A Survey on Deep Learning Techniques in Natural Language Processing and AI Alignment

www.surveyx.cn

Abstract

This survey paper delves into the transformative role of deep learning in Natural Language Processing (NLP) and AI alignment, emphasizing the profound impact of neural network architectures on modeling complex patterns in data. The integration of deep learning techniques has not only advanced NLP capabilities, such as automatic speech recognition and sentiment analysis, but also significantly contributed to AI alignment by ensuring systems adhere to human values. The paper highlights the challenges of interpretability and data requirements, advocating for probabilistic models to enhance robustness and transparency. Furthermore, the survey explores the integration of physiological data with large language models (LLMs), enhancing empathic AI capabilities and suggesting future research directions in hybrid architectures and interdisciplinary approaches. The importance of refining deep neural networks to align better with cognitive functions and exploring novel stimuli modalities is underscored, alongside the potential of the cerebellum's role in managing temporal feedback. By addressing these challenges, the survey aims to contribute to the development of more efficient, interpretable, and aligned AI systems, fostering advancements in deep learning, NLP, and AI alignment.

1 Introduction

1.1 Significance of Deep Learning in NLP and AI Alignment

Deep learning has revolutionized Natural Language Processing (NLP) by providing advanced tools for modeling complex linguistic patterns. Techniques such as Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory networks (LSTMs) have notably improved the processing and interpretation of sequential data, particularly in tasks like offensive language detection on social media, where traditional methods have struggled [1]. Additionally, deep learning's application in creative fields, exemplified by the generation of Urdu poetry, demonstrates its versatility [2].

Beyond NLP, deep learning plays a crucial role in AI alignment, ensuring that AI systems reflect human values and ethical standards. Its historical significance is underscored by its dominance in shaping scientific progress and epistemic culture within AI research [3]. However, the complexity of these models necessitates the incorporation of Explainable AI (XAI) techniques to promote transparency and trustworthiness, especially as AI systems are increasingly used in sensitive applications [4].

The challenges of deep learning, such as the requirement for extensive datasets and computational power, have led to the exploration of probabilistic models that integrate uncertainty, addressing limitations of traditional approaches [5]. This shift highlights the growing recognition of uncertainty in AI predictions, aligning with the broader objectives of AI alignment.

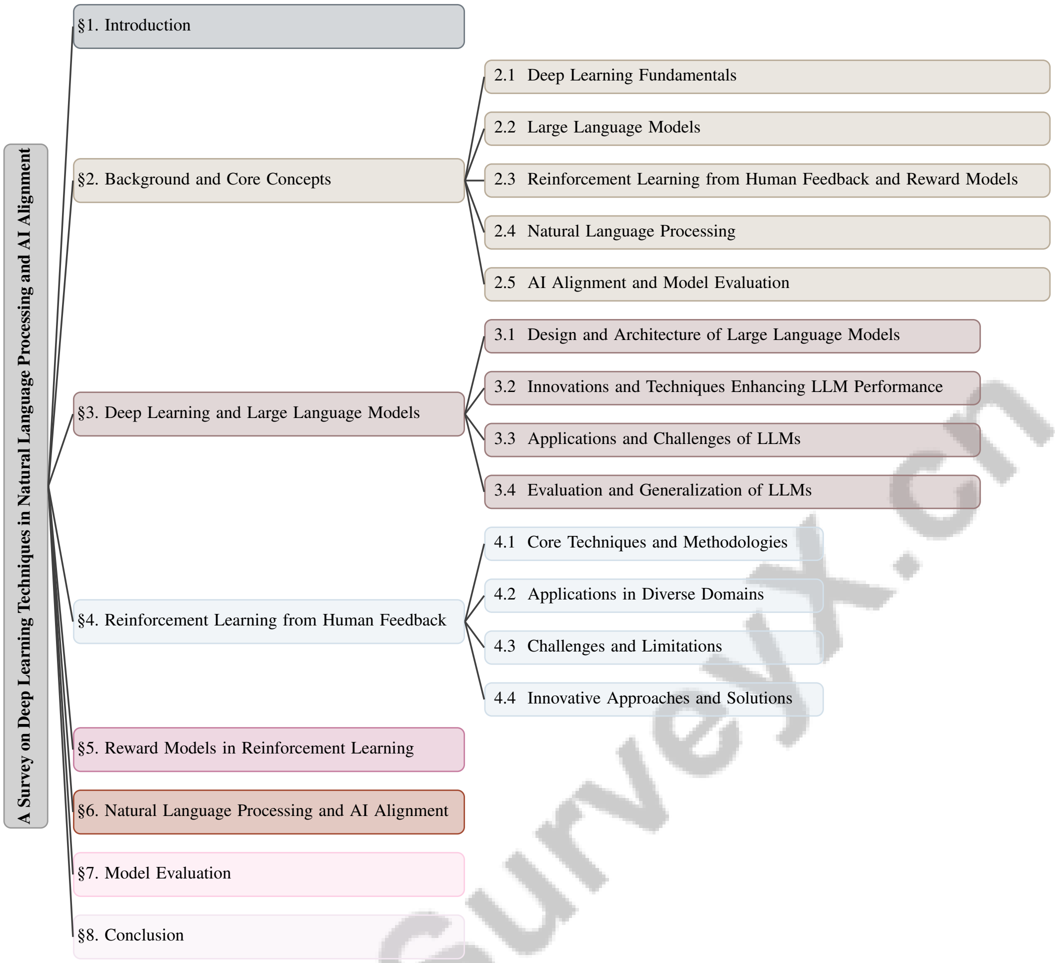

A Survey on Deep Learning Techniques in Natural Language Processing and AI Alignment
§1. Introduction
§2. Background and Core Concepts
2.1 Deep Learning Fundamentals
2.2 Large Language Models
2.3 Reinforcement Learning from Human Feedback and Reward Models
2.4 Natural Language Processing
2.5 AI Alignment and Model Evaluation
§3. Deep Learning and Large Language Models
3.1 Design and Architecture of Large Language Models
3.2 Innovations and Techniques Enhancing LLM Performance
3.3 Applications and Challenges of LLMs
3.4 Evaluation and Generalization of LLMs
§4. Reinforcement Learning from Human Feedback
4.1 Core Techniques and Methodologies
4.2 Applications in Diverse Domains
4.3 Challenges and Limitations
4.4 Innovative Approaches and Solutions
§5. Reward Models in Reinforcement Learning
§6. Natural Language Processing and AI Alignment
§7. Model Evaluation
§8. Conclusion

Figure 1: chapter structure

1.2 Motivation for the Survey

This survey aims to provide a thorough overview of deep learning methodologies, reflecting their rapid growth and significant impact across various fields [6]. Despite their transformative potential, deep learning's adoption in business analytics remains limited, where it could greatly enhance data-driven decision-making [7]. By examining the practical utility and effectiveness of deep learning models in real-world scenarios, this survey seeks to bridge this gap [8].

Additionally, the survey promotes engagement between the machine learning community and classical Japanese literature through the Kuzushiji datasets, fostering model comparisons and driving advancements in related research [9]. This initiative emphasizes the importance of benchmarking in AI's evolution, as demonstrated by the transition from Symbolic AI to the current era dominated by Deep Learning [3]. Benchmark datasets are essential for evaluating model performance and fostering innovation [10].

Another motivation is to address the critical issue of explainability in machine learning and deep learning models, particularly in software engineering, where transparency is vital for practical deployment [4]. This survey also tackles the selection of appropriate loss functions and performance metrics, which are crucial for effective model training and evaluation [11].

The survey further investigates modeling and processing uncertainty in deep learning applications, an essential area as AI systems are deployed in complex environments [5]. By exploring the relationship between deep neural networks and brain recordings, the survey aims to enhance understanding of both AI models and brain function, highlighting the interdisciplinary nature of AI research [12].

Ultimately, the survey aspires to advance the development of AI systems that are efficient, interpretable, and aligned with human values. Through a comprehensive analysis of challenges and advancements in deep learning, neural information retrieval, and neuro-symbolic AI, it aims to illuminate future research pathways that address existing limitations in interpretability and reasoning, thereby contributing significantly to AI research across various domains [13, 14, 15, 16].

1.3 Relevance to Current AI Research

This survey is pertinent to ongoing AI research, addressing critical challenges and opportunities presented by deep learning technologies across various domains. Deep learning's potential is particularly evident in natural language processing, where it has transformed the ability to model and interpret complex linguistic patterns [17]. However, significant challenges remain, including the black-box nature of AI models, which complicates understanding and diminishes trust in critical applications [4]. This survey explores these challenges, emphasizing the need for enhanced transparency and trustworthiness in AI systems.

Moreover, the survey is relevant to AI alignment research, especially in predicting performance in configurable software systems, which requires accurate modeling capabilities [18]. It also addresses inefficiencies in training due to the indiscriminate inclusion of features, underscoring the need for tailored approaches in AI model development [19].

The evolution of AI research from a diverse field to a monoculture dominated by deep learning raises concerns about scientific progress and autonomy [3]. This survey advocates for a balanced approach that incorporates diverse methodologies and perspectives.

Furthermore, integrating brain data into AI model training presents a promising avenue for enhancing cognitive process understanding and improving AI performance [12]. This interdisciplinary approach underscores the survey's relevance in advancing AI research by bridging artificial and natural intelligence.

1.4 Structure of the Survey

This survey is structured to thoroughly explore deep learning techniques in the context of Natural Language Processing (NLP) and AI alignment. The organization reflects the evolution of AI, categorized into three stages: the Era of Symbolic AI, the Benchmarking Era, and the Deep Learning Era, each characterized by distinct organizational and epistemic features [3]. The survey begins with an introduction outlining the significance of deep learning in NLP and AI alignment, followed by discussions on the motivation for the survey and its relevance to current AI research.

The second section provides foundational knowledge on deep learning, large language models, reinforcement learning from human feedback, reward models, NLP, AI alignment, and model evaluation. Subsequent sections focus on specific areas such as the design, architecture, and innovations in large language models, as well as the role of reinforcement learning from human feedback.

The survey also examines the conceptual framework and challenges of reward models in reinforcement learning, followed by an exploration of the intersection between NLP and AI alignment. The penultimate section discusses model evaluation, highlighting its importance in AI alignment and NLP.

Finally, the conclusion summarizes key findings and discusses implications for future research and developments in deep learning, NLP, and AI alignment. This structured approach facilitates a comprehensive analysis of current advancements and emerging trends in fields such as neural information retrieval, financial analysis, information security, and legal document classification, highlighting both challenges and promising future research directions [13, 20, 21, 22].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Deep Learning Fundamentals

Deep learning, a pivotal component of artificial intelligence, employs multilayered neural networks to model complex patterns in diverse datasets. This approach mimics human cognition by transforming raw data into increasingly abstract representations, significantly advancing fields like natural language

AI-generated, for reference only.

processing and image recognition [17]. Key architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have notably improved tasks like offensive language detection and speech-to-text conversion.

The backpropagation algorithm is central to parameter optimization, though challenges like the vanishing gradient problem can hinder deep network training. Innovations such as the Fast-Forward Network (FFNet) architecture offer parallel data paths to enhance gradient flow and training efficiency [23]. Mini-batch stochastic gradient descent (SGD) balances computational efficiency and convergence speed [24].

Generalization in deep learning is crucial, as traditional uniform convergence bounds often fail to explain overparameterized models' capabilities [25]. Insights into deep networks' geometry, particularly through gradient flow dynamics on invariant manifolds, improve understanding of generalization behavior [26]. Techniques like meta-learning and data augmentation boost performance in resource-limited scenarios [27].

Interpretability remains a challenge due to deep learning models' opacity. Frameworks categorizing interpretability methods into post-hoc and intrinsic approaches enhance understanding [28]. This is vital in applications like software vulnerability detection, where evaluating learning-based models' effectiveness across vulnerability types is crucial [29].

Energy efficiency in deep learning is increasingly important, with frameworks analyzing energy consumption across NLP models, emphasizing efficiency in algorithm design and hardware utilization [30]. Analog deep learning methods offer advantages in energy efficiency and speed, suitable for rapid processing and low power consumption applications [31].

Methodological rigor in deep learning experiments is essential, as under-specification and flaws can compromise findings, especially in software engineering [32]. Integrating domain-specific knowledge, such as tailoring molecular representations in chemical tasks using language models, exemplifies deep learning techniques' versatility [19].

Incorporating scientific theory principles could enhance deep learning models' extrapolation capabilities. Bayesian neural network interpretations are gaining attention for properties like uncertainty estimation, model robustness, and regularization [33]. Deep belief networks (DBNs) and RNNs continue to evolve, managing high-dimensional problems and capturing memory dependencies.

Innovations in training algorithms, hyperparameter optimization, and scalable architectures are key to overcoming technical challenges and advancing applications across domains. The ISBE method normalizes raw scores and propagates errors, simplifying training without Cross-Entropy [34]. The OntoEnricher method illustrates supervised sequential deep learning models' application in enhancing information security ontologies [22]. Adversarial robustness advancements are critical for addressing safety and reliability in machine learning applications [35].

Deep learning's emulation of brain processes, particularly the cerebellum's role in managing temporal feedback, highlights neural networks' potential in tackling complex learning problems [36]. Exploring both biological and artificial neural networks is crucial for enhancing deep learning capabilities and applications [12].

2.2 Large Language Models

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) by processing and generating human-like text with extensive datasets and advanced architectures. Transformer-based models exhibit exceptional proficiency in language understanding and generation, advancing tasks like text classification, sentiment analysis, and machine translation [37]. LLMs outperform traditional methods by capturing intricate text patterns [38].

LLMs' evolution is linked to neural network architecture advancements, enhancing performance in sequential tasks such as speech recognition and language modeling [38]. Empirical comparisons underscore the importance of selecting appropriate methods based on dataset availability [39]. Domain-specific pretrained language models meet specialized fields' demands, offering improvements through tailored datasets and architectures [40].

A core strength of LLMs is zero-shot learning, enabling models to generalize to new tasks with minimal data. This adaptability is crucial for advancing NLP applications across diverse tasks [3].

Integrating symbolic reasoning with neural networks, as seen in the Neural-Symbolic Stack Machine (NeSS), enhances compositional generalization capabilities [3].

LLMs' significance extends beyond traditional NLP tasks into areas like recommender systems, enhancing performance through deep learning techniques [37]. They also improve security vulnerability detection by capturing vulnerabilities' full context, addressing existing methods' challenges [41].

Inference optimization is critical for LLMs, focusing on enhancing efficiency and speed during deployment. Techniques like knowledge distillation transfer knowledge from multiple models to a single, compact model, optimizing LLMs by reducing complexity while maintaining accuracy [20]. These advancements underscore LLMs' pivotal role in propelling AI research and applications across domains, highlighting their future significance in NLP and AI.

2.3 Reinforcement Learning from Human Feedback and Reward Models

Reinforcement learning from human feedback integrates human evaluators into AI models' learning loop, refining behaviors through iterative interactions with dynamic environments. This approach leverages reinforcement learning's adaptability, enriched by human insights, for tasks requiring strategic planning and decision-making [42]. Deep learning architectures enhance large datasets' processing and intricate patterns' identification, improving performance in unpredictable settings [43].

A significant challenge in reinforcement learning is maintaining robustness in fluctuating conditions, especially when human coders' behavior annotation can lead to inefficiencies and inaccuracies [43]. Innovative frameworks like the Generic Reinforced Explainable Framework with Knowledge Graph (REKS) combine reinforcement learning with knowledge graphs to enhance session-based recommendation models' explainability [42].

Reward models guide AI behavior by assigning value to actions and outcomes, shaping the learning process with success and failure criteria. Hybrid architectures merging deep reinforcement learning with symbolic reasoning exemplify multi-faceted AI approaches' potential to enhance problem-solving capabilities [44].

Optimizing learning rates in reinforcement learning, particularly in meta-learning contexts, highlights adaptive strategies' importance that adjusts to varying conditions [45]. Task interference poses a challenge when a single neural network learns multiple unrelated tasks, raising questions about a unified model managing multiple objectives without performance compromise [46]. Innovative approaches are needed to compartmentalize learning processes and minimize cross-task interference, ensuring effective balance across goals.

Reinforcement learning from human feedback, combined with strategic reward model use, advances AI systems capable of acquiring complex behaviors. Integrating human insights and leveraging sophisticated architectures promise enhanced adaptability and efficacy across applications. Managing computational demands and ensuring robustness in variable environments remain critical research and development areas. A primary deep learning challenge is substantial data and computational resource needs, which can be a barrier for researchers and practitioners seeking effective model implementation [6].

2.4 Natural Language Processing

Natural Language Processing (NLP) focuses on enabling machines to understand, interpret, and generate human language. Deep learning techniques have significantly advanced NLP, facilitating nuanced linguistic structure modeling and comprehension. Models like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) showcase creative potential in generating Urdu poetry [2]. The Structured Deep Neural Network (DNN) enhances recognition accuracy by capturing utterances' global structure [47].

Deep learning has enabled sophisticated NLP applications, including text classification, language modeling, machine translation, and sentiment analysis. These tasks benefit from carefully selected loss functions and metrics, crucial for optimizing performance and aligning with human values [11]. CNNs improve sentiment analysis accuracy, particularly in resource-constrained environments with limited labeled datasets.

A significant NLP challenge involves processing low-resource languages, often lacking substantial research and monolingual datasets. Addressing these challenges is critical for enhancing language model quality across linguistic contexts. Tokenization plays a vital role, as demonstrated by research on languages like Assamese, where innovative approaches improve performance [48].

Integrating NLP with AI alignment is essential for developing systems adhering to human values and intentions. The OntoEnricher method employs NLP to extract vulnerabilities, threats, and controls from unstructured text, contributing to AI systems capable of processing and analyzing multimodal data [22]. Ensuring interpretability in NLP models is crucial, particularly where understanding predictions is vital for maintaining AI alignment with human values.

NLP continues to evolve, driven by deep learning innovations and AI alignment principles integration. Addressing challenges like low-resource language processing and enhancing interpretability through deep learning and symbolic reasoning combination, NLP systems stand to make substantial strides in accurately understanding and processing human language. Neuro-symbolic AI techniques adoption can lead to improved reasoning capabilities, out-of-distribution generalization, and learning from limited data, fostering robust and versatile AI applications across domains [49, 50, 14].

2.5 AI Alignment and Model Evaluation

AI alignment ensures AI systems' objectives and behaviors align with human values and intentions, a challenge compounded by deep learning models' complexity and opacity [4]. These "black box" models necessitate robust evaluation frameworks to assess efficacy, reliability, and generalization capabilities [6]. Traditional measures like parameter counts are often inadequate for capturing deep networks' generalization behavior, highlighting the need for alternative theoretical approaches [8].

Model evaluation is crucial in AI alignment, providing metrics and methodologies to assess performance and generalization abilities. The shortcomings of uniform convergence in explaining overparameterized models' intricacies highlight the necessity for innovative evaluation techniques [4]. This is particularly important where integrating physiological data, such as the Physiology-Driven Empathic LLM (EmLLM), enhances models' empathetic responses to users' psychological states [41].

Understanding the generalization gap, reflecting the difference between training and unseen data performance, is vital for evaluating a model's ability to generalize beyond its training environment [36]. Effective evaluation requires metrics providing insights into predictive accuracy and computational efficiency, ensuring models are effective and efficient [12]. Metrics selection is guided by evaluating models' generalization of agent behavior understanding to new scenarios, reflecting human-like reasoning [6].

Deep learning models' complexity necessitates evaluation metrics accounting for model complexity and representation quality, as existing methods struggle to capture intricate activation distributions and redundancy within large language models (LLMs) [4]. Understanding local and global information mapping relations is crucial for defining learning and improving evaluation frameworks [41].

Challenges in model evaluation include existing methods' limitations for estimating mutual information, often suffering from intractability and high variance, indicating the need for innovative techniques accurately capturing dependencies and interactions within AI models [36]. Integrating causal reasoning into evaluation is essential for developing explanations reflecting genuine cause-and-effect relationships rather than spurious correlations, aligning with broader AI alignment goals [4].

Neural network width's impact on adversarial training and model robustness is another important consideration, influencing natural accuracy and perturbation stability, crucial for ensuring reliable AI systems [6]. High dimensionality of configuration spaces and performance data sparsity are essential considerations in AI alignment and evaluation, affecting models' adaptability to new domains and scenarios [12].

Energy consumption and computational efficiency are critical in model evaluation, highlighting differences between architectures and resource requirements [41]. Adapting models trained on multiple labeled source domains to an unlabeled or sparsely labeled target domain underscores managing domain shifts effectively [4]. Existing benchmarks often rely on rule-based systems that do not scale well and are limited by expert knowledge, reducing adaptability to new domains [6].

AI-generated, for reference only.

AI alignment and model evaluation are interrelated fields requiring continuous innovation and refinement to address challenges like interpretability, real-time data integration, and deploying explainable AI models in critical applications. As neural networks and large language models advance, ongoing research is essential to enhance AI systems' effectiveness and transparency across domains, from information retrieval to financial analysis and automated grading [51, 20, 13, 52, 4]. Developing robust evaluation frameworks and metrics ensures AI systems are effective and aligned with human values, paving the way for trustworthy and reliable AI applications.

3 Deep Learning and Large Language Models

To fully grasp the transformative impact of large language models (LLMs) on artificial intelligence, it is essential to explore their design and architecture. This examination reveals the mechanisms that enable LLMs to process and generate human-like text with remarkable proficiency, thereby advancing natural language processing capabilities. As illustrated in Figure 2, the hierarchical structure of Deep Learning and LLMs is depicted, highlighting key areas such as design and architecture, innovations and techniques that enhance performance, applications and challenges, and evaluation and generalization. Each category is further broken down into specific components, showcasing the complexity and multifaceted nature of LLMs, which underscores their significance in the field of artificial intelligence.

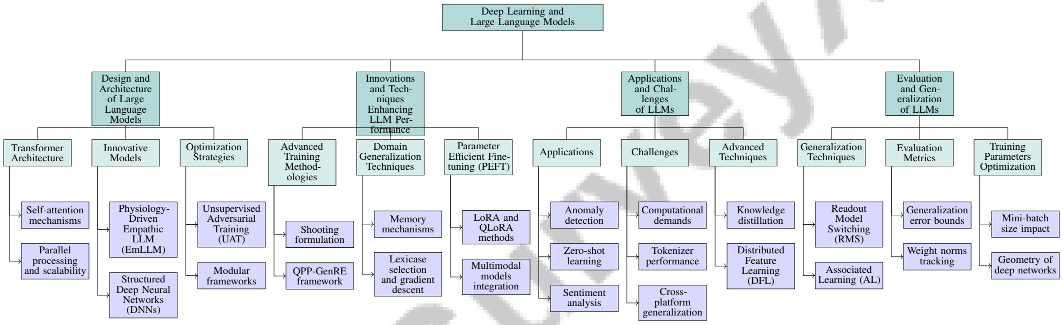

Deep Learning and Large Language Models
Design and Architecture of Large Language Models
Innovations and Techniques Enhancing LLM Performance
Applications and Challenges of LLMs
Evaluation and Generalization of LLMs
Transformer Architecture
Innovative Models
Optimization Strategies
Advanced Training Methodologies
Domain Generalization Techniques
Parameter Efficient Fine-tuning (PEFT)
Applications
Challenges
Advanced Techniques
Generalization Techniques
Evaluation Metrics
Training Parameters Optimization
Self-attention mechanisms
Parallel processing and scalability
Physiology-Driven Empathic LLM (EmLLM)
Structured Deep Neural Networks (DNNs)
Unsupervised Adversarial Training (UAT)
Modular frameworks
Shooting formulation
QPP-GenRE framework
Memory mechanisms
Lexicase selection and gradient descent
LoRA and QLoRA methods
Multimodal models integration
Anomaly detection
Zero-shot learning
Sentiment analysis
Computational demands
Tokenizer performance
Cross-platform generalization
Knowledge distillation
Distributed Feature Learning (DFL)
Readout Model Switching (RMS)
Associated Learning (AL)
Generalization error bounds
Weight norms tracking
Mini-batch size impact
Geometry of deep networks

Figure 2: This figure illustrates the hierarchical structure of Deep Learning and Large Language Models, highlighting key areas such as design and architecture, innovations and techniques enhancing performance, applications and challenges, and evaluation and generalization. Each category is further broken down into specific components, showcasing the complexity and multifaceted nature of LLMs.

3.1 Design and Architecture of Large Language Models

LLMs are predominantly based on the transformer architecture, utilizing self-attention mechanisms to manage dependencies across extensive text sequences efficiently. This design is crucial for navigating complex linguistic structures, facilitating parallel processing and scalability [41]. The self-attention mechanism is adept at capturing intricate language patterns, supporting advanced natural language processing tasks.

A notable innovation is the Physiology-Driven Empathic LLM (EmLLM), which integrates deep learning models to predict psychological states from physiological data, enhancing empathic interactions and improving human-computer interaction [41]. Advanced neural network structures, like Structured Deep Neural Networks (DNNs), are utilized for applications such as phoneme recognition, capturing both sequential and contextual information [47].

Optimization is key in LLM development. Unsupervised Adversarial Training (UAT) uses unlabeled data to enhance classifier robustness against adversarial attacks, improving model reliability [53]. Moreover, modular frameworks decompose neural networks into components with local objectives, allowing efficient updates and integration of symbolic reasoning with neural architectures [41].

The architecture of LLMs, characterized by transformer-based self-attention mechanisms, advanced neural networks, and innovative optimization strategies, enhances performance across tasks like legal document classification and information retrieval [13, 21]. These elements collectively drive advancements in natural language processing and AI applications.

As illustrated in Figure 3, this figure illustrates the key components of Large Language Model (LLM) design and architecture, highlighting the use of transformer-based self-attention mechanisms, innovative models like EmLLM and Structured DNN, and optimization strategies including Unsupervised Adversarial Training and modular frameworks. Understanding these intricacies is vital for advancing artificial intelligence applications. The examples highlight critical aspects such as anomaly detection, data preprocessing, and comparative analysis of machine learning quantifiers, showcasing the multifaceted nature of LLM design and architecture [54, 55, 14].

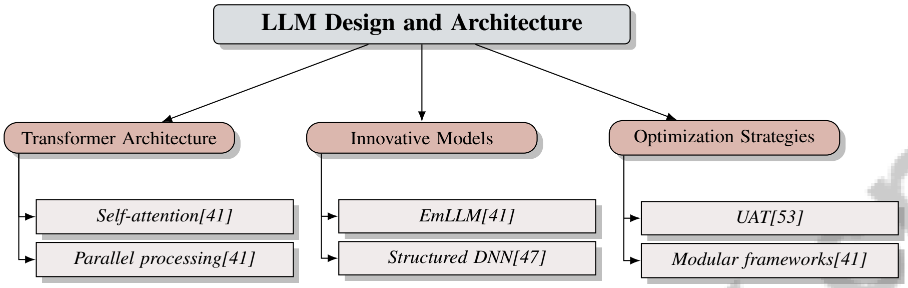


Figure 3: This figure illustrates the key components of Large Language Model (LLM) design and architecture, highlighting the use of transformer-based self-attention mechanisms, innovative models like EmLLM and Structured DNN, and optimization strategies including Unsupervised Adversarial Training and modular frameworks.

3.2 Innovations and Techniques Enhancing LLM Performance

Recent advancements in LLMs are driven by innovative techniques that enhance performance across applications. The shooting formulation optimizes deep neural networks (DNNs) by parameterizing them through initial conditions, simplifying optimization [56]. This approach facilitates efficient convergence and improved performance.

Advanced training methodologies have been crucial for LLM performance. Frameworks like QPP-GenRE use LLMs to generate relevance judgments, predicting information retrieval metrics without separate models for each measure, thus streamlining evaluation [37, 57].

Domain generalization techniques, incorporating memory mechanisms for text classification, significantly enhance generalization across unseen domains through a multi-source meta-learning framework [22, 58]. Lexicase selection combined with gradient descent improves generalization in deep learning models.

Parameter Efficient Fine-tuning (PEFT) methods, like LoRA and QLoRA, reduce computational demands while enhancing LLM capabilities in tasks such as automatic grading and legal document classification [52, 21, 55]. Fine-tuned LLMs achieve remarkable accuracy, demonstrating the importance of refining training strategies.

In multimodal models, integrating text and image data into a unified latent space enhances LLMs' discriminative capabilities, improving recommendations and addressing challenges in understanding complex data [55, 21, 59, 37]. This integration allows LLMs to effectively process multimodal information, expanding applicability in fields requiring textual and visual data synthesis.

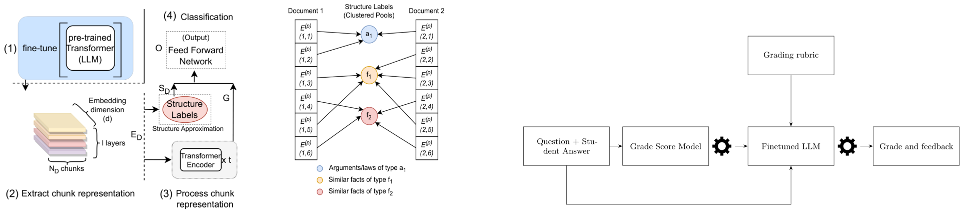
These innovations and techniques collectively contribute to LLMs' evolution, enhancing applications in AI, including personalized recommendation systems, automated grading, and information retrieval. Ongoing progress is propelled by innovative approaches that leverage LLMs to process diverse data types, improving accuracy and relevance across domains [52, 21, 13, 37].

As shown in Figure 4, recent innovations in fine-tuning pre-trained Transformers and applying LLMs in educational contexts significantly enhance performance. These examples underscore LLMs' transformative potential in offering practical solutions across diverse fields [21, 52].

3.3 Applications and Challenges of LLMs

LLMs are pivotal in various applications, processing and generating human-like text with accuracy and contextual relevance. The LLMAD framework exemplifies precise anomaly detections with

AI-generated, for reference only.



(a) Fine-tuning a pre-trained Transformer for structure labeling[21]

(b) A flowchart illustrating grading student answers using a grading rubric and a fine-tuned language model (LLM).[52]

Figure 4: Examples of Innovations and Techniques Enhancing LLM Performance

comprehensive explanations, fostering user trust and operational efficiency [54]. Such capabilities are crucial in domains requiring robust anomaly detection and interpretability.

In NLP, LLMs demonstrate versatility across tasks. Direct Feedback Alignment (DFA) validates transformer architectures' scalability in real-world scenarios [60]. LLMs excel in zero-shot learning, surpassing traditional algorithms, with models like GPT-4 achieving superior accuracy [55]. This generalization ability enables adaptation to new challenges with minimal data.

Innovative training methodologies enhance sentiment analysis, with a two-stage approach improving prediction across domains [61]. Despite successes, LLMs face challenges like computational demands in training and deployment. Techniques like knowledge distillation mitigate these challenges, as shown by reduced error rates in ASR systems [62]. Tokenizer performance, crucial for low-resource languages, emphasizes optimized preprocessing techniques [48].

Conversational systems highlight LLMs' role in enhancing user interactions. Data-driven turn-taking improves conversation fluidity, as shown by benchmarks refining multiparty models [63]. In cross-platform hate speech detection, LLMs disentangle platform-specific features, improving generalization and detection accuracy [64].

In business analytics, LLMs outperform traditional models in prediction accuracy and operational effectiveness, handling complex data-driven tasks [7]. Challenges remain in optimizing models for specific applications to ensure consistent performance.

Distributed Feature Learning (DFL) combines neural networks and symbolic reasoning, enhancing performance on complex tasks [65]. This synergy advances LLM capabilities beyond traditional NLP applications, tackling more intricate challenges.

While LLMs revolutionize natural language processing, ongoing research is essential to address computational, interpretability, and domain-specific challenges. Enhancing architectures and training methodologies broadens LLM applications, improving effectiveness in automatic grading, recommendation systems, and legal document classification. Techniques like PEFT and hierarchical frameworks optimize performance while reducing costs, achieving high accuracy in tasks like essay scoring and legal judgment prediction [52, 21, 20, 37].

3.4 Evaluation and Generalization of LLMs

Evaluating and generalizing LLMs are crucial for assessing performance across diverse tasks and datasets. Table 2 provides an in-depth examination of representative benchmarks used in the evaluation and generalization of large language models, highlighting the diversity of tasks, dataset sizes, and performance metrics considered in current research. LLMs' generalization capacity is often measured by performance on unseen data, essential for robust language understanding [70]. Techniques like Readout Model Switching (RMS) enhance adaptability in evaluations across tasks and data sizes.

Generalization in deep learning models, including LLMs, is influenced by alignment between training data characteristics and the learned hypothesis. Even with data memorization, models can achieve good generalization if elements are appropriately aligned [71]. Exploring complexity measures, such as sharpness and norms, provides insights into neural networks' generalization behavior [72]. The

AI-generated, for reference only.

Benchmark	Size	Domain	Task Format	Metric
DLBENCH[66]	760,000	Image Classification	Neural Network Training	Accuracy, Elapsed Time
GGP[67]	13,500	Machine Learning	Generalization Gap Prediction	R2, L1 Loss
ZSL-LLM[55]	4,000	Sentiment Analysis	Text Classification	F1 Score, Accuracy
ChatGPT-TextDistinction[68]	10,000	Natural Language Processing	Text Classification	Accuracy, MCC
AUTONLU[49]	1,000,000	Natural Language Processing	Sequence Labeling	F1-score
MPTT[63]	120,787	Conversational Systems	Turn-Taking Prediction	Accuracy
XAI-BM[69]	80	Image Classification	Attribution	Completeness, Compactness
CCID[43]	13,450	Behavior Analysis	Behavior Classification	UAR

Table 1: Table ef presents a comprehensive overview of various benchmarks utilized in the evaluation and generalization of large language models (LLMs). It details the size, domain, task format, and performance metrics of each benchmark, illustrating the diversity and complexity of tasks that LLMs are assessed on. This table serves as a critical reference for understanding the breadth of datasets and evaluation criteria employed in current LLM research.

Benchmark	Size	Domain	Task Format	Metric
DLBENCH[66]	760,000	Image Classification	Neural Network Training	Accuracy, Elapsed Time
GGP[67]	13,500	Machine Learning	Generalization Gap Prediction	R2, L1 Loss
ZSL-LLM[55]	4,000	Sentiment Analysis	Text Classification	F1 Score, Accuracy
ChatGPT-TextDistinction[68]	10,000	Natural Language Processing	Text Classification	Accuracy, MCC
AUTONLU[49]	1,000,000	Natural Language Processing	Sequence Labeling	F1-score
MPTT[63]	120,787	Conversational Systems	Turn-Taking Prediction	Accuracy
XAI-BM[69]	80	Image Classification	Attribution	Completeness, Compactness
CCID[43]	13,450	Behavior Analysis	Behavior Classification	UAR

Table 2: Table ef presents a comprehensive overview of various benchmarks utilized in the evaluation and generalization of large language models (LLMs). It details the size, domain, task format, and performance metrics of each benchmark, illustrating the diversity and complexity of tasks that LLMs are assessed on. This table serves as a critical reference for understanding the breadth of datasets and evaluation criteria employed in current LLM research.

Activation Variance-Sparsity Score (AVSS) quantifies each layer's contribution, optimizing model evaluation [73].

Experiments show Associated Learning (AL) achieves test accuracies comparable to end-to-end backpropagation training, despite most components not directly receiving residual signals [74]. This indicates alternative training methodologies can achieve robust generalization, reducing computational complexity while maintaining performance.

Generalization error is a key metric, with recent studies deriving tight upper bounds for approximation and generalization errors [75]. These bounds offer a theoretical framework for understanding model performance without scale invariance assumptions.

Evaluation methods tracking weight norms and generalization bounds as dataset size varies provide insights into empirical test errors and theoretical predictions [25]. This approach underscores the need for robust evaluation frameworks accounting for model complexity and dataset characteristics.

The geometry of deep networks influences generalization capabilities. Analyzing convergence rates and gradient flows provides a deeper understanding of mechanisms driving performance [26]. Mini-batch size impacts training outcomes, with smaller sizes leading to lower variance in stochastic gradient estimators and improved results [24]. Optimizing training parameters enhances model generalization.

Evaluating and generalizing LLMs involve a multifaceted approach considering various metrics and methodologies. Refining evaluation techniques and understanding factors influencing generalization help develop robust, reliable language models for diverse applications. Neural networks can be interpreted as forms of Super Space, where activation functions map linear spaces into infinite dimensions, providing a theoretical basis for understanding LLMs' expansive capacity [76].

4 Reinforcement Learning from Human Feedback

4.1 Core Techniques and Methodologies

Reinforcement learning from human feedback (RLHF) integrates human evaluators with AI systems to refine agent behaviors through iterative feedback, enhancing strategic decision-making capabilities. A key challenge in RLHF is managing high-dimensional function estimation, effectively addressed by deep learning models adept at processing complex data interactions [41]. The Physiology-Driven Empathic LLM (EmLLM) exemplifies RLHF's potential by predicting psychological states and improving human-computer interaction through physiological data integration [41].

Direct Feedback Alignment (DFA) is a prominent technique in RLHF, replacing traditional backpropagation with direct global error signal projection via random feedback matrices. This method aligns with RLHF principles by facilitating efficient error correction and model training without the computational burden of conventional backpropagation [41]. DFA is particularly advantageous in multitask scenarios, enhancing detection and explanation capabilities through dialogue-based data processing and multitask fine-tuning [47].

Curriculum learning is crucial in RLHF, structuring task presentations to improve learning outcomes by distinguishing relevant features from irrelevant ones. This structured approach is vital for developing AI systems that adapt to dynamic environments [56]. The shooting formulation optimizes neural networks by parameterizing them through Hamiltonian particle ensembles' initial conditions, enhancing model adaptability [56].

Advanced techniques like Unsupervised Adversarial Training (UAT) combine supervised and unsupervised strategies to bolster adversarial robustness, enabling models to learn from both labeled and unlabeled data [53]. This dual approach is crucial for enhancing AI systems' resilience against adversarial attacks, a critical consideration in RLHF.

Recursive autoencoders for feature extraction automate learning processes, significantly improving predictive accuracy over traditional methods. This automation is essential for efficiently handling complex data interactions inherent in RLHF, allowing models to distill meaningful patterns from human feedback [41]. Additionally, optimizing mini-batch sizes affects variance and convergence rates during training, underscoring the importance of refining training processes for enhanced model performance [41].

An innovative causality-guided disentanglement method enhances robustness and generalization capabilities by distinguishing invariant from platform-specific features, ensuring learning processes are not confounded by irrelevant variations [41]. RLHF focuses on merging human insights with sophisticated learning architectures, enhancing interpretability and performance of intelligent agents. This integration addresses deep reinforcement learning's "black box" challenges by leveraging human knowledge to improve learning efficiency while adhering to human-centric principles such as transparency, fairness, and accountability [77, 78, 79, 80]. By employing symbolic learning, direct feedback mechanisms, and curriculum learning strategies, RLHF enhances AI systems' adaptability and effectiveness across diverse applications.

As illustrated in Figure 5, this figure categorizes core techniques and methodologies in RLHF into reinforcement learning, learning strategies, and optimization techniques. The methodologies presented highlight the integration of human insights, advanced learning strategies, and optimization techniques to enhance AI systems' adaptability and performance. The examples further emphasize the nuanced integration of human feedback in complex evaluations, the role of multilayer perceptrons in refining outputs, and the systematic categorization and utilization of human feedback, underscoring RLHF's potential to refine AI systems through structured human input [16, 13, 21].

4.2 Applications in Diverse Domains

RLHF has broad applications across various domains, integrating human insights with adaptive learning strategies to enhance AI performance. In robotics, RLHF enhances object detection, navigation, and control tasks, with deep learning models effectively processing complex sensory data and adapting to dynamic environments [77]. These advancements highlight RLHF's potential to improve robotic autonomy and efficiency in unstructured settings.

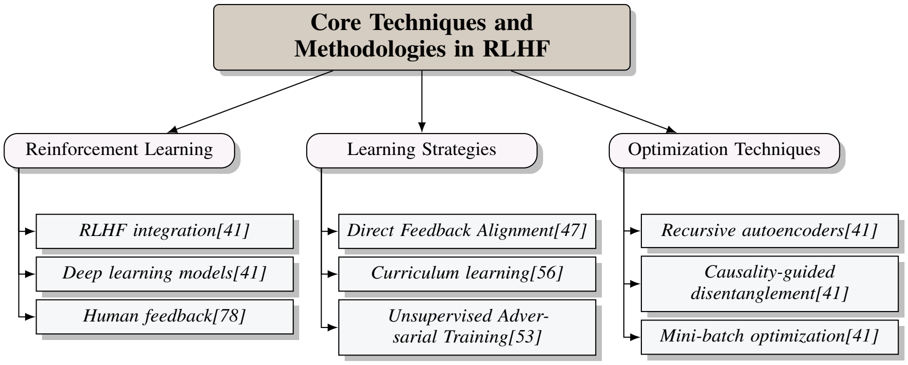

Core Techniques and Methodologies in RLHF
Reinforcement Learning
Learning Strategies
Optimization Techniques
RLHF integration[41]
Deep learning models[41]
Human feedback[78]
Direct Feedback Alignment[47]
Curriculum learning[56]
Unsupervised Adversarial Training[53]
Recursive autoencoders[41]
Causality-guided disentanglement[41]
Mini-batch optimization[41]

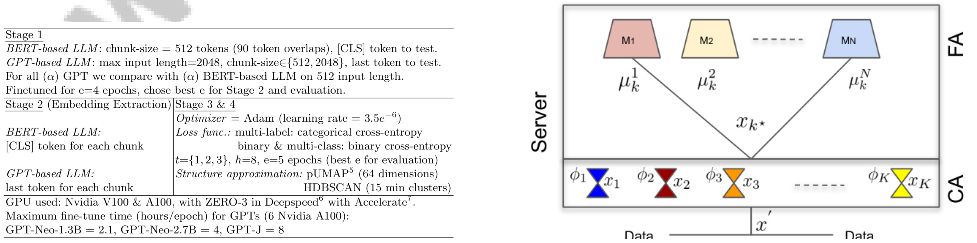
Figure 5: This figure illustrates core techniques and methodologies in Reinforcement Learning from Human Feedback (RLHF), categorized into reinforcement learning, learning strategies, and optimization techniques. These methodologies highlight the integration of human insights, advanced learning strategies, and optimization techniques to enhance AI systems' adaptability and performance.

In business analytics, RLHF optimizes decision-making processes and resource allocation, showcasing its transformative potential through data-driven insights and strategic planning [7]. By integrating RLHF techniques, businesses can develop adaptive analytical models aligned with organizational goals and market dynamics.

Healthcare is a pivotal sector where RLHF can drive advancements in personalized medicine and patient care through multimodal data integration from electronic health records (EHRs) and clinical notes. Large language models trained on extensive EHR data improve chronic disease prediction accuracy, facilitating tailored treatment plans and enhancing communication between healthcare providers and patients [20, 21, 81, 22, 43]. Incorporating human feedback into learning algorithms allows healthcare systems to customize treatment plans, enhancing outcomes and patient satisfaction.

In intelligent tutoring systems, RLHF enhances educational experiences by adapting instructional strategies based on student feedback. This personalization fosters increased engagement and academic achievement by creating tailored learning paths that accommodate individual styles and paces. Techniques like curriculum learning strategically organize educational content, aligning with students' cognitive processes while incorporating insights from studies on large language models (LLMs) and multimodal data integration [82, 52, 21, 22, 37].

The diverse applications of RLHF across various domains demonstrate its versatility and potential for innovation. By leveraging human insights and advanced methodologies, AI systems enhance capabilities in robotics, business analytics, healthcare, and education. The integration of large language models and frameworks like retrieval-augmented generation further enhances these systems by providing contextual understanding and real-time data integration, addressing challenges such as interpretability and the need for up-to-date information [13, 20, 22].



Stage 1
BERT-based LLM: chunk-size = 512 tokens (90 token overlaps), [CLS] token to test.
GPT-based LLM: max input length=2048, chunk-size$\in\{512, 2048\}$, last token to test.
For all (α) GPT we compare with (α) BERT-based LLM on 512 input length.
Finetuned for e=4 epochs, chose best e for Stage 2 and evaluation.

Stage 2 (Embedding Extraction)	Stage 3 & 4
	Optimizer = Adam (learning rate = $3.5e^{-6}$)
BERT-based LLM:	*Loss func.:* multi-label: categorical cross-entropy
[CLS] token for each chunk	binary & multi-class: binary cross-entropy
	t={1, 2, 3}, h=8, e=5 epochs (best e for evaluation)
GPT-based LLM:	*Structure approximation:* pUMAP[5] (64 dimensions)
last token for each chunk	HDBSCAN (15 min clusters)

GPU used: Nvidia V100 & A100, with ZERO-3 in Deepspeed[6] with Accelerate[7].
Maximum fine-tune time (hours/epoch) for GPTs (6 Nvidia A100):
GPT-Neo-1.3B = 2.1, GPT-Neo-2.7B = 4, GPT-J = 8

(a) Training and Evaluation of GPT-based LLMs on a Large Text Corpus[21]

(b) Server and Client Communication in a Distributed System[51]

Figure 6: Examples of Applications in Diverse Domains

As depicted in Figure 6, RLHF enhances machine learning models' performance across diverse domains by incorporating human intuition and expertise. The examples illustrate the importance of fine-tuning with human feedback in large language models and optimizing communication protocols in distributed systems, showcasing RLHF's versatile applications and the significant enhancement of machine learning models' performance across various technological landscapes [21, 51].

4.3 Challenges and Limitations

Method Name	Computational Challenges	Data Dependencies	Theoretical Understanding
UAT[53]	Computational Demands	Unlabeled Data	Limited Theoretical Insights
AL[74]	Training Time Complexity	Diverse Datasets	Theoretical Understanding Weak
AXAI[83]	Large Explainer Trees	Local Training Samples	Limited Theoretical Insights

Table 3: This table presents a comprehensive analysis of various learning methods in the context of Reinforcement Learning from Human Feedback (RLHF), highlighting their computational challenges, data dependencies, and the extent of theoretical understanding. The table provides insights into the limitations faced by methods such as Unsupervised Adversarial Training (UAT), Associated Learning (AL), and Explainable AI (AXAI) in terms of computational demands, data requirements, and theoretical insights.

Table 3 provides a detailed examination of the computational challenges, data dependencies, and theoretical understanding associated with different learning methods, as discussed in the context of RLHF challenges and limitations. RLHF faces several challenges and limitations impacting its implementation and efficacy. A primary challenge is modeling uncertainty, essential for accurately capturing data and parameter variabilities, compounded by the computational difficulties of Bayesian inference [5]. Additionally, the computational demands of training deep networks and the need for model interpretability and generalization across diverse datasets continue to pose challenges [8].

The reliance on high-quality unlabeled data for Unsupervised Adversarial Training (UAT) presents another limitation. If the distribution of unlabeled data significantly differs from labeled data, model robustness may be compromised [53]. Existing benchmarks often inadequately address the combined challenges of data scarcity and noise robustness, limiting spoken language understanding (SLU) models' practical applications [45].

The theoretical understanding of Associated Learning (AL) remains limited, with performance varying across different network structures, particularly in complex architectures like ResNet. This variability underscores the need for further exploration and validation of AL methodologies for diverse neural network architectures [74]. Moreover, the narrow focus on scaling and predictive accuracy within AI research may undermine the diversity of approaches, hindering long-term scientific innovation [3].

The complexity and ambiguity of human language complicate establishing reliable evaluation criteria, affecting feedback quality that guides learning processes. This ambiguity can lead to inconsistent interpretations and responses, impacting RLHF training outcomes [83]. Additionally, the inherent locking in neural networks, where later computations must complete before feedback can be applied to earlier layers, presents a challenge in credit assignment, affecting training efficiency [36].

Finally, the heavy programming demands for future RLHF experiments may hinder validation and experimentation, limiting advancements in theoretical insights and practical applications [76]. Addressing these challenges and limitations is crucial for enhancing RLHF's robustness and applicability across diverse domains.

4.4 Innovative Approaches and Solutions

Method Name	Learning Optimization	Framework Development	Evaluation Metrics
N2P[84]	-	Theoretical Framework	Mean Squared Error
DF[85]	Curriculum-aware Algorithms	Geometric And Structured	Standardized Metrics
BS[86]	Balancing Strategy	Geometric Frameworks	Predictive Performance
BNN+LV[44]	Active Learning	Robust Framework	Log-likelihood Measures

Table 4: Overview of innovative methodological approaches in reinforcement learning with human feedback (RLHF), highlighting key aspects such as learning optimization strategies, framework development, and evaluation metrics. The table presents various methods including N2P, DF, BS, and BNN+LV, detailing their contributions to enhancing neural network operations, robustness, and performance evaluation.

Innovative approaches in RLHF are vital for overcoming existing challenges and enhancing AI systems' performance. One promising direction is integrating curriculum-aware algorithms that consolidate synapses at curriculum change points, optimizing the learning process compared to traditional methods [82]. This approach can significantly enhance RLHF systems' adaptability and efficiency by structuring learning to align with human cognitive patterns.

Exploring geometric frameworks for deep learning networks offers another innovative avenue. Extending these frameworks to non-linear networks and investigating noise effects on training dynamics can provide deeper insights into model behavior and improve robustness [26]. Such understanding is crucial for developing RLHF systems that can operate effectively in noisy and unpredictable environments.

Future research should also focus on improving the integration of linguistic features and employing deep learning techniques to develop more accurate evaluation metrics [16]. Such advancements would enable precise assessments of model performance, particularly in complex language processing tasks.

The NN2Poly framework represents an innovative solution for enhancing neural network operations' efficiency, particularly regarding polynomial representations [84]. Future research should explore its application to larger networks and non-tabular data types, such as images or audio, to expand its utility across various domains.

DeepFeature proposes a novel method for examining internal feature maps in deep neural networks (DNNs) to identify vulnerabilities that can be addressed to enhance overall performance [85]. This approach is critical for improving RLHF systems' robustness and reliability by systematically identifying and mitigating potential weaknesses.

Developing standardized evaluation metrics for explainability in reinforcement learning is another crucial area for future research. By refining XRL taxonomies and enhancing human knowledge integration, researchers can establish frameworks for evaluating RLHF systems' transparency and interpretability more effectively [80].

Incorporating structured representations of uncertainties, as demonstrated by methods like Balance-SHAP, can improve model interpretability by adjusting class distributions in SHAP's background and explanation data [86]. This approach enables informed decision-making and efficient data collection amid complex noise [44].

Overall, these innovative approaches and solutions present promising pathways for overcoming RLHF challenges. By incorporating advanced methodologies and emphasizing components such as curriculum learning, geometric frameworks, and model evaluation, researchers can significantly improve RLHF systems' adaptability and effectiveness. This enhancement not only fosters interpretability and reasoning in complex tasks, such as financial analysis—where real-time data and textual information integration is crucial—but also leverages structured learning sequences to optimize performance across diverse applications. Utilizing frameworks like Stock-Chain and datasets such as AlphaFin allows for benchmarking and refining these systems to handle data scarcity and information hallucination challenges, ultimately leading to more robust outcomes across various domains [20, 82]. Table 4 provides a comprehensive overview of the innovative approaches and solutions in reinforcement learning with human feedback, focusing on learning optimization, framework development, and evaluation metrics.

5 Reward Models in Reinforcement Learning

5.1 Conceptual Framework of Reward Models

The conceptual framework of reward models in reinforcement learning (RL) is pivotal for guiding agents in optimal decision-making by assigning values to actions and outcomes. This framework enhances RL systems' adaptability and effectiveness by integrating diverse data modalities, improving interpretability, and providing contextual explanations for decision-making processes [20, 22, 87, 21, 37]. A critical element is contextual information integration, which differentiates between normal fluctuations and genuine anomalies, refining decision-making through historical data and expert knowledge [22]. The OntoEnricher method exemplifies how reward models enhance interpretability by assigning value within enriched ontologies [22].

Heterogeneous relational reasoning, using entity type embeddings and graph neural networks (GNNs), encodes neighborhood information, reducing action space and enhancing reasoning capabilities [69]. This allows reward models to provide nuanced evaluations, improving RL agents' performance. Explainability is crucial, categorized into agent model-explaining, reward-explaining, state-explaining, and task-explaining methods [69], promoting transparency and interpretability essential for trust and reliability.

Robust evaluation metrics ensure consistent performance across environments. Metrics for zero-shot learning scenarios evaluate model performance while considering class imbalance and dataset challenges [9]. The use of F1 scores in named entity recognition tasks highlights the importance of metrics that reflect model performance [10]. Integrating short-term session information with long-term knowledge graph data enhances RL systems' adaptability [57].

The conceptual framework of reward models in reinforcement learning incorporates contextual information, enhances relational reasoning through GNNs, emphasizes explainability, and integrates robust evaluation metrics, improving RL systems' trust and applicability in real-world scenarios [88, 80].

5.2 Challenges in Designing Reward Models

Designing effective reward models in reinforcement learning faces challenges, especially in deep learning applications. High computational demands and limited real-world dataset samples can lead to overfitting [89]. Ensuring robustness with limited data is complicated by uniform regularization parameters across networks of varying widths, resulting in suboptimal performance [90].

Accurately predicting performance in configurable software systems is challenging due to numerous variables and configurations [18]. Potential biases in ground truth explanations derived from training data can lead to discrepancies between expected and actual outcomes [69]. Lack of benchmarks incorporating contextual factors limits reward models' applicability in educational contexts [39].

As illustrated in Figure 7, the key challenges in designing reward models for reinforcement learning can be categorized into three main areas: computational constraints, data limitations, and contextual modeling needs. Each category highlights significant issues such as high computational demands, limited data samples, and the necessity for context-sensitive modeling. Analog systems, while efficient, are susceptible to noise, limiting their use in precision-required scenarios [31]. Designing reward models that integrate and manage noise while maintaining accuracy remains significant.

These challenges involve computational constraints, data limitations, and the need for accurate, context-sensitive modeling. Addressing these issues is essential for developing robust reinforcement learning systems capable of effective performance across diverse applications [91].

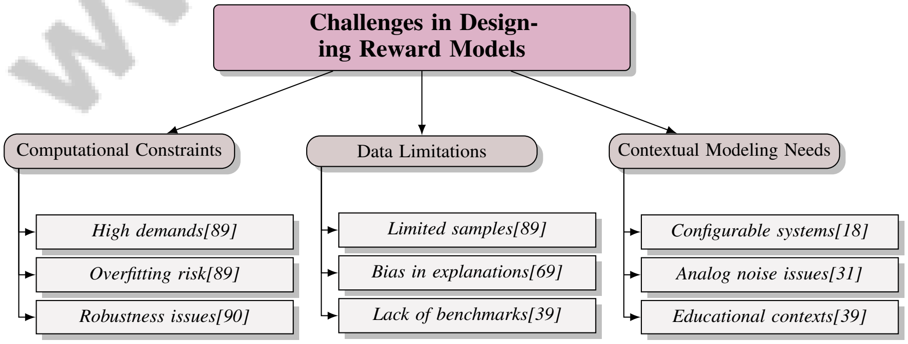


Figure 7: This figure illustrates the key challenges in designing reward models for reinforcement learning, categorized into computational constraints, data limitations, and contextual modeling needs. Each category highlights significant issues such as high computational demands, limited data samples, and the necessity for context-sensitive modeling.

AI-generated, for reference only.

5.3 Evaluation Metrics for Reward Models

Evaluating reward models in reinforcement learning is crucial for assessing their effectiveness in guiding agent behavior. Standard classification metrics like accuracy and the Matthews Correlation Coefficient (MCC) are used to evaluate model performance [68]. Accuracy measures the proportion of correctly predicted instances, but may not reflect effectiveness in imbalanced class distributions. In low-resource settings, precision, recall, and F1-score offer nuanced performance insights, crucial for optimizing model selection and training strategies, particularly in applications like financial analysis [51, 20, 27].

The MCC provides a balanced measure by considering true and false positives and negatives, addressing class imbalance effectively. This robustness ensures reliable model output interpretations, crucial in applications like clinical decision-making and financial forecasting [92, 93, 86, 94]. Evaluation frameworks must consider application domain specifics. In education, metrics assessing student engagement and learning outcomes are vital, aligning with curriculum learning principles [52, 82, 27, 57]. In business analytics, metrics reflecting resource allocation efficiency provide comprehensive reward model performance evaluation.

Evaluation metrics should align with application objectives and constraints, ensuring effective performance measurement in areas like activity recommendation and machine translation evaluation [51, 52, 16, 95, 57]. Employing a combination of standard and domain-specific metrics ensures thorough reward model assessments, facilitating robust reinforcement learning system development.

5.4 Applications and Case Studies

Reward models in reinforcement learning demonstrate versatility across applications. Zero-shot learning with large language models (LLMs) generalizes across tasks without extensive retraining, applied to datasets like COVID-19-related tweets and economic texts, showcasing adaptability in diverse textual formats [55]. In creative applications, Kuzushiji datasets serve as benchmarks for classification and generative modeling, highlighting reward models' potential in innovation [9].

In education, reward models enhance learning experiences by integrating contextual knowledge and historical data, tailoring experiences to individual needs and improving outcomes and engagement. This aligns with curriculum learning principles, emphasizing structured information presentation [13, 68, 82]. In business analytics, reward models optimize decision-making and resource allocation, improving predictive performance and operational efficiency compared to traditional approaches [20, 95, 94, 7]. Aligning model predictions with organizational goals enhances operational efficiency and strategic planning.

These applications illustrate reward models' capabilities in driving innovation and improving performance across domains. By harnessing contextual information and employing sophisticated modeling techniques, reward models enhance AI applications' capabilities, particularly in information security, where enriched ontologies and neural representations enable effective anomaly detection and threat intelligence [13, 22].

6 Natural Language Processing and AI Alignment

The alignment of artificial intelligence (AI) systems with human cognitive processes is a pivotal research area, focusing on bridging computational models and human reasoning. This domain emphasizes neuro-symbolic AI, which integrates deep learning with symbolic reasoning to enhance interpretability, generalization, and reasoning capabilities in tasks like natural language processing (NLP). Researchers aim to develop models that emulate human-like reasoning by examining various knowledge representation and reasoning methods. A human-centric AI approach prioritizes transparency, fairness, and social good, fostering collaborative relationships between humans and machines in cognitive tasks [65, 14, 79, 78]. As AI technologies evolve, designing systems that embody human values and decision-making processes is crucial, addressing technical and ethical challenges to ensure AI operates in interpretable, transparent, and trustworthy ways.

The following subsection delves into methodologies and approaches that align AI systems with human reasoning, integrating human expertise, symbolic reasoning, and complexity measures to develop AI models resonating with human cognitive frameworks.

AI-generated, for reference only.

6.1 Aligning AI Systems with Human Reasoning

Aligning AI systems with human reasoning involves enhancing interpretability and understanding of AI models in congruence with human cognitive processes. Incorporating human expertise in model training ensures AI systems maintain accuracy while aligning with human intuition and decision-making [22]. This is crucial in nuanced applications like NLP and information retrieval, where capturing semantic relationships is essential [50]. Platforms like AUTONLU democratize AI model development, enabling non-experts to create models aligned with human reasoning [49]. This adaptability allows AI systems to match users' understanding, enhancing engagement [63].

Research highlights the promise of integrating symbolic reasoning with neural networks to improve interpretability and performance [14]. This hybrid approach leverages the strengths of both paradigms, aligning AI systems with human cognitive processes. Heterogeneous relational reasoning enhances interpretability and action space efficiency, providing clear decision-making pathways [64]. Exploring complexity measures, such as sharpness and norms, reveals insights into AI models' generalization capabilities, critical for aligning AI systems with human reasoning [72]. These measures elucidate interactions between optimization processes and model complexity, contributing to AI systems that generalize effectively across diverse tasks and environments [96].

Benchmarks distinguishing between human-generated and AI-generated text are crucial for transparency and accountability in AI systems [68]. They ensure AI-generated content aligns with human values and expectations, fostering trust and reliability in AI applications [97]. Aligning AI systems with human reasoning requires integrating human expertise, symbolic reasoning, and complexity measures, prioritizing interpretability, adaptability, and transparency to develop AI systems resonating with human cognitive processes and values [41].

6.2 Challenges in Goal Alignment

Aligning AI goals with human intentions presents challenges due to biases and limitations in datasets and methodologies used in AI development. Biases in training datasets of large language models (LLMs) can hinder AI systems' generalizability, leading to outcomes misaligned with human intentions [55]. Human validation introduces biases and inconsistencies in feedback loops, complicating goal alignment [79]. The complexity of data sources can obscure essential patterns necessary for effective goal alignment.

The scarcity of comprehensive datasets for specific languages, like Persian, exacerbates alignment difficulties. Challenges in processing complex sentences in these languages highlight limitations in current models to capture human language nuances, vital for aligning AI systems with human intentions [98]. Addressing these challenges requires high-quality datasets, methodologies to mitigate biases, and improved strategies for integrating human feedback consistently and unbiasedly. This is crucial in big data contexts, where volume, velocity, and veracity present unique complexities. Advancements in LLMs and human-centric machine learning demonstrate that overcoming these challenges can enhance performance in tasks like automated grading and recruitment while ensuring fairness and transparency in AI-driven decision-making [51, 78, 52, 60, 79]. Addressing these challenges is essential for developing AI systems aligned with human values, achieving desired outcomes across diverse applications.

6.3 Data Efficiency and Evaluation Metrics

Data efficiency and evaluation metrics are critical for ensuring AI systems operate effectively and align with human values. AI models' ability to learn from limited data while maintaining high performance is essential for real-world deployment, where data may be scarce or costly. Integrating semi-supervised learning (SSL) techniques with kernel methods has provided insights into wide neural networks' behavior, highlighting potential data efficiency improvements through advanced mathematical frameworks [99].

Theoretical foundations of infinite-width limits in neural networks offer pathways for developing data-efficient learning algorithms, emphasizing AI systems' interplay with human values. This facilitates creating AI applications prioritizing utility, social good, fairness, privacy, transparency, and accountability—key tenets of a Human-Centric Machine Learning approach. Addressing these aspects minimizes misalignment risks due to data limitations and ensures AI-driven decision-making processes do not perpetuate biases, as seen in automated recruitment algorithms and multimodal data integration [51, 14, 79, 78].

Evaluation metrics are vital in AI alignment, assessing performance and reliability. Metrics must reflect specific goals and constraints of the application domain, capturing human values and intentions nuances. Metrics encompassing the complexity and variability of real-world data are crucial for evaluating AI systems' robustness and adaptability, particularly in cybersecurity, where dynamic threat landscapes necessitate sophisticated anomaly detection and reasoning capabilities. Advanced deep learning architectures, like Bidirectional LSTMs, enhance security ontologies by accurately extracting contextual information from unstructured text, improving AI-driven security analytics' interpretability and reliability. In financial analysis, integrating real-time data with machine learning approaches can enhance prediction accuracy and reasoning transparency, addressing interpretability challenges and integrating diverse information sources [20, 13, 22, 4, 57].

By prioritizing data efficiency and establishing robust evaluation metrics, researchers can significantly enhance AI systems' alignment with human values. This approach improves AI interactions' effectiveness and trustworthiness and addresses fairness, transparency, and accountability concerns in AI-driven decision-making. Integrating human-centric principles, such as privacy and social good, into AI development can mitigate biases, ensuring these systems operate beneficially and equitably for all users [78, 13, 79, 16, 22]. This approach enhances AI model performance while fostering greater confidence in their ability to operate safely and ethically across diverse applications.

6.4 Integration of Human Expertise

Integrating human expertise is pivotal for enhancing AI alignment, ensuring AI systems are developed and deployed in ways consistent with human values and intentions. Human expertise is crucial for identifying and addressing limitations in current AI models, including the need for interpretability and efficiency in monolithic networks. Future research should explore identifying modular components within these networks, leveraging human expertise to enhance interpretability and efficiency, vital for aligning AI systems with human reasoning and decision-making processes [100].

In NLP, particularly for text-to-speech (TTS) systems, human expertise is essential for improving robustness and developing methods suitable for low-resource languages. These efforts address existing gaps, ensuring AI systems effectively process and generate speech across diverse linguistic contexts [101]. By integrating linguistic and cultural knowledge, human experts can guide AI models' development to be culturally aware and capable of handling different languages' nuances.

Interdisciplinary collaborations, exemplified by the Kuzushiji dataset, significantly contribute to the field by making classical Japanese literature accessible to machine learning applications, fostering collaboration between AI researchers and classical literature experts [9]. These collaborations underscore the importance of human expertise in enriching AI systems with domain-specific knowledge, ultimately enhancing alignment with human values and cultural contexts.

The integration of human expertise is essential for advancing AI alignment, providing insights and knowledge to ensure AI systems are developed ethically, interpretable, and aligned with human values. By fostering collaboration between AI researchers and domain experts, the field can leverage advanced techniques like deep learning and neuro-symbolic AI to enhance aligned and trustworthy AI systems development, addressing challenges like contextual understanding in NLP and automating model selection for diverse applications, ultimately leading to more robust and interpretable AI solutions [13, 22, 14, 51].

6.5 Interpretable and Trustworthy AI Systems

Developing interpretable and trustworthy AI systems is essential for ensuring AI models align with human values and can be reliably deployed across diverse applications. A significant challenge in achieving this goal is the 'black box' nature of deep learning models, which often obscures decision-making processes and complicates interpretability [6]. This limitation emphasizes the need for methodologies enhancing transparency and providing clear insights into model predictions.

Interpretable AI systems benefit from advanced techniques elucidating AI decisions' rationale. Methodologies employing tree-based local explanations, like AraucanaXAI, enhance understanding by detailing relationships between input variables and model outputs. Implementing approaches

AI-generated, for reference only.

integrating deep learning with symbolic reasoning is vital for fostering trust in AI systems, as these methods enhance systems' ability to operate in alignment with human reasoning and expectations, ensuring transparency, accountability, and fairness in decision-making processes [68, 78, 13, 22, 14].

Additionally, integrating task-specific representations, exemplified by models like MolTailor, improves interpretability by focusing on relevant features for specific tasks, enhancing predictive performance and aligning AI systems with human values. This focus aids in understanding model behavior and contributes to more effective decision-making processes [7].

Trustworthy AI systems also require advancements in inference optimization, particularly regarding long context lengths impacting alignment with human values. Promoting energy-efficient algorithms and equitable access to computational resources is essential for fostering trust and sustainability in AI development [37]. Furthermore, exploring clusters generated by methods like MESc and developing explanation algorithms are promising avenues for future research aimed at improving large language models' interpretability.

In deep learning and reinforcement learning contexts, enhancing autonomy, efficiency, and safety in applications like unmanned aerial systems (UAS) is critical for building trust in AI technologies. However, limitations of analog deep learning methods, particularly in scalability, highlight the need for ongoing research and development to ensure these systems are efficient and trustworthy [31].

The pursuit of interpretable and trustworthy AI systems necessitates a comprehensive strategy synergizing domain expertise, tailored task representations, and sophisticated optimization techniques, especially in enhancing information security through advanced methodologies like deep learning for ontology enrichment and neural information retrieval. This multifaceted approach leverages contextual knowledge to improve anomaly detection and threat intelligence while addressing challenges in extracting relevant concepts from unstructured data, fostering the development of robust AI systems capable of delivering high accuracy and reliability in critical applications [13, 22]. By addressing these challenges, researchers can develop AI models that are effective and aligned with human values, fostering greater trust and reliability in AI applications.

7 Model Evaluation

7.1 Comparative Analysis of Evaluation Techniques

Evaluating machine learning and deep learning models is crucial for determining their effectiveness and efficiency. Traditional metrics, like CrossEntropy, are commonly used in classification tasks, yet the ISBE method has shown superior accuracy and reduced computational time with the MNIST dataset [34]. This suggests that alternative evaluation techniques can significantly enhance performance in specific scenarios.

In language modeling, adopting novel architectures and training schemes has led to substantial improvements. For instance, evolving recurrent networks have achieved a 6-point reduction in perplexity compared to standard LSTM models [38]. This highlights the importance of exploring innovative strategies to optimize model performance.

Selecting suitable evaluation metrics is critical for accurately assessing model effectiveness. Metrics such as accuracy, precision, recall, and F1-score must be chosen based on the task and dataset characteristics. In contexts with imbalanced datasets, relying solely on accuracy can be misleading, necessitating a broader range of metrics for a comprehensive evaluation [57, 51, 13, 27].

To optimize models for real-world applications, advanced evaluation frameworks that assess both computational efficiency and predictive accuracy are essential. This approach overcomes the limitations of traditional single-scalar metrics, which can obscure the complexities of information retrieval measures and hinder interpretability. Frameworks like QPP-GenRE and Stock-Chain, which decompose query performance prediction and integrate retrieval-augmented generation, exemplify robust evaluation methodologies [20, 57]. Continuous refinement of evaluation techniques will aid in developing AI systems that are reliable and aligned with human values.

7.2 Role of Evaluation in AI Alignment and NLP

Evaluation plays a pivotal role in AI alignment and NLP, offering insights into the effectiveness and reliability of AI systems. In AI alignment, evaluation methodologies ensure models adhere to ethical standards and human values, with careful selection of loss functions and metrics tailored to specific tasks [11]. This precision is crucial for models that must withstand adversarial attacks, which are assessed through varying amounts of unlabeled data [53].

In NLP, evaluation techniques enhance language understanding and processing capabilities. For example, the accuracy of generated outputs, such as Urdu poetry, improves through iterative evaluations of different epochs in the GRU model [2]. Simplifying network structures, as demonstrated by the particle-based shooting approach, can achieve competitive performance while reducing computational demands [56].

Assessing generalization performance and representation diversity in NLP models is vital for developing systems that adapt to new linguistic contexts, ensuring robustness and effectiveness across diverse applications [13, 22, 51]. Incorporating physiological data into models like EmLLM further emphasizes the need for comprehensive evaluation frameworks that predict human-like responses, reinforcing alignment with human values.

Evaluation is integral to AI alignment and NLP, equipping researchers with tools to enhance model performance. Employing a range of evaluation techniques, including manual and automatic methods, ensures AI systems align with human values while enhancing NLP applications. This thorough approach fosters trust and reliability in AI technologies, addressing challenges in reasoning, generalization, and interpretability. Leveraging state-of-the-art models and parameter-efficient fine-tuning methods can further improve AI performance across various NLP tasks, contributing to a robust and ethically aligned AI landscape [52, 13, 16, 22, 14].

7.3 Future Directions in Model Evaluation

Future directions in model evaluation will focus on enhancing human-machine collaboration, particularly in data matching and less-than-supervised learning environments [79]. This approach aims to improve synergy between human insights and machine learning models, leading to more effective evaluation processes. Exploring infinite-width limits in neural networks could extend findings to deep networks and various loss functions in semi-supervised learning, providing a rigorous understanding of ReLU activations and their implications for model performance [99].

Explainable reinforcement learning (XRL) faces challenges, especially in integrating human knowledge optimally into XRL methods. Establishing universally accepted evaluation metrics that effectively assess transparency and interpretability remains a significant hurdle [80]. Addressing these gaps is essential for advancing evaluation techniques that ensure AI systems are interpretable and aligned with human values.

Ongoing research should refine evaluation metrics to account for the complexity and variability of real-world data, ensuring models are robust and adaptable across diverse applications. By focusing on explainability, evaluation methods, and automated model selection, researchers can develop robust frameworks for evaluating AI models. This approach addresses transparency challenges associated with AI's black-box nature, especially in high-stakes applications like software engineering and information retrieval, while enhancing AI systems' reliability and trustworthiness. Integrating advanced evaluation techniques and automated selection processes will facilitate broader AI adoption, particularly in resource-constrained environments, leading to improved decision-making and outcomes across various domains [13, 4, 51, 16].

8 Conclusion

This survey highlights the profound impact of deep learning technologies in transforming various domains, with a particular focus on natural language processing (NLP) and AI alignment. The integration of deep learning techniques has significantly advanced the capacity of AI systems to model intricate patterns and solve complex problems. The evolution of structured deep neural networks (DNNs) presents promising advancements in automatic speech recognition, indicating potential benefits from exploring the use of multiple phoneme states to enhance model efficacy.

Addressing uncertainty within deep learning models remains crucial, as probabilistic approaches offer pathways to increased robustness and interpretability. Despite challenges in data requirements and interpretability, deep learning continues to be a cornerstone technology with vast potential for future applications across multiple sectors.

The fusion of physiological data with large language models (LLMs) marks a significant stride in enhancing empathic AI capabilities, propelling new research directions in deep learning, NLP, and AI alignment. This development underscores the dynamic, interdisciplinary nature of deep learning research and advocates for hybrid architectures that integrate deep learning with diverse AI methodologies to tackle complex real-world challenges.

Future investigations should focus on refining DNN models to better reflect brain activity, exploring novel stimulus modalities to deepen cognitive function understanding, and enhancing model robustness through hybrid architectures. Additionally, insights into the cerebellum's role in temporal feedback and credit assignment could improve learning efficiency and reduce errors in tasks reliant on temporal feedback.

References

[1] Andrei-Bogdan Puiu and Andrei-Octavian Brabete. Towards nlp with deep learning: Convolutional neural networks and recurrent neural networks for offensive language identification in social media, 2019.

[2] Muhammad Shoaib Farooq and Ali Abbas. Urdu poetry generated by using deep learning techniques, 2023.

[3] Bernard J. Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution, 2024.

[4] Sicong Cao, Xiaobing Sun, Ratnadira Widyasari, David Lo, Xiaoxue Wu, Lili Bo, Jiale Zhang, Bin Li, Wei Liu, Di Wu, and Yixin Chen. A systematic literature review on explainability for machine/deep learning-based software engineering research, 2025.

[5] Daniel T. Chang. Probabilistic deep learning with probabilistic neural networks and deep probabilistic models, 2021.

[6] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5):91, 2023.

[7] Mathias Kraus, Stefan Feuerriegel, and Asil Oztekin. Deep learning in business analytics and operations research: Models, applications and managerial implications, 2019.

[8] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. A review of deep learning with special emphasis on architectures, applications and recent trends, 2019.

[9] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[10] Ege Kesim and Aysu Deliahmetoglu. Named entity recognition in resumes, 2023.

[11] Juan Terven, Diana M. Cordova-Esparza, Alfonso Ramirez-Pedraza, Edgar A. Chavez-Urbiola, and Julio A. Romero-Gonzalez. Loss functions and metrics in deep learning, 2024.

[12] Subba Reddy Oota, Zijiao Chen, Manish Gupta, Raju S. Bapi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey), 2024.

[13] Ye Zhang, Md Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. Neural information retrieval: A literature review, 2017.

[14] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promise in natural language processing? a structured review, 2022.

[15] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, 51(5):1–36, 2018.

[16] Lifeng Han. Machine translation evaluation resources and methods: A survey, 2018.

[17] Nicholas G. Polson and Vadim O. Sokolov. Deep learning, 2018.

[18] Jingzhi Gong. Pushing the boundary: Specialising deep configuration performance learning, 2025.

[19] Haoqiang Guo, Sendong Zhao, Haochun Wang, Yanrui Du, and Bing Qin. Moltailor: Tailoring chemical molecular representation to specific tasks via text prompts, 2024.

[20] Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework, 2024.

[21] Nishchal Prasad, Mohand Boughanem, and Taoufiq Dkaki. Exploring large language models and hierarchical frameworks for classification of large unstructured legal documents, 2024.

[22] Lalit Mohan Sanagavarapu, Vivek Iyer, and Raghu Reddy. A deep learning approach for ontology enrichment from unstructured text, 2021.

[23] Ahmed Ibrahim, A. Lynn Abbott, and Mohamed E. Hussein. Input fast-forwarding for better deep learning, 2017.

[24] Xin Qian and Diego Klabjan. The impact of the mini-batch size on the variance of gradients in stochastic gradient descent, 2020.

[25] Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning, 2021.

[26] Govind Menon. The geometry of the deep linear network, 2024.

[27] Hongseok Choi and Hyunju Lee. Exploiting all samples in low-resource sentence classification: Early stopping and initialization parameters, 2024.

[28] Lindsay Munroe, Mariana da Silva, Faezeh Heidari, Irina Grigorescu, Simon Dahan, Emma C. Robinson, Maria Deprez, and Po-Wah So. Applications of interpretable deep learning in neuroimaging: a comprehensive review, 2024.

[29] Chao Ni, Liyu Shen, Xiaodan Xu, Xin Yin, and Shaohua Wang. Learning-based models for vulnerability detection: An extensive study, 2024.

[30] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020.

[31] Aditya Datar and Pramit Saha. The promise of analog deep learning: Recent advances, challenges and opportunities, 2024.

[32] Sira Vegas and Sebastian Elbaum. Pitfalls in experiments with dnn4se: An analysis of the state of the practice, 2023.

[33] Jack K Fitzsimons, Sebastian M Schmon, and Stephen J Roberts. Implicit priors for knowledge sharing in bayesian neural networks, 2019.

[34] Wladyslaw Skarbek. Cross entropy in deep learning of classifiers is unnecessary – isbe error is all you need, 2023.

[35] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

[36] Joseph Pemberton, Ellen Boven, Richard Apps, and Rui Ponte Costa. Cortico-cerebellar networks as decoupling neural interfaces, 2021.

[37] Jiahao Tian, Jinman Zhao, Zhenkai Wang, and Zhicheng Ding. Mmrec: Llm based multi-modal recommender system, 2024.

[38] Aditya Rawal and Risto Miikkulainen. From nodes to networks: Evolving recurrent neural networks, 2018.

[39] Shalini Pandey, George Karypis, and Jaideep Srivastava. An empirical comparison of deep learning models for knowledge tracing on large-scale dataset, 2021.

[40] Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. Pretrained domain-specific language model for general information retrieval tasks in the aec domain, 2022.

[41] Poorvesh Dongre, Majid Behravan, Kunal Gupta, Mark Billinghurst, and Denis Gračanin. Integrating physiological data with large language models for empathic human-ai interaction, 2024.

[42] Wannita Takerngsaksiri, Rujikorn Charakorn, Chakkrit Tantithamthavorn, and Yuan-Fang Li. Pytester: Deep reinforcement learning for text-to-testcase generation, 2024.

[43] Sandeep Nallan Chakravarthula, Haoqi Li, Shao-Yen Tseng, Maija Reblin, and Panayiotis Georgiou. Predicting behavior in cancer-afflicted patient and spouse interactions using speech and language, 2019.

[44] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning, 2018.

[45] Shang-Wen Li, Jason Krone, Shuyan Dong, Yi Zhang, and Yaser Al-onaizan. Meta learning to classify intent and slot labels with noisy few shot examples, 2020.

[46] Jun Yu, Yutong Dai, Xiaokang Liu, Jin Huang, Yishan Shen, Ke Zhang, Rong Zhou, Eashan Adhikarla, Wenxuan Ye, Yixin Liu, Zhaoming Kong, Kai Zhang, Yilong Yin, Vinod Namboodiri, Brian D. Davison, Jason H. Moore, and Yong Chen. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras, 2024.

[47] Yi-Hsiu Liao, Hung-Yi Lee, and Lin shan Lee. Towards structured deep neural network for automatic speech recognition, 2015.

[48] Sagar Tamang and Dibya Jyoti Bora. Performance evaluation of tokenizers in large language models for the assamese language, 2024.

[49] Nham Le, Tuan Lai, Trung Bui, and Doo Soon Kim. Autonlu: An on-demand cloud-based natural language understanding system for enterprises, 2020.

[50] Sarvesh Patil. Deep learning based natural language processing for end to end speech translation, 2018.

[51] Vivek Sharma, Praneeth Vepakomma, Tristan Swedish, Ken Chang, Jayashree Kalpathy-Cramer, and Ramesh Raskar. Expertmatcher: Automating ml model selection for users in resource constrained countries, 2019.

[52] Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. Investigating automatic scoring and feedback using large language models, 2024.

[53] Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

[54] Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection, 2024.

[55] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. Large language models are zero-shot text classifiers, 2023.

[56] François-Xavier Vialard, Roland Kwitt, Susan Wei, and Marc Niethammer. A shooting formulation of deep learning, 2020.

[57] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. Query performance prediction using relevance judgments generated by large language models, 2024.

[58] Yuxuan Hu, Chenwei Zhang, Min Yang, Xiaodan Liang, Chengming Li, and Xiping Hu. Learning to generalize unseen domains via multi-source meta learning for text classification, 2024.

[59] Dan Sun, Yaxin Liang, Yining Yang, Yuhan Ma, Qishi Zhan, and Erdi Gao. Research on optimization of natural language processing model based on multimodal deep learning, 2024.

[60] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures, 2020.

[61] Pratik Kayal, Mayank Singh, and Pawan Goyal. Weakly-supervised deep learning for domain invariant sentiment classification, 2019.

[62] Mun-Hak Lee and Joon-Hyuk Chang. Knowledge distillation from language model to acoustic model: a hierarchical multi-task learning approach, 2021.

[63] Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. Learning multi-party turn-taking models from dialogue logs, 2019.

[64] Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. Causality guided disentanglement for cross-platform hate speech detection, 2023.

[65] Emile van Krieken. Optimisation in neurosymbolic learning systems, 2024.

[66] Andres Gomez Tato. Evaluation of machine learning fameworks on finis terrae ii, 2018.

[67] Scott Yak, Javier Gonzalvo, and Hanna Mazzawi. Towards task and architecture-independent generalization gap predictors, 2019.

[68] Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. Distinguishing human generated text from chatgpt generated text using machine learning, 2023.

[69] Rafaël Brandt, Daan Raatjens, and Georgi Gaydadjiev. Precise benchmarking of explainable ai attribution methods, 2023.

[70] Yazhe Li, Jorg Bornschein, and Marcus Hutter. Evaluating representations with readout model switching, 2024.

[71] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8), 2017.

[72] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[73] Zichen Song, Yuxin Wu, Sitan Huang, and Zhongfeng Kang. Avss: Layer importance evaluation in large language models via activation variance-sparsity analysis, 2024.

[74] Yu-Wei Kao and Hung-Hsuan Chen. Associated learning: Decomposing end-to-end backpropagation based on auto-encoders and target propagation, 2021.

[75] Yoshikazu Terada and Ryoma Hirose. Fast generalization error bound of deep learning without scale invariance of activation functions, 2019.

[76] John Chiang. Activation functions not to active: A plausible theory on interpreting neural networks, 2023.

[77] Jithin Jagannath, Anu Jagannath, Sean Furman, and Tyler Gwin. Deep learning and reinforcement learning for autonomous unmanned aerial systems: Roadmap for theory to deployment, 2020.

[78] Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment, 2023.

[79] Avigdor Gal and Roee Shraga. Human's role in-the-loop, 2022.

[80] Yunpeng Qing, Shunyu Liu, Jie Song, Huiqiong Wang, and Mingli Song. A survey on explainable reinforcement learning: Concepts, algorithms, challenges, 2023.

AI-generated, for reference only.

[81] Jun-En Ding, Phan Nguyen Minh Thao, Wen-Chih Peng, Jian-Zhe Wang, Chun-Cheng Chug, Min-Chen Hsieh, Yun-Chien Tseng, Ling Chen, Dongsheng Luo, Chi-Te Wang, Pei fu Chen, Feng Liu, and Fang-Ming Hung. Large language multimodal models for 5-year chronic disease cohort prediction using ehr data, 2024.

[82] Luca Saglietti, Stefano Sarao Mannelli, and Andrew Saxe. An analytical theory of curriculum learning in teacher-student networks, 2022.

[83] Enea Parimbelli, Giovanna Nicora, Szymon Wilk, Wojtek Michalowski, and Riccardo Bellazzi. Tree-based local explanations of machine learning model predictions, araucanaxai, 2021.

[84] Pablo Morala, Jenny Alexandra Cifuentes, Rosa E. Lillo, and Iñaki Ucar. Nn2poly: A polynomial representation for deep feed-forward artificial neural networks, 2023.

[85] Dong Huang, Qingwen Bu, Yahao Qing, Yichao Fu, and Heming Cui. Feature map testing for deep neural networks, 2023.

[86] Mingxuan Liu, Yilin Ning, Han Yuan, Marcus Eng Hock Ong, and Nan Liu. Balanced background and explanation data are needed in explaining deep learning models with shap: An empirical study on clinical decision making, 2022.

[87] Huizi Wu, Hui Fang, Zhu Sun, Cong Geng, Xinyu Kong, and Yew-Soon Ong. A generic reinforced explainable framework with knowledge graph for session-based recommendation, 2022.

[88] Mandana Saebi, Steven Krieg, Chuxu Zhang, Meng Jiang, and Nitesh Chawla. Heterogeneous relational reasoning in knowledge graphs with reinforcement learning, 2020.

[89] Michael Widrich, Bernhard Schäfl, Hubert Ramsauer, Milena Pavlović, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, and Günter Klambauer. Modern hopfield networks and attention for immune repertoire classification, 2020.

[90] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34:7054–7067, 2021.

[91] Siwei Lai. Word and document embeddings based on neural network approaches, 2016.

[92] G. Jogesh Babu, David Banks, Hyunsoon Cho, David Han, Hailin Sang, and Shouyi Wang. A statistician teaches deep learning, 2021.

[93] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

[94] Stefan Feuerriegel and Ralph Fehrer. Improving decision analytics with deep learning: The case of financial disclosures, 2018.

[95] Prerna Agarwal, Avani Gupta, Renuka Sindhgatta, and Sampath Dechu. Goal-oriented next best activity recommendation using reinforcement learning, 2022.

[96] Robert Balkin, Hector D. Ceniceros, and Ruimeng Hu. Stochastic delay differential games: Financial modeling and machine learning algorithms, 2023.

[97] Kanishk Gandhi, Gala Stojnic, Brenden M. Lake, and Moira R. Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others, 2022.

[98] Ali Nazarizadeh, Touraj Banirostam, and Minoo Sayyadpour. Sentiment analysis of persian language: Review of algorithms, approaches and datasets, 2022.

[99] Maximilian Fleissner, Gautham Govind Anil, and Debarghya Ghoshdastidar. Infinite width limits of self supervised neural networks, 2025.

[100] Atish Agarwala, Abhimanyu Das, Brendan Juba, Rina Panigrahy, Vatsal Sharan, Xin Wang, and Qiuyi Zhang. One network fits all? modular versus monolithic task formulations in neural networks, 2021.

[101] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis, 2021.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.