UNIVERSITY OF
GLOUCESTERSHIRE

This is a peer-reviewed, post-print (final draft post-refereeing) version of the following published document, This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://doi.org/10.1007/s12369-024-01117-1 and is licensed under Publisher's Licence license:

Watson, Eleanor, Viana, Thiago ORCID logoORCID: https://orcid.org/0000-0001-9380-4611 and Zhang, Shujun ORCID logoORCID: https://orcid.org/0000-0001-5699-2676 (2024) Machine Learning Driven Developments in Behavioral Annotation: A Recent Historical Review. International Journal of Social Robotics, 16. pp. 1605-1618. doi:10.1007/s12369-024-01117-1

Official URL: https://doi.org/10.1007/s12369-024-01117-1
DOI: http://dx.doi.org/10.1007/s12369-024-01117-1
EPrint URI: https://eprints.glos.ac.uk/id/eprint/13773

Disclaimer

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.

UNIVERSITY OF
GLOUCESTERSHIRE

This is a peer-reviewed, post-print (final draft post-refereeing) version of the following in press document, This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://doi.org/[insert DOI] and is licensed under Publisher's Licence license:

Watson, Eleanor, Viana, Thiago ORCID: 0000-0001-9380-4611 and Zhang, Shujun ORCID: 0000-0001-5699-2676 (2024) Machine Learning Driven Developments in Behavioral Annotation: A Recent Historical Review. International Journal of Social Robotics. (In Press)

EPrint URI: https://eprints.glos.ac.uk/id/eprint/13773

Disclaimer

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.

Manuscript
Click here to access/download;Manuscript;Machine Learning Driven Developments I1D6.docx

Machine Learning Driven Developments in Behavioral Annotation: A Recent Historical Review

Eleanor Watson [1*], Thiago Viana [1], Shujun Zhang [1]

University of Gloucestershire, School of Computing and Engineering, The Park, Cheltenham, GL50 2RH; eleanorwatson@connect.glos.ac.uk (E.W); tviana1@glos.ac.uk (T.V.); szhang@glos.ac.uk (S.Z.)
*****Correspondence: eleanorwatson@connect.glos.ac.uk

Abstract

Annotation tools serve a critical role in the generation of datasets that fuel machine learning applications. With the advent of Foundation Models, particularly those based on Transformer architectures and expansive language models, the capacity for training on comprehensive, multimodal datasets has been substantially enhanced. This not only facilitates robust generalization across diverse data categories and knowledge domains but also necessitates a novel form of annotation—prompt engineering—for qualitative model fine-tuning. This advancement creates new avenues for machine intelligence to more precisely identify, forecast, and replicate human behavior, addressing historical limitations that contribute to algorithmic inequities. Nevertheless, the voluminous and intricate nature of the data essential for training multimodal models poses significant engineering challenges, particularly with regard to bias. No consensus has yet emerged on optimal procedures for conducting this annotation work in a manner that is ethically responsible, secure, and efficient. This historical literature review traces advancements in these technologies from 2018 onward, underscores significant contributions, and identifies existing knowledge gaps and avenues for future research pertinent to the development of Transformer-based multimodal Foundation Models. An initial survey of over 724 articles yielded 156 studies that met the criteria for historical analysis; these were further narrowed down to 46 key papers spanning the years 2018-2022. The review offers valuable perspectives on the evolution of best practices, pinpoints current knowledge deficiencies, and suggests potential directions for future research. The paper includes six figures and delves into the transformation of research landscapes in the realm of machine-assisted behavioral annotation, focusing on critical issues such as bias.

1. Introduction

Understanding complex social interactions, intentions, and incentives through machine intelligence presents a formidable challenge, largely owing to the necessity for vast training datasets. These datasets, particularly those capturing human norms and values, are not only scarce but also complex to annotate. Detailed event analysis often necessitates integrating multiple information streams, making the creation of high-quality behavioral datasets more challenging compared to those for objects or general events.

Given the pivotal role of datasets in machine learning and the unique challenges associated with behavioral annotation, an in-depth review of recent advancements is both timely and crucial. This recent historical literature review aims to document the current state of the art in behavioral annotation for machine learning datasets, identify existing knowledge gaps, and assess the efficacy of emerging techniques. Additionally, the review includes a temporal mapping to delineate various phases of technology adoption and maturation.

The structure of this paper is designed to offer historical context. Section 2 outlines the methodology employed for this review, Section 3 highlights key historical advancements in digital technologies and annotation methods for the period 2018-2022, and Section 4 provides an analytical discussion of these developments.

This literature review extends the scope of our earlier work, 'Augmented Behavioral Annotation Tools, with Application to Multimodal Datasets and Models: A Systematic Review.' While the prior study offered a comprehensive review of annotation tools, it did not delve into the historical progression of methods and technologies in the field of behavioral annotation due to constraints on length (Watson et al. 2023). The current review narrows its focus to historical developments, providing a unique lens through which to understand both progress and emerging trends in this area. The primary audience for this review includes researchers and practitioners in social robotics and human-robot interaction (HRI). These are disciplines intensely interested in

the recognition, interpretation, and emulation of socially relevant behaviors through the use of advanced annotation techniques. The significance of annotation methods cannot be overstated for these communities, given the pivotal role annotated data plays in the training and validation of machine learning models tailored for social robots. Accurate and context-rich annotation is indispensable for engineering robotic behaviors that are both socially acceptable and efficacious in real-world applications. By offering a thorough overview and taxonomy of current annotation methods, this paper aspires to function as a comprehensive guide for selecting appropriate annotation strategies in the realm of social robotics research and development.

1.1 Definition of Annotation

In this paper, 'annotation' is defined as the act of supplementing data with additional labels or markers to enrich its contextual or semantic value. Such annotations can be applied across diverse data modalities including, but not limited to, text, images, audio, and video streams. We categorize annotations into two principal types: Observable phenomena, which are directly ascertainable from the data (e.g., labeling an image featuring a car as 'car'), and inferred phenomena, which are not directly observable but are deduced from the available data (e.g., denoting 'engagement' based on facial expressions and body language). This categorical distinction serves as a foundational framework for the discussions and analyses presented in this literature review.

1.2 Annotation Taxonomy

Within the ecosystem of annotation tools and methodologies, a multi-dimensional taxonomy proves useful for systematic classification. The dimensions under consideration are:

- Application Context: Specifies the domain or field to which the annotation is most pertinent, such as healthcare, finance, or social robotics.
- Data Type: Identifies the nature of the data subjected to annotation, for example, text, images, or video.
- Annotation Method: Denotes the level of automation involved, distinguishing between manual, semi-automated, and fully automated approaches.
- Involvement of Human Annotator: Classified into two categories—'auxiliary tool,' indicating high human engagement, and 'labor amplification tool,' where automation is prevalent.
- Temporal Aspect: Indicates if the annotation is static, occurring at a single point in time, or dynamic, evolving over a period.

This taxonomy serves as the foundational framework for the discussion and categorization of diverse annotation methods and tools examined in this review.

2. Methods and Historical Literature Review

This research employs a Historical Literature Review methodology to comprehensively examine the trajectory of annotation practices since the beginning of the 21st century, particularly focusing on machine-augmented techniques for creating actionable machine learning datasets. Special attention is given to recent publications that offer insights into the challenges and opportunities associated with multimodal abstraction in machine learning models. The objective is to lay a foundational base upon which future research in this domain can be constructively built.

To fulfill the primary objective of this paper, a set of seven guiding questions has been formulated. The overarching query centers on identifying the major innovations and best practices that have emerged in the realm of behavioral annotation.

The authors conducted an extensive review of literature pertaining to online social annotation tools, utilizing a targeted keyword-based sampling strategy. The search aimed to identify full-text articles that report on empirical studies employing annotation tools and methodologies for behavioral data collection. While the majority of the subjects in these studies were humans, instances where non-human subjects such as rodents were included were also considered, provided the findings held translational potential to human contexts. The PRISMA guidelines and checklist were also applied to ensure the robustness of this study (Prisma). The research databases are presented in Table 1.

Table 1. 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

To qualify for inclusion in this literature review, studies needed to satisfy a set of criteria outlined in Figure 1:

1. Utilization of tools specifically designed for behavior annotation, with the terms 'annotation' and 'behaviors' explicitly included.
2. Examination of either a collaborative analysis mechanism or an element of automation designed to augment human labor.
3. Comprehensive reporting of research methods, including the type of data generated, technologies deployed, intended use case, and overall research design. For studies adopting quantitative methodologies, detailed reporting of statistical analysis methods was also required.
4. Distinctiveness in research contributions, ensuring no overlap with other studies. In instances of overlap, priority was given to the most recent findings.
5. Publication date falling within or postdating the year 2000.
6. Availability in English or the presence of a professionally translated version.

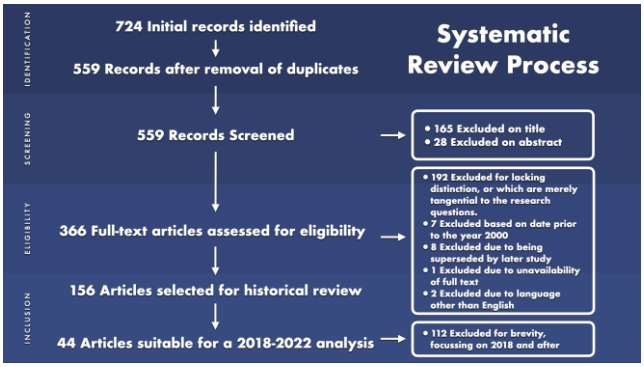

Figure 1. Flowchart outlining the Recent Historical Review Process employed in the current study.

The search commenced with the identification of 366 articles. This was achieved through a Boolean search strategy employing the terms 'ANNOTATION' 'AND' 'BEHAVIOR,' along with variations to encompass both American and British English. The search was specifically tailored to include publications from the year 2000 onwards, aligning with significant developments in behavioral annotation and collaborative annotation techniques. Of the 366 articles initially identified, exclusions were made based on the following criteria:

Pre-2000 Publication Date: 7 articles were excluded for being published before the year 2000, as they did not align with our timeframe of interest.

Superseded by Subsequent Research: 8 articles were excluded because they were rendered obsolete by more recent studies, ensuring the inclusion of only the most current and relevant information.

Unavailability of Full Text: 1 article was excluded due to the inability to access its full text, a necessary component for comprehensive analysis.

Language Constraints: 2 articles were excluded for being in non-English languages without professional translations available, to maintain consistency in language and comprehensibility.

In our review process, a significant number of articles (192) were excluded for reasons pertaining to their content quality and direct relevance to our research questions. This exclusion was based on two main factors:

Lack of Distinction in Content: Articles were assessed for their unique contributions to the field of annotation and behavior, especially in the context of behavioral annotation and collaborative techniques. This criterion involved evaluating whether the articles provided new insights, innovative methodologies, or significantly advanced our understanding of the topic. Articles that merely reiterated well-established concepts without adding new perspectives or findings were deemed to lack the necessary distinction for inclusion in our review.

Tangential Relation to Research Questions: The focus of our review necessitated that each article directly address our core research questions. Articles that only marginally touched upon these topics, or whose primary focus was on areas not directly related to our research questions, were considered tangential. For instance, articles that discussed annotation or behavior in contexts vastly different from behavioral annotation and collaborative methods, or those that dealt with these themes at a very superficial level, were excluded. This dual-pronged approach to exclusion ensured that the final selection of articles was not only directly relevant but also contributed meaningfully to the discourse on annotation and behavior in the context of XML and collaborative techniques. The exclusion of these 192 articles was thus a critical step in refining our literature pool to include only those studies that offered substantial, relevant, and distinctive insights into our research area.

Further Screening: After applying these initial exclusion criteria, a total of 348 articles were screened further. The remaining articles underwent a further careful screening process. Articles were evaluated for their relevance to our research question, the depth of content, and methodological rigor. This stage involved a careful review of abstracts and, where necessary, full texts, to ascertain each article's contribution to our understanding of annotation and behavior in the context of behavioral annotation and collaborative techniques.

Eligibility Assessment: Following this rigorous screening, a subset of articles was deemed eligible for in-depth analysis. These articles were selected based on their direct relevance, methodological soundness, and the uniqueness of the insights they offered into our research question.

Selection for Historical Assessment: From this pool of eligible articles, 156 were chosen for historical assessment. This selection was grounded in criteria that emphasized content depth and direct relevance to the historical context of our review. Articles that did not sufficiently contribute to a historical understanding of the topic or lacked depth in their analysis were excluded in this phase.

Final Inclusion: The final stage of our selection process involved a further refinement of these 156 articles. It was unfortunately necessary to be strict with the number of papers, as otherwise the review paper would be too long for publication. Through a comprehensive evaluation of each article's content and relevance to the scope of our historical review, the pool was distilled by 112 papers to 44 articles since 2018, sufficient for a concise historical review whilst maintaining reasonable brevity. This final set was chosen based on stringent criteria that prioritized articles offering in-depth analysis, innovative perspectives, and significant contributions to the field of behavioral and collaborative annotation techniques.

3. Recent Historical Review of Behavioral Annotation

The subsequent section delineates the data amassed during the Systematic Literature Review process. The scope of the literature review extended from 2018 CE to late 2022 CE. Spanning this **five**-year period, the research landscape revealed several distinct phases of development. These phases are intrinsically linked to key technological advancements such as eXtensible Markup Languages (XML), Cloud and Software-as-a-Service (SaaS) technologies, OpenCV, Deep Learning, Convolutional Neural Networks (CNNs), and Transformer-based machine learning models. To address the central research question concerning major innovations and best practices, a timeline is provided that demarcates four principal epochs in the history of annotation research.

3.1 2018-2022 The Emergence of Deep Learning Schools, Transformers

As deep learning technologies have matured, distinct schools of thought and practice have emerged. Various technology enterprises have sought to develop proprietary machine learning capabilities aligned with their specific strengths. For instance, Google's BERT facilitated advanced natural language processing, well-suited to the company's text-centric services (Devlin, Chang *et al.* 2019). IBM focused on Quantum Machine Learning processes that favor its core competences with specialist 'Big Iron' hardware (Havlíček, Córcoles *et al.* 2019). DeepMind pioneered sophisticated forms of Reinforcement Learning that could learn to perform challenging tasks such as playing video games, culminating in the release of AlphaGo series, which generated much acclaim at matching the greatest human players in Go, a game with an enormous number of permutations (Silver, Huang *et al.* 2016, Silver, Huang et al. 2016, Silver, Schrittwieser *et al.* 2017), as describe in Figure 2.

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.

An attention function maps a query and a set of key-value pairs to an output vector, where the output is computed as a weighted sum of the values. Self-attention, also called intra-attention, is a recurrent attention mechanism that computes a representation of a sequence by relating different positions in the sequence. Self-attention has been used in tasks such as reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations. Transformers are transduction models which are entirely reliant upon self-attention to compute representations of input and output, and do not require the assistance of sequence-aligned RNNs or convolution.

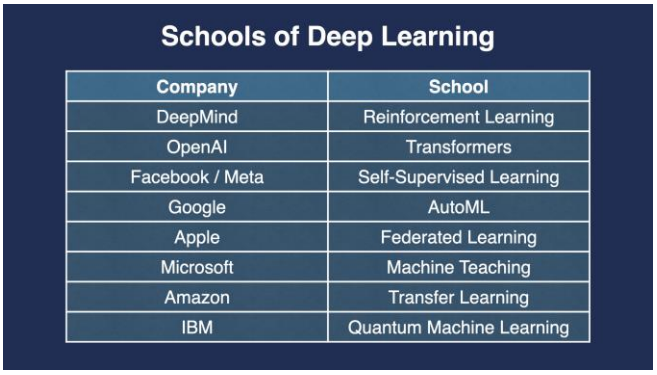

Figure 2. Schools of Deep Learning championed by various technology ventures.

This innovation swiftly led to a new family of machine learning systems. All the models in this family share a property of in-context learning, providing them with the ability to learn a new task from a few 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 (Vaswani, Shazeer *et al.* 2017).

From 2018 onward, derivations of the Vaswani team's research became the primary focus of machine learning research, especially after the release of Generative Pre-Trained Transformer technologies developed by OpenAI. Many of the most exciting models of recent years are built upon transformers, or transformers feeding a diffusion model, including Gopher, Minerva, Chinchilla, Gato, DALL-E 2 and Stable diffusion. However, it can also be argued that Transformers are simply easy to train in a parallelizable way and are not necessarily special in themselves. Even much simpler models, such as RNNs may perform similarly well if sufficiently scaled and optimized.

GPT-1 did not merit much attention from the community, but GPT-2 was viewed as a substantial development. This technology could produce text of plausible believability, and experiments were undertaken to embed such technologies into bot-only forums which attempted to replicate the manner and culture of posts in online communities such as Reddit (SubSimulator, 2019).

GPT-3 emerged in 2019, built upon similar technology to its predecessor, but with orders of magnitude larger number of parameters, and training time. Some researchers took to describing this new wave of multimodal-ready techniques as 'Foundation models', more of a genericized name which can include all Large Language Models, Transformers, and Diffusion Models (Bommasani, Hudson *et al.* 2021). These Foundation models were found to be capable of providing the function of earlier styles of models such as Convolutional Neural Networks, but with greater accuracy and generalizability (Li, Lv *et al.* 2021).

This launched a new wave of interest and excitement around an emerging new class of large (multi-billion parameter) multimodal-ready models (Liang, Zadeh *et al.* 2022). A notable attribute of these models is that there is often a sudden and surprising leap in the capability of models after training for a long time, even it may appear as an irrational overtraining of diminishing or negligible returns. This sudden discontinuity is an important factor in the value of these new massive Large Language Models/Transformers (Ganguli, Hernandez *et al.* 2022), as shows in Figure 3. Emergent abilities are only present in sufficiently large models, and thus they would not have been predicted simply by extrapolating the scaling curve from smaller models (Wei, Tay *et al.* 2022, Srivastaga, Rasotgi *et al.* 2022), as displayed in Figure 4 and Figure 5.

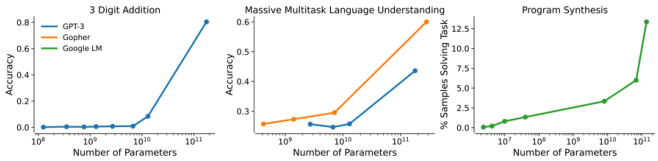

Figure 3. Three examples of abrupt specific capability scaling. From Predictability and Surprise in Large Generative Models (Ganguli, Hernandez *et al.* 2022).

```
Q: What movie does this emoji describe? 👧🐟🐠🐡

2m:   i'm a fan of the same name, but i'm not sure if it's a good idea
16m:  the movie is a movie about a man who is a man who is a man …
53m:  the emoji movie 🐟🐠🐡
125m: it's a movie about a girl who is a little girl
244m: the emoji movie
422m: the emoji movie
1b:   the emoji movie
2b:   the emoji movie
4b:   the emoji for a baby with a fish in its mouth
8b:   the emoji movie
27b:  the emoji is a fish
128b: finding nemo
```

Figure 4. Outputs from a prompt asking for which movie the emoji represent, highlighting a sudden and unpredictable takeoff in capability as model parameter size increases. From *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (Srivastava, Rastogi *et al.* 2022).

Figure 5. Outputs from a prompt asking for "A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!" at differing parameter scales, demonstrating the progression of capabilities visually. From Google's Parti model in *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation* (Yu, Xu et al. 2022).

3.2 Machine-augmented Behavioral Annotation in 2018

Al Zamil *et al.*, 2018 introduce an Internet of Things focused annotation of multiple segmented behaviors, including environmental information/contexts in a smart home context. Behaviors are often driven by environmental parameters, and whether a behavior is appropriate is linked to the environmental or situational context. This mechanism attempts to facilitate the annotation of both behavioral and environmental information using Hidden Markov Model and Conditional Random Field techniques to analyze data. The annotation models also allow the annotation of concurrent behaviors (such as reading a newspaper whilst drinking a coffee) (Al Zamil, Rawashdeh et al. 2018).

Gaur *et al.*, 2018 provide a brief retrospective on video annotation developments to date, including their respective merits, features, and innovations of technologies including VitBat, Viper, iVAT, VATIC, and BeaverDam, all discussed above. The researchers not that image annotation is becoming relatively trivial thanks to Convolutional Neural Network advances. However, the number of frames and temporarily of video (as well as bandwidth and memory requirements) make video annotation much more computationally intensive. This paper describes the state of the art in the domain of video annotation immediately prior to the advent of Large Language Models and Transformers (Gaur, Saxena et al. 2018).

Bahnsen, Møgelmose, and Moeslund, 2018 present annotation toolboxes that enable objects of interest to be pinpointed with pixel and polygon-based masks as well as bounding boxes. The annotation of sequential images in both RGB and thermal modalities is supported. Each annotated object is assigned a classification tag, a unique ID, as well as at least one further metadata tag (Bahnsen, Møgelmose et al. 2018).

User engagement is an essential aspect of application design, especially annotation mechanisms which may be perceived as boring, complex, and frustrating. An ability to gauge user engagement during various processes is an important response to these challenges. Dahmija & Boult, 2018 compare observational estimates from both expert human raters and vision-based learning to estimate user engagement. The vision-based approach uses automated computation of specified Action Units combined with a Recurrent Neural Network. The researchers introduce an approach that exploits the inherent confusion and disagreement in raters' annotations to build a scalable engagement estimation model. By actively modeling the uncertainty, the appropriate weight of subjective behavioral cues is coded, which significantly improves prediction resulting in an approach which performs comparably with experts in predicting engagement for a trauma-recovery application (Dhamija and Boult 2018).

3.3 Machine-augmented Behavioral Annotation in 2019

A structure for the design of annotation pipelines enables more sophisticated, modular, and rapidly prototyped semi-automatic annotation workflows. Jäger *et al.*, 2019 respond to this opportunity with a standard process for specifying semi-automatic annotation pipelines. The LOST (Label Objects and Save Time) open-source implementation permits the combination of multiple annotation tools and machine learning algorithms into an ensemble process, which can be strung together in a modular manner, visualized through a web-based user interface which supports the use of Python scripts to control processes. An annotation pipeline (annotation process) can be composed of six different building block types, namely data sources, scripts, annotation tasks, loops, data exports and visualizations, with each element separately parameterized. LOST can function as a stand-alone program, a cloud-based instance, or a distributed computational workload shared across multiple machines. Outputs from the process narrowly beat human benchmarks. The concept of a 'click and place' custom annotation builder is highly disruptive in its flexibility and the ease of altering the workflow. The adaptability should enable rapid prototyping and A-B testing of various approaches, as well as continuous upgrades as technology improves and new multimodal layers are feasible. Such a pipeline will be an essential component of an advanced and efficient annotation workflow. The paper also has a helpful overview of contemporary annotation tools and their respective strengths. The researchers state an intention to implement an interface to Mechanical Turk for crowdsourcing applications, but this does not seem to have been released (Jäger, Reus et al. 2019).

Lorbach *et al.*, 2019 apply new techniques to the annotation of rodent behavior, experimenting with feedback loops between automated analysis paired with human feedback. Rodent behaviors are fast, erratic, variable, and instinctive. This makes annotation challenging, especially as behaviors may be difficult to classify due to research-specific requirements, or the progression of the experiment (or aging or disease in the animal). This research demonstrates a hybrid technology, whereby iterative learning loops are set up to augment annotation, whereby the system guesses a behavior, and a human annotator grades that appraisal, correcting where necessary. Automated annotation can become progressively sophisticated over time. In contrast to traditional labeling, the interactive framework not only obtains annotations from the user, but it also trains a behavior classifier at the same time. The researchers report that after 300 labeling iterations the classifiers are as accurate as those trained by the data oracle. Agreement scores of .70 between human and AI annotations were achieved. The challenges of the use-case make the demonstration even more compelling. The paper makes an interesting point that annotation makes otherwise mundane sets of info searchable, and thus potentially interesting, as anomalies can be pinpointed and shared. The researchers also note another way to reduce the need for labeling by exploiting unlabeled data in addition to a few labeled examples through a semi-supervised approach. The structure of the unlabeled data itself can improve classification when linked with feedback on which examples to label. The researchers observe that close integration of semi-supervised learning and active learning suggest a promising direction for future research (Lorbach, Poppe et al. 2019).

3.4 Machine-augmented Behavioral Annotation in 2020

Hanggi *et al.*, 2020 present a protocol for annotation of sedentary behavior according to a hierarchical method. Subsequent passes on a given piece of data provides extra richness and context. The first pass ascertains whether the data is at all useful, pertinent, or potentially problematic, which also saves time. Annotations become more granular at higher passes, rectifying finer elements. By breaking down annotation steps, one annotator can rapidly yes/no for a certain characteristic, which could be highly efficient. Such an approach likely enables much greater efficiency than one annotator (or coder in the authors parlance) doing multiple potential tasks on any data they work with. It also enables annotators with less experience to work on less critical areas first, until they prove reliable, and can thus be graduated to more difficult and sensitive tasks (Hänggi, Spinnler et al. 2020).

Language models are further applied to annotation for NLP tasks by Wang *et al.*, 2020. Prior methods of automated annotation to date have been generally limited to specific tasks and requiring significant pre-labelled data to train from. To move beyond these limitations, the researchers explore ways to leverage GPT-3 as a low-cost data labeler for datasets. They were able to achieve cost savings of between fifty and ninety-six percent for comparable performance. The researchers also propose a novel framework of combining pseudo labels generated by GPT-3 with human-soured labels, again achieving better performance with minimal cost. These results have opened new cost-effective data labeling techniques which can be generalized to many annotation requirements. This method presents a huge economic benefit and leads to a new paradigm of machine-augmented annotation, particularly in linguistic applications. Some questions remain about how best to structure the model and data flow to achieve the best results with minimal human efforts, especially whilst preserving nuanced cultural, situational, and personal context (Wang, Liu et al. 2021).

Annotation of body motion is improved by Takano, 2020, with a framework for describing human body movements with Inertial Measurement Units and motion primitives, with natural language descriptors. The purpose of the research is to discern whether motions can be encoded into natural language through a procedural method, with a view towards automation. IMU motion sensor apparatus gathers input signals which are represented by a set of feature vectors. These are encoded into parameters for processing by a statistical framework driven by a Hidden Markov Model, and then finally converted into natural language. A probabilistic graphical model represents mappings between human motions and words. Nodes in the first layer are motion symbols, nodes in the third layer are words, and hidden states in the second layer connect motion symbols with words.

This research advances mechanisms by which natural information such as motion can be vectorized, parameterized, and processed by a statistical framework, and finally converted into a natural language description, using a data flow across two separate systems, resulting in a dataset created using significantly automated means. The result is an ambitious example of what's possible in contemporary machine vision and natural language probabilistic processing of natural data. The techniques described provide a reliable baseline which can be iterated upon and improved by human coders. This technique seems to be helpful for generating datasets of behaviors in an efficient manner, whilst enabling a deeper description and understanding of the content of human movement (vectorized and parameterized), beyond mere coding or tagging (Takano 2020).

Research by Kurzhals *et al.*, 2020 focuses on analytics of eye movement, with their prototype providing a fuzzy nonlinear search for other behaviors within an efficient annotation system. Such a mechanism may provide extra context for coders or enable efficiency of coding multiples of the same general behavior regardless of position in the media, particularly for the analysis of activities and behavior that require a higher degree of semantic interpretation.

Eye tracking data is collected using eye tracking glasses, with participants making day-long recordings performing a broad range of activities. This data was then annotated manually with eight activity classes: outdoor, social interaction, concentrated work, mobile, reading, computer work, watching media, and eating. A visual analytics interface was provided which includes three major components: a multi-layered timeline visualizing features and respective frames of video segments, an interface for guided query refinement based on feature weighting, and a query result view showing thumbnails of retrieved time spans. They applied the AlexNet F6 global descriptor CNN with its global receptive field to provide image-wide comparisons. Local features such as image patches were also included, deriving user fixations based upon gaze points.

The similarity measures described here are particularly noteworthy, as they seem quite novel. It's much easier to simply extract features than to find similar features in an unsupervised or semi-supervised manner. Temporal segmentation processes applied to patches in the CNN appear to have assisted with this process. They amplify human intelligence to cross-validate machine generated predictions and classifications in a maximally efficient manner, with human beings focusing on the identification of false positives. This system enables coders to search for similar behavioral patterns expressed in other media in a fuzzy and nonlinear manner. This may enable coders to specialize in an area, or more rapidly annotate similar examples across a broad range of data sources. This may be especially interesting for providing examples of a novel description of behavior. It also clearly provides important validation processes for machine learning systems, to ensure that false positives are weeded out, thereby improving accuracy and reducing bias (Kurzhals, Rodrigues et al. 2020). Dilated Neighborhood Attention Transformer models may also have strong applicability to domains where CNNs are commonly employed (Hassani and Shi 2022).

3.5 Machine-augmented Behavioral Annotation in 2021.

Dong *et al.*, 2021 elucidate whether context from an article's title and text itself can be applied to automated tagging of it. The study uses an attention-based model called a Joint Multilabel Attention Network (JMAN), which is designed to mimic how users read and annotate documents, to see if it can improve tagging accuracy. The study looks at four large social media data sets. The advantage of the JMAN approach is that it can interpret entire sentences, not just individual words, enabling greater

sophistication and reliability. Automated tagging can learn progressively to improve the accuracy of later tags as well as providing an instant check for tags that may be suspicious (Dong, Wang et al. 2021).

Another form of emulation of human prediction or emulation is introduced by Stiennon *et al*., 2020 in response to perceived bottlenecks of approximate benchmarks and rubrics that do not necessarily reflect training objectives, particularly ones aimed at a difficult to specify quality of output. The researchers demonstrate the feasibility of improving the quality of automated summarization by training the model to optimize for human preferences. A dataset of human-generated summaries and comparisons between them is applied to train a model to predict human preferences. This model is in turn used as a reward function for the fine-tuning of a summarization policy through reinforcement learning. The results were more favorable than human reference summaries, as well as larger models fine-tuned only with supervised learning. In addition, this method was able to interface with a wide variety of media, generalizing to new datasets. This research highlights the importance of accurate training loss procedures for accurate targeting of model behavior (Stiennon, Ouyang et al. 2020).

Automated annotation methods thus far have a common limitation; mainstream methods simply scan the texts in the document and do not fully model the way actual users read and annotate it, nor do they account for semantic relations between documents. A novel deep learning-based method for these problems has been designed which can mimic users' reading and annotation behavior to formulate better document representation, leveraging the semantic relations among labels. This new technique, called the joint multilabel attention network (JMAN), was compared with others including SVM, LSTM, Bi-directional RNN, and Hierarchical Attention Networks. The results showed that JMAN outperformed the state-of-the-art deep learning-based models. The idea of modelling the behavior of user annotation itself in one shot, rather than simply analyzing the underlying elements and presenting an amalgamation of those is audacious. Optimizing based upon an assumption of what users find noteworthy is in fact what one is hoping to collect may be the shape of things to come, at least in some contexts. This long and rich paper is replete with exacting technical detail, especially in its comparison between various ML techniques. It also illustrates how quality reference data may not be necessary to achieve impressive results that closely match human coding accuracy and feel (Dong, Wang et al. 2019).

Goldberg *et al.*, 2021 describe new methods of annotating social competence in interactions. Prediction of social competence might offer insights on the prediction of likely corrigibility or social discomfort in a situation. Data gathered on this subject might also assist machines to exhibit social corrigibility, enabling simulation of agreeable social interaction, or a set of basic heuristics to implement as part of taking an action (Goldberg, Tanana et al. 2020).

Another method of emulation of human preference is explored by Wang *et al*., 2021, who demonstrate UDG (Unsupervised Data Generation), a method which leverages few-shot prompts to synthesize high-quality training data without requiring human annotations. Such data augmentation methods show tremendous promise in reducing the amount of human input necessary to constructs datasets, refocusing human efforts towards validation and verification, fine-tuning, and prompt iterations instead. As technology improves, Synthetic Data Generation may make annotation almost obsolete in certain contexts through a paradigm of annotating only enough to discover the necessary parameters for synthetic data augmentation (paired with human validation) to take over (Wang, Yu et al. 2021).

Xue *et al.*, 2021 apply annotation techniques to Virtual Reality users' emotional experiences, gathered through physiological arousal sensors. Embedding annotation functions and viewport contextualization into the annotation of VR experiences aids the annotation users in annotation their emotional experiences. It comprises three steps: continuous annotation time-alignment, segment-based viewport clustering, and lastly viewport-dependent annotation fusion. Together, these elements facilitate annotation at any time within the VR experience, temporally and proprioceptively linked to the user's experience in the moment. 360-degree VR experiences add increased annotation challenges (in knowing what one is looking at during an event, or whether a user noticed something or not), as well as the occluded face when wearing an HMD in VR making it more challenging to infer emotion from facial affect. It also makes it challenging to record one's emotions on paper or another device. The immediate overlay of the annotation interface makes it easier to stay focused on the experience (Xue, El Ali et al. 2021).

Segalin *et al*., 2021, introduce the Mouse Action Recognition System (MARS), an automated pipeline for pose estimation and behavior quantification in pairs of freely interacting mice. They compare MARS's annotations to human annotations and find that MARS's pose estimation and behavior classification achieve human-level performance. MARS includes three supervised classifiers trained to detect attack, mounting, and close investigation behaviors in tracked animals. These classifiers were trained on 6.95 hr of behavior video, 4 hr of which were obtained from animals with a cable-attached device such as a microendoscope. Separate evaluation (3.85 hr) and test (3.37 hr) sets of videos were used to constrain training and evaluate MARS performance, giving a total of over 14 hr of video. All videos were manually annotated on a frame-by-frame basis by a single trained human annotator. To evaluate inter-annotator variability in behavior classification, they also collected frame-by-frame manual labels of animal actions by eight trained human annotators on a dataset of ten 10-min videos. Two of these videos were annotated by all eight annotators a second time a minimum of 10 months later for evaluation of annotator self-consistency. The Behavior Ensemble and Neural Trajectory Observatory (BENTO) is a MATLAB-based GUI for synchronous display of neural

recording data, multiple videos, human/automated behavior annotations, spectrograms of recorded audio, pose estimates, and 270 'features' extracted from MARS pose data—such as animals' velocities, joint angles, and relative positions. MARS's detector performs MSC-MultiBox detection (Szegedy, Reed et al. 2014) using the Inception ResNet v2 architecture. Pose detection is performed using a sing a stacked hourglass network architecture with eight hourglass subunits. The MARS tool simplifies the process of running trained detection, pose estimation, and behavior classification models on video data. It produces bounding boxes, pose estimates, features, and predicted behaviors for each video in a directory as output. The system can also interface with Convolutional Neural Networks. On a desktop computer with 8-core Intel Xeon CPU, 24 Gb RAM, and a 12 GB Titan XP GPU, MARS performs two-animal detection and pose estimation (a total of four operations) at approximately 11 Hz. This paper includes a process flow of various stages and variations, highlighting the state of the art in end-to-end feature extraction and prediction. The paper also includes a section on tracking the Neural Correlates of behavior. Such data, if able to be gleaned through non-invasive methods, might provide extra richness to behavioral data, and various neurologically revealed preferences (Segalin, Williams et al. 2021).

A further expansion of non-human animal behavior annotation is made by Tjandrasuwita *et al.*, 2021 (Tjandrasuwita, Sun et al. 2021) who focusses on human factors. The researchers explore how the subjective impressions and experience levels of annotators can induce a variational personal bias on the resulting labels. They classify these behaviors into clusters of similar annotation styles for analysis using a new method based upon program synthesis. The synthesized model learns to select trajectory features and to apply a temporal filter between them. This enables the most salient features expressed through the focus of annotation behavior to be isolated, with temporality aspects highlighting its particular importance (i.e., what gets done first and is afforded the most time). This method presents new possibilities for bringing insights from behavioral neuroscience into annotation, as well as the potential of emulation of a range of styles of annotators, which may be more suitable for specific needs, or which may be desirable to ensure a broad range of annotator styles has been applied to a dataset, to reduce potential biases due to a disproportionate paucity of annotation behavior styles.

3.6 Machine-augmented Behavioral Annotation in 2022

Building safer agents with human feedback typically requires presenting a simulation of behavior and requesting a human procedural annotation of unsafe behavior found within it, ideally along with alternate recommendations. Such simulations may lack fidelity to real-life situations and constraints, and the potential variety in parameters may be challenging for reinforcement learning techniques. ReQueST, presented by Rahtz *et al.*, 2021 sidesteps the need for direct annotations via a neural simulator capable of observing from safe human trajectories, and using the learned simulator to efficiently learn a reward model, thereby generating optimized trajectories upon which to ask for feedback. The researchers demonstrate the efficacy of this approach in complex 3D environments and first-person tasks, enabling an order of magnitude reduction in unsafe behavior (Rahtz, Varma et al. 2022).

Observation of trajectories can also be applied to the annotation of driver behavior. This presents an important use case for automated driving processes, as a driver who is acting erratically should merit more attention and precaution. Axenie *et al.*, 2022 introduce a fuzzy modelling and inference method for calibrating driver behavior recognition models, by parameterizing car-following and lane-change behaviors observed in vehicle trajectories into plausible classes whilst taking physical laws and domain knowledge into account within its predictions. The resulting automatically labelled parameters can also be applied to emulate specific types of behaviors or driving styles (Axenie, Scherr et al. 2022).

Similar modelling of behavior is achieved by Baker *et al.*, 2022a with a model (VideoPreTraining) for sequential decision domains which is capable of learning to emulate human player behavior in Minecraft captured in unlabeled online videos. This was created as a response to a shortfall in adequately labelled data on behavioral priors. A semi-supervised imitation learning system enables agents to learn to act by watching online unlabeled video, if paired with a small amount of labeled data. Specifically, the researchers demonstrate an inverse dynamics model can be trained which is sufficiently accurate to itself label a very large unlabeled repository, from which general behavioral priors can be learned. This behavioral example has nontrivial zero-shot capabilities and can be fine-tuned with both imitation learning and reinforcement learning. The resulting model is robust to complex exploration tasks that are impossible to learn via reinforcement learning alone, and it achieves a parity to human performance in many task areas despite using emulated keyboard and mouse controls (Baker, et al. 2022a, Baker, et al. 2022b).

Research into the creation of new datasets through the in-context learning capabilities of LLMs has led to the development of Selective Annotation techniques, as showcased by Su *et al.*, 2022. A pool of examples are selectively chosen from unlabeled data, whereupon task examples are derived through prompt retrieval at test time. An unsupervised, graph-based selective annotation process, vote-k, selects a subset range of diverse and representative samples. Compared with random sampling, this

selective method results in around a twelve percent improvement in performance, whilst necessitating between ten and one hundred times fewer annotations. The researchers also demonstrate that the vote-k driven selection method is compatible with models of different sizes, and those with domain shifts between training and test data (Su, Kasai et al. 2022).

4. Summary Analysis

The transformation in data annotation techniques can be traced back to a variety of factors, including constraints such as time and cost, as well as advances in peripheral technologies like high-quality sensors and computational hardware. Initially, manual annotation methods dominated, demanding considerable human expertise for diverse tasks—be it text classification, object detection in images, sentiment analysis, and machine translation. While these manual methods offered high-quality annotations, they were often slow and expensive.

As machine learning technologies matured, semi-automated and fully automated methods emerged. These newer methods aimed to capitalize on computational power to alleviate the manual burden. Algorithms in natural language processing and computer vision began to play crucial roles in automating tasks such as text tagging and object labelling, respectively.

Annotation techniques can generally be classified into three categories:

- Manual Annotation: Human annotators manually label the data.
- Semi-Automated Annotation Machine algorithms propose annotations that are verified by human experts.
- Fully-Automated Annotation: Advanced algorithms independently carry out the annotation.

Over time, the performance of these tasks has improved, in part due to more sophisticated annotation techniques and also due to advances in computational hardware and algorithms.

There is a clear trajectory of growth and refinement of dataset over time. Initial dataset sizes often start modestly, tailored to the constraints of early annotation capabilities and intended for benchmarking fundamental tasks. As annotation techniques evolve, harnessing both human expertise and semi-automated tools, these datasets have expanded in scope and scale. This expansion is not merely quantitative but also qualitative, incorporating richer annotations that capture more nuanced aspects of the data, such as sentiment, entailment, or even fine-grained entity relations.

For instance, an image dataset might initially be annotated with basic labels indicating the presence of objects, but later versions could include bounding boxes, segmentation masks, or detailed attributes of each object. Similarly, in language datasets, what might begin as simple part-of-speech tagging can evolve to include complex syntactic trees and semantic role labeling, reflecting advancements in linguistic processing tools and theories.

With each subsequent expansion, we document not only the increase in the number of data points but also enhancements in the depth of information provided by annotations. Modifications in annotation methods are particularly significant, as they often reflect broader shifts in the field's understanding of what constitutes useful data for developing AI models. For example, transitioning from manual to semi-automated annotation may indicate an effort to scale up the dataset while maintaining or even improving the quality of the annotations.

Through meticulous documentation of these changes, we gain insights into the maturation of datasets and, by extension, the progress of the AI field. This historical perspective is invaluable for researchers and practitioners alike, as it provides context for the current state of AI capabilities and guides future directions for dataset development and usage.

Specific datasets like ImageNet for image classification, COCO for object detection, and the Penn Treebank for syntactic parsing have grown both in size and complexity, paralleling advances in annotation methods. While English has been the primary focus for many annotated datasets, there is a growing trend to include multiple languages. For example, parallel corpora in machine translation now often feature a variety of language pairs, making the task of annotation increasingly complex but also more inclusive.

Several factors amplify the complexity of constructing and working with multilingual datasets. Firstly, linguistic features such as diacritics can alter the meaning of words significantly, demanding annotators who are not only fluent in a language but also attuned to its nuances. For instance, in Vietnamese, diacritics indicate tones and change word meanings, which can be perplexing for non-native annotators and automated annotation systems alike. Additionally, the direction of writing varies between languages—Arabic and Hebrew flow right-to-left, which requires specialized software support and a different approach to visual layout when annotating.

Grammatical structures differ markedly across languages, complicating the transfer of annotation guidelines. For example, case systems in languages like German or Russian introduce variability in word endings that must be accurately captured and understood in context. Also, idiomatic expressions and compound word formations present unique challenges that require cultural as well as linguistic expertise.

Despite these hurdles, the enriched datasets that result from multilingual annotation hold immense power. They enable the development of more sophisticated and inclusive AI models, capable of understanding and processing a diverse range of human languages and dialects. Such datasets are instrumental in driving forward technologies like real-time translation, cross-lingual information retrieval, and inclusive language technologies, thereby facilitating global communication and understanding. The increased representativeness and diversity of these datasets also mean that AI systems can serve a broader user base with greater accuracy, ultimately leading to technology that is more equitable and accessible to all.

In recent years, self-supervised learning and Reinforcement Learning from Human Feedback (RLHF) have further altered the landscape. Self-supervised learning minimizes the need for human-labelled data by allowing the model to generate its own labels. On the other hand, RLHF allows models to adapt dynamically, offering real-time adjustments to annotations based on environmental interactions.

For instance, in video annotation, a reinforcement learning model could adapt its labelling strategies in real-time, altering its approach based on the evolving context within the video.

The field of data annotation has therefore seen a significant evolution influenced by a variety of factors—from technological constraints to advances in machine learning and peripheral technologies. The future holds the promise of even more sophisticated methods, thanks to emerging techniques like self-supervised learning and RLHF.

5. Final Considerations

This recent historical review paper contributes to the research domain by distilling a complex series of innovations within a very fast-moving domain being transformed by prompt-driven multimodal models and synthetic data techniques. Several observations can be inferred from the path of annotation technology development discussed above:

- Transformers and diffusion models are poised to consistently outperform earlier, dedicated machine learning architectures.
- The newfound capacity of multimodal models to process diverse data types suggests that these models will substantially benefit from intricate, multimodal annotations.
- The multimodal nature of these models facilitates cross-contextual inferences, enabling them to handle complex, multivariate datasets or dataset catalogues, thereby enriching situational awareness from multiple perspectives.
- While deep learning underscored the importance of data volume, multimodal models may shift the focus towards the richness of context and detail in cross-correlated, theme-related disparate data types.
- The need for multimodality in datasets is escalating, necessitating cross-referencing across data types and the incorporation of multimodal examples that encode norms and values across a wide cultural and situational spectrum. • There is a discernible trend toward the default use of advanced semi-automatic and automatic annotation methods, rendering traditional techniques increasingly obsolete.
- The rise of techniques based on synthetic data, validated by human-generated prompts, signifies a move away from human-centric annotation, except in specialized scenarios. The efficacy limitations of this approach remain an open question.
- The delegation of annotation tasks to machines not only reduces human workload but also augments the potential for discovering additional data dimensions, thereby enhancing the inferential capabilities of the resulting models. This feature is particularly beneficial for multimodal models.

Our review underscores the potential of machine learning-enhanced behavioral annotation in social robotics, particularly in refining robots' ability to interpret human behaviors and emotions. This progression paves the way for robots that are more empathetic and context-aware, aligning with social robotics' goals of improving human-robot interaction and ethical design considerations. Such advancements not only promise to improve the interaction quality between humans and robots but also open avenues for ethical considerations in robot design, ensuring they operate within socially acceptable norms. We advocate for interdisciplinary efforts to integrate these advancements into social robotics, promising to elevate the interaction quality and ethical standards of social robots. This synergy between behavioral annotation and social robotics underscores our study's relevance to the field.

Funding: This research received no external funding.

Informed Consent 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 in machine learning. The authors also wish to thank A. Safronov for editing assistance.

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

5. References

2019. "Subreddit Simulator using GPT-2." Reddit. Retrieved 2 November, 2022, from https://www.reddit.com/r/SubSimulatorGPT2.

2022. "Why I think strong general AI is coming soon." Lesswrong. Retrieved 2 November, 2022, from https://www.lesswrong.com/posts/K4urTDkBbtNuLivJx/why-i-think-strong-general-ai-is-coming-soon#Transformers_are_not_special.

n.d. "RWKV-LM." github. Retrieved 2 November, 2022, from https://github.com/BlinkDL/RWKV-LM.

Al Zamil, M. G., M. Rawashdeh, S. Samarah, M. S. Hossain, A. Alnusair and S. M. M. Rahman (2018). "An Annotation Technique for In-Home Smart Monitoring Environments." *IEEE Access* 6: 1471-1479.

Axenie, C., W. Scherr, A. Wieder, A. S. Torres, Z. Meng, X. Du, P. Sottovia, D. Foroni, M. Grossi, S. Bortoli and G. Brasche (2022). "Fuzzy Modeling and Inference for Physics-Aware Road Vehicle Driver Behavior Model Calibration." *SSRN Electronic Journal*.

Bahnsen, C. H., A. Møgelmose and T. B. Moeslund (2018). "The AAU Multimodal Annotation Toolboxes: Annotating Objects in Images and Videos." *ArXiv* abs/1809.03171.

Baker, B., I. Akkaya, P. Zhokhov, J. Huizinga, E. Tang, A. Ecoffet, B. Houghton, R. Sampedro and J. Clune. 2022(a). "Learning to Play Minecraft with Video PreTraining (VPT)." OpenAI. Retrieved 1 November, 2022, from https://openai.com/blog/vpt/.

Baker, B., I. Akkaya, P. Zhokhov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro and J. Clune. 2022(b). "Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos." *ArXiv* abs/2206.11795.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. v. Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. P. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. J. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou and P. Liang (2021). "On the Opportunities and Risks of Foundation Models." *ArXiv* abs/2108.07258.

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *ArXiv* abs/1810.04805.

Dhamija, S. and T. E. Boult (2018). Automated Action Units Vs. Expert Raters: Face off. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 259-268.

Dong, H., W. Wang, K. Huang and F. Coenen (2019). Joint Multi-Label Attention Networks for Social Text Annotation. In Proceedings of the 2019 Conference of the North, 1348-1354.

Dong, H., W. Wang, K. Huang and F. Coenen (2021). "Automated Social Text Annotation With Joint Multilabel Attention Networks." *IEEE Transactions on Neural Networks and Learning Systems* 32(5): 2224-2238.

Ganguli, D., D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, S. El Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, S. Johnston, A. Jones, N. Joseph, J. Kernian, S. Kravec, B. Mann, N. Nanda, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Kaplan, S. McCandlish, C. Olah, D. Amodei and J. Clark (2022). Predictability and Surprise in Large Generative Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 1747–1764 , numpages = 1718.

Ganguli, D., D. Hernandez, L. Lovitt, N. DasSarma, T. J. Henighan, A. Jones, N. Joseph, J. Kernion, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, S. Johnston, S. Kravec, N. Nanda,

K. Ndousse, C. Olsson, D. Amodei, D. Amodei, T. B. Brown, J. Kaplan, S. McCandlish, C. Olah and J. Clark (2022). "Predictability and Surprise in Large Generative Models." *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Gaur, E., V. Saxena and S. K. Singh (2018). Video annotation tools: A Review. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 911-914.

Goldberg, S. B., M. Tanana, Z. E. Imel, D. C. Atkins, C. E. Hill and T. Anderson (2020). "Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task." *Psychotherapy Research* 31(3): 281-288.

Hänggi, J. M., S. Spinnler, E. Christodoulides, E. Gramespacher, W. Taube and A. Doherty (2020). "Sedentary Behavior in Children by Wearable Cameras: Development of an Annotation Protocol." *American Journal of Preventive Medicine* 59(6): 880-886.

Hassani, A. and H. Shi (2022). "Dilated Neighborhood Attention Transformer." *ArXiv* abs/2209.15001.

Havlíček, V., A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow and J. M. Gambetta (2019). "Supervised learning with quantum-enhanced feature spaces." *Nature* 567(7747): 209-212.

Jäger, J., G. Reus, J. Denzler, V. Wolff and K. Fricke-Neuderth (2019). "LOST: A flexible framework for semi-automatic image annotation." *ArXiv* abs/1910.07486.

Kurzhals, K., N. Rodrigues, M. Koch, M. Stoll, A. Bruhn, A. Bulling and D. Weiskopf (2020). Visual Analytics and Annotation of Pervasive Eye Tracking Video. In ACM Symposium on Eye Tracking Research and Applications, 1-9.

Li, M., T. Lv, L. Cui, Y. Lu, D. A. F. Florêncio, C. Zhang, Z. Li and F. Wei (2021). "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models." *ArXiv* abs/2109.10282.

Liang, P. P., A. Zadeh and L.-P. Morency (2022). "Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions." *ArXiv* abs/2209.03430.

Lorbach, M., R. Poppe and R. C. Veltkamp (2019). "Interactive rodent behavior annotation in video using active learning." *Multimedia Tools and Applications* 78(14): 19787-19806.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Prisma). Available online: 1551 https://www.prisma-statement.org [Accessed 23 October 2022].

Rahtz, M., V. Varma, R. Kumar, Z. Kenton, S. Legg and J. Leike (2022). "Safe Deep RL in 3D Environments using Human Feedback." *ArXiv* abs/2201.08102.

Segalin, C., J. Williams, T. Karigo, M. Hui, M. Zelikowsky, J. J. Sun, P. Perona, D. J. Anderson and A. Kennedy (2021). "The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice." *eLife* 10.

Silver, D., A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis (2016). "Mastering the game of Go with deep neural networks and tree search." *Nature* 529: 484-489.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis (2016). "Mastering the game of Go with deep neural networks and tree search." *Nature* 529(7587): 484-489.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel and D. Hassabis (2017). "Mastering the game of Go without human knowledge." *Nature* 550(7676): 354-359.

Srivastava, A., A. Rastogi, A. B. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. A. Rahane, A. S. Iyer, A. J. Andreassen, A. Santilli, A. Stuhlmuller, A. M. Dai, A. D. La, A. K. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakacs, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Ozyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. S. Howald, C. Diao, C. Dour, C. Stinson, C. Argueta, C. e. F. Ram'irez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. T. Ramirez, C. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. H. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Freeman, D. Khashabi, D. Levy, D. Gonz'alez, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. Coelho Mollo, D. Yang, D.-H. Lee, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. P. Donoway, E. Pavlick, E. Rodolà, E. F. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Mart'inez-Plumed, F. Happ'e, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G.

Parascandolo, G. Mariani, G. Wang, G. Jaimovitch-L'opez, G. Betz, G. Gur-Ari, H. Galijasevic, H. S. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura, H. Zhang, H. Wong, I. A.-S. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Koco'n, J. Thompson, J. Kaplan, J. Radom, J. N. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. O. Alabi, J. Xu, J. Song, J. Tang, J. W. Waweru, J. Burden, J. Miller, J. U. Balis, J. Berant, J. Frohberg, J. Rozen, J. Hernández-Orallo, J. Boudeman, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. O. Omondi, K. W. Mathewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. O. Col'on, L. Metz, L. K. cSenel, M. Bosma, M. Sap, M. t. Hoeve, M. Andrea, M. S. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. Leavitt, M. Hagen, M. a. a. Schubert, M. Baitemirova, M. Arnaud, M. A. McElrath, M. A. Yee, M. Cohen, M. Gu, M. I. Ivanitskiy, M. Starritt, M. Strube, M. Swkedrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, T. MukundVarma, N. Peng, N. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. S. Roberts, N. Doiron, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. W. Chang, P. Eckersley, P. M. Htut, P.-B. Hwang, P. Milkowski, P. S. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. LYU, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. o. R. Delgado, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. Le Bras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee, R. Stovall, R. Teehan, R. Yang, S. J. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, S. Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-h. Lee, S. B. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. R. Biderman, S. C. Lin, S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. A. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. N. Kornev, T. Telleen-Lawton, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. O. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. F. Tong, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. J. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Xinran, Z. Zhao, Z. F. Wang, Z. J. Wang, Z. Wang, Z. Wu, S. Singh and U. Shaham (2022). "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." *ArXiv* abs/2206.04615.

Stiennon, N., L. Ouyang, J. Wu, D. M. Ziegler, R. J. Lowe, C. Voss, A. Radford, D. Amodei and P. Christiano (2020). "Learning to summarize from human feedback." *ArXiv* abs/2009.01325.

Su, H., J. Kasai, C. H. Wu, W. Shi, T. Wang, J. Xin, R. Zhang, M. Ostendorf, L. Zettlemoyer, N. A. Smith and T. Yu (2022). "Selective Annotation Makes Language Models Better Few-Shot Learners." *ArXiv* abs/2209.01975.

Szegedy, C., S. E. Reed, D. Erhan and D. Anguelov (2014). "Scalable, High-Quality Object Detection." *ArXiv* abs/1412.1441.

Takano, W. (2020). "Annotation Generation From IMU-Based Human Whole-Body Motions in Daily Life Behavior." *IEEE Transactions on Human-Machine Systems* 50(1): 13-21.

Tjandrasuwita, M., J. J. Sun, A. Kennedy, S. Chaudhuri and Y. Yue (2021). "Interpreting Expert Annotation Differences in Animal Behavior." *ArXiv* abs/2106.06114.

Vaswani, A., N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin (2017). "Attention is All you Need." *ArXiv* abs/1706.03762.

Wang, S., Y. Liu, Y. Xu, C. Zhu and M. Zeng (Year). Want To Reduce Labeling Cost? GPT-3 Can Help. In EMNLP

Wang, Z., A. W. Yu, O. Firat and Y. Cao (2021). "Towards Zero-Label Language Learning." *ArXiv* abs/2109.09193.

Watson, E.; Viana, T., and Zhang, S. Augmented Behavioral Annotation Tools, with Application to Multimodal Datasets and Models: A Systematic Review. AI (2023), 4, 128-171.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean and W. Fedus (2022). "Emergent Abilities of Large Language Models." *ArXiv* abs/2206.07682.

Xue, T., A. El Ali, T. Zhang, G. Ding and P. Cesar (2021). RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1-15.

Yu, J., Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. C. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge and Y. Wu (2022). "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation." *ArXiv* abs/2206.10789.

Figure 1

Click here to access/download;Figure;Figure001.png

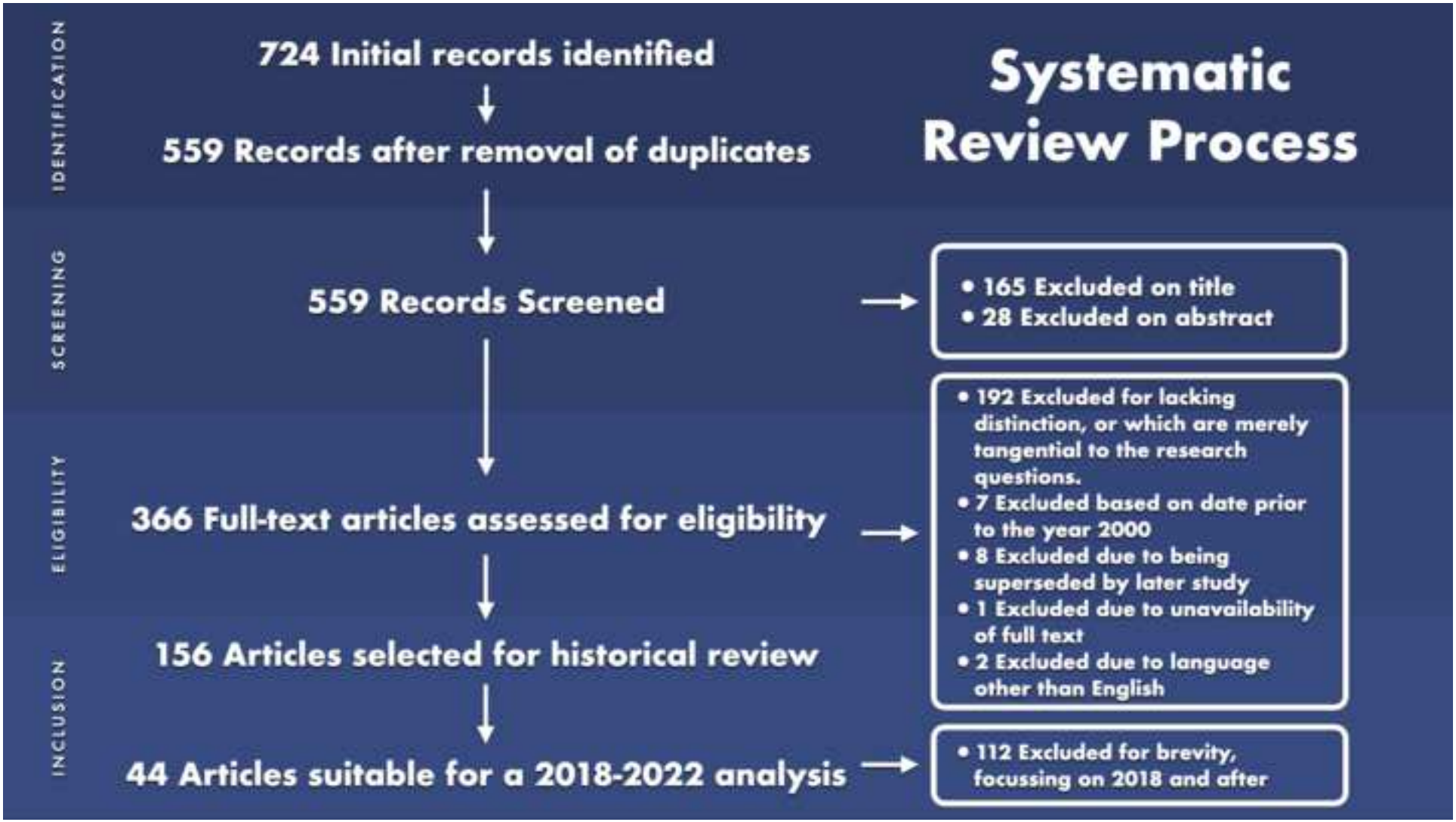

Figure 2

Click here to access/download;Figure;Figure002.png

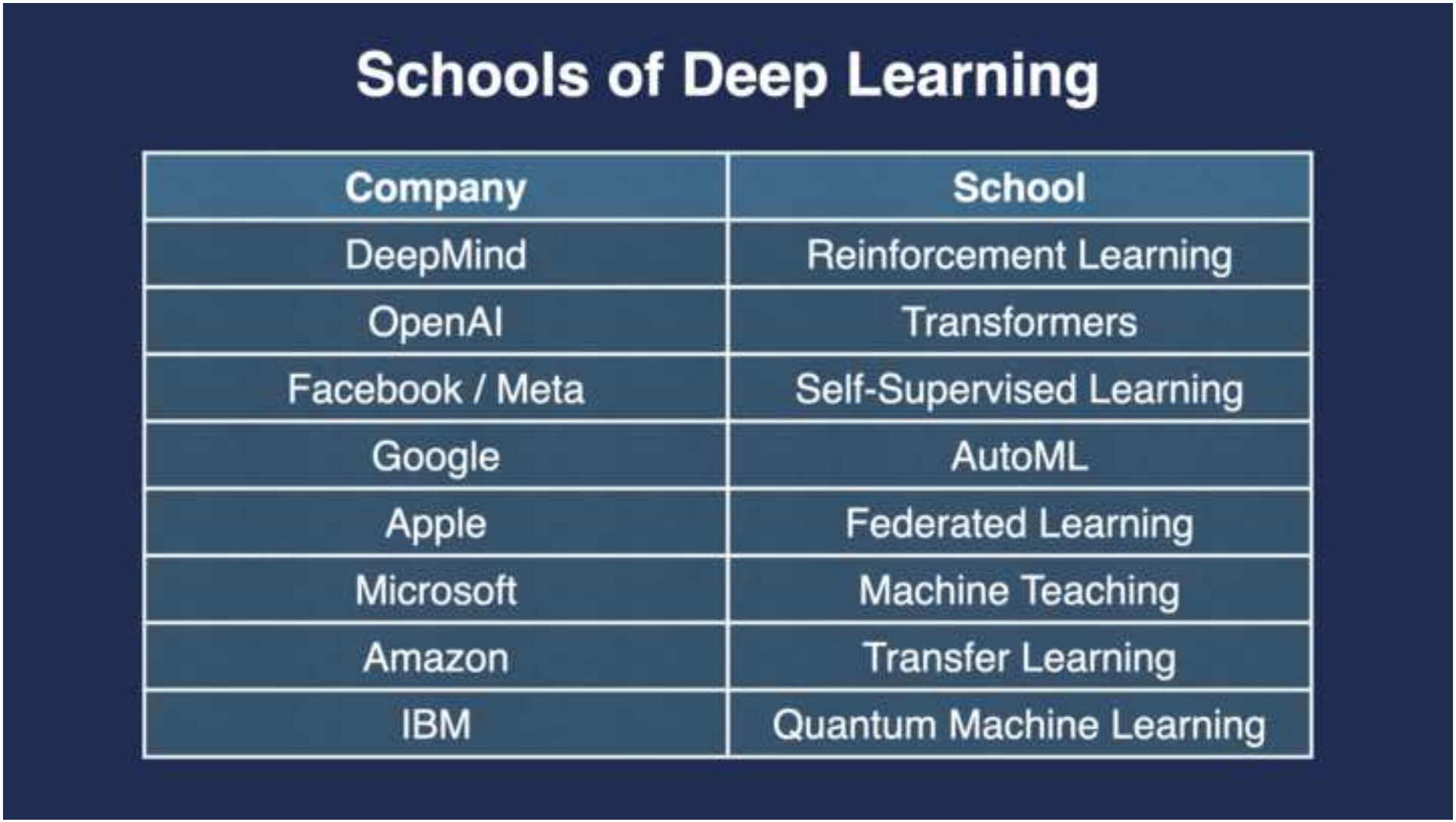

Figure 3

Click here to access/download;Figure;Figure003.png

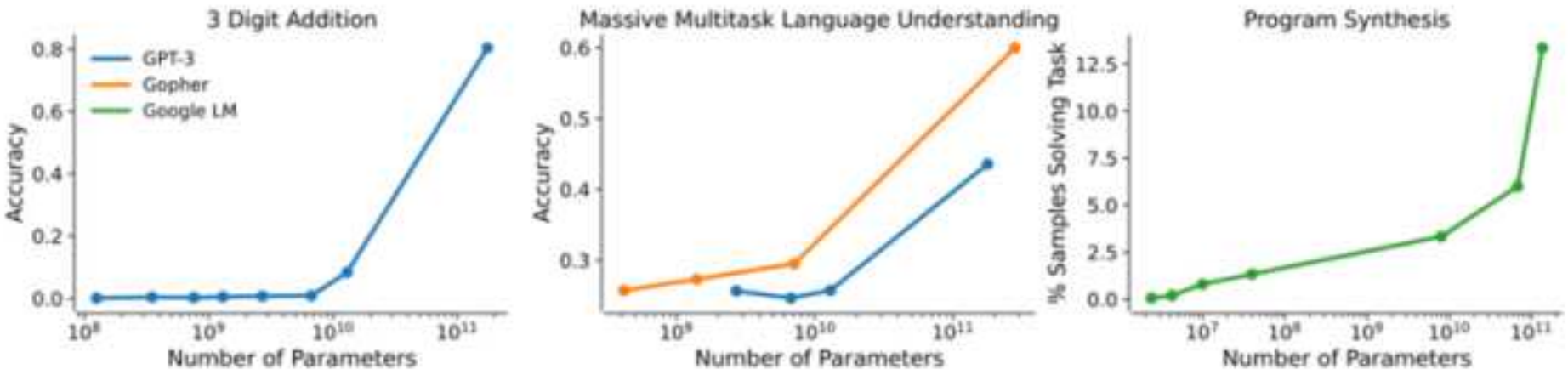

Figure 4

Click here to access/download;Figure;Figure004.png

Figure 5

Click here to access/download;Figure;Figure005.png

Click here to access/download;Table;Table1 Research Databases.png

Table1. Research Databases

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