Modeling by Estimating Gradients of the Data Distribution. Available online: https://yang-song.net/blog/2021/score/ (accessed on 23 October 2022).

352. Sohl-Dickstein, J.N.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *arXiv* **2015**, arXiv:1503.03585.
353. Liu, N.; Li, S.; Du, Y.; Torralba, A.; Tenenbaum, J. Compositional Visual Generation with Composable Diffusion Models. *arXiv* **2022**, arXiv:2206.01714.
354. Search Engine: You. Available online: https://you.com/ (accessed on 23 October 2022).
355. Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J.; et al. A Generalist Agent. *arXiv* **2022**, arXiv:2205.06175.
356. Gato as the Dawn of Early Agi. Available online: https://www.lesswrong.com/posts/TwfWTLhQZgy2oFwK3/gato-as-the-dawn-of-early-agi (accessed on 23 October 2022).
357. Why I Think Strong General Ai Is Coming Soon. Available online: https://www.lesswrong.com/posts/K4urTDkBbtNuLivJx/why-i-think-strong-general-ai-is-coming-soon (accessed on 23 October 2022).
358. Huang, J.; Gu, S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; Han, J. Large Language Models Can Self-Improve. *arXiv* **2022**, arXiv:2210.11610.
359. Sheng, A.; Padmanabhan, S. Self-Programming Artificial Intelligence Using Code-Generating Language Models; OpenReview 2022. Available online: https://openreview.net/forum?id=SKat5ZX5RET (accessed on 23 October 2022).
360. Laskin, M.; Wang, L.; Oh, J.; Parisotto, E.; Spencer, S.; Steigerwald, R.; Strouse, D.; Hansen, S.; Filos, A.; Brooks, E.; et al. In-Context Reinforcement Learning with Algorithm Distillation. *arXiv* **2022**, arXiv:2210.14215.
361. Fawzi, A.; Balog, M.; Huang, A.; Hubert, T.; Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Ruiz, F.R.; Schrittwieser, J.; Swirszcz, G.; et al. Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning. *Nature* **2022**, *610*, 47–53. [CrossRef]
362. Strassen, V. Gaussian Elimination Is Not Optimal. *Numer. Math.* **1969**, *13*, 354–356. [CrossRef]
363. Kauers, M.; Moosbauer, J. The Fbhhrbnrssshk-Algorithm for Multiplication in Z5 × 52 Is Still Not the End of the Story. *arXiv* **2022**, arXiv:2210.04045.
364. The Bitter Lesson. Available online: http://www.incompleteideas.net/IncIdeas/BitterLesson.html (accessed on 23 October 2022).
365. Lee, K.-H.; Nachum, O.; Yang, M.; Lee, L.; Freeman, D.; Xu, W.; Guadarrama, S.; Fischer, I.; Jang, E.; Michalewski, H.; et al. Multi-Game Decision Transformers. *arXiv* **2022**, arXiv:2205.15241.
366. Stephen Wolfram Writings: Games and Puzzles as Multicomputational Systems. Available online: https://writings.stephenwolfram.com/2022/06/games-and-puzzles-as-multicomputational-systems/ (accessed on 23 October 2022).
367. Cui, Z.J.; Wang, Y.; Shafiullah, N.; Pinto, L. From Play to Policy: Conditional Behavior Generation from Uncurated Robot Data. *arXiv* **2022**, arXiv:2210.10047.
368. Du, N.; Huang, Y.; Dai, A.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A.; Firat, O.; et al. Glam: Efficient Scaling of Language Models with Mixture-of-Experts. *arXiv* **2022**, arXiv:2112.06905.
369. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.; et al. Opt: Open Pre-Trained Transformer Language Models. *arXiv* **2022**, arXiv:2205.01068.
370. Facebook Research: Chronicles of OPT Development. Available online: https://github.com/facebookresearch/metaseq/tree/main/projects/OPT/chronicles (accessed on 23 October 2022).
371. How Much of Ai Progress Is from Scaling Compute? *And How Far Will It Scale?* Available online: https://www.metaculus.com/notebooks/10688/how-much-of-ai-progress-is-from-scaling-compute-and-how-far-will-it-scale/ (accessed on 23 October 2022).
372. Micikevicius, P.; Stosic, D.; Burgess, N.; Cornea, M.; Dubey, P.; Grisenthwaite, R.; Ha, S.; Heinecke, A.; Judd, P.; Kamalu, J.; et al. Fp8 Formats for Deep Learning. *arXiv* **2022**, arXiv:2209.05433.
373. The First Posit-Based Processor Core Gave a Ten-Thousandfold Accuracy Boost. Available online: https://spectrum.ieee.org/floating-point-numbers-posits-processor (accessed on 23 October 2022).
374. Mosaic Llms (Part 2): GPT-3 Quality for <$500 k. Available online: https://www.mosaicml.com/blog/gpt-3-quality-for-500k (accessed on 23 October 2022).
375. Yang, G.; Hu, E.; Babuschkin, I.; Sidor, S.; Liu, X.; Farhi, D.; Ryder, N.; Pachocki, J.; Chen, W.; Gao, J. Tensor Programs V: Tuning Large Neural Networks Via Zero-Shot Hyperparameter Transfer. *arXiv* **2022**, arXiv:2203.03466.
376. Nagarajan, A.; Sen, S.; Stevens, J.; Raghunathan, A. Axformer: Accuracy-Driven Approximation of Transformers for Faster, Smaller and More Accurate Nlp Models. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 18–23 July 2022; pp. 1–8.
377. Stelzer, F.; Röhm, A.; Vicente, R.; Fischer, I.; Yanchuk, S. Deep Neural Networks Using a Single Neuron: Folded-in-Time Architecture Using Feedback-Modulated Delay Loops. *Nat. Commun.* **2021**, *12*, 5164. [CrossRef] [PubMed]
378. Kirstain, Y.; Lewis, P.; Riedel, S.; Levy, O. A Few More Examples May Be Worth Billions of Parameters. *arXiv* **2021**, arXiv:2110.04374.
379. Schick, T.; Schütze, H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. *arXiv* **2021**, arXiv:2001.07676.
380. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D.; Hendricks, L.; Welbl, J.; Clark, A.; et al. Training Compute-Optimal Large Language Models. *arXiv* **2022**, arXiv:2203.15556.
381. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.; Singhal, S.; Som, S.; et al. Image as a Foreign Language: Beit Pretraining for All Vision and Vision-Language Tasks. *arXiv* **2022**, arXiv:2208.10442.

382. New Scaling Laws for Large Language Models. Available online: https://www.lesswrong.com/posts/midXmMb2Xg37F2Kgn/new-scaling-laws-for-large-language-models (accessed on 23 October 2022).
383. Trees Are Harlequins, Words Are Harlequins. Available online: https://nostalgebraist.tumblr.com/post/680262678831415296/an-exciting-new-paper-on-neural-language-model (accessed on 23 October 2022).
384. Understanding Scaling Laws for Recommendation Models. Available online: https://threadreaderapp.com/thread/1563455844670246912.html (accessed on 23 October 2022).
385. Jurassic-X: Crossing the Neuro-Symbolic Chasm with the Mrkl System. Available online: https://www.ai21.com/blog/jurassic-x-crossing-the-neuro-symbolic-chasm-with-the-mrkl-system (accessed on 23 October 2022).
386. Introducing Adept. Available online: https://www.adept.ai/post/introducing-adept (accessed on 23 October 2022).
387. Hugging Face: Transformers. Available online: https://github.com/huggingface/transformers (accessed on 23 October 2022).
388. Democratizing Access to Large-Scale Language Models with Opt-175b. Available online: https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/ (accessed on 23 October 2022).
389. Why Tool Ais Want to Be Agent Ais. Available online: https://www.gwern.net/Tool-AI (accessed on 23 October 2022).
390. Wunderwuzzi's Blog: GPT-3 and Phishing Attacks. Available online: https://embracethered.com/blog/posts/2022/gpt-3-ai-and-phishing-attacks/ (accessed on 23 October 2022).
391. Wu, Y.; Jiang, A.; Li, W.; Rabe, M.; Staats, C.; Jamnik, M.; Szegedy, C. Autoformalization with Large Language Models. *arXiv* **2022**, arXiv:2205.12615.
392. Fei, N.; Lu, Z.; Gao, Y.; Yang, G.; Huo, Y.; Wen, J.; Lu, H.; Song, R.; Gao, X.; Xiang, T.; et al. Towards Artificial General Intelligence Via a Multimodal Foundation Model. *Nat. Commun.* **2022**, *13*, 3094. [CrossRef] [PubMed]
393. Caccia, M.; Mueller, J.; Kim, T.; Charlin, L.; Fakoor, R. Task-Agnostic Continual Reinforcement Learning: In Praise of a Simple Baseline. *arXiv* **2022**, arXiv:2205.14495.
394. Fan, L.; Wang, G.; Jiang, Y.; Mandlekar, A.; Yang, Y.; Zhu, H.; Tang, A.; Huang, D.-A.; Zhu, Y.; Anandkumar, A. Minedojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. *arXiv* **2022**, arXiv:2206.08853.
395. My Bet: Ai Size Solves Flubs. Available online: https://astralcodexten.substack.com/p/my-bet-ai-size-solves-flubs (accessed on 23 October 2022).
396. What Does It Mean When an Ai Fails? A Reply to Slatestarcodex's Riff on Gary Marcus. Available online: https://garymarcus.substack.com/p/what-does-it-mean-when-an-ai-fails (accessed on 23 October 2022).
397. Somewhat Contra Marcus on Ai Scaling. Available online: https://astralcodexten.substack.com/p/somewhat-contra-marcus-on-ai-scaling (accessed on 23 October 2022).
398. Fitzgerald, M.; Boddy, A.; Baum, S. 2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Technical Report 20-1. 2020. Available online: https://gcrinstitute.org/papers/055_agi-2020.pdf (accessed on 23 October 2022).
399. Metaculus: Date Weakly General Ai Is Publicly Known. Available online: https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/ (accessed on 23 October 2022).
400. Superglue Leaderboard Version: 2.0. Available online: https://super.gluebenchmark.com/leaderboard/ (accessed on 23 October 2022).
401. Roy, R.; Raiman, J.; Kant, N.; Elkin, I.; Kirby, R.; Siu, M.; Oberman, S.; Godil, S.; Catanzaro, B. Prefixrl: Optimization of Parallel Prefix Circuits Using Deep Reinforcement Learning. In Proceedings of the 2021 58th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 5–9 December 2021; pp. 853–858.
402. Kelly, B.T.; Malamud, S.; Zhou, K. The Virtue of Complexity in Return Prediction. *Natl. Bur. Econ. Res. Work. Pap. Ser.* **2022**, *30217*, 21–90. [CrossRef]
403. Are You Really in a Race? *The Cautionary Tales of Szilárd and Ellsberg*. Available online: https://forum.effectivealtruism.org/posts/cXBznkfoPJAjacFoT/are-you-really-in-a-race-the-cautionary-tales-of-szilard-and (accessed on 23 October 2022).
404. The Time Is Now to Develop Community Norms for the Release of Foundation Models. Available online: https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models (accessed on 23 October 2022).
405. Agi Ruin: A List of Lethalities. Available online: https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities (accessed on 23 October 2022).
406. Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; Batra, D. Deal or No Deal? *End-to-End Learning of Negotiation Dialogues. arXiv* **2017**, arXiv:1706.05125.
407. Ought, Inc. Interactive Composition Explorer. Available online: https://github.com/oughtinc/ice (accessed on 23 October 2022).
408. Shu, T.; Bhandwaldar, A.; Gan, C.; Smith, K.; Liu, S.; Gutfreund, D.; Spelke, E.; Tenenbaum, J.; Ullman, T. Agent: A Benchmark for Core Psychological Reasoning. *arXiv* **2021**, arXiv:2102.12321.
409. Aligned AI: The Happy Faces Benchmark. Available online: https://github.com/alignedai/HappyFaces (accessed on 23 October 2022).
410. Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; Irving, G. Alignment of Language Agents. *arXiv* **2021**, arXiv:2103.14659.
411. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and Social Risks of Harm from Language Models. *arXiv* **2021**, arXiv:2112.04359.

412. Glaese, A.; McAleese, N.; Trkebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. Improving Alignment of Dialogue Agents Via Targeted Human Judgements. *arXiv* **2022**, arXiv:2209.14375.
413. Xie, C.; Cai, H.; Song, J.; Li, J.; Kong, F.; Wu, X.; Morimitsu, H.; Yao, L.; Wang, D.; Leng, D.; et al. Zero and R2D2: A Large-Scale Chinese Cross-Modal Benchmark and a Vision-Language Framework; 2022.; arXiv 2022, arXiv: 2205. 0386.
414. Nvidia Omniverse Replicator Generates Synthetic Training Data for Robots. Available online: https://developer.nvidia.com/blog/generating-synthetic-datasets-isaac-sim-data-replicator/ (accessed on 23 October 2022).
415. Starke, S.; Zhang, H.; Komura, T.; Saito, J. Neural State Machine for Character-Scene Interactions. *ACM Trans. Graph.* **2019**, *38*, 1–14. [CrossRef]
416. Liu, R.; Wei, J.; Gu, S.S.; Wu, T.-Y.; Vosoughi, S.; Cui, C.; Zhou, D.; Dai, A. Mind's Eye: Grounded Language Model Reasoning through Simulation. *arXiv* **2022**, arXiv:2210.05359.
417. Mitrano, P.; Berenson, D. Data Augmentation for Manipulation. *arXiv* **2022**, arXiv:2205.02886.
418. Karpas, E.D.; Abend, O.; Belinkov, Y.; Lenz, B.; Lieber, O.; Ratner, N.; Shoham, Y.; Bata, H.; Levine, Y.; Leyton-Brown, K.; et al. Mrkl Systems: A Modular, Neuro-Symbolic Architecture That Combines Large Language Models, External Knowledge Sources and Discrete Reasoning. *arXiv* **2022**, arXiv:2205.00445.
419. Ling, H.; Kreis, K.; Li, D.; Kim, S.; Torralba, A.; Fidler, S. Editgan: High-Precision Semantic Image Editing. *arXiv* **2021**, arXiv:2111.03186.
420. Fedus, W.; Dean, J.; Zoph, B. A Review of Sparse Expert Models in Deep Learning. *arXiv* **2022**, arXiv:2209.01667.
421. Rajbhandari, S.; Li, C.; Yao, Z.; Zhang, M.; Aminabadi, R.; Awan, A.; Rasley, J.; He, Y. Deepspeed-Moe: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation Ai Scale. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 18332–18346.
422. Kittur, A.; Yu, L.; Hope, T.; Chan, J.; Lifshitz-Assaf, H.; Gilon, K.; Ng, F.; Kraut, R.; Shahaf, D. Scaling up Analogical Innovation with Crowds and Ai. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 16654. [CrossRef] [PubMed]
423. Wang, C.-Y.; Yeh, I.-H.; Liao, H. You Only Learn One Representation: Unified Network for Multiple Tasks. *arXiv* **2021**, arXiv:2105.04206.
424. Meng, K.; Bau, D.; Andonian, A.; Belinkov, Y. *Locating and Editing Factual Associations in GPT*; 2022.
425. Meet Loab, the Ai Art Woman Haunting the Internet. Available online: https://www.cnet.com/science/what-is-loab-the-haunting-ai-art-woman-explained/ (accessed on 23 October 2022).
426. Weng, L. Learning with Not Enough Data Part 1: Semi-Supervised Learning. Available online: https://lilianweng.github.io/posts/2021-12-05-semi-supervised/ (accessed on 23 October 2022).
427. Davis, K.M.; Torre-Ortiz, C.; Ruotsalo, T. Brain-Supervised Image Editing. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
428. Machado, K.; Kank, R.; Sonawane, J.; Maitra, S. A Comparative Study of Acid and Base in Database Transaction Processing. *Int. J. Sci. Eng. Res.* **2017**, *8*, 116–119.
429. Lesswrong: Comment by User Gwern. Available online: https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do?commentId=CWKFyJYfgoZfP9955 (accessed on 23 October 2022).
430. Bostrom, N. Base Camp for Mt. Ethics DRAFT version 0.9 2022. Available online: https://nickbostrom.com/papers/mountethics.pdf (accessed on 23 October 2022).
431. Wang, Z.; Yu, A.; Firat, O.; Cao, Y. Towards Zero-Label Language Learning. *arXiv* **2021**, arXiv:2109.09193.
432. Ge, X.; Zhang, K.; Gribizis, A.; Hamodi, A.; Sabino, A.; Crair, M. Retinal Waves Prime Visual Motion Detection by Simulating Future Optic Flow. *Science* **2021**, *373*, 6553. [CrossRef] [PubMed]
433. Import Ai 269: Baidu Takes on Meena; Microsoft Improves Facial Recognition with Synthetic Data; Unsolved Problems in Ai Safety. Available online: https://jack-clark.net/2021/10/11/import-ai-269-baidu-takes-on-meena-microsoft-improves-facial-recognition-with-synthetic-data-unsolved-problems-in-ai-safety/ (accessed on 23 October 2022).
434. Touvron, H.; Cord, M.; Jegou, H. Deit Iii: Revenge of the Vit. *arXiv* **2021**, arXiv:2204.07118.