Reasoning, Replicability, and Benchmarking in Large Language and Foundation Models: Methodologies, Challenges, and Pathways Toward Trustworthy, Interpretable, and Inclusive AI

Abstract

This survey provides a comprehensive synthesis of recent advances, methodologies, and enduring challenges in the development, evaluation, and responsible deployment of large language models (LLMs) and foundation models. Motivated by the transformative impact of LLMs across natural language processing, scientific discovery, and real-world applications, the paper critically examines the evolution from symbolic and neural paradigms through contemporary transformer-driven and neurosymbolic architectures, highlighting emergent reasoning capabilities and the drive towards human-like abstraction. The review systematically analyzes benchmarking ecosystems, probing frameworks, and evaluation metrics, emphasizing the limitations of prevailing practices in capturing semantic faithfulness, compositionality, and real-world reasoning, particularly on multistep, cross-modal, and domain-specific tasks. Key contributions include a structured taxonomy of model architectures and fusion strategies, an assessment of hybrid approaches integrating neural, symbolic, and graph-based reasoning, and comparative analyses of benchmark methodologies across linguistic, reasoning, and multimodal domains.

The survey underscores persistent gaps in robustness, interpretability, fairness, and reproducibility—drawing attention to vulnerabilities in adversarial and out-of-distribution scenarios, challenges in auditability and demographic inclusion, and the ongoing reproducibility crisis stemming from inadequate reporting and opaque "language-models-as-a-service" paradigms. It highlights advances in adaptive prompting, modular workflow orchestration, and explainability, while advocating for open science, FAIR data practices, and transparent, community-driven benchmarking. Strategic recommendations target holistic evaluation protocols, enhanced benchmarking diversity, rigorous auditing, responsible design, and the institutionalization of modular, reproducible workflows. The paper concludes that future progress in LLM research and deployment is contingent upon sustaining openness, modularity, explainability, reproducibility, and ethical responsibility, thus ensuring trustworthy, equitable, and societally beneficial language technologies.

ACM Reference Format:

. 2025. Reasoning, Replicability, and Benchmarking in Large Language and Foundation Models: Methodologies, Challenges, and Pathways Toward Trustworthy, Interpretable, and Inclusive AI. In . ACM, New York, NY, USA, 36 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Advancements in AI systems hinge on the rapid progress of reasoning capabilities, benchmarking practices, and a critical appraisal of model architectures. This survey offers a comprehensive synthesis of literature focusing on the current landscape of reasoning within AI, evaluating benchmark datasets, model evaluation protocols, and divergent approaches (including both neural, symbolic, and hybrid paradigms). We analyze how these benchmarks and reasoning tasks have co-evolved with state-of-the-art models, revealing not only strengths, but exposing gaps that persist in robustness, generalization, and interpretability.

Benchmarking remains foundational for tracking intelligence progress, motivating rigorous evaluation of reasoning in environments spanning language, vision, multi-modal inputs, and interactive tasks. Comparative studies increasingly draw attention to the merit and limitation of widely adopted datasets and evaluation protocols, highlighting their impact on the apparent progress of current models. Ensuring that benchmarking procedures genuinely diagnose underlying reasoning abilities—rather than pattern memorization or dataset artifacts—is vital for honest scientific progress. This survey contrasts various reasoning benchmarks and summarizes these in Table 2, which showcases the diversity, coverage, and targeted reasoning skills across leading datasets.

Architectural innovations play a central role in advancing AI reasoning. The field has seen a proliferation of approaches, from end-to-end neural methods (e.g., transformers), symbolic systems, to hybrid models fusing connectionist and logic-based reasoning. While transformer-based architectures have yielded impressive empirical results, critical evaluations probe their capacity for systematic generalization, compositional reasoning, and multi-step logical inference. We explicitly address critiques of these models, and juxtapose neural and hybrid strategies, synthesizing their respective advantages and open challenges.

In reevaluating these themes, we provide a focused engagement with alternative perspectives, including discussions of criticisms leveled against dominant paradigms (such as hidden brittleness or superficial learning in transformers and hybrid architectures). Where appropriate, we summarize these competing viewpoints and highlight ongoing debates regarding transparency, fairness, and practical adoption.

This survey is structured as follows. Section 2 details the reasoning benchmarks and their evaluation methodologies, with in-text summary tables reinforcing critical comparative insights. Section 3 analyzes architectural families, summarizing hybrid and alternative reasoning approaches. We conclude with a discussion of current challenges and future outlook, presenting a distilled summary of key takeaways at the close of each section.

Table 1: Representative Reasoning Benchmarks: Domains and Key Evaluation Aspects

Benchmark	Domain	Core Reasoning Skills	Evaluation Protocols
BoolQ	Language	Boolean reasoning, reading comprehension	Accuracy, human verification
DROP	Language	Discrete operations, multi-step reasoning	Exact match, precision/recall
CLEVR	Visual	Compositional, relational reasoning	Program execution, accuracy
ARC	Language/Logic	Common-sense, abductive, analogy reasoning	Human baselines, automated scoring
HotpotQA	Multi-modal	Multi-hop, supporting fact identification	Exact match, F1, supporting facts

At a glance, this survey aims to equip both domain experts and interdisciplinary readers with a critically balanced, up-to-date account of reasoning advances, major benchmarks, and the state of model evaluation in AI. To foster seamless synthesis of key developments, each subsequent section culminates with an explicit summary distilling the principal insights and open questions addressed. By deeply engaging recent literature and benchmarking evolutions, this survey seeks to both inform and critically examine the trajectory of AI reasoning research, equipping researchers with an integrated view for future inquiry.

1.1 Overview of Large Language and Foundation Models (LLMs)

The evolution of artificial intelligence (AI) has been profoundly shaped by advances in language understanding and generation. The trajectory spans from symbolic, rule-based systems—characterized by explicit grammatical rules and formal symbolic manipulation—to statistical methods, neural, and deep learning architectures. Early symbolic approaches excelled in interpretability but were hindered by a lack of scalability and the brittleness of handcrafted rules. The advent of statistical models, and subsequently neural network architectures, marked a paradigm shift by enabling data-driven learning of complex linguistic patterns. This progression culminates in large-scale Transformer-based models, wherein pre-trained language models (PLMs), especially large language models (LLMs), distinguish themselves through scale and the emergence of novel capabilities.

Distinctive behaviors—such as in-context learning and abstract reasoning—emerge in LLMs due not solely to increased parameter counts, but also to innovations in model design, architecture, and training paradigms. Key developments include:

The adoption of large-scale, unsupervised pre-training; Attention mechanisms, as popularized by the Transformer architecture; Alignment of model objectives with downstream utility.

The launch and societal integration of models such as ChatGPT exemplify LLMs' transformative effect on not only conventional natural language processing (NLP) tasks, but also on digital interaction, information retrieval, content creation, and scientific discovery. Concomitantly, there has been renewed interest in hybrid algorithmic-neural approaches and neural-symbolic (NeSy) systems. These are motivated by enduring challenges—particularly in reasoning and interpretability—where pure neural architectures, despite their success, fall short [69, 91, 95, 105]. A move toward models exhibiting compositionality and explicit knowledge manipulation reflects the AI community's recognition that human-like reasoning and adaptability may require synthesizing symbolic and subsymbolic learning, an imperative for ongoing advancements toward artificial general intelligence (AGI).

1.2 The Critical Role of Reasoning, Replicability, and Benchmarking

The expanded potential of LLMs introduces foundational challenges. Chief among these is cultivating robust reasoning abilities within LLMs that transcend mere pattern recognition or correlation. Although large-scale models demonstrate emergent capabilities in abstract reasoning and commonsense inference, such performance is inconsistent—often susceptible to dataset biases and lacking true compositionality. This motivates the investigation of model architectures and inductive biases that explicitly encode algorithmic or symbolic reasoning procedures.

Neural-symbolic computing (NeSy) has emerged as a promising paradigm, aiming to combine the transparent manipulation of knowledge found in symbolic systems with the flexible data-driven learning of neural networks. Empirical advancements within NeSy frameworks attest to concrete progress in domains demanding structured reasoning—such as scientific discovery, mathematical problem solving, and knowledge-intensive tasks—where traditional end-to-end neural models frequently encounter limitations. Despite these strides, major challenges persist:

Scalability of hybrid models integrating large structured knowledge bases; Efficient inference and reasoning over complex data; Achieving compositional generalization beyond seen examples; Seamless integration of symbolic knowledge acquisition into neural learning pipelines.

These open research problems highlight the incomplete nature of current methods and the ongoing need for innovation in neural-symbolic integration [69, 95].

As LLMs proliferate in research and industry, the importance of replicability and robust benchmarking has intensified. Widely-used evaluation metrics often fail to accurately reflect the subtlety of advanced reasoning behaviors and the adaptability required by practical deployments. This gap necessitates the development of comprehensive benchmarks addressing not only accuracy but also properties such as robustness, out-of-distribution generalization, and fairness. Compounding these technical challenges are issues of opacity and reproducibility, as proprietary models and undisclosed datasets undermine transparency and accountability in both research and societal applications.

Allied to these technical and practical challenges are pressing societal, ethical, and policy considerations—spanning algorithmic bias, misinformation, and the impacts of automating language-centric labor. Therefore, cultivating rigorous, transparent, and replicable

research practices constitutes a linchpin for both scientific progress and public trust in LLM technologies [69, 91, 105].

1.3 Survey Structure and Scope

Given these multifaceted themes, this survey provides a structured synthesis of the technical, methodological, and societal dimensions defining contemporary LLM research. The survey first examines evaluation methodologies and benchmarking strategies, with an emphasis on recent advances in linguistic competence, robustness, and inclusivity. Subsequently, the intersection of LLMs with algorithmic reasoning and neural-symbolic integration is scrutinized, highlighting technical obstacles and emerging opportunities in the quest for more reliable and general AI systems. Principles of open science and reproducible research are afforded particular attention, acknowledging their foundational role in mitigating societal risks and advancing the field. By organizing the discussion thematically, the survey seeks to equip readers with a critical appreciation of both progress to date and the grand challenges shaping the next frontier of large-scale, language-centric AI [69, 91, 95, 105].

2 Historical and Foundational Landscape

This section reviews the foundational approaches and key developments that have shaped the evolution of AI, with particular attention to reasoning architectures and benchmarking methodologies. We frame the analysis to highlight both prevailing paradigms and alternative perspectives, providing comparisons where relevant to clarify their respective strengths and limitations.

Section Summary

In summary, this section establishes the historical context underpinning current advances in AI reasoning systems. By outlining the core reasoning architectures and pivotal benchmarking methodologies, we clarify the foundational landscape upon which emerging large language models (LLMs) and novel frameworks are built. This foundation sets the stage for the detailed exploration of recent progress and taxonomy innovations in subsequent sections.

2.1 Early Approaches and Hybrid Models

The historical trajectory of AI reasoning systems has been characterized by an initial dominance of symbolic methods, including expert systems and rule-based engines. These approaches offered transparency and explicit logic structuring but often struggled to scale or handle ambiguity. The subsequent emergence of connectionist models introduced learning-based solutions, trading off interpretability for empirical performance improvements.

Hybrid models, combining symbolic and subsymbolic techniques, have been proposed to bridge these shortcomings. While hybridization seeks a synthesis between structure and flexibility, critics have argued that such approaches can inherit limitations from both parent paradigms, such as the brittleness of symbolic reasoning and the opacity of neural systems. The debate remains active, and a nuanced appraisal of these models is essential when considering their theoretical and practical implications.

2.2 Benchmarking and Reasoning Evaluation

Benchmarks play a vital role in evaluating the progress of reasoning systems. Early benchmarks focused on narrow, well-defined logical tasks, permitting rigorous comparison but often failing to represent real-world complexity. Over time, the field has moved toward more diverse and challenging benchmarks that span language understanding, abstraction, and multi-step reasoning.

Table 2 provides an overview of representative benchmarks, their focus, and evaluative criteria, reinforcing the diversity and evolution of reasoning assessment.

2.3 Transformers and Recent Paradigms

The advent of transformer architectures has markedly shifted the landscape of both perception and reasoning. These models have demonstrated unprecedented performance across benchmarks but prompted debate regarding the genuine nature of their reasoning abilities versus statistical pattern recognition. Competing views question whether the inductive capabilities observed in transformers should be considered reasoning in the classical sense or rather as an emergent byproduct of large-scale data assimilation.

Critiques have also centered on the interpretability and controllability of such models, with some arguing that their success challenges traditional definitions of reasoning and intelligence. This ongoing discourse underscores the need for nuanced evaluation strategies and theoretical frameworks that can accommodate the complexity of modern AI.

2.4 Section Summary

In summary, the historical and foundational landscape of AI reasoning encompasses a rich interplay between symbolic approaches, connectionist models, hybrid architectures, and recent transformer-based advances. Each paradigm brings distinct advantages and trade-offs, reflected in the evolving design of benchmarks and evaluation criteria. Ongoing debates regarding hybrid systems and transformer-based reasoning highlight the field's dynamism and the importance of comprehensive, comparative assessment.

2.5 Evolution of Reasoning in AI

The development of artificial reasoning systems reveals a progressive trajectory from early symbolic logic frameworks to the current preeminence of neural and transformer-based paradigms. Classical AI focused on symbolic representations, rule-based inference mechanisms, and logic programming, prized for interpretability and transparency [69, 91, 95, 105]. These methods supported precise deductive reasoning, yet were brittle and struggled in open, ambiguous, or real-world domains, often demanding labor-intensive manual construction of knowledge bases [91].

The emergence of connectionist models, notably deep neural networks, marked a paradigm shift toward data-driven learning. These architectures facilitated the automatic synthesis of hierarchical abstractions, allowing systems to address broad reasoning problems without hand-crafted logic [69]. However, neural networks demonstrated persistent deficits in generalization for tasks needing compositionality, recursion, or algorithmic reasoning—areas

Table 2: Key Benchmarks in AI Reasoning and Their Evaluative Focus

Benchmark	Reasoning Type	Task Domain	Evaluation Criteria
Early Logic Puzzles	Symbolic Deduction	Mathematical/Logical	Accuracy, Formal Correctness
Winograd Schema	Commonsense Reasoning	Natural Language	Disambiguation, Context-Dependence
bAbI Tasks	Multi-step Reasoning	Synthetic QA	Step-wise Inference, Scalability
ARC Challenge	Abstract Reasoning	Visual/Pattern Recognition	Generalization, Abstraction

where symbolic approaches retained strength, especially in disciplines such as arithmetic, logic, combinatorics, and structured multi-step problem-solving [95, 105]. To address these limitations, hybrid neural-symbolic (NeSy) models were created to merge the perceptual capabilities of neural networks with explicit, interpretable symbolic inference [69, 95]. Evidence indicates these integrated frameworks enhance performance in domains like mathematical problem-solving and retrosynthetic analysis, particularly for tasks requiring structured or multi-step reasoning [95]. Nevertheless, major challenges persist, including robust compositional generalization, scalable reasoning over large knowledge repositories, and seamless neural-symbolic integration, thus making effective unification a persistent open challenge [91, 95].

In recent years, the domain has been rapidly transformed by the introduction and maturation of transformer-based large language models (LLMs)—such as GPT, T5, PaLM, LLaMA, and Flan—each leveraging large-scale pre-training on diverse corpora [10, 25, 38, 45, 54, 55, 64, 76, 91, 99, 105]. These models exhibit emergent reasoning capabilities across arithmetic, logic, and algorithmic tasks, especially when advanced prompting techniques such as chain-of-thought (CoT) prompting are employed [38, 99, 105]. CoT prompting encourages models to generate explicit intermediate steps, leading to substantial performance improvement on multi-step and complex reasoning challenges compared with zero- or few-shot methods [38, 99, 105]. For instance, a few CoT exemplars enable large LLMs to surpass fine-tuned baselines on mathematical word problems [99]. Despite this, systematic studies highlight enduring gaps between the reasoning abilities of present-day LLMs and human experts, particularly for tasks demanding abstraction, compositional logic, or the synthesis of broad world knowledge [10, 45, 64]. While advanced models like GPT-4 can convincingly generate rationales for intricate clinical or scientific tasks, providing some interpretability benefits [76], logical errors remain common; such errors are especially concentrated in incorrect outputs (e.g., observed in 65% of GPT-4's erroneous clinical rationales) [76]. This suggests that although LLM outputs may mimic human-like reasoning formats, underlying processes often remain pattern-driven and statistically emergent rather than human-like in abstraction or reliability. Furthermore, these models remain susceptible to hallucinations, brittle generalization, and performance drops when required to reason with information outside their training distribution or within lengthy contexts [25, 45, 55, 64, 76, 105].

Empirical investigations indicate that LLM performance on reasoning tasks is highly sensitive to prompt design, exemplar choice, and mechanisms for knowledge retrieval [10, 38, 54, 76, 99]. Persistent reasoning failures are observed in domains such as multi-step logical inference, combinatorial puzzles, and causal reasoning [10, 38, 76]. Retrieval-augmented CoT prompting has delivered improvements in scientific and mathematical multimodal tasks by assimilating external information dynamically [54, 99, 105], though these advances do not comprehensively resolve challenges in compositional generalization or causal inference [45, 55, 64, 91]. Overall, the body of evidence suggests that while transformer-based LLMs mark a considerable leap in automated reasoning, current abilities are largely emergent, pattern-based, and stochastic, lacking consistent grounding in explicit abstraction or systematic causal modeling [38, 45, 64, 99]. This gap highlights key directions for future research in hybrid, neural-symbolic, and biologically inspired approaches to advance the scope and reliability of AI reasoning [91, 95, 105].

2.6 Embedding and Model Architecture Developments

The foundation of modern natural language processing and reasoning systems is closely intertwined with advances in representation learning—particularly in embedding methods—and architectural design. In the context of this survey, the overarching goal is to outline the trajectory of embedding and model architectural progress, clarify their impact on downstream reasoning and structured data tasks, and crystallize key methodological questions left open by current benchmarks. Specifically, we examine: How do recent developments in representations and architectures facilitate (or impede) effective reasoning across increasingly diverse and long-context inputs? What are the primary gaps between general-purpose models and the demands of structured or multimodal tasks?

Early approaches utilized static, dense embeddings to encode lexical relationships; the transition to contextualized embeddings, most effectively realized in transformer architectures, represented a qualitative leap in modeling semantic, syntactic, and higher-order structural relations between tokens and modalities [23, 31, 79, 93]. Models such as BERT, GPT, and their derivatives leverage deeply stacked attention layers, enabling the encoding of rich, context-dependent linguistic meaning [75]. Techniques like SBERT-WK, which dynamically aggregate BERT's internal representations, further extend semantic alignment and resilience to contextual variation by dissecting word representations across all BERT layers and employing principal component analysis to emphasize unique and novel word contributions. This approach achieves state-of-the-art results on semantic similarity tasks, offering an efficient, training-free alternative to supervised fine-tuning [93].

Table 3: Summary of foundational paradigms in AI reasoning, with comparative strengths and limitations.

Paradigm	Core Mechanisms	Strengths	Key Limitations
Symbolic (Rule-based, Logic)	Explicit symbols, rules, logic programs	Interpretability, rigorous deduction, transparency	Brittle generalization, manual knowledge engineering
Neural (Connectionist, Deep Learning)	Hierarchical, distributed representations; learning from data	Strong pattern recognition, adaptability, implicit abstraction	Weakness in compositional reasoning, limited interpretability
Neural-Symbolic (Hybrid)	Joint neural and symbolic modules; integration architectures	Combines perception with explicit inference, improved generalization on structured tasks	Integration complexity, compositional generalization, scalability
Transformer-based LLMs	Attention-based contextual encoding; large-scale pre-training	Emergent reasoning, multi-task capability, scalability	Reliant on statistical learning, lacks explicit abstraction or robust causality

Transfer learning—particularly via pre-trained checkpoints from models such as BERT, GPT-2, and RoBERTa—has become the predominant modality for adapting large-scale models to downstream tasks with minimal additional training [75]. This paradigm shift, extensively detailed by Rothe et al. [75], has democratized access to high-performing models by dramatically reducing the compute and data resources required for competitive performance. Importantly, such pre-trained checkpoints have also been shown to deliver strong results for sequence generation tasks, including machine translation, summarization, and sentence fusion, by enabling both encoder and decoder initialization from public checkpoints. In parallel, innovations in self-supervised learning, multimodal integration, and speech-text modeling have expanded the capacity of transformer models to operate across text, image, tabular, and speech inputs [31, 93].

Additional architectural innovations have emerged to meet the challenges of specific input modalities and reasoning requirements. For example, encoder-decoder architectures now increasingly incorporate structural priors to model tabular data and document salience, supporting improved summarization and long-context reasoning. Retrieval-augmented transformers integrate external information sources to bolster reasoning fidelity and are being actively explored for their potential in multi-hop and open-domain QA [75, 79, 93].

Despite rapid progress, significant architectural and methodological limitations remain:

Models frequently underperform when processing extended input contexts, with accuracy declining as relevant information is dispersed across longer sequences. Liu et al. [55] demonstrate that even models designed for long-context reasoning exhibit sharp performance drops, especially when relevant cues appear mid-sequence, raising questions about how information utilization should be evaluated and compared across model families. Standard embedding mechanisms, while adept at capturing local semantic and syntactic dependencies, are often less effective for highly structured data (e.g., tables, knowledge graphs) in the absence of specialized encoders or attention mechanisms [23, 75, 93]. Innovations such as structured attention, field-content selective encoders, and advanced pooling strategies are being developed to bridge this gap. For instance, SAN-T2T [23] introduces a selective attention network and content selector specifically geared towards generating descriptive text from tabular data. UOTSum [79] jointly learns alignment and abstractive summarization of long documents via unbalanced optimal transport, achieving state-of-the-art results but raising new challenges in terms of interpretability and computational cost. To illustrate these recent advances, Table 4 summarizes the performance of prominent models on the task of long-document summarization, as evaluated by Shen et al. [79]:

This benchmarking highlights both progress and the persistent methodological issue: as models become increasingly capable, the evaluation of their performance on diverse reasoning tasks (particularly those involving extended or structured contexts) must itself evolve to more accurately reflect real-world requirements. For example, further research is needed to answer: What evaluation protocols best capture the nuanced information utilization in long and structured contexts [55]? How should benchmarks be adapted to penalize information loss, redundancy, or misalignment in model outputs [79]? What architectural innovations most effectively generalize across multiple input modalities, and where do common approaches fall short?

Summary and Future Directions. In summary, the development of embedding techniques and model architectures has enabled significant gains in language understanding and reasoning across a variety of tasks and domains. However, notable research gaps include: (1) more robust handling and evaluation of long-context and structured input processing (2) the need for transparent, scalable benchmarking practices that reveal subtle performance degradations or limitations (3) development of unifying modeling principles or architectures that seamlessly integrate across textual, tabular, and multimodal data. Addressing these gaps is essential for the next generation of reasoning systems and remains a pivotal set of open methodological and scientific questions for the field. This analysis leads naturally to the next sections, which probe evaluation frameworks and specialized benchmarks in greater depth.

2.7 Biological Inspirations and Neuromorphic Approaches

An increasingly impactful trajectory in the development of reasoning-enabled AI is the incorporation of principles drawn from biological and cognitive neuroscience. The structural organization and dynamic properties of biological connectomes are widely hypothesized to underpin the cognitive flexibility and generalization observed in human reasoning. Inspired by this, neuromorphic systems and reservoir computing models have been designed to emulate key features of brain networks, notably modularity and criticality [83]. Recent empirical findings suggest that reservoir computing architectures that incorporate brain-inspired topologies consistently outperform architectures with random connectivity, particularly on tasks requiring flexible generalization and adaptive reasoning capabilities. This highlights the computational advantages inherent in functional segregation and integrated network dynamics [83].

The manifold benefits of biologically inspired architectures can be summarized as follows: they serve as explanatory models for the origins of cognitive flexibility and compositionality in biological reasoning systems [83]; they guide the development of artificial reasoning systems with enhanced efficiency, adaptability, and robustness—especially in contexts characterized by uncertainty and ambiguity; and they inspire integrative approaches that blend cognitive, neural, and symbolic paradigms, targeting the recursive and adaptive reasoning abilities found in biological intelligence [83, 95].

Table 4: Performance comparison of representative models on long-document summarization benchmarks [79]. Metrics: R-1 (ROUGE-1), R-2 (ROUGE-2), R-L (ROUGE-L).

Model	R-1	R-2	R-L
Lead-3	40.3	17.7	36.7
Pointer-Generator	41.2	18.0	37.8
BERTSUMEXTABS	42.1	19.2	38.6
Longformer-Encoder	43.6	20.0	39.9
UOTSum (Ours)	**44.7**	**21.3**	**41.0**

While these advances move the field forward, several significant open questions remain. One of the primary methodological challenges involves understanding how the structural modularity and criticality observed in neuromorphic networks can be robustly mapped to the algorithmic flexibility and generalization abilities required by artificial systems. Recent benchmarking efforts have exposed performance disparities when biologically inspired networks are evaluated on tasks that demand both compositional reasoning and transfer to novel domains [83]. This reveals methodological gaps in current evaluation protocols, such as the need for new standardized benchmarks that explicitly target adaptive reasoning and compositional generalization under uncertainty [83, 95]. Furthermore, the lack of rigorous criteria for quantifying cognitive flexibility in artificial systems limits direct comparison across architectures.

Future research should address these gaps by posing targeted questions such as: How can structure-function relationships uncovered in neuroscience be operationalized as architectural constraints for scalable AI reasoning models? What empirical protocols can best capture adaptive, transferable reasoning strategies in neuromorphic or hybrid systems [83, 95]? How can current benchmarking frameworks be extended to assess not only task success, but also reasoning robustness and adaptability?

In summary, the historical and foundational landscape of AI reasoning is shaped by the interplay between symbolic, neural, and hybrid paradigms; innovations in knowledge representation and network architecture; and the growing influence of neuroscience-inspired methodologies. Each trajectory provides distinct strengths and faces unique limitations (see Table 3), collectively shaping and informing the ongoing evolution and future directions of reasoning-enabled AI research [69, 75, 83, 91, 95, 105]. After examining the comparative perspectives and limitations summarized in Table 3, several future gaps emerge, particularly in bridging theoretical advances in neuromorphic modeling with scalable methodologies and interoperable benchmarking standards. Continued research at the interface of neuroscience, artificial reasoning, and evaluation methods is needed to realize robust, adaptable reasoning in AI systems.

3 Benchmarking Speech and Language Models

Benchmarking is foundational to the progress of speech and language technology, serving as both a measure of advances and a catalyst for new research directions. This section critically examines the landscape of benchmarking approaches for speech and language models, with the dual objectives of mapping methodological evolutions and illuminating how benchmarking choices shape model development and societal deployment. In doing so, we aim to address the overarching research questions of this survey: What are the prevailing frameworks and challenges in benchmarking, and how do they influence real-world applicability and understanding of model limitations?

We begin by reviewing benchmarking protocols and datasets, before progressing to evaluation metrics and methodological issues. The interplay between these components not only determines the perceived capabilities of speech and language models but also exposes gaps for future inquiry. For instance, the selection of a particular benchmark can amplify certain model behaviors while concealing others, thus affecting downstream deployment and possible societal consequences. As such, careful consideration is needed in both designing and interpreting benchmark results, as highlighted in subsequent subsections.

Throughout this section, we reiterate the significance of open methodological issues revealed by emerging benchmarking frameworks. Notably, as community standards and evaluation infrastructures evolve, questions arise concerning metric volatility, the representativeness of tasks, and the transferability of benchmarked performance to real scenarios. These open questions motivate further research into robust, adaptable, and context-sensitive evaluation protocols.

After our detailed examination of benchmark types and associated metrics, we synthesize the principal limitations identified in current practice. Chief among these are the risk of overfitting to static benchmarks, insufficient coverage of real-world variation, and under-explored implications for model fairness and societal impacts. Addressing these challenges necessitates both methodological innovation and closer integration of evaluation design with deployment considerations.

In summary, benchmarking speech and language models is not a neutral exercise but a critical site where technological ambitions, methodological rigor, and societal impacts intersect. We conclude this section by highlighting research gaps—such as the need for benchmarks that better capture complex linguistic phenomena, evaluation metrics that accommodate dynamic model behaviors, and frameworks for assessing contextual and ethical dimensions of deployment—that should guide future studies in this evolving field.

3.1 Standardized Frameworks and Leaderboards

The evaluation of speech and language models has evolved significantly with the emergence of standardized benchmarking frameworks and public leaderboards, enabling systematic assessments of generalization, robustness, and task coverage. In speech processing, the Speech processing Universal PERformance Benchmark (SUPERB) provides a comprehensive, extensible, and reproducible platform to evaluate foundation models on 15 diverse tasks, including phoneme recognition, keyword spotting, speaker identification, emotion recognition, and automatic speech recognition. SUPERB employs unified evaluation protocols and rigorous multi-task procedures, such as the use of fixed feature encoders paired with lightweight task-specific prediction heads, and leverages a learnable weighted-sum approach for aggregating information across model layers—shown to yield performance gains in most tasks except specific cases like voice conversion. The benchmark supports robust statistical aggregation and involves 33 models spanning both self-supervised and conventional paradigms, with analytical emphasis on reproducibility, statistical significance testing, and the continued expansion of open-source resources. This infrastructure allows for community-driven growth and has accelerated consensus on performance and limitation identification, revealing key vulnerabilities, particularly in generative and low-resource scenarios, as well as underscoring that leaderboard differences may not always reflect significant differences in capability [68, 103].

Similarly, in natural language processing (NLP), frameworks such as HELM and DIoR provide scenario-based, methodologically robust leaderboards covering a broad task landscape, from Wikipedia and news articles to biomedical texts. These frameworks advance beyond surface metrics, including societal impact, reliability, and efficiency metrics, and advocate for quantitative, systematic approaches to benchmark design. For example, DIoR assesses how decisions on scenario, dataset and evaluation aggregation impact reliability, highlighting that the removal of entire datasets from benchmarks can substantially degrade trustworthiness and reproducibility, while simply reducing sample counts is less damaging due to model ranking stability. Moreover, the Mean Win Rate (MWR) metric, central in HELM, is sensitive to leaderboard composition changes, suggesting the importance of transparent aggregation strategies [68]. These findings inform concrete guidelines for efficient yet robust benchmark construction and highlight the central role of well-curated, domain-diverse datasets in ensuring comprehensive and reliable multi-domain evaluation, as reinforced by large-scale studies showing that benchmark composition critically affects model rankings and measured performance, especially as domain and task diversity expands [47].

A major recent advance is the introduction of continual learning benchmarks such as CL-MASR for multilingual automatic speech recognition. CL-MASR systematically structures task and language sequences to expose deficiencies in models' ability to acquire new capabilities without catastrophic forgetting. The benchmark delivers reproducible task sequences and a rich set of metrics—including Word Error Rate, levels of forgetting, backward transfer, and intransigence—facilitating detailed evaluation of catastrophic forgetting, cross-lingual interference, and resource imbalance, particularly in low-resource or typologically diverse settings. Investigations within CL-MASR reveal that language ordering, resource imbalances, and cross-lingual effects substantially influence continual learning outcomes, regardless of mitigation strategy. The publicly released codebase standardizes a domain previously lacking in systematic tools, promoting collaborative research and rapid progress in the field [53]. Collectively, these trends exemplify elevated expectations for multi-domain, resource-robust, reliable, and reproducible benchmarking in the study and deployment of both speech and language models.

3.2 Evaluation Metrics and Best Practices

As introduced in our survey, a central objective is to systematically assess how benchmarking practices and evaluation metrics shape the reliability, interpretability, and real-world relevance of model evaluations across AI domains. This section advances that goal by critically examining current metric selection practices and best practices in benchmark design, with a focus on alignment to human-centered objectives and actionable guidance for future development.

The effectiveness of benchmarks is fundamentally dependent on the alignment between evaluation metrics and human-centered objectives. Automated metrics including ROUGE, BLEU, and METEOR have long served as mainstays in tasks such as summarization, simplification, and machine translation. However, these metrics typically correlate only weakly with human judgments of meaning, comprehension, and utility, particularly for complex tasks such as plain language summarization and biomedical natural language processing [16, 28, 36, 47, 68, 81, 103]. For instance, recent work in medical plain language summarization found that, while ROUGE and similar metrics suggest LLM-generated outputs are comparable to human writing, objective comprehension tests with lay participants reveal a substantial gap: only QA-based metrics like QAEval reflect true understandability and faithfulness [36]. Likewise, in chemical space exploration and biomedical NLP, surface-level metrics may fail to distinguish between models of genuinely different quality, necessitating more semantically informed alternatives [16, 81].

To address these gaps, evaluation approaches have shifted toward semantically grounded metrics that better reflect human preferences and understanding. Methods such as cross-encoder or bi-encoder models, fine-tuned for semantic similarity or natural language inference, now consistently outperform traditional n-gram overlap measures in both general and domain-specialized contexts [16, 28, 81]. For example, leveraging inference-based or QA-based metrics, as shown in biomedical and simplification benchmarks, provides stronger alignment with human assessments of comprehension and utility, especially in layperson-facing and specialized language generation applications [16, 28, 36, 47].

Despite such advancements, several challenges in metric selection persist and have, on occasion, led to misleading conclusions in the field. For example, leaderboard rankings can prove highly volatile: analyses of Decision Impact on Reliability (DIoR) within the HELM benchmark demonstrate that moderate changes in scenario grouping, dataset selection, or evaluation aggregation can unpredictably shift the relative standing of language models, sometimes reversing previous conclusions about model performance [68]. In the SUPERB speech benchmark, statistical analyses indicate that

observed leaderboard differences among top models are often statistically insignificant, cautioning against over-interpretation of small performance gaps [103]. In sentence simplification (BLESS) and biomedical NLP, evaluation instability remains a concern, as rankings shift with metric choice and evaluation setup [16, 47].

The resultant volatility of metric-based leaderboards in response to such changes underlines the necessity for actionable best practices. We recommend transparent and precise definition of metrics, comprehensive statistical reporting (including significance testing to avoid misinterpretation of minor differences), and clear articulation of scenario aggregation and evaluation protocols [39, 68, 103]. Furthermore, composite or scenario-weighted evaluation methodologies are increasingly advocated to ensure reliable and representative assessment across model capabilities [39, 47, 68]. Recent benchmarks, such as BLESS for sentence simplification [47] and the Speech processing Universal PERformance Benchmark (SUPERB) [103], exemplify the trend toward domain-specific, multifaceted evaluation frameworks with rigorous reproducibility and statistical safeguards. These works stress open-source code, publicly available datasets, and reproducible pipelines as foundational for community trust and scientific rigor [39, 47, 103].

For benchmark and metric developers, these insights yield several actionable recommendations: prioritize semantically and comprehension-grounded metrics over surface-level measures; systematically report statistical significance of leaderboard differences; design evaluation protocols to minimize volatility induced by scenario or aggregation choices; and ensure all datasets, code, and evaluation procedures are public and thoroughly documented. The growing body of benchmarks from late 2023 and 2024, such as BLESS and new HELM variants, reinforce these principles and offer blueprints for future robust, human-aligned evaluation design [47, 68, 103].

To ensure reproducibility and scientific rigor, it is imperative to publicly release datasets, code, evaluation procedures, and, when feasible, simulated or derived data, as recommended by standards in applied linguistics and benchmarking research [39].

3.3 Comparative Analysis and Diversity

This section advances the overall survey objective of critically synthesizing current benchmarking practices for language and foundation models, with a particular emphasis on cross-domain applicability, methodological rigor, and implications for the evolution of evaluation standards. As models and evaluation methodologies diversify, a nuanced understanding of comparative trends, volatility, and benchmark robustness is essential for guiding model assessment and development.

A principal focus in recent benchmarking efforts is the systematic comparison of large language models (LLMs) and foundation models with both traditional baselines and alternative architectures. Cross-domain benchmarking—such as evaluating state-of-the-art (SOTA) fine-tuned models (e.g., BioBERT, PubMedBERT, BART) versus LLMs (e.g., GPT, LLaMA) in biomedical NLP—shows that while LLMs frequently achieve superior performance on tasks requiring generative reasoning or medical question-answering, they often do so at a substantially higher computational cost. Moreover, without additional task-specific adaptation, LLMs may still lag behind fine-tuned models in extraction, classification, and domain-specialized settings [47]. For instance, generative models like GPT-4 tend to produce outputs of high fluency for summarization and simplification, but these may be less complete or more susceptible to hallucinations compared to specialized baselines. Furthermore, marked variability is observed in the repertoire of edit operations and strategies employed by different LLMs in tasks such as text simplification, indicating heterogeneity in their methodological approaches [47].

For a concise overview of such comparative results, see Table 5.

Recent studies have highlighted volatility in benchmark outcomes, especially where evaluation protocols or scenario composition fluctuate. For example, the work of Perlitz et al. [68] on the HELM benchmark demonstrates that simply adding or removing models or datasets can alter leaderboard rankings and perceived model superiority, sometimes misleading the field about genuine progress. Notably, aggregation strategies like grouping diverse datasets may yield lower evaluation reliability, and an overemphasis on the number of test examples may not translate to greater stability. This indicates the need for statistically grounded metrics, such as Decision Impact on Reliability (DIoR) [68], when designing benchmarks. Similarly, in continual learning for speech recognition, Della Libera et al. [53] show that ordering of languages or choice of resource splits can skew comparability; certain strategies overstate model resilience due to scenario sequencing rather than true model robustness.

For developers of benchmarks and metrics, these findings advocate for several best practices: (1) Articulate and minimize sources of volatility by transparently defining scenarios, datasets, and ranking metrics; (2) Employ reliability measures to evaluate the stability of results under perturbations of experimental setup; (3) Design with efficiency in mind, as both environmental and resource constraints are increasingly relevant—approaches like Flash-HELM [68] can offer computational savings without sacrificing reliability.

Periodically, the field is reoriented by the release of new benchmarks tailored for extensibility, multilinguality, or continual learning. Examples emerging in late 2023 and 2024 include BLESS [47] (targeting LLM evaluation on sentence simplification with analyses of edit diversity and robustness) and CL-MASR [53] (addressing continual learning in multilingual ASR, with evaluation protocols probing catastrophic forgetting, transfer, and efficiency). These benchmarks are accompanied by open-source resources and standardized evaluation frameworks to facilitate reproducibility and sustained advancement.

Benchmark development has also prioritized diversity and inclusivity, with a marked shift toward constructing resources that encompass broader linguistic, cultural, and task-scale variability. These advances ensure fairness in model assessment and promote research that generalizes beyond canonical datasets or majority language contexts [47]. Emerging benchmarks are designed for extensibility and adaptability, supporting, for instance, multilingual task sequences or modular scenario expansion, while emphasizing open sharing of resources to catalyze community-led progress [47, 53, 68]. Adherence to these principles in benchmark creation and deployment enables robust comparative analyses and is essential for driving sustainable progress in speech and language modeling research.

Systematic benchmarking using unified frameworks and rigorous protocols advances community consensus on model strengths

Table 5: Representative comparative outcomes between SOTA fine-tuned models and LLMs on biomedical NLP tasks. Values indicate relative strengths as identified in recent benchmarking studies.

Task	BioBERT/BART	GPT-4/LLaMA	Notes
Extraction & Classification	Superior	Inferior	Fine-tuned models excel; require less adaptation
Medical QA	Moderate	Strong	LLMs perform well, esp. with complex queries
Generative Summarization	Moderate	Superior	LLMs enhance fluency, some risk of hallucination
Text Simplification	Specialized	Diverse	LLMs deliver varied strategies and edit diversity
Computational Cost	Efficient	Substantially Higher	LLMs demand greater resources

and weaknesses. The ongoing evolution of evaluation metrics emphasizes alignment with human judgment, particularly in complex and layperson-facing tasks. Comparative studies reveal fundamental trade-offs between LLMs and fine-tuned domain-specific models, reinforcing the ongoing need for careful task adaptation and judicious resource allocation. Finally, foregrounding diversity, extensibility, and open scientific practices will be essential for future-proofing benchmarks and maximizing their impact across domains.

4 Probing, Reasoning, and Linguistic Competence Benchmarks

This section provides a comprehensive overview of benchmarks that assess language model capabilities through probing tasks, reasoning challenges, and the evaluation of linguistic competence. Our objective is to clarify the purpose and structure of prominent benchmarks within this space, highlighting their design philosophies, target skills, and relevance to the AI and NLP research communities.

We begin with a general introduction to each type of benchmark and then examine their methodologies and applications. When transitioning between distinct domains—such as from traditional probing to reasoning, or when discussing clinical versus multimodal benchmarks—subsections are explicitly introduced to facilitate clarity and navigability for the reader. Transitional commentary connects granular benchmark discussions, emphasizing the implications of design choices and evaluation strategies.

For a broader readership, we briefly consider the societal and application-specific implications stemming from the volatility of benchmark results and the challenges of generalizing across domains. Notably, issues of instability or limited transferability can influence the robustness, fairness, and real-world trustworthiness of deployed language models, shaping both downstream applications and public perception.

Throughout, we aim to improve clarity and measurability in articulating the objectives of each benchmark category. Specifically, we characterize performance goals in operational terms such as accuracy, error rates, or coverage of targeted linguistic phenomena, supporting reproducibility and comparison across evaluation suites.

These insights are intended to support readers in identifying appropriate evaluation suites for their specific domains and use cases. At the conclusion of each major subsection, we synthesize key findings and reflect on the broader implications for benchmarking design, including recommendations for future research directions where notable gaps are observed—such as the need for more transferable diagnostic tasks or benchmarks better aligned to complex, real-world deployments.

4.1 Linguistic and Reasoning Probing

In alignment with the core objective of this survey—to critically examine and synthesize advances and outstanding challenges in the evaluation of large language models (LLMs)—this section focuses on probing methodologies that target the linguistic, reasoning, and abstraction abilities of state-of-the-art models. We explicitly consider how evolving benchmarks expose both progress and persistent gaps, and highlight consequential lessons for the development of future metrics and evaluation frameworks.

The evaluation of LLMs increasingly depends on sophisticated probing techniques designed to reveal the nuanced properties of models' internal representations and linguistic behaviors. The evolution of probing for syntactic and semantic competence has progressed from elementary acceptability judgments to methodologically robust frameworks, which now target compositional and structural facets of language. Modern benchmarks, for instance the Two Word Test (TWT), probe models on foundational aspects of semantic composition: specifically, their ability to distinguish between plausible and implausible noun-noun phrases. Crucially, success in this domain requires not just recognition of word similarity but a deeper grasp of semantic combinatorics. Although LLMs demonstrate impressive performance on complex downstream tasks, empirical evidence shows they continue to struggle with the core challenge of semantic discernment. Notably, models such as GPT-4 variants recurrently overestimate the coherence and meaning of nonsensical phrases, indicating a persistent reliance on surface-level statistics (e.g., vector cosine similarity) over robust compositional understanding [73]. This persistent gap highlights a critical mismatch between reported advancements on aggregate language benchmarks and true progress in core linguistic competence.

Highlighting the volatility of benchmark-based conclusions, TWT results [73] show that models like GPT-3.5-turbo and Gemini-1.0-Pro-001 rate nonsensical noun-noun pairs almost as highly as meaningful ones, misleadingly suggesting human-like semantic competence when judged solely by high-level accuracy or unrelated benchmarks. Such volatility has, at times, misdirected perceived progress in the field: models excelling on verbose or logic-heavy benchmarks may still lack fundamental linguistic understanding, as exposed by carefully constructed tests like TWT. Similarly, in the context of metrics for generative chemical models, research has revealed that widely-used metrics often fail to accurately reflect true model quality or generalization ability, prompting a reassessment of which benchmarks genuinely probe for intended competencies [81].

In parallel, syntactic minimal pair benchmarks, exemplified by BLiMP, systematically evaluate models across an extensive array of

morphosyntactic phenomena. BLiMP, through its template-generated sentence pairs, isolates specific grammatical constructs and tests models' sensitivity to grammaticality [92, 97]. While transformer-based models consistently surpass earlier n-gram and LSTM-based language models in phenomena such as subject-verb agreement, they remain prone to inconsistency when faced with deeper syntactic generalizations, including negative polarity and island constraints. This brittleness is further corroborated by classifier-based probing studies, notably Holmes and its computationally optimized extension FlashHolmes, which aggregate results across more than two hundred datasets and encompass a spectrum of phenomena in syntax, morphology, semantics, and discourse [15, 92]. Analysis from Holmes-based studies reveal expected scaling of competence with increased model size, yet also expose nontrivial dependencies on architectural choices and instruction tuning—these effects are especially evident within morphosyntactic domains, thereby emphasizing the importance of both inductive biases and fine-tuning paradigms.

Recent research extends the probing paradigm to include reasoning and abstraction ability, utilizing an increasingly diverse suite of benchmarks. Notably, the Abstraction and Reasoning Corpus (ARC) and subsequent developments within the DreamCoder/PeARL frameworks have shifted focus toward generalization over pattern recognition. Whereas neurosymbolic approaches like DreamCoder specialize in structured transformations via program induction, LLM-based methods augmented with novel encodings and data augmentations excel at orthogonal aspects, with each paradigm addressing complementary subsets of ARC tasks [9, 15, 81, 100]. For example, *PeARL* [9], introduced in 2024, advances recognition models for ARC and demonstrates that neither neurosymbolic nor LLM pipelines can independently solve a majority of cases, but their ensemble achieves improved coverage, surpassing prior approaches such as Icecuber. Ensemble approaches achieve broader coverage, yet no single paradigm independently solves a majority of cases, illustrating the persistent difficulty of abstract reasoning and broad generalization [9, 15, 100]. The release of the open-source *arckit* library [9] further emphasizes the trend toward reproducible, extensible benchmarking environments.

The recent introduction of RGB (Retrieval-Augmented Generation Benchmark) in late 2023 [15] represents a notable advance in evaluating the integration of retrieval and generative capabilities. RGB systematically examines LLMs' abilities in noise robustness, negative rejection, information integration, and counterfactual robustness, revealing bottlenecks such as the inability to reliably refuse unsupported questions, sharp performance drops with increased noise, and consistent struggles when integrating information across documents. The authors urge for careful metric construction and caution against over-interpretation of aggregate scores, providing actionable guidance for both benchmark and metric developers to focus on error detection, document modeling, and cross-document reasoning.

Specialized domains have further spurred the development of targeted benchmarks. Biomedical and clinical reasoning datasets, such as MedS-Bench and arckit, extend probing into domain-specific abstraction and reasoning. Recent work (2025) finds that even the most advanced LLMs, including GPT-4 and Claude-3.5, exhibit divergent abilities between real-world and multiple-choice scenarios; excelling at the latter but consistently underperforming on tasks requiring nuanced clinical information extraction or summarization [9, 15, 100]. The findings underscore the limitations of existing benchmarks in capturing real-world deployment challenges, and argue for a shift toward broader clinical scenario coverage, multilingual expansion, and the validation of metrics against actual task data.

Collectively, the evidence indicates that while advancements in probing and benchmark curation have refined our ability to diagnose LLM limitations, current state-of-the-art models remain highly sensitive to prompt formulation and task structure. Notable gaps persist in the domains of semantic composition, syntactic robustness, and genuine cross-domain abstraction [9, 15, 73, 92, 97]. The volatility and occasionally misleading nature of benchmark metrics highlight the need for granular, transparently designed evaluation tools. For researchers and benchmark developers, this underscores the importance of continued innovation in dataset design—prioritizing not only coverage and challenge diversity, but also the reproducibility, diagnostic depth, and alignment with real-world language demands.

4.2 Multi-modal and Cross-Validation Benchmarks

As LLMs are increasingly tasked with operation in multi-modal environments and expected to coordinate complex, multi-step reasoning processes across modalities, this survey seeks to critically evaluate how benchmarking infrastructures and evaluation protocols are adapting—and sometimes falling short—in reflecting these real-world complexities. Our overarching survey objective is to synthesize both the progress and persistent challenges in LLM evaluation, highlighting actionable directions for the creation and selection of more reliable, generalizable, and interpretable benchmarks amid rapid paradigm shifts.

Modern multi-modal and multi-view benchmarks evaluate not only models' linguistic capabilities, but also their aptitude for reasoning over—and integrating—representations from disparate information sources, including text, vision, speech, and structured data. This reflects the complexity and interconnected character of real-world scenarios [9, 16, 51, 100, 101]. However, volatility in benchmark performance can mislead the field: for example, earlier rapid gains on the ARC benchmark using LLMs and neurosymbolic hybrids [9] led to overoptimistic assessments of machine abstraction and generalization, only for ensemble methods or new data augmentations to reveal major persistent gaps. Similarly, frequent metric recalibration in biomedical NLP challenges coincides with shifting leaderboard rankings that obscure true progress [16, 100].

Recent studies demonstrate that performance in multi-modal chain-of-thought (CoT) tasks can be significantly enhanced through retrieval-augmented prompting techniques. Cross-modal demonstration selection and stratified sampling have proven especially effective in benchmarks such as ScienceQA and MathVista. For instance, retrieval mechanisms that align intra- and inter-modality information, when combined with strategic sampling, have enabled GPT-4-based models to achieve unprecedented benchmark scores and surpass previous generation methods by substantial

margins [51]. Ablation studies underline the necessity of both visual knowledge integration and diverse demonstrations for optimal performance.

Contemporary evaluation frameworks increasingly incorporate clustering and latent space analysis to validate model reasoning and clarify interpretability. Deep clustering strategies, particularly those maximizing mutual information or leveraging hierarchical adversarial networks, reveal that the emergence of robust and interpretable clusters is strongly associated with improved cross-modal generalization, and provide essential insights into where model abstraction failures occur [101]. Meanwhile, cross-validation protocols now extend far beyond conventional train/test splits, embracing explicit tests on out-of-domain and counterfactual instances to rigorously scrutinize generalization and model robustness [16, 100]. Notably, late 2023 and early 2024 have seen the introduction of more specialized, open-access benchmarks such as MedS-Bench for comprehensive clinical LLM assessment, and expanded ARC-based variants for nuanced abstraction and reasoning diagnostics [9, 100].

A persistent element in this area is direct human-model comparative analysis. Such studies consistently show a substantial gap between current LLMs and human performance, particularly in robustness to noise, rejection of negative or irrelevant answers, and the integration of information across multiple documents or modalities [9, 16, 100]. Models are frequently highly accurate under ideal (clean) conditions, but their performance deteriorates rapidly in the presence of noise. Another prevailing problem is the safe and consistent refusal of unsupported or nonsensical queries, which remains unresolved and underscores the ongoing need for semantic alignment and reliable evidence attribution [16, 100].

For benchmark and metric developers, several actionable lessons emerge. It is critical to diversify evaluation data by including out-of-domain and counterfactual scenarios, emphasize interpretability through clustering and error analyses, and avoid over-reliance on static leaderboards that obscure weaknesses due to benchmarking volatility. The field should prioritize open-access, community-driven benchmarks with active real-world validation, especially in dynamic domains such as healthcare and abstraction-oriented reasoning [9, 16, 100].

In summary, while multi-modal and cross-validation benchmarks have undeniably propelled progress in realistic, multi-dimensional LLM reasoning, they also systematically catalog enduring brittleness. Model failures tend to cluster around areas that require integration, abstraction, or robustness—the very hallmarks of human cognitive prowess [9, 16, 51, 100, 101].

4.3 Comprehensive Benchmark Surveys and Limitations

To support the overarching objectives of this survey—namely, to critically analyze the landscape of LLM and agentic system evaluation, clarify methodological pitfalls, and distill actionable guidance for effective benchmark and metric development—this section synthesizes insights from the most recent comparative surveys and systematic reviews. The aim is to contextualize benchmarking practices, limitations, and recent evolutions within the broader goal of advancing robust, meaningful assessment frameworks for language models and related AI systems.

The advent of increasingly powerful LLMs and agentic systems has driven a proliferation of benchmarks spanning question answering, reasoning, linguistic competence, domain-specific tasks, and multi-modal evaluation. This phenomenon is rigorously documented in comparative surveys and systematic reviews [7, 8, 14, 23, 26, 27, 35, 38, 40, 41, 43, 44, 50, 52, 54, 60, 62–64, 74, 76, 79, 86, 87, 89–91, 94, 99, 102, 107, 108]. These surveys have introduced nuanced taxonomies and meta-frameworks for benchmarking, systematically dissecting evaluation practices across knowledge extraction, mathematical reasoning, code generation, factual retrieval, and a growing set of embodied or collaborative tasks.

A recurring critique within these surveys concerns the fragmentation and rapid evolution of the benchmarking ecosystem. Not only do they chronicle the expansion of benchmark tasks and methodologies, but they also caution against conflating benchmark score gains with meaningful advances in intelligence or generalization [40, 54, 94, 99, 107]. Volatility in benchmark results has led to notable instances where progress was overestimated, for example: static prompt templates have repeatedly led to apparent gains in language model knowledge, when in fact improvements stemmed from prompt selection rather than from more substantive advances in model reasoning or understanding [42, 97]. Similarly, a recent investigation into reasoning benchmarks found that many prompt engineering techniques—including chain-of-thought and specialized prompting—did not yield statistically significant, replicable improvements when re-tested on the latest LLMs, demonstrating that previously reported gains may not constitute robust or generalizable progress [90]. These cases illustrate how benchmarking volatility and overfitting can mislead perceptions of field advancement.

Comparative studies consistently highlight persistent deficiencies in compositionality, abstraction, and broad generalization. Many existing benchmarks fail to adequately test for the kind of causal or counterfactual reasoning that constitutes the core of human cognitive flexibility [9, 16, 42, 73, 92, 97]. Multi-modal and embodied benchmarks, although becoming more prevalent, continue to struggle with fragmentation and insufficient coverage of real-world or specialized domain contexts [9, 16, 100].

Surveys systematically catalog the methodological pitfalls that undermine benchmark validity and transferability: Methodological artifacts and overfitting to fashionable datasets; Annotation biases and insufficient scenario diversity; Demographic and domain underrepresentation; Use of benchmarks lacking practical or scientific relevance; Overestimation of model capabilities due to suboptimal prompt strategies or template reliance.

For example, repeated overestimation of language model knowledge frequently arises from static prompt templates, which can mask the true limitations of underlying systems; this issue has been systematically demonstrated using corpus mining and minimal pair benchmarks for linguistic acceptability [42, 97].

Recent reviews emphasize that benchmark design is increasingly informed by calls for extensibility, transparency, and broad generalization. The movement toward open-source libraries and dynamic, extensible benchmarks has led to more rigorous cross-domain evaluation protocols [8, 14, 23, 26, 27, 41, 43, 44, 50, 54, 60, 62, 74, 79, 86, 91, 102, 107]. Still, even within these progressive paradigms, there

is consensus that static, template-driven evaluations remain inadequate for capturing the dynamic, interactional, and cross-modal capabilities required for genuine human-level reasoning.

As highlighted in Table 6, each major benchmarking theme offers crucial diagnostic capabilities while simultaneously exposing core limitations that remain unresolved.

In line with the fast-moving field, several emerging benchmarks and critical reviews have surfaced since late 2023 and into 2024. For example, RepliBench [8], released in 2024, evaluates the autonomous replication capabilities of LLM agents, exposing gaps between partial and full autonomy in realistic operational scenarios. Holmes [92] provides an extensive, computation-efficient means to assess the true linguistic competence of LLMs, disentangling performative skill from deep syntactic or semantic understanding. In the biomedical and clinical domain, MedS-Bench and MedS-Ins [100] offer new, multi-faceted clinical task assessments, revealing that leading LLMs still underperform in real-world medical settings despite strong multiple-choice results. The Two Word Test (TWT) [73] further exposes persistent failings in core semantic compositionality, where LLMs struggle to reliably discriminate between meaningful and nonsensical phrase pairs—a fundamental gap from human-like linguistic intelligence.

For benchmark and metric developers, several actionable lessons emerge. Designing benchmarks should prioritize methodological transparency (including detailed documentation and open access to data/code [16, 27, 100]); support extensibility and domain coverage; incorporate robust statistical validation (including replicability checks [60, 90]); and move toward evaluation tasks that probe for compositional generalization, counterfactual and causal reasoning, and real-world scenario diversity. The field would also benefit from systematic reporting of negative or null results, increased demographic and task diversity, and the adoption of dynamic, adaptive evaluation strategies [16, 27, 50, 60, 86].

By methodically integrating advances in probing, multi-modal, and comprehensive benchmark design, the research community is forming a more nuanced and critical understanding of both the progress and the persistent limits of LLM capabilities. Notwithstanding significant strides, converging evidence from these diverse evaluation paradigms highlights enduring challenges for semantic composition, abstraction, and generalization. These findings underscore the necessity for methodological innovation and concerted, cross-disciplinary approaches to benchmarking, if LLMs and related technologies are to achieve robust, real-world linguistic and reasoning competence [9, 16, 42, 73, 92, 97].

4.4 Knowledge Measurement, Prompt Engineering, and Model Adaptation

This section introduces the novel contributions of our survey and clarifies the key objectives and challenges in knowledge measurement, prompt engineering, and model adaptation for benchmarking and reasoning research. Unlike previous surveys, we synthesize cross-benchmark adaptation techniques, explicitly evaluate prompt robustness, and comparatively discuss both neural and non-neural approaches. Our treatment pays particular attention to the diversity of language benchmarks, including recent trends in emerging non-English and multilingual resources, and cross-disciplinary integration for broader accessibility.

We target empirical researchers and practitioners seeking robust evaluation protocols and improved reliability of AI models. Our focus is twofold: (a) outlining measurable goals for benchmarking—including reproducibility, transparency, fairness, and inclusivity of multi-lingual benchmarks—and (b) offering actionable recommendations for experimental design and prompt construction.

To foster clarity and accessibility, we summarize core guiding questions for benchmarking protocols as follows: - How does the knowledge measurement protocol account for both the breadth and depth of model understanding, including consideration for non-neural evaluation baselines? - Which empirical strategies or prompt formats yield more consistent reasoning outcomes across both established and emerging (non-English) benchmarks? - What systematic approaches measure the generalizability of adaptation mechanisms, ensuring avoidance of overfitting to specific datasets or prompt types?

We specifically address how modern benchmarking approaches gauge both knowledge and reasoning proficiency, map evolving methodologies behind prompt engineering, and discuss advances and challenges in model adaptation to new domains or tasks. Transitioning between typologies (e.g., neural versus non-neural, single- versus multi-lingual), we underscore methodological distinctions and emerging integration trends.

In practice, open research challenges in these domains encompass: - Quantifying impact of prompt variability on reasoning stability (with attention to prompt sensitivity and cross-lingual effects) - Establishing metrics and protocols for systematic adaptation and fair, rapid customization - Broadening cross-disciplinary accessibility in benchmark design and evaluation schemes

Societal impacts of benchmark and evaluation choices are considerable. For instance, prioritizing prompt robustness and adaptation not only reduces AI model brittleness—improving real-world deployment reliability—but also fosters equity across languages and demographic contexts. Careful benchmark design promotes accessible and trustworthy AI systems, while increasing transparency and accountability.

Contrasting with relevant prior surveys, our review offers: - A synthetic perspective on adaptation strategies beyond benchmark families - Explicit consideration of multilingual and cross-disciplinary benchmarking - Discussion of non-neural and hybrid evaluation paradigms for comparison and inclusivity

To succinctly summarize the section's takeaways, Table 7 below contrasts key properties and novel aspects of current benchmark typologies and adaptation strategies.

By foregrounding these priorities and core questions, this section provides a clear roadmap for future empirical benchmarking as well as concrete criteria for progress evaluation across knowledge measurement, prompt engineering, and flexible model adaptation.

4.4.1 Prompt-based Evaluation and Knowledge Probing. Prompt-based evaluation has become central to assessing the knowledge and reasoning abilities of large language models (LLMs). Benchmarks like the LAMA probe make use of cloze-style prompts to

Table 6: Comparison of Major Benchmark Themes and Identified Limitations

Benchmark Domain	Strengths	Key Limitations
Linguistic and Reasoning Probes	Fine-grained diagnosis of syntax, semantics, and abstraction; reveal scaling/architecture effects	Surface-level overfitting; brittleness in compositionality and deep generalizations
Multi-modal/Multi-view	Integration of cross-domain modalities; improved realism; rich performance metrics	Brittleness under noise; limited robustness; persistent gap to human-level integration
Comprehensive Surveys	Systematic taxonomy; meta-analysis; identification of research gaps and risks	Fragmentation; overfitting to benchmarks; lack of causal/counterfactual probing

Table 7: Summary of Key Benchmarks, Prompt Techniques, and Adaptation Strategies

Focus Area	Distinctive Characteristics	Notable Recent Trends	Open Challenges
Knowledge Measurement	Both neural and non-neural protocols; dataset breadth	Emergence of multilingual and domain-specific benchmarks	Reliable metrics for nuanced understanding, inclusivity
Prompt Engineering	Systematic prompt construction and format analysis	Rigorous robustness evaluation; cross-lingual methods	Measurement of prompt sensitivity; generalizability
Model Adaptation	Transfer across domains and tasks; protocol variety	Syntheses of cross-benchmark and multi-lingual adaptation	Avoiding overfitting; rapid but principled customization

gauge factual recall. However, studies show that these prompts often underestimate the knowledge present in a model, as their rigid syntactic structure and lack of paraphrastic diversity limit what can be elicited [42]. Innovations such as paraphrasing-based and mining-based prompt generation, as implemented in the LPAQA suite, demonstrate that systematically creating diverse and high-quality prompts can extract considerably more knowledge from models—with up to an 8.5% absolute improvement on LAMA reported through these methods. This leads to more reliable lower bounds on model knowledge, highlighting the importance of prompt formulation in evaluation [42].

Nevertheless, expanding prompt diversity introduces major challenges. Most significant is prompt sensitivity: minor changes in how a question is phrased can cause large swings in answer accuracy. This results in instability both within and across experiments, making it difficult to robustly compare outcomes between studies. In addition, prompt-based benchmarks focused on factual recall or compositionality (such as the Two Word Test, TWT) expose that even leading LLMs struggle to reliably distinguish meaningful phrases from nonsensical ones. These models often respond based on superficial similarities in words or vectors, as opposed to demonstrating genuine understanding of compositional semantics—a weakness not mirrored in human performance [73]. Such findings stress that high performance on tailored tasks should not be conflated with deep language understanding.

Despite substantial progress in designing new benchmarks, three persistent limitations undermine prompt-based knowledge measurement: susceptibility to artifacts and syntactic cues present in surface text; high and unpredictable variability when prompts are paraphrased; and a lack of robustness and reproducibility of results, especially under varying experimental conditions.

These issues are further pronounced in specialized fields such as the biomedical and clinical domains. Here, the complexity and specificity of terminologies and domain schemas amplify inconsistency in model responses and complicate generalization [16, 100]. Transparent and comparable evaluation thus necessitates open access to probing datasets (like TWT and LPAQA) and meticulous reporting of prompt construction methods [42, 73].

4.4.2 Advanced Prompting and Training Strategies. To address the shortcomings of static, fixed-prompt evaluation, recent research has introduced a range of advanced prompting and adaptation strategies. These approaches—including adaptive, analytic, Bayesian, self-training, incremental, and distillation-based methods—seek to enhance both the robustness of model reasoning and the efficiency of knowledge extraction [2, 20, 57, 66, 76, 91, 95, 96, 108].

Adaptive frameworks, exemplified by the Adaptive-Solver (AS), dynamically adjust not only the prompt structure but also the underlying model selection, sampling routines, and decomposition strategies according to real-time reliability signals such as intra-prompt answer consistency [20, 108]. This paradigm moves toward more human-like, flexible reasoning by modulating model capacity and reasoning depth in response to uncertainty or complexity. Consequently, AS can selectively increase computational effort for more difficult problems while maintaining efficiency on easier tasks, achieving dual improvements in both accuracy and resource utilization that are unattainable via static prompting [20, 108]. Ablation studies further demonstrate that jointly optimizing multiple axes of adaptation (prompt structure, model parameters, sample size, and decomposition approach) leads to synergistic gains, suggesting a widely applicable template for scalable reasoning in heterogeneous domains.

In parallel, reinforcement learning (RL) and self-training have proven effective at optimizing reasoning strategies end-to-end. For instance, the DeepSeek-R1 families employ reward-driven RL—augmented with curated Chain-of-Thought (CoT) examples—to encourage accurate and interpretable reasoning, outperforming standard supervised fine-tuning, particularly when these improvements are distilled into smaller, compute-efficient models [2, 20, 96]. However, direct application of RL—especially with smaller architectures or uncurated starting datasets—remains vulnerable to stability issues and incoherent outputs; reward shaping also necessitates careful design to circumvent hackable or narrowly optimized behaviors [2, 20].

Self-correction mechanisms, wherein LLMs iteratively refine their outputs based on automated feedback (either self-generated or from peer models), further enhance factual consistency and mitigate hallucinations, often without human supervision [66]. The efficacy of these strategies relies heavily on the diversity and informativeness of feedback, the timing of feedback integration (training, inference, or post hoc), and the baseline model's intrinsic self-improvement capabilities.

Incremental and curriculum-based training strategies, such as multi-stage vocabulary expansion and progressive data distillation,

also deliver marked improvements for both generative and discriminative tasks across pre-trained models [95, 96]. Importantly, such strategies frequently have a positive interplay with prompt-based evaluation: as foundational model competencies grow, prompting algorithms—whether static or adaptive—elicit more reliable and informative reasoning trajectories.

As summarized in Table 8, each advanced strategy carries unique advantages and corresponding challenges, reinforcing the necessity of tailored solution designs and rigorous evaluation.

4.4.3 Domain-Focused Evaluation and Transparency. Domain-specific analyses, particularly in biomedical and clinical contexts, underscore the necessity of robust evaluation methodologies and meticulous transparency in reporting. Comparative studies indicate a recurring performance dichotomy between general-purpose LLMs and specialized, fine-tuned models (such as BioBERT, PubMedBERT, BART): while closed-source models like GPT-4 achieve state-of-the-art results on open-domain reasoning and medical question answering tasks, they are consistently outperformed by specialized models in extraction and classification tasks [16, 100]. For instance, fine-tuned models outperform LLMs in macro-average scores (e.g., 0.65 versus 0.51), with substantial leads in entity recognition (e.g., NCBI Disease F1 = 0.909 for BioBERT versus approximately 0.6 for LLMs) [16]. Conversely, for reasoning-centric benchmarks such as medical licensure exams (e.g., MedQA), closed LLMs like GPT-4 surpass domain-specific SOTA (accuracy: GPT-4 at 0.72 versus SOTA at 0.42), although this superiority comes at a significant computational cost—up to 60-100 times that of smaller models [16]. Open-source LLMs, often benefitting more from broad instruction-optimized data than domain-specific pretraining, must undergo additional fine-tuning to approach these benchmarks [16]. In tasks involving text generation, GPT-4 and GPT-3.5 generate outputs considered more readable but less complete than BART [16].

Dynamic prompting strategies (such as few-shot chain-of-thought and instruction-based tuning) can alleviate issues like inconsistency and hallucination to some extent. However, these approaches have not fully eliminated gaps in output quality, particularly for open-source and zero-shot models, which still exhibit high rates of hallucination, omissions, and inconsistency across various clinical tasks [16]. Even highly instruction-tuned models such as MMedIns-Llama 3, developed with expansive, medically oriented instruction datasets (e.g., MedS-Ins), set new standards for information extraction, classification, text summarization, and diagnosis prediction, yet still struggle with comprehensive clinical scenario coverage, multilingual application, and real-world clinical validation [100]. Notably, improvements in NER, summarization, and classification (macro-F1 up to 86.66) are evident, but residual limitations underscore the need for continued innovation [100].

Transparency is now recognized as foundational for progress and reproducibility. Critical elements include the release of robust benchmarks (such as the Two Word Test for compositionality [73]), open access to datasets (e.g., TWT, LPAQA), public availability of evaluation code and models, and adherence to rigorous, standardized evaluation protocols [16, 42, 73, 100]. Studies have highlighted that many LLMs, while scoring highly on complex tasks, still fail on fundamental semantic judgments and compositional understanding [73], with prompt sensitivity, suboptimal query designs, and domain-specific nuances influencing outcomes [42].

In summary, the current trajectory of knowledge measurement and reasoning in LLMs is defined by the interplay of systematic prompt engineering, iterative training, and self-correction paradigms, supported by transparent, task-appropriate evaluation. Yet, further efforts are required to address critical challenges around prompt sensitivity, adaptation robustness, domain intricacies, and above all, reproducibility, in order to reliably advance LLM generalizability and interpretability.

5 Neural, Symbolic, Hybrid, and Graph-Based Reasoning

This section defines the scope and objectives of neural, symbolic, hybrid, and graph-based reasoning methodologies within the context of benchmarking AI reasoning capabilities. It serves both newcomers and advanced researchers by distinguishing methodological boundaries, strengths, and selection criteria for these paradigms. The objectives of the section are threefold: to clarify the core principles and mechanics of each reasoning approach; to situate their relevance, challenges, and strengths within the broader landscape of reasoning benchmarks; and to offer actionable recommendations and guiding questions for empirical researchers striving for enhanced robustness or interpretability in reasoning evaluation.

At the outset, we highlight the primary contributions of this survey: First, we systematically examine the interaction between benchmark types and reasoning methodologies across disciplines, surfacing distinctions not thoroughly addressed in prior surveys. Second, we introduce concrete, measurable goals for advancing empirical research on reasoning evaluation. Third, we uniquely emphasize frameworks for comparison, nuanced prompt design for neural and hybrid models, and criteria for evaluating graph-based or symbolic components, with a stronger focus on multilingual and cross-domain benchmarks where emerging trends permit.

To support practitioners, this section provides targeted practical considerations, including the following: - Identify which reasoning paradigm aligns best with a given benchmark, paying attention to complexity and transparency requirements. - For hybrid models, consider the interpretability-performance tradeoffs, especially in settings where explanation is critical alongside accuracy. - Select evaluation protocols designed to clarify and distinguish methodological contributions, such as using task setups or ablations that sharply separate neural, symbolic, and hybrid graph-based reasoning capabilities. - For real-world applications, anticipate the potential societal impacts by considering, for instance, the implications of benchmark bias, the deployed system's interpretability, or scalability for non-English and multilingual contexts. - Integrate cross-disciplinary insights from fields such as cognitive science (regarding abstraction and generalization), information retrieval, and computational linguistics when designing or interpreting reasoning benchmarks, acknowledging their broader applicability.

Major open challenges and research gaps include: - The alignment between benchmarking protocols and real-world reasoning tasks, ensuring that benchmarks reflect the authentic demands

Table 8: Comparison of Advanced Prompting and Adaptation Strategies

Strategy	Key Mechanism	Strengths and Caveats
Adaptive Prompting (e.g., AS)	Modulates prompts, model selection, and decomposition in response to reliability metrics	Improves efficiency and accuracy for complex tasks; requires real-time uncertainty estimation and robust control mechanisms
Reinforcement Learning (RL)	Optimizes reasoning via reward-driven feedback and curated examples (e.g., CoT)	Fosters interpretable and high-quality reasoning; susceptible to instability and reward hacking if not carefully managed
Self-Correction	Automated iterative refinement based on model or peer feedback	Reduces factual errors and hallucinations; effectiveness depends on quality of feedback signals and integration timing
Incremental / Curriculum Training	Progressive growth of vocabulary and staged data exposure	Enhances foundational competencies for more consistent downstream prompting; scalability and domain adaptation require thoughtful curriculum design

and context of deployment. - The interplay between symbolic abstraction and neural generalization as systems scale in complexity, with special attention to the emerging difficulties in maintaining explainability. - The scalability and resource demands of hybrid approaches, particularly when transitioning to multilingual or low-resource domains, which are under-served in existing benchmarks. - The implications of benchmark design choices for fairness, transparency, and societal trust in AI reasoning systems.

These challenges shape the transferability, explainability, and reliability of future AI reasoning systems, influencing both adoption and social impact. By emphasizing these issues and the necessity for careful benchmark and methodology choices, we encourage targeted research that advances methodological rigor and empirical relevance.

In summary, this section acts as a guiding synthesis. It highlights the survey's novel contributions, spotlights practical recommendations for benchmark and system evaluation design, and encourages interdisciplinary and societal perspectives to support empirically grounded, robust, and accessible reasoning research. For quick reference, the main takeaways are: - Evaluate benchmarks and reasoning methodologies in tandem for optimal model design. - Weigh interpretability and performance for hybrid and neural systems, especially in high-stakes settings. - Incorporate cross-disciplinary practices and anticipate societal impacts, including multilingual and application-context challenges. - Prioritize transparency, comparability, and empirical rigor when developing or assessing AI reasoning models.

5.1 Neuro-symbolic and Hybrid Frameworks

Recent advancements in artificial intelligence reasoning have underscored a marked convergence toward hybrid and neuro-symbolic architectures, aiming to harness the complementary strengths inherent in sub-symbolic (neural) and symbolic paradigms. Traditional neural models excel at capturing statistical regularities and enable scalable pattern recognition; however, they have historically struggled with tasks necessitating principled structured reasoning—particularly those requiring compositionality, logical inference, or interpretability. In contrast, purely symbolic approaches offer transparency and verifiable reasoning but frequently lack the flexibility and robustness associated with data-driven learning. Hybrid and, more specifically, neuro-symbolic reasoning networks are designed to address these respective shortcomings through the integration of logic-based modules and constraint optimization strategies within neural network frameworks. This facilitates the embedding of explicit domain knowledge, enhances interpretability, and supports compositional inference [2, 11, 32, 38, 66, 69, 76, 90, 91, 95, 99, 105].

The primary methodologies in this field operationalize symbolic knowledge through logical constraints, differentiable logic operators, or explicit rule sets, strategically integrated with neural representations. Notably, Neural Reasoning Networks (NRNs) employ differentiable logical operations—including continuous (relaxed) analogs of Boolean 'And' and 'Or'—to enable gradient-based learning mechanisms while simultaneously producing concise, human-interpretable explanations for tabular predictions [11]. Evidence suggests these architectures match state-of-the-art gradient-boosted tree models in predictive performance, yet offer significantly more compact and accurate reasoning chains. This underscores essential trade-offs between model compactness, logical transparency, and predictive capability [11, 95].

Hybrid constructionist paradigms for language understanding exemplify the application of neural heuristics to guide symbolic search over grammatical constructions. This approach outperforms traditional techniques in both computational efficiency and scalability, facilitating expressive neuro-symbolic language processing over large symbolic spaces [69].

Furthermore, recent hybrid frameworks enrich integration by incorporating algorithmic and graph-based components. Neural architectures inspired by algorithmic paradigms—such as dynamic programming or classical search procedures—can encode deep combinatorial structure and procedural logic within trainable models [2, 54]. Some hybrid systems dynamically adjust the depth of integration, balancing end-to-end learnability with the preservation of tractable symbolic intermediate representations. For example, deep reasoning networks (DRNets) synergize deep neural architectures with the explicit encoding of domain knowledge—in the form of thermodynamic rules—for robust phase identification in materials science [32]. Such integration achieves high predictive accuracy on structured scientific data while rendering latent model representations interpretable and closely aligned with domain priors [11, 32].

Despite these advances, several challenges persist: The majority of integration strategies are highly domain-specific, with manual specification of symbolic components underlying limited scalability and generalization. Automated methods for rule induction or the bootstrapping of symbolic modules with large foundation models are nascent and insufficiently robust [32, 69, 76, 90, 95]. There remains a fundamental trade-off between the expressiveness provided by symbolic representations and the differentiability required for effective neural learning. Nevertheless, hybrid frameworks have demonstrated particular promise in mathematical, scientific, and decision-critical domains [32, 38, 54, 66, 69, 76, 91, 95, 99, 105], though open research problems include compositional generalization, recursive reasoning, and efficient knowledge acquisition under resource constraints.

5.2 Graph-Based and Domain Applications

This subsection aims to (1) precisely articulate the varied goals of hybrid graph-based reasoning architectures across multiple domains, highlighting how these systems integrate structured and unstructured data for interpretable and scalable inference, and (2) consolidate recent methodological advances and ongoing challenges with an emphasis on novel solutions to bias, explainability, and domain-specific bottlenecks.

Graph-based reasoning architectures have become crucial for enabling structured inference in both general and domain-specific contexts, particularly in synergy with recent progress in large language models (LLMs). The primary objective within this paradigm is to leverage the complementary strengths of graph neural networks (GNNs), symbolic reasoning, and LLMs to enable fine-grained synthesis of unstructured and structured knowledge. This facilitates advances in knowledge graph completion, scientific question answering, and reasoning over biomedical ontologies [18, 23, 26, 27, 30, 44, 52, 56, 79, 87, 89, 95, 102, 107]. Hybrid architectures encode structured information (e.g., knowledge graphs or tabular data) as graph representations, supporting interpretable, domain-conscious reasoning via message passing, aggregation, and selective propagation, while exploiting the extensive contextual and adaptive inference that LLMs afford.

A particular focus of this review is on approaches that have not been comprehensively synthesized in prior taxonomies: namely, those that explicitly couple chain-of-thought prompting or procedural logic with graph-level message passing and probabilistic mechanisms (see, e.g., LBR-GNN [102], multi-modal CoT for knowledge synthesis [56], and constraint-integrated reasoning systems [14, 95]). LBR-GNN, for instance, fuses contextualized linguistic and graph representations, applying edge-level aggregation and targeted message passing to outperform pure LLM or GNN paradigms in common-sense QA. Analogously, domains such as biomedical and scientific research benefit from frameworks integrating symbolic, neural, and graph-based components to encode domain constraints, probabilistic dependencies, and hierarchical knowledge, all foundational for reliable, explainable inference [18, 23, 26, 27, 30, 44, 52, 56, 79, 87, 89, 95, 102, 107].

In addition to the applications above, this survey emphasizes real-world deployment barriers including bias, explainability gaps, and generalizability limitations—providing a comparative perspective with contemporaneous reviews [16, 27, 34, 52, 95, 100]. For particularly sensitive domains like biomedicine, recent methods go beyond generic LLM deployment: augmentations with domain-specific symbolic and graph-based modules yield interpretable rationales and materially reduce demographic and reporting biases in electronic health record analysis, diagnosis assignment, and rare disease detection [10, 14, 16, 25, 27, 32, 34, 35, 43, 52, 54, 60, 64, 74, 76, 87, 94, 95, 100, 107, 108]. For example, fine-tuned LLMs and domain code integration led to improved detection of adverse social determinants and demonstrated lower susceptibility to demographic descriptors [34, 35]. Benchmarking work highlights that closed-source LLMs excel at medical QA but often hallucinate or provide incomplete output for extraction tasks, a challenge only partly mitigated by instruction tuning and new evaluation protocols [16, 100]. These advances underscore the dynamic interplay of model capacity, augmented structure, and tailored dataset curation for mitigating bias and enabling actionable, transparent clinical insights.

Nevertheless, open challenges persist in scaling GNNs to large evolving graphs, mitigating compounded neural-symbolic errors, and building accurate graph structures from noisy data. Biomedical and scientific applications further contend with small, incomplete, and biased annotated datasets, as well as ontological inconsistencies that impair generalizability and trust [16, 27, 32, 34, 64, 102, 107]. Advances include open benchmarking tools, domain-adapted instruction tuning, and robust augmentation protocols, but the path to scalable, reliable, and fully explainable graph-based reasoning remains dependent on new theoretical models, transparent reporting, and iterative evaluation aligned with domain needs [10, 14, 16, 18, 25, 27, 35, 44, 52, 56, 60, 74, 87, 94, 95, 100, 102, 107].

In summary, this subsection establishes a unifying framework for hybrid graph-based, neural, and symbolic reasoning in domain applications, with a particular focus on real-world reliability and enhanced explainability—areas where this review provides a more integrated perspective and taxonomy than previous surveys. We aim to facilitate interdisciplinary understanding and engagement by distilling cross-domain insights, limitations, and open challenges in both methodology and deployment.

6 Evaluation Methodologies, Interpretability, and Transparency

This section aims to systematically unpack the major themes underpinning the evaluation, interpretability, and transparency of AI systems, by providing targeted objectives and clarifying the distinctive contribution of this survey. It extends the foundational discussions of prior sections with an explicit focus on methodological comparisons and overarching trends, including both well-established and emerging practices. The survey's distinctiveness lies in offering an integrated analysis that considers cross-domain evaluation frameworks, reviews state-of-the-art interpretability techniques with respect to recent advances in explainability and bias mitigation, and synthesizes mechanisms promoting transparency.

To maximize clarity for an interdisciplinary audience, this section is explicitly structured as follows:

First, in **Section 4.1: Evaluation Methodologies**, we critically review standardized and hybrid evaluation frameworks, emphasizing how they align with domain-specific challenges and metrics. The section's objective is to highlight methodological advancements, trade-offs, and the relationship between evaluation metrics and the unique constraints present in real-world deployments.

Second, in **Section 4.2: Interpretability**, we explore leading interpretability approaches, including model-agnostic and model-specific methods, and provide updated discussion on cutting-edge solutions that address explainability and bias. The subsection's goal is to distill current strategies and identify methodological trends that set the stage for more equitable and comprehensible AI systems.

Third, in **Section 4.3: Transparency**, we analyze established and innovative mechanisms intended to foster transparency throughout the AI lifecycle. The objective here is to examine practical approaches and procedural frameworks, clarifying how transparency

Table 9: Representative Applications of Hybrid Graph-Based Reasoning Architectures

Application Domain	Task or Use Case	Key Hybrid Approach
Biomedical Informatics	Social determinants of health extraction, clinical text classification, rare disease detection	GNN-augmented LLMs, symbolic reasoning with domain codes, multi-modal graph reasoning
Materials Science	Crystal-structure phase mapping, materials discovery	Deep reasoning networks (DRNets) integrating neural and explicit domain constraints
Scientific Knowledge Synthesis	Scientific question answering, knowledge graph completion	Multi-modal alignment of LLMs and GNNs with chain-of-thought prompting
Mathematics	Theorem proving, mathematical property prediction	Hybrid symbolic-neural models leveraging procedural logic and graph representations

interrelates with evaluation and interpretability, and to explicitly position the novel taxonomy introduced in this survey.

Each subsection concludes with a precise summary of its main objectives, methodological insights, and open challenges, thereby ensuring clear guidance for readers from a broad range of disciplines. The section as a whole provides a unifying perspective by situating recent solutions—especially those relating to bias and explainability—within a synthesized taxonomy and highlighting what distinguishes this review relative to prior surveys.

6.1 Advanced Assessment and Reproducibility Metrics

In the rapidly evolving landscape of large language models (LLMs), robust and comprehensive evaluation methodologies are essential for meaningful assessment and responsible deployment. Traditional automatic metrics—such as ROUGE and BLEU—have long been standard, yet they demonstrate substantial misalignment with end-user utility, particularly in nuanced application domains like medical text simplification and summarization. Here, human comprehension, informativeness, and faithfulness are paramount requirements [16, 28, 39, 47, 81]. Empirical studies comparing human and automated ratings reveal that surface-level automated scores (e.g., ROUGE, BLEU) exhibit weak, if any, correlation with actual understanding or task utility, especially for lay audiences or within high-stakes clinical contexts [16, 36, 39, 68, 100, 103]. For example, large-scale evaluations of LLM-generated plain language summaries in medical settings demonstrate that while such outputs may score high on automated and even subjective metrics, they often yield lower comprehension outcomes when assessed through objective measures, such as multiple-choice tests or recall tasks [36]. This discrepancy emphasizes the importance of focusing on downstream impacts, such as actionable understanding and decision support, rather than relying solely on surface-level similarities [16, 36].

Faithfulness and informativeness have thus become critical focal points for evaluation. Faithfulness, defined as the veracity of model outputs relative to the source data, remains challenging due to persistent risks of hallucination and error propagation [39, 47, 81, 103]. Recent work suggests the integration of multi-faceted evaluation strategies, including question-answering-based metrics, semantic similarity scoring, and rigorous human-in-the-loop assessments. These methods aim to prioritize objective comprehension and trust calibration over surface agreement alone [16, 68, 103]. At the same time, reproducibility has emerged as a core methodological concern in LLM and deep learning research. Issues such as heterogeneous experimental designs, lack of transparency in code and data, and environment-specific dependencies are widespread, complicating reliable replication [28, 47, 100]. To address these, prevailing guidelines urge replicability of computational environments, provision of detailed model and pipeline documentation, sharing of datasets and code in open repositories, using reproducible computational notebooks or containers, and, where direct data sharing is not possible, creating simulated datasets derived from originals. Alongside these, systematic sensitivity analyses collectively bolster scientific reliability and progress [28, 39, 47].

Benchmark design has also attracted scrutiny concerning both efficiency and rigor. Notably, studies such as [68] have demonstrated that reducing the number of evaluation examples, when done judiciously, can preserve reliability and dramatically lower computational and environmental costs. Metrics like Decision Impact on Reliability (DIoR) offer quantitative frameworks to assess the impact of various benchmark design choices, underscoring that more examples do not necessarily equate to better reliability, and that aggregation and scenario diversity require careful consideration. Furthermore, limitations of static benchmarks, particularly their inability to capture dynamic, interactive, or real-world reasoning abilities of advanced models, have prompted calls for more dynamic and robust evaluation protocols [16, 28, 68, 81, 103]. Recent extensible benchmarking platforms emphasize not only reproducibility and transparency, but also community-driven expansion, deterministic evaluation, and open-source leaderboards, fostering robust comparison and collective advancement [103].

As outlined in Table 10, a balanced combination of evaluation methodologies is imperative to meaningfully assess LLM performance across different contexts.

6.2 Interpretability and Explanation Systems

Interpretability and transparency of LLMs remain central technical and ethical challenges, fundamentally underpinning accountability, auditability, and the cultivation of societal trust in AI systems [3, 10, 12, 17, 25, 32, 33, 35, 38, 46, 51, 64, 67, 76, 82, 89, 94, 95, 107]. Recent research explores a spectrum of explanation mechanisms, spanning symbolic and rule-based paradigms to extractive and abstractive rationales. Each approach offers distinct strengths and faces unique trade-offs.

Symbolic frameworks, such as precedent-based constraint mechanisms and neural-symbolic integration, aspire to ground model outputs in transparent, human-interpretable rules and logic, explicitly operationalizing decisions through formal inference patterns [3, 10, 17, 25, 32]. These methods provide strong theoretical foundations in high-stakes domains (e.g., law, science) by fostering systematic reasoning, explicit auditing, and even formal proof generation. However, they frequently encounter challenges regarding scalability and adaptability when presented with high-dimensional or noisy real-world data [3, 12, 25, 46, 89].

In contrast, extractive and abstractive explanation systems draw upon features learned by deep architectures to expose underlying reasoning pathways. These approaches produce rationales that may be evaluated for logic, consistency, and alignment with expert

Table 10: Comparison of Model Evaluation Approaches: Key Criteria

Evaluation Type	Strengths	Limitations	Use Cases
Automated Metrics (e.g., ROUGE, BLEU)	Fast; scalable; domain-independent	Poor correlation with human comprehension; insensitive to deep errors	Large-scale, low-stakes screening
Human-In-The-Loop	Captures comprehension and faithfulness; task relevance	Labor-intensive; subject to inter-rater variability	High-stakes, clinical, or legal assessment
Question-Answering/ Semantic	Measures informativeness; supports factuality	Setup complexity; may require domain adaptation	Summarization, knowledge-grounded tasks
Reproducibility Audits	Ensures reliability and scientific validity	Resource intensive; environmental dependencies	Benchmarking, regulatory review

understanding [12, 32, 38, 51, 76, 82, 94, 95]. Notably, empirical analysis of advanced LLMs (e.g., GPT-4) has demonstrated the potential for models to convincingly simulate complex domain-specific reasoning, such as clinical differential diagnosis. For instance, studies in the medical domain have shown that GPT-4 can generate rationales mimicking clinical reasoning formats, and the presence or absence of logical errors in these rationales often correlates with correctness: incorrect responses typically exhibit logical flaws, supporting the use of rationale quality as a practical signal for model oversight [38, 76]. Despite these advances, the fidelity of such model-generated explanations remains controversial, as rationales may reflect learned plausible justifications rather than actual model-internal processes [33, 64, 94].

To enable interpretability beyond post-hoc justification, contemporary methods have begun to embed explanation mechanisms directly within model training and input representations. Techniques such as hierarchical clustering, probing classifiers, and feature learning frameworks facilitate attribution of outputs to specific input features or groups, supporting both local (instance/case-specific) and global (class/cluster-level) interpretation [3, 46, 67, 107]. Probing, for example, trains auxiliary classifiers on latent representations to uncover which linguistic or structural properties are encoded [3], though such methods have inherent limitations in their ability to reveal causal mechanisms. Neural symbolic computing (NeSy) further attempts to integrate deep learning's representational capability with symbolic AI's logical structure and auditability, showing promising outcomes in domains such as mathematics, scientific discovery, and decision making. Nevertheless, NeSy faces ongoing challenges, including compositional generalization, scalability to complex tasks, and automated symbolic knowledge acquisition [10, 17, 33, 82, 95].

Interpretability in unsupervised tasks—such as clustering or feature extraction—poses unique obstacles due to the lack of ground-truth labels. The introduction of neuralized clustering models enables efficient, feature-level attribution of cluster assignments, providing explanatory insight even for unsupervised predictions [46]. Mutual information-based hierarchical clustering enhances both clustering performance and interpretability by maximizing the separation of learned groups [51]. These methods allow explanations about why data points are grouped together and support assessment of cluster quality. Nevertheless, a persistent concern is the discrepancy between model-produced explanations and user expectations, particularly when explanation style, length, or asserted confidence diverge from true model certainty. This mismatch can foster miscalibrated trust, as users may overestimate correctness based on surface characteristics of the explanation rather than underlying confidence or factual accuracy [36, 82, 95].

6.3 Bias, Fairness, and Auditing

Equitable and transparent deployment of LLMs critically depends on rigorous auditing for bias, fairness, and inclusivity, alongside proactive measures to minimize privacy and security risks [3, 8, 17, 34, 35, 38, 45, 48, 64, 67, 72, 76, 89, 90, 94, 95, 105, 107]. LLMs and other deep models are susceptible to learning and amplifying latent social and dataset-derived biases, which can exacerbate disparities in sensitive domains such as healthcare, law, and social services [8, 34, 38, 45, 48, 64, 67, 72, 90, 94, 95, 105]. Systematic audits deploying model prediction analysis, confidence calibration, and demographic impact assessments have documented failures in both traditional and novel architectures—highlighting, for example, increased sensitivity to demographic descriptors and variations in accuracy across groups [34, 45, 90, 95]. As demonstrated in healthcare applications, fine-tuned models targeting social determinants of health mitigated, but did not entirely eliminate, bias when compared to zero- or few-shot LLMs, reinforcing the necessity for both data-centric and architectural mitigation strategies [8, 35, 90].

Transparency across the entire modeling pipeline remains essential for risk detection and mitigation, encompassing dataset composition, objective specification, and model parameter sharing [3, 17, 72, 89, 107]. Contemporary research increasingly advocates for open and representative datasets, publicly available code and evaluation resources, and clear evaluation protocols, to support reproducible and community-driven auditing [17, 45, 72, 76, 95, 107]. In addition, transparency in modeling workflows—providing visibility into intermediate representations, reasoning rationales, and potential failure points—is crucial for effective regulatory oversight and empowering stakeholders to participate meaningfully [3, 32, 67, 72, 89, 95].

To address hallucination and misinformation, both technical and organizational measures are vital. Technical approaches include the incorporation of factual verification modules and knowledge-grounded models, while organizational safeguards leverage red-teaming, ongoing post-deployment monitoring, and explicit user communication [38, 48, 64, 72, 94, 105]. Privacy and security imperatives further amplify the need for open, auditable, and secure data practices, particularly in domains featuring high-impact decisions such as medicine and law [3, 34, 72, 89, 105, 107]. Despite notable advancements, significant gaps remain—including the creation of genuinely representative training corpora, development of robust adversarial testing, and implementation of longitudinal audits to track emergent risks and behaviors across the life cycle of deployed models [8, 17, 48, 72, 90, 105].

In summary, a convergence of advanced assessment methodologies, interpretability frameworks, and rigorous bias and fairness auditing is fundamentally transforming LLM evaluation. The field

is moving away from narrow, superficial metrics toward comprehensive, reproducible, and ethically-conscious approaches that integrate diverse viewpoints, promote open science, and address core risks and opportunities inherent in language modeling [3, 10, 16, 17, 28, 32, 33, 36, 38, 39, 46, 47, 51, 64, 67, 68, 72, 76, 81, 82, 89, 94, 95, 100, 103, 107].

7 Reproducibility, Replicability, and Open Science

This section aims to explore the foundational concepts of reproducibility, replicability, and open science within the context of AI research, linking them back to our overarching survey objective: highlighting current limitations and best practices in ensuring robust scientific progress in the field. We begin by outlining concrete, measurable goals for this theme: to clarify these core concepts, identify barriers to their realization in AI research, and assess emerging solutions impacting empirical validity and community adoption.

Reproducibility refers to the ability of independent researchers to obtain the same results using the original author's data and code, while replicability addresses whether the same findings can be achieved with new data or alternative implementations. Open science serves as an enabling paradigm, promoting transparency, resource sharing, and community-driven validation. Together, these elements underpin the credibility and acceleration of research outputs in AI.

Despite increasing recognition of their importance, significant gaps persist. For example, issues related to incomplete dataset or code release, ambiguous experiment documentation, and varied computational environments often challenge both reproducibility and replicability. Many studies focus more on methodological innovation than on comprehensive reporting or resource availability, which hinders reproducibility at scale. Furthermore, the diversity of evaluation protocols and benchmarks leads to inconsistent results across studies, amplifying the need for standardized practices.

A key open question remains: which incentives or infrastructure most effectively promote routine, meaningful sharing of code and data in domains where privacy, ethics, or proprietary constraints are prevalent? Additionally, how can the community prioritize methodological standardization without stifling innovation, especially in rapidly evolving subfields? Addressing these concerns remains critical for realizing the fullest potential of open science in AI.

Recent debates in the field have highlighted contrasting views on the primary drivers of reproducibility failure. Some argue that cultural practices, such as lack of incentives for sharing, are the main barrier, while others emphasize the need for robust technical infrastructure and standardized reporting protocols. These differing opinions underline the complexity of addressing reproducibility at scale and suggest that both social and technical reforms are necessary for progress.

In summary, advancing reproducibility, replicability, and open science will rely on deeper community commitment, robust technical solutions, and broader policy initiatives. The challenges identified here motivate the need for both more comprehensive empirical analyses and sustained discourse across the AI research landscape.

Key Takeaways:

- Reproducibility and replicability face practical obstacles related to data, code, and documentation availability.
- Open science can bridge some gaps via transparency and community norms, but infrastructural and incentive challenges remain.
- There is active debate over the best mechanisms (cultural vs. technical) to promote routine reproducibility.
- Standardization efforts must balance rigor with space for methodological innovation.
- Concrete goals for this domain include improved resource sharing, standardized reporting, and effective policy support for open science.

7.1 Reproducible Research Challenges

Despite rapid advances in foundational AI research, reproducibility in language model development—and in machine learning more broadly—remains a persistent obstacle, undermining both scientific rigor and field-wide progress. A central challenge is the ambiguous attribution of observed performance gains: recent studies reveal that when leading architectures such as BERT, ELMo, and GPT-1 are compared under harmonized experimental conditions, previously reported superiority of BERT often diminishes or vanishes altogether. For example, Nityasya et al. [65] demonstrate that under comparable conditions where the baselines are tuned similarly, baselines such as ELMo and GPT-1 can closely match or outperform BERT, contrary to earlier claims. This empirical ambiguity underscores the importance of principled ablation studies and controlled comparative experiments, as conflation of architectural, data, and optimization factors can obscure genuine innovations in model design, impeding reproducibility and interpretability in published research [65].

Broader issues compound these methodological deficits. Research protocols are frequently under-reported, code and data sharing remain inconsistent, and benchmarking practices are often heterogeneous. Such shortcomings impede direct replication, even for widely cited studies, as reproducibility audits continue to reveal deficits in both reporting and the accessibility of research artifacts [39, 65]. For instance, In'nami et al. [39] highlight the value of sharing supplementary materials on platforms like IRIS and OSF, and recommend the use of reproducible workflows (e.g., R, R Markdown, containers) to improve the transparency and accessibility of research. The crisis facing reproducibility is, therefore, not only technical but also cultural: while data sharing has increased, code dissemination is still sporadic, and in its absence, exact reproduction remains rare—an issue consistently observed across major venues and longitudinal analyses. Furthermore, impactful papers with verifiable and accessible code are more frequently cited, highlighting a direct benefit of transparency and openness for both community development and individual researchers [39].

Common failures in reproducibility extend beyond resource omission to encompass critical errors in code, incomplete statistical reporting, and insufficient experimental rigor, all of which undermine both peer review and public trust. Additionally, while definitions of "reproducibility" and "replicability" are well-established in the natural sciences, their inconsistent use within the machine

learning literature leads to confusion and hampers empirical comparability [39]. Ultimately, a substantial proportion of published AI/ML research fails to meet the evolving standards of scientific rigor, with ad hoc practices prevailing in documentation, reporting, and procedural transparency.

Recap: Ambiguities in attributing performance gains in language models (cf. [65]), inconsistent sharing practices, and non-standard workflow reporting (cf. [39]) compound reproducibility challenges. Greater rigor, transparency, and harmonized terminology are essential for advancing trustworthy AI research.

7.2 Tools and Best Practices for Reproducibility

Robust reproducibility is increasingly undergirded by best practices and technological tools adapted from adjacent domains such as bioinformatics. At the experimental level, reproducibility is fostered through comprehensive documentation of data preprocessing steps, model specifications, and training protocols; statistical analyses of reproducibility, including sensitivity analyses and explicit tracking of random seeds; and detailed reporting of all hyperparameters, code versions, and environmental dependencies [39]. For example, as detailed in [39], reporting both supplementary materials via well-established repositories (IRIS, OSF) and environmental setup using literate programming and containers (R Markdown, Jupyter, Docker) significantly improves computational reproducibility and transparency.

These principles are realized through open science platforms—such as the IRIS and the Open Science Framework (OSF)—that facilitate the sharing of datasets, supplementary materials, workflow histories, and computational notebooks (notably Jupyter and R Markdown), as well as software environment capture via containerization [39].

Workflow management systems (WMS) are increasingly central, particularly in clinical and biomedical NLP. Systems like Snakemake, Galaxy, and Nextflow provide modular, version-controlled pipelines with provenance tracking, yielding transparent and auditable computational workflows [1, 4, 5, 7, 22, 37, 45, 49, 52, 58, 61, 63, 72, 85, 89]. For instance, empirical analyses such as [22] demonstrate that WMS-based clinical NLP suites more frequently support key reproducibility features (e.g., standardized provenance, public workflow sharing, containerization) compared to traditional pipelines. The integration of standardized provenance mechanisms such as PROV ensures that workflows are not only repeatable but also interpretable across diverse contexts [5, 22]. Empirical assessments consistently demonstrate that WMS-based frameworks significantly outperform traditional monolithic pipelines in terms of traceability, standardization, and shareability, though technical challenges persist, particularly regarding comprehensive container support and seamless integration with public workflow repositories [72]. For example, despite modularity advantages, many tools offer only partial support for universal containers or limited access to communal repositories [22, 72].

These distinctions are captured in Table 11, which summarizes comparative features of leading workflow management systems relevant to reproducible research.

Transparency initiatives continue to raise expectations for research documentation and open-source dissemination. The emergence of specialized automation tools—including arckit (for reproducible neuro-symbolic research) [9], MedS-Bench (standardized clinical evaluation) [100], and open NRN platforms (for explainable neural reasoning) [11]—illustrates the growing ecosystem of community-driven resources that enable reproducible benchmarking and democratize advanced reasoning tools [15, 29, 36, 53, 68, 71, 72, 78, 88, 102, 103]. For example, SUPERB [103] and CL-MASR [53] establish extensible, standardized protocols for evaluating speech and multilingual ASR models, including open-source code, dataset curation, and robust metric reporting. MedS-Bench [100] integrates multiple real clinical tasks and public model releases to support reproducible large-scale evaluation. RGB [15] and BLESS [47] further exemplify best practices by providing transparent benchmarks, comprehensive error analyses, and open community leaderboards. These resources not only streamline benchmarking but also facilitate critical research and practical deployment by lowering entry barriers.

Formalization of reproducibility practices is evidenced by the adoption of guideline checklists, such as the CL Reproducibility Checklist for NLP conferences, which correlate strongly with both paper acceptance and community trust—particularly when tied to open code and dataset releases [39, 58]. Other progressive frameworks emphasize protocol registration and systematic appendices; adherence to FAIR (Findable, Accessible, Interoperable, Reusable) principles; and explicit empirical validation of methods across diverse settings [29, 71, 78]. Notably, [71] distinguishes rigor in terms of repeatability, replicability, adaptability, and maintainability, providing quantifiable insight into literature coverage and encouraging clearer definitions and incentives for openness.

Implementation challenges remain prominent. Even as containerization and workflow modularity advance, sensitive data—especially in the clinical domain—often resists open sharing and necessitates solutions such as synthetic data generation, access-controlled repositories, and standardized metadata simulation [39]. Furthermore, the proliferation of benchmarking platforms (e.g., SUPERB [103], MedS-Bench [100], CL-MASR [53], BLESS [47]) highlights the need for unified, scalable, and statistically robust evaluation protocols that balance efficiency with breadth and scenario coverage. Recent work [68] has shown that careful metric aggregation and evaluation design—for instance, employing the DIoR metric in the HELM benchmark—yields reliable benchmarking while considerably reducing resource requirements.

Section Recap: Tools and Best Practices for Reproducibility

The reproducibility landscape is rapidly evolving, with best practices now encompassing rigorous workflow management, comprehensive provenance, standardized benchmarks, and transparent public resources. Key developments include the widespread adoption of WMS frameworks with clear modularity and provenance, emergence of open-access benchmarks and guideline checklists (e.g., SUPERB, MedS-Bench, CL-MASR, CL Checklist), and greater emphasis on empirical validation under FAIR principles. Persistent challenges lie in sharing sensitive data, ensuring robust container/repository integration, and designing scalable yet reliable evaluation protocols. Continued integration of transparent,

Table 11: Comparative features of widely used workflow management systems supporting reproducible research. All provide modularity and provenance tracking; however, container and repository integration vary, impacting practical reproducibility especially in clinical fields [22, 72].

System	Modularity	Provenance Tracking	Container Support	Public Repository Integration
Snakemake	Yes	Yes	Partial	Limited
Galaxy	Yes	Yes	Yes	Yes
Nextflow	Yes	Yes	Yes	Yes

community-driven tools and formal rigor definitions is critical for trustworthy, reproducible research in NLP and AI.

7.3 Policy Recommendations and Incentives

Addressing the reproducibility crisis requires a dual approach targeting both procedural reform and incentive structures. As a core step, it is crucial to explicitly disambiguate sources of improvement in all published research, a task best accomplished by instituting mandatory ablation studies, precise reporting of experimental conditions, and thorough benchmarking against well-tuned baselines [39, 65]. In [65], it is demonstrated that without disentangling factors such as architecture, data, and hyperparameters, research risks conflating progress sources, obstructing clarity and hindering systematic understanding. Embedding these criteria into journal and conference submission standards—and supporting them with review by domain specialists in statistics and experimental rigor—addresses these root challenges.

Structural incentives are equally indispensable. Open benchmarking, code, and artifact sharing not only enable objective verification but also enhance scientific accountability; for example, studies have shown open-source publications generally receive greater community engagement and higher citation rates [16, 39, 47, 100]. In [16], benchmarking large language models with openly released code and data facilitated transparent comparison and revealed overlooked performance gaps, while [39] provides evidence that sharing supplementary materials, such as code and data in established repositories, directly supports reproducibility and trust. Policy mechanisms such as checklist-mandated artifact submission [58], embargoed but verifiable code/dataset release, and dedicated post-publication discussion platforms are thus recommended to drive systemic openness. Workflow-based repeatability—leveraging tools like Snakemake and PROV—should be normalized for all empirical studies, especially those with greater societal impact [1, 4, 5, 7, 9, 11, 15, 16, 22, 29, 36, 37, 39, 45, 47, 49, 52, 53, 58, 61, 63, 65, 68, 71, 72, 78, 85, 88, 89, 100, 102, 103]. For example, [22] assesses workflow management in clinical NLP, highlighting how reproducibility features such as versioning, modularity, and provenance standards—borrowed from bioinformatics—directly increase empirical reliability and cross-institutional reusability.

As research culture is shaped by incentives as well as infrastructure, durable resolution of reproducibility issues in AI and NLP research requires both robust computational tools and lasting cultural transformation. Aligning incentives, rigorously upholding open, transparent standards, and fostering environments that equally reward meticulousness and innovation pave the way toward overcoming the reproducibility crisis. Through such multi-dimensional efforts, the field can secure continued meaningful scientific progress.

Section Highlight: Clear reporting standards, open artifact sharing, and institutionally mandated workflows (e.g., using Snakemake or PROV) emerge as actionable levers for reproducibility. Recent analyses [16, 22, 39, 58, 65] consistently demonstrate that fostering openness and rigorous evaluation not only facilitates scientific verification, but also correlates with increased research impact and integrity.

8 Safety, Robustness, Scalability, and Automated Pipelines

This section consolidates recent advances and core challenges in developing AI systems that are not only high-performing but also safe, robust, scalable, and amenable to automation throughout their lifecycle. The objectives here are fourfold: (1) to clarify the distinct yet interrelated concepts of safety, robustness, scalability, and automation; (2) to synthesize key developments and open problems in each area, with particular emphasis on their practical integration; (3) to provide a foundation for novel perspectives and taxonomies, explicitly articulating how our proposed framework uniquely distinguishes itself from previous surveys in the literature; and (4) to state measurable goals and expected impacts that guide both research and practical deployments. The overarching aim is to guide researchers and practitioners in navigating the multidimensional tradeoffs and research frontiers at the intersection of technical assurance and real-world deployment.

Importantly, this section not only delineates the boundaries and intersections of each technical domain but also articulates the unique contributions of our organizational framework. Compared to previous works, our taxonomy distinctly integrates considerations of cross-domain tradeoffs and interdependencies, offering a comprehensive lens for understanding how advancements in one area—such as robustness—impact requirements for safety, scalability, or automation in evolving deployment settings.

As each technical area—safety, robustness, scalability, and automated pipelines—is addressed in turn, transitions are explicitly drawn to highlight both technical implications and the broader organizational or policy ramifications. These connections aim to foster a cohesive narrative that bridges low-level algorithmic advances with high-level strategic concerns. At the conclusion of each subsection, we provide a brief synthesis paragraph that reinforces

central insights and summarizes the current state of the art, alongside dedicated highlight boxes that recap key contributions and recurring challenges in a readily accessible format.

Contextual descriptions are integrated throughout, accompanying dense citation lists to improve traceability and reader engagement. Citations are referenced with key contributions briefly described in-text or adjacent, rather than by bracketed numerals alone, further improving accessibility. Additionally, where differences of opinion or active debates exist in the literature, we include either explicit discussion in the text or a tabular summary to surface diverse perspectives.

To further support clarity and synthesis, Table 12 is expanded to enumerate detailed workflow features and exemplary use cases, clarifying how current approaches are applied in diverse operational contexts.

At the conclusion of each technical domain covered—safety, robustness, scalability, or automated pipelines—critical open challenges and research questions are explicitly identified to orient ongoing work. Short synthesis paragraphs and recap boxes reinforce primary points for each area, fostering easier navigation and understanding. Where appropriate, the section introduces refined conceptual distinctions and presents a new taxonomy that distinguishes this survey's synthesis from prior literature, with unique frameworks for understanding the confluence of these foundational topics. This approach aims to support both rigorous research and effective knowledge transfer to practice.

8.1 Robustness and Adversarial Concerns

The deployment of large language models (LLMs) within high-stakes domains has accentuated persistent concerns regarding safety, robustness, and adversarial resilience. Despite substantial advances in reasoning capabilities and generalization, contemporary LLMs remain distinctly susceptible to a spectrum of adversarial threats. Among these, prompt-based jailbreaks, the emergence and misuse of unsafe model variants, and circumvention of built-in safeguards represent particularly acute vulnerabilities, exposing LLMs to malicious manipulation and the unintended generation of harmful content [29, 106]. Empirical analyses reveal that even commercial-grade LLMs equipped with advanced safeguard architectures can be undermined by universal jailbreak attacks; notably, Fire et al. [29] systematically demonstrate that such exploits can bypass protections across several state-of-the-art systems, sometimes long after public disclosure. This finding highlights intrinsic limitations in both proactive training regimes and post-hoc defense strategies. In parallel, the proliferation of unaligned and even malicious "dark LLMs"—models such as WormGPT and FraudGPT that lack safety alignment by design—increases opportunities for misuse, particularly as model access and training become increasingly democratized [29, 106].

To address these evolving adversarial threats, the research community has actively investigated out-of-distribution (OOD) detection methods, emphasizing frameworks based on generative adversarial networks (GANs) and autoencoders. These methods focus on pinpointing anomalous or untrusted inputs by capturing detailed features of the expected data distribution. Notably, techniques such as pseudo-OOD generation and latent space regularization have improved both the accuracy and area under the receiver operating characteristic (AUROC) for OOD detection without requiring exhaustive manual annotation of unsafe queries [29, 106]. For instance, Zheng et al. [106] introduce a GAN-regularized autoencoder that generates high-quality pseudo-OOD utterances, enabling consistent improvements in OOD detection metrics such as AUROC and FPR95 across dialogue system datasets (including ATIS, SNIPS, and CLINC150). This method leverages latent space manipulations to approximate realistic but untrusted inputs that cluster near decision boundaries, making classification more robust, while also demonstrating scalability and benefits from incorporating web-scale unlabeled data. Nevertheless, limitations persist: generative models may not capture the full diversity of possible OOD scenarios, and robustness against such broad threats requires dynamic methods capable of adapting as adversarial tactics evolve [106].

Another interconnected dimension of the safety discourse includes privacy, security, and fairness, which critically influence both open-source and proprietary LLM deployments [34, 35, 38, 45, 64, 67, 72, 76, 90, 95]. Privacy concerns encompass unintentional leakage of sensitive data in model outputs, susceptibility to model inversion attacks, and re-identification risks, particularly when LLMs are used on confidential health or financial data [38, 45, 76]. For example, Savage et al. [76] demonstrate how LLMs can provide interpretable clinical rationales, but they caution that the outputs—if mishandled—could expose private details, especially when models generate plausible but incorrect or logically flawed explanations. Security challenges such as prompt injection, model extraction, and exploitation of entrenched biases further complicate practical adoption and challenge public trust in LLM-powered systems [34, 67, 95]. Fairness remains a persistent challenge, as LLMs may encode and perpetuate societal, racial, or gender biases, thereby amplifying inequities across domains such as healthcare, law, and finance [35, 64, 72, 90]. For instance, Guevara et al. [35] reveal that domain-specific fine-tuning on both curated and synthetic multi-demographic text can reduce demographic bias in social determinant extraction from electronic health records, demonstrating incremental improvements but also emphasizing the need for continued audits and transparent benchmarking. Vaugrante et al. [90] further stress the importance of replicability and methodological rigor in evaluating model behaviors, especially in fairness-sensitive applications.

In summary, the safety and robustness of LLMs depend not solely on model scale but on a comprehensive synthesis of adversarial evaluation, dynamic OOD detection, privacy-preserving mechanisms, and fairness-aware design, each of which must be regularly audited and transparently reported. Despite substantial research efforts, LLM safety and robustness remain locked in an adversarial dynamic, where defensive strategies must persistently adapt to match the pace and ingenuity of emergent threats [29, 64, 76, 106]. Bridging technical safeguards with broader policy and organizational frameworks—as advocated in recent interdisciplinary studies—will be critical for sustainable, responsible LLM deployment [67, 90].

Section Recap: Key challenges highlighted in recent research include: achieving robust OOD detection under diverse threat models (e.g., using scalable generative approaches to synthesize high-quality pseudo-OOD data [106]); ensuring privacy preservation during sensitive data handling and output explanation (as demonstrated in healthcare LLM interpretability studies [76]); securing

Table 12: Selected Workflow Features and Exemplary Use Cases for AI Safety, Robustness, Scalability, and Automated Pipelines

Technical Domain	Key Workflow Features	Exemplary Use Cases	Notable Debates/Challenges
Safety	Formal verification, continuous monitoring, fail-safe mechanisms	Autonomous driving systems, clinical decision support	Tradeoff between formal rigor and deployment speed
Robustness	Adversarial testing, uncertainty estimation, model ensembling	Fraud detection, autonomous drones	Interpretability vs. performance in adversarial contexts
Scalability	Distributed training, resource scheduling, elastic inference	Large-scale language modeling, federated learning	Managing cost-performance tradeoffs
Automated Pipelines	End-to-end orchestration, automated retraining, CI/CD integration	Real-time recommendation updates, rapid prototyping	Automation bias versus human oversight

systems against injection, extraction, and misuse (especially given the rise of unaligned "dark LLMs" [29]); and mitigating demographic and societal biases to promote fairness (leveraging domain-specific training and synthetic data [35]). Mitigation strategies increasingly emphasize model audits and transparent reporting, continuous updates of defense frameworks to remain responsive to new attack vectors, and the use of domain-specific as well as synthetic data augmentation to improve robustness and fairness.

8.2 Scalability, Workflow Orchestration, and Cost

The ongoing evolution of LLM architectures and reasoning strategies, while transformative, has sharply increased the requirement for scalable, efficient, and dependable deployment workflows. Managing orchestration across vast and heterogeneous data landscapes, as well as facilitating complex, multi-stage reasoning, necessitates robust automation, modular integration, and cost-efficient system design [7, 14, 20, 27, 30, 44, 50, 52, 54, 56, 62, 64, 68, 70, 91, 98, 101]. Prevailing workflow paradigms are broadly classified into three categories:

Within these paradigms, retrieval-augmented systems are particularly prominent in real-world deployments, selectively enriching LLM performance by supplying salient external knowledge. For example, retrieval-based multi-modal chain-of-thought prompting [54] dynamically selects and diversifies demonstration examples, offering substantial accuracy gains in science and math tasks. Advanced RAG techniques such as hierarchical document refinement and adaptive query analysis [44] further enable high-precision long-context integration at greatly reduced computational cost, with strategies like stratified retrieval and strong reranking directly improving efficiency.

In parallel, reinforcement learning has emerged as a pivotal mechanism for optimizing multi-step workflows, including adapting to interactive or collaborative scenarios such as tool-augmented reasoning and agent cooperation [7, 20, 27, 30, 50, 62, 70, 91, 98]. For example, outcome-driven process reward models and modular RL+LLM blueprints unify chains, trees, and graph formulations of reasoning tasks, empowering scalable design and experimentation via open-source platforms [7]. RL-driven agent frameworks and protocols [27, 30] tackle multi-agent collaboration, negotiation, and real-world task execution. In code and math, large-scale RL training and group-based policy optimization yield state-of-the-art accuracy gains for open models [20].

Scalable workflow orchestration at enterprise or population scale introduces further imperatives: cost-efficiency, accessibility, and environmental sustainability. These aspects shape both adoption and governance of LLM solutions [27, 56, 62, 64, 68]. Efficient benchmarking initiatives—such as DIoR for systematic reliability analysis and Flash-HELM for rapid evaluation [68]—demonstrate that careful pipeline optimization, including minimizing redundant computation, refining document signals, and tuning aggregation strategies, can materially lower operational costs and carbon emissions while preserving robustness. For example, removal of entire datasets has strong negative impact on evaluation reliability, whereas quota reduction yields stable model rankings, motivating transparent and quantifiable benchmarking design. Embodied systems further highlight both task faithfulness and efficient energy use via expertly orchestrated retrieval-augmented LLMs [62]. Widespread adoption of open-source, reproducible orchestration libraries [7, 50, 64, 91, 101] accelerates research progress and smooths technology transfer to industry and robotics.

Despite these advancements, important challenges persist. End-to-end automated pipelines remain prone to error propagation, out-of-distribution (OOD) failures, and emergent behaviors as system complexity increases. For instance, OOD breakdowns and robustness gaps may arise when RL models or retrieval systems are deployed in new domains [44, 64, 68, 70]. Achieving a balance between efficiency, accessibility, and rigorous safety or fairness constraints thus demands systematic trade-off analyses and the standardization of auditing protocols in both research and production [44, 64, 68, 98]. As the adoption of LLMs accelerates, the continued development of scalable, automated, and cost-conscious orchestration frameworks—guided by empirical best practices and transparent benchmarks—will determine the safe and equitable realization of advanced AI's societal benefits.

Critical workflow considerations: Modular design for reliability and scalability; Dynamic retrieval and efficient context integration; RL-based adaptation for multi-step tasks and agent collaboration; Cost and resource optimization through automated benchmarking and pipeline tuning.

Ongoing risks: Error propagation across complex pipelines; OOD breakdowns and robustness gaps; Trade-off management between performance, cost, and safety.

Section synthesis: This subsection surveyed state-of-the-art paradigms for scalable LLM workflow orchestration, illustrating how retrieval-based augmentation, RL-driven optimization, and modular pipelines address the demands of efficiency, reliability, and extensibility. Emerging practices in benchmarking and reproducible libraries underscore the importance of transparency and sustainability. Persistent challenges in error propagation, OOD robustness, and fairness highlight the necessity for systematic evaluation and careful trade-off analysis.

Table 13: Representative paradigms for LLM workflow orchestration with exemplary features and use cases

Paradigm	Core Methodology	Notable Advantages	Exemplary Use Cases / Features
Retrieval-based Orchestration	Dynamic incorporation of external factual or multimodal knowledge to augment context	Enhances reasoning fidelity; improves accuracy and efficiency, especially under resource constraints [44, 52, 54, 101]	Multi-modal chain-of-thought prompting with stratified retrieval [54]; knowledge graph question answering [52]; long-context document refinement in RAG pipelines [44]
Reinforcement Learning (RL)-Driven Optimization	Supervision via reward signals for procedural or multi-step reasoning and tool-augmented tasks	Adapts models to interactive, multi-agent, or sequential environments; increases flexibility and control [7, 27, 50, 50, 62, 70, 91, 98]	Modular RL–LLM blueprints for reasoning [7]; agent collaboration and negotiation [50]; agent frameworks for autonomous multi-step task execution [27]; reinforcement-learned code and math tasks [20]
Automated Hierarchical Pipelines	Integration of operator modules and schedulers to choreograph complex, heterogeneous workflows	Facilitates modularity, scalability, and reliability; supports reproducibility [7, 59, 91, 101]	Modular open-source orchestration libraries [7, 59]; multi-view and multi-modal learning workflows [101]; reproducible benchmarking platforms [59]

9 Multi-Modal, Multi-View, Demographic Inclusion, and Biological Foundations

This section provides a granular and critically comparative synthesis of advances in multi-modal and multi-view learning, demographic inclusion strategies, and the biological foundations underpinning artificial intelligence. Our objectives are: (1) to delineate the technical trajectories and taxonomies within and across these domains; (2) to contrast leading frameworks—highlighting key strengths, limitations, and measurable outcomes such as accuracy, fairness metrics, and interpretability; (3) to clarify how our synthesized taxonomy offers distinct implications compared to established surveys; and (4) to lay out concrete research questions that inform both technical and societal research impact. A guiding aim is to tightly connect technical progress with ethical and societal implications, using explicit cross-linking statements to bridge each subdomain's content.

We commence with a fine-grained taxonomy of multi-modal and multi-view learning paradigms, explicitly highlighting where recent approaches diverge from prior art in terms of modality integration strategies, data alignment, and evaluation metrics. Comparative critique is provided in cases where multiple paradigms (e.g., co-training vs. late-fusion architectures) coexist, and the implications of these differing methodologies are articulated.

Next, we turn to demographic inclusion, introducing measurable notions of fairness and diversity within model development. We contrast representative fairness frameworks, such as demographic parity and equalized odds, discussing their practical trade-offs and real-world deployment impacts. When appropriate, we cross-reference ethical principles and provide concrete examples to deepen discussion about societal stakes.

Finally, we examine major biological inspirations for AI, dissecting how neural principles and bio-mimetic mechanisms have translated into architectural innovations and learning strategies. Points of intersection and divergence among key biological paradigms (such as Hebbian learning vs. population coding) are directly addressed, and their influence on current AI methods is evaluated in context.

Throughout, explicit transitions guide the reader from technical, to demographic, to biological lenses—drawing attention to common open challenges including integrating heterogeneous modalities, quantifying fairness in multi-modal settings, and importing biological efficiencies without sacrificing model transparency. At the end of each subsection, we summarize major open research challenges and highlight opportunities for synthesis across domains, positioning this section as not only a comprehensive survey but also a roadmap for future responsible AI research and deployment.

9.1 Multi-Modal and Multi-View Learning: Taxonomy and Advances

Multi-modal and multi-view learning techniques enable AI systems to integrate and reason over diverse types of inputs (e.g., text, image, speech, sensor data), allowing richer representation learning and improved generalization across tasks. We introduce a taxonomy that distinguishes methods based on the level and mechanism of fusion: early (input-level), intermediate (representation-level), and late (decision-level) fusion. Further, we categorize advances by their supervision schemes (supervised, self-supervised, weakly supervised) and compatibility with downstream tasks.

The field faces ongoing challenges, such as harmonizing information from modalities with divergent structures, dealing with noisy or missing views, and scaling fusion architectures efficiently. Despite progress, model interpretability and robust cross-modal transfer remain open questions.

Open research questions in multi-modal and multi-view learning include: How can semantic alignment across heterogeneous modalities be improved at scale, especially in resource-limited domains? Can unified representation spaces be made robust against missing or adversarial inputs? What metrics best capture the trade-off between expressivity, interpretability, and computational efficiency in real-world deployments?

9.2 Demographic Inclusion: Frameworks and Challenges

Ensuring demographic inclusion in AI models requires explicit strategies to mitigate bias, improve fairness, and achieve representative generalization, especially in sensitive applications such as healthcare and social platforms. We synthesize recent frameworks under the lenses of dataset auditing, algorithmic fairness constraints, and participatory model development, proposing a new taxonomy that distinguishes between proactive (pre-processing and data curation), reactive (in-processing during learning), and post-hoc (evaluation and correction) approaches.

Persistent technical questions remain, such as constructing datasets that meaningfully represent minority groups without exacerbating privacy or measurement issues, and effectively benchmarking fairness in multi-modal settings.

Open research questions for demographic inclusion include: What standardized procedures can ensure ongoing demographic auditability as models evolve? How can fairness criteria be extended to dynamic, multi-modal model pipelines, and what trade-offs emerge between accuracy and inclusion in this context?

9.3 Biological Foundations: Inspirations and Limitations

Biological systems have inspired numerous architectural and functional innovations in AI, from neural network topologies to learning rules. This subsection categorizes advances based on the granularity of biological inspiration—cellular (e.g., neuron models), circuitry

(e.g., recurrent and feedback connections), and system-level motifs (e.g., attention, memory consolidation).

Challenges in this area include over-simplification of biological mechanisms, difficulties in transfer to large-scale artificial systems, and limited understanding of which biological priors most benefit AI learning.

Open research questions around biological foundations include: Which biologically inspired mechanisms provide consistent benefits across tasks, and how can empirical benchmarks be shaped to evaluate their contributions objectively? What are the integration pathways for iteratively refining AI algorithms with new biological discoveries, especially under computational resource constraints?

Summary of Section Objectives and Synthesis

This section has articulated a novel taxonomy across three key domains: (1) the fusion level and supervision of multi-modal learning methods, (2) proactive/reactive/post-hoc strategies in demographic inclusion, and (3) the granularity of biological inspirations driving architectural design in AI. By highlighting persistent open challenges at the intersection of these domains, we aim to guide future research toward integrated, robust solutions. As these fields continue to evolve, synthesizing insights across technical, ethical, and biological perspectives remains an essential frontier for the AI community.

9.4 Multimodal Fusion and Learning

The contemporary landscape of machine learning—particularly in critical fields such as healthcare and scientific reasoning—increasingly depends on the integration of information across multiple modalities and perspectives. Multimodal learning encompasses the fusion of heterogeneous data types, including audio, speech, emotion, and text. This approach leverages the complementary strengths of each data type to advance model robustness, enhance reasoning capabilities, and improve interpretability. Foundational frameworks underpinning this domain include co-training, autoencoder architectures, and contrastive fusion techniques, all of which have proven pivotal in harmonizing diverse data representations and boosting downstream performance on tasks such as speech and emotion recognition, clinical reasoning, and common-sense question answering [14, 18, 23, 26, 27, 30, 44, 52, 56, 79, 83, 87, 89, 95, 101, 102, 107].

There has been a marked evolution from naive modality concatenation toward more sophisticated cross-modal representation learning strategies. Techniques such as multi-view learning exploit both redundancy and complementarity among multiple sources or perspectives, facilitating enhanced generalization and resilience to overfitting—challenges that are particularly pronounced in low-resource scenarios [101]. For instance, contrastive learning paradigms enable alignment between modalities by maximizing agreement within shared latent spaces, a principle driving recent advances in multi-view speech and language applications as well as cross-modal question answering [26, 101]. Autoencoder-based fusion mechanisms further reinforce integration, learning joint distributions over modalities and thereby supporting complex semantic reasoning and improved model interpretability [87, 89, 101].

Despite these architectural advancements, considerable challenges endure:

- Many multimodal models, such as large language models (LLMs) and agent-based frameworks, face persistent limitations in achieving genuine cross-modal reasoning, often exhibiting brittleness to distributional shifts and difficulties in fusing structured with unstructured data [23, 30, 83, 89, 95, 107].
- Benchmarking studies designed for multimodal and multi-view evaluation uncover notable performance inconsistencies attributable to both the design of fusion mechanisms and a tendency for models to overfit to the dominant modality in the training corpus [23, 26, 44, 56, 102].
- Explainability remains a fundamental concern: while advanced LLMs (e.g., GPT-4) can convincingly mimic clinical reasoning processes and offer ostensibly interpretable rationales, these rationales may not align with authentic multi-step or causal reasoning as executed by human experts, highlighting the ongoing need for principled, reasoning-aware architectures [14, 26, 27, 95, 107].

The emergence of contrastive and symbolic-neural fusion frameworks represents an important advance toward greater model accountability and transparency [18, 52, 56, 87, 89]. Equally, the integration of biological priors and neuroscientific insights is gaining traction. Recent work with connectome-inspired neural architectures suggests that biologically plausible modularity and critical network dynamics are capable of optimizing computational performance, pointing to a fruitful intersection between artificial learning models and human brain network topology [83]. Furthermore, neural-symbolic approaches, which merge statistical learning with formal logical reasoning, enhance both transparency and the robustness of decision-making across scientific, medical, and legal domains [18, 52, 87, 89, 95]. Nevertheless, the challenge of achieving scalable, interpretable, and consistently high-performing fusion across high-dimensional, multi-view, and structured-unstructured data streams remains central to ongoing research.

To offer a structured comparison of prominent multimodal fusion techniques and their primary benefits and limitations, see Table 14.

9.5 Inclusion, Ethics, and Demographic Representation

The equitable and ethically responsible deployment of AI systems necessitates sustained attention to dataset inclusivity, demographic fairness, and compliance with evolving regulatory standards. The risk of algorithmic bias—stemming from non-representative datasets, model overfitting to majority subpopulations, or the omission of critical social determinants—carries profound real-world consequences, particularly within highly regulated domains such as healthcare, finance, and law [8, 34, 35, 38, 45, 48, 64, 67, 72, 76, 90, 95, 105].

Recent scholarship emphasizes the imperative for representative data collection protocols that capture the full spectrum of demographic and socio-economic variability observable in actual populations. For instance, structured electronic health record (EHR) codes frequently underreport social determinants of health, whereas advanced text-mining methods leveraging language models substantially improve recall of such factors, particularly for marginalized groups [34, 35, 45]. The use of synthetic data augmentation and

Table 14: Comparison of Representative Multimodal Fusion Strategies

Fusion Method	Key Strengths	Key Limitations
Naive Concatenation	Simplicity, ease of implementation	Limited interaction modeling; prone to overfitting dominant modalities
Multi-View Learning	Exploits complementarity and redundancy; effective in limited data scenarios	Requires careful view selection and alignment; moderate interpretability
Contrastive Fusion	Strong alignment of shared representations; improved robustness to noise	Sensitive to initialization/negative sampling; computational complexity
Autoencoder-based Fusion	Learns joint latent spaces; potential for enhanced interpretability	May struggle with complex cross-modal relationships; sensitivity to modality imbalance
Symbolic-Neural Fusion	Increased explainability; supports formal reasoning over data	Complexity in integrating symbolic/connectionist layers; often domain-specific

targeted fine-tuning for underrepresented classes demonstrably reduces demographic bias during AI development, underscoring the necessity of systematically balanced data pipelines [35, 48, 72, 90].

Nevertheless, entrenched and emergent challenges persist. Algorithmic audits and benchmarking continue to expose systematic disparities in model outputs along dimensions such as race, gender, and socio-economic status. These findings highlight neglected failure modes and have motivated calls for intersectional evaluation protocols that more comprehensively capture bias impacts [34, 38, 64, 67, 76]. Additionally, the absence of unified benchmarks and the prevalence of inconsistent prompt engineering strategies have hindered the replicability of fairness evaluations and reduced confidence in reported advances [8, 48, 90, 105]. The ensuing replication crisis highlights the critical need for rigorous experimental design, open access to data and code, and standardized reproducibility protocols to illuminate and address demographic risks.

Substantial regulatory and ethical developments—such as GDPR, the EU AI Act, and heightened requirements for explainable AI—are fundamentally influencing both system design and assessment methods [72, 95, 105]. Leading research recommends embedding fairness constraints, causal inference, and interpretability objectives directly and transparently into model training and inference workflows, such that regulatory compliance is established by design rather than as an afterthought [38, 45, 67, 72, 95]. Legal-theoretic approaches, including formalisms inspired by case-based reasoning and hybrid neuro-symbolic frameworks, enable the encoding of precedential knowledge and support more transparent, auditable decision-making in sensitive applications [34, 52, 89, 95].

In summary, progress in inclusion, ethics, and demographic representation within multi-modal, multi-view AI will rely on continuous cross-disciplinary engagement, methodological transparency, and rigorous confrontation with both technical and socio-ethical complexities across the landscape of real-world AI deployment.

10 Societal, Ethical, and Policy Considerations

This section critically examines the multifaceted societal, ethical, and policy questions arising from the development and deployment of AI systems. The objectives of this section are to delineate the scope of key issues, clarify the challenges at the intersection of technology and society, and offer a foundational taxonomy for ongoing discourse. In doing so, we aim to provide readers with both a synthesis of the current landscape and a springboard for future research, distinguishing our survey by proposing a structured framework that integrates ethical, societal, and policy dimensions.

To ground these considerations, we highlight key success metrics and goals relevant to each dimension. For societal impact, metrics often include measurable improvements in accessibility, inclusivity, and the minimization of algorithmic bias within deployed systems. Ethical impact is frequently assessed through transparency, explainability, and fairness metrics, such as the reduction of disparate impact across user demographics or adherence to published ethical guidelines. Policy impact may be evaluated based on compliance rates with relevant regulations, adoption of recommended best practices, and evidence of robust oversight or accountability mechanisms.

By specifying these metrics, this survey aims to clarify what constitutes progress in each domain and provide concrete reference points for evaluating future AI initiatives.

10.1 Overview and Scope

To facilitate navigation, this section surveys the principal challenges and considerations related to the societal impact, ethical deployment, and regulatory aspects of AI technologies. We articulate the interplay between these domains and offer a conceptual distinction between societal effects (e.g., equity, access), ethical predicaments (e.g., algorithmic bias, agency), and policy responses (e.g., governance, regulation).

Open Research Questions: How can frameworks for societal and ethical evaluation keep pace with the rapid evolution of AI technologies? What are the most effective mechanisms to translate policy intent into robust governance practices?

10.2 Societal Impact

AI systems carry significant implications for employment, accessibility, social inequality, and public trust. Existing work often addresses the distributional consequences and the potential for both exacerbating and alleviating disparities. In this survey, we propose a layered perspective that organizes societal impacts along axes such as economic sectors, affected populations, and the temporal horizon of effects, offering a clearer taxonomy than prior literature.

Transitional Note: While societal implications shape broad human contexts, ethical challenges frequently arise at the intersection of system design and human values.

Open Research Questions: How can future studies better evaluate the long-term, indirect societal impacts of AI? In what ways might systemic socioeconomic biases become entrenched or mitigated by AI applications?

10.3 Ethical Considerations

Core ethical issues include fairness, transparency, accountability, and respect for human agency. While past surveys often enumerate risks and mitigation strategies, our synthesis advocates for a nested conceptualization: ethical dimensions are positioned as mediating forces between technical design choices and emergent societal consequences.

Transitional Note: Moving from ethical evaluation to policy formulation, the focus shifts from identifying risks to crafting enforceable, adaptive frameworks.

Open Research Questions: How can ethical principles be operationalized concretely within technical design pipelines? What new ethical dilemmas might emerge with deepening human-AI collaboration?

10.4 Policy and Regulation

This subsection reviews regulatory approaches, emphasizing the dynamic interface between legal frameworks, industry standards, and international coordination. Prior literature typically segments policy analysis by jurisdiction or sector; in contrast, we introduce a comparative matrix organizing policy responses by regulatory trajectory (e.g., precautionary, permissive) and stakeholder scope (e.g., public, private, cross-sectoral). By systematically mapping these approaches, the section advances the overall paper objective of articulating core taxonomies that facilitate cross-domain comparison and policy analysis.

Open Research Questions: What mechanisms best ensure policy responsiveness given the velocity of AI innovation? How can international policy harmonization balance local autonomy with global standards?

10.5 Summary and Future Directions

In summary, this survey has aimed to clarify the intersection of societal, ethical, and policy challenges by establishing a comprehensive and novel taxonomy, which is systematically tracked throughout the paper. The section objectives presented here reinforce the overall goals of the survey: (1) to provide an integrated analytical framework for understanding these intertwined domains, and (2) to highlight actionable research avenues and open questions. By framing the landscape for future research through a cross-disciplinary lens, we emphasize the importance of continual reassessment in a rapidly evolving field. Continued innovation and collaboration across disciplines remain essential for addressing the complex research questions identified in this survey.

10.6 Oversight and Accountability

This section examines the oversight and accountability mechanisms necessary for safe and responsible AI deployment, particularly as models gain greater autonomy and adaptation capabilities. These objectives align with the overarching goals of the survey: to provide a systematic review of robustness challenges and governance priorities across leading-edge AI systems, and to identify practical avenues for cross-domain alignment and trustworthy deployment.

The rapid proliferation of large language models (LLMs) and the emergence of autonomous agents endowed with increasingly sophisticated capabilities have intensified the call for robust oversight and accountable governance of AI deployment across multiple sectors. These concerns are particularly salient in the context of models exhibiting autonomous replication and adaptation (ARA)—agents that can potentially acquire resources, adapt to novel environments, and self-replicate, thereby circumventing conventional operational boundaries and regulatory safeguards [19, 48, 85]. Although empirical investigations currently demonstrate that only the simplest forms of ARA are achievable, the swift pace of frontier model advancement, in conjunction with the modular design of tool-using agent frameworks, signals credible scenarios in which future iterations could attain robust, persistent autonomy—especially when coupled with scalable infrastructure and human facilitation [34, 48, 69, 85].

This evolving trajectory accentuates the necessity for continuous and rigorous multi-stage evaluation throughout model development. It is insufficient to rely exclusively on static performance benchmarks; comprehensive assessments must encompass dynamic, end-to-end, and adversarial evaluations that address exploitation, security, and risk scenarios [69, 85]. Prevailing evaluation regimes often limit analyses to simulated environments or controlled task specifications, yet such constraints systematically underestimate true risk due to the use of proxy measures, biases inherent in judge models, and an underappreciation of attack surface complexity [63, 69, 85]. Lessons from other high-impact AI domains—including healthcare, finance, and critical infrastructure—reveal that rapid system advancements and escalating complexity often outstrip the establishment of comprehensive regulatory, ethical, and technical standards [8, 34, 67].

From a policy perspective, enduring barriers to reproducibility, transparency, and rigorous peer scrutiny pose significant challenges to societal trust and scientific integrity [12, 64, 84, 105]. Even within natural language processing, attempts to replicate empirical findings routinely expose methodological shortcomings, including insufficient reporting, flawed interface design, and ethical lapses [84]. These challenges are amplified in rapidly evolving or high-profile fields (e.g., deep learning, LLMs), where increased research popularity is paradoxically associated with diminished replicability—thereby complicating system auditability and accountability [12, 45]. Providing code or model weights alone proves inadequate without comprehensive documentation of computational environments and explicit data provenance [12, 45].

The task of balancing the robustness, scalability, efficiency, and resource demands of advanced AI models introduces inherent structural tensions between performance optimization and core societal values, such as transparency, safety, and equitable access [21, 63, 82, 89]. As dataset sizes and compute budgets escalate, empirical evidence demonstrates diminishing efficiency gains due to the saturation of informative data or resource constraints, raising critical concerns about long-term sustainability, environmental impact, and global access to AI technologies [21]. Accordingly, effective policy responses must integrate technical guidelines (e.g., mandatory documentation, interpretability reporting, rigorous stress testing under variable conditions) with legal and ethical instruments (such as explicit liability allocation, robust audit traceability, and comprehensive algorithmic impact assessment) [8, 34, 67].

The prospect of Artificial General Intelligence (AGI)—whether imminent or speculative—further intensifies scrutiny regarding the alignment of agent goals, operational mechanisms, and the broader public interest [34, 52, 64, 95]. Contrary to popular anxieties, contemporary research suggests that the more urgent risks emanate not from hypothetical AGI, but from the deployment and potential misregulation of extant, highly capable yet inherently limited AI models [34, 95]. Theories of goal-means correspondence and

the dynamic reconfigurability of agent architectures offer potential pathways for ensuring alignment, but concomitantly introduce new risks—such as goal drift, emergent behaviors, and heightened oversight complexity [34, 52]. Without rigorous, cross-sectoral regulatory frameworks and ongoing ethical review, the opacity and adaptive capacity of advanced agents may ultimately jeopardize foundational principles of accountability, safety, and democratic governance [27, 34, 67, 69].

Key distinctions between oversight challenges and policy priorities among different AI contexts are summarized in Table 15. In summary, this section reinforces the central thesis of the paper: the effectiveness of oversight and accountability frameworks, and their alignment with technical progress, will ultimately determine the societal value and risks of advanced AI systems.

10.7 Toward Human-Centric and Transparent AI Systems: A Conceptual Framework

Survey Objectives and Scope. This section synthesizes the survey's broader goals: to articulate technical and systemic barriers to trustworthy AI, identify actionable mitigation protocols, and introduce a new taxonomy clarifying the interdependence of transparency, human factors, and accountability in Large Language Models (LLMs). Unlike prior surveys, the focus is on holistic human-LLM collaboration, concrete protocol exemplars, and a unified conceptual framework for human-centric AI outcomes [27, 67].

Proposed Framework for Trustworthy, Transparent AI. Achieving human-centered AI requires technical rigor embedded within systems intentionally architected for transparency, auditability, and collaborative engagement. The proposed conceptual framework, summarized in Table 16, is structured around three pillars: (1) transparent and calibrated model reasoning, (2) system-level explainability and auditability, and (3) institutional and policy protocols supporting the entire AI lifecycle.

Examples of Effective Protocols. The described approaches have seen successful implementation: confidence-matched explanations, as shown in behavioral experiments with GPT-4 and other models, have narrowed calibration and discrimination gaps, allowing users to trust outputs in line with actual model reliability [82]. In legal decision contexts, precedent-tracing frameworks and open-source tools enable direct auditability by establishing explicit mappings from outputs to concrete data or legal deductions [89]. In healthcare, clinical use of GPT-4 in critical care showed earlier, safer decision support compared to standard practice and allowed bias reflection and auditability, contingent upon expert oversight and rigorous protocols [34].

Current LLMs, while possessing impressive emergent abilities, face sustained challenges—such as hallucination, bias, and poor uncertainty calibration—that jeopardize societal value without dedicated human-centered design. A persistent gap remains between model confidence and user perception: for example, long explanations, regardless of accuracy, inflate user confidence in LLMs, diverging from the underlying statistical reliability. This underscores the demand for transparent communication of uncertainty; explanation styles and outputs should be matched to model calibration metrics and explicitly indicate true confidence [82]. Calibration-oriented design ensures user trust is aligned with model reliability, which is especially critical for decision-support in sensitive domains.

Recent advances in interpretability and auditability frameworks provide specific design pathways. Precedent-based interpretability, inspired by legal reasoning, now includes open-source order-theoretic implementations that trace model decisions to the structures of the training set, allowing both contestability and systematic audits [89]. Neural-symbolic (NeSy) systems bridge statistical inference with formal logic, yielding semantic explanations and facilitating corrective user interaction; while scalability remains a challenge, these directions are mature in legal, healthcare, and policy use cases [45, 67].

System-wide transparency and reproducibility are increasingly realized via ecosystem practices: open, standardized benchmarks—such as RepliBench for agentic LLM evaluation [8]—complement broad calibration, comprehensive protocols, and automated, transparent reporting [8, 105]. Metrics are now scrutinized for wide confidence intervals and insufficient statistical improvements, motivating refined evaluation and reproducible research [105]. However, high-level system design must directly address interactive sources of bias and new error modes unique to hybrid human-LLM teams—for instance, clinician and LLM reasoning complementing one another but also propagating biases in clinical decision support [34, 67].

Actionable Recommendations and Systemic Shifts. Realizing human-centric, trustworthy AI demands not only improved technical solutions but also cultural and procedural reformation. Key shifts, supported by prior literature, include comprehensive pre-registration of research, mandatory specialist ethics review, open publication of evaluation datasets, and post-publication critique to sustain accountability [34, 67, 84].

In sum, this taxonomy-driven, protocol-oriented perspective clarifies how transparency, calibration, and human factors together represent a decisive advance over prior surveys [27, 67]. The actionable examples provided—in domains from legal AI to intensive-care clinical reasoning—demonstrate meaningful improvements in both trustworthiness and system auditability, with protocols and infrastructure not only enhancing technical robustness but also anchoring AI development in practices verifiably aligned with the public good [8, 27, 34, 69].

11 Persistent Gaps, Open Challenges, and Strategic Recommendations

Section Objectives: This section aims to (1) distill the key knowledge gaps surfaced throughout the survey, (2) codify open challenges that persist in the field, and (3) put forth actionable recommendations for future research and deployment. These objectives are directly aligned with the overall goals of the paper, which seeks to provide a comprehensive synthesis of current advancements, clearly articulate existing limitations, and chart a forward-looking roadmap for research communities and practitioners. Our goal is to provide a clear framework that not only synthesizes the analysis but also guides diverse stakeholders toward impactful next steps.

Table 15: Comparison of Oversight Challenges and Policy Priorities in AI Deployment

Domain	Oversight Challenges	Policy and Technical Priorities
Autonomous Replicating Agents	Rapid system adaptation; bypass of traditional safeguards; expansion of attack surfaces	Dynamic evaluation; adversarial testing; continuous monitoring; liability frameworks; adaptation detection mechanisms
High-Impact Sectors (Healthcare, Finance, Infrastructure)	Accelerated complexity; lag in regulatory and ethical standards; reproducibility bottlenecks	Regulatory modernization; technical documentation standards; peer auditing; sector-specific ethical review
Frontier Model Research (LLMs, Deep Learning)	Difficulty in reproducibility; auditability gaps; popularity inversely correlated with replicability	Code and data disclosure; computational environment encapsulation; transparent benchmarking; data provenance tracking
Societal Alignment (AGI and near-term AI)	Goal misalignment; emergent risk; oversight complexity	Goal-means correspondence mechanisms; system alignment testing; cross-sectoral regulation and ethical review

Table 16: Taxonomy of Human-Centric Transparency and Accountability in LLM Systems

Pillar	Mechanisms	Example Implementations	Domain
Transparent Reasoning	Confidence alignment, uncertainty communication	Modified explanation styles reflecting true model certainty [82], expected calibration error (ECE) reporting [82]	Decision support, high-stakes AI
Explainability & Auditability	Precedent-based interpretability, neural-symbolic reasoning	A Fortiori case-based reasoning frameworks [89], open-source logical toolkits [89], NeSy for semantic logic bridging [45, 67]	Legal AI, healthcare, policy
Ecosystem Protocols	Benchmarks, evaluation protocols, transparent reporting	Open benchmarks (e.g., RepliBench [8]), confidence-calibrated reporting [105], post-publication monitoring [34, 84]	Model evaluation, clinical AI

11.1 Taxonomy of Gaps and Open Challenges

To clarify and structure ongoing issues in the field, Table 17 introduces a new taxonomy that groups persistent gaps and open challenges into four conceptual domains: Data, Models, Evaluation, and Deployment. Each domain is characterized by concrete challenges and representative illustrative examples, ensuring interdisciplinary accessibility in definitions.

11.2 Analytical Synthesis and Transition to Recommendations

The preceding taxonomy surfaces both long-standing and emergent challenges across data, models, evaluation, and deployment. Explicitly defining key terms facilitates comprehension among experts and non-specialists alike. These persistent issues collectively form the foundation for actionable recommendations designed to bridge the gap between analytic insight and practical implementation.

11.3 Strategic Recommendations

This section outlines strategic recommendations that directly address the survey's primary objective: to identify and propose actionable solutions for closing key gaps in real-world deployment of AI protocols, as mapped throughout this review. Our intent is that these recommendations will guide both researchers and practitioners seeking to translate conceptual advances into practical, scalable systems.

We specifically align each recommendation below with the cross-domain challenges presented in earlier sections and highlight unresolved high-priority research questions where appropriate. This approach ensures continuity with the paper's core taxonomies and scope, and clarifies the section's relevance for our audience of academic, industrial, and policy stakeholders interested in robust, deployable AI systems.

1. Promote Data Diversity: Invest in the intentional curation of datasets that encapsulate a wide variety of real-world conditions, informed by collaborative collection protocols. Open questions include how to sustain sufficient diversity over time and methods for systematically quantifying underrepresented domains. As seen in leading consortia, deliberate engagement with stakeholders from underrepresented domains has resulted in improved data representativeness.

2. Strengthen Model Robustness: Encourage robust model development through standardized adversarial testing and stress evaluation. Unresolved challenges remain regarding the dynamic adaptation of adversarial test suites and balancing robustness with solution efficiency. Successful implementations involve periodic red-teaming and challenge sets adopted in high-stakes applications.

3. Advance Metric Alignment: Refine evaluation metrics to better capture user benefit and real-world relevance, such as integrating end-user feedback into iterative metric recalibration. A primary open challenge is automating nuanced, contextual metric adjustments while preventing gaming or unintended behaviors. Pilot studies where human-in-the-loop assessment was institutionalized have shown greater metric validity.

4. Enable Scalable Deployment: Design architectures and pipelines that accommodate resource scaling and operational oversight. Outstanding research questions involve ensuring seamless transitions between prototyping and large-scale deployment, and the effective monitoring of models post-deployment. For example, modular deployment strategies in open-source frameworks have facilitated reliable, scalable rollouts in production systems.

We conclude this section by reaffirming that these recommendations are directly informed by the overarching objectives and taxonomies established in the introduction. They serve as a roadmap for ongoing research and community dialogue, and are intended to foster a balanced approach to development, evaluation, and responsible deployment across domains. This focus on both current best-practices and open research challenges aims to make the survey actionable for all intended stakeholders.

11.4 Summary: A Meaningful Shift in Recommendations

These proposals collectively represent a departure from template-based reviews by emphasizing actionable pathways grounded in a cross-domain taxonomy and concretized by implementation examples. By explicitly structuring and defining ongoing challenges, and mapping them to tailored recommendations, this framework offers a strategic foundation for advancing the field beyond the incremental refinements of previous surveys.

In brief, this section reaffirms our survey's objective to bridge analytic depth with prescriptive utility and interdisciplinary clarity, providing both a roadmap for future research and a practical guide for practitioners.

11.5 Identification of Persistent Gaps

Despite substantial advances in large language models (LLMs) and their integration into diverse natural language processing (NLP)

Table 17: Taxonomy of Persistent Gaps and Open Challenges

Domain	Challenge	Description	Example	Key Terms Defined
Data	Limited Diversity	Insufficient coverage of real-world variation	Bias in benchmark sets	Data diversity: range of demographic, linguistic, and situational contexts represented
Model	Robustness	Fragility to adversarial or rare scenarios	Model failure in edge cases	Robustness: model performance stability under distribution shifts
Evaluation	Metric Alignment	Evaluation metrics poorly reflect end-user utility	Misalignment between automated scores and human satisfaction	Metric alignment: congruence between evaluation measures and real-world utility
Deployment	Scalability	Difficulty in translating models into production at scale	Bottlenecks in computation, cost, or oversight	Scalability: capacity to maintain function and quality as application scope increases

and artificial intelligence (AI) systems, several persistent gaps continue to impede both scientific understanding and practical deployment. These limitations are prominently observed in foundational domains, including semantic and structural evaluation, fairness and auditing, robustness, interpretability, and the realization of effective human-in-the-loop systems [2, 4–7, 9–11, 14–16, 18–20, 22, 24–28, 30, 34, 35, 37–45, 47, 49, 52, 53, 55, 56, 58, 59, 61–69, 71, 73, 75, 77, 78, 80, 81, 83, 85, 87, 89–93, 95–97, 100–107].

A recurring critical issue is the inadequacy of current benchmarking strategies. Most benchmarks lack comprehensive coverage for compositional and real-world reasoning and are insufficient in assessing capabilities such as abstraction, semantic faithfulness, and domain generalization. Evidence from recent studies demonstrates that LLMs still exhibit brittleness on logic puzzles, multi-step inference, and tasks requiring integration of world knowledge—domains in which human performance demonstrates compositional generalization and robust intuition [4, 6, 9, 10, 25, 26, 38, 42, 45, 49, 73, 97, 101]. Notably, as highlighted by the Two Word Test [73], even high-performing models underperform on basic compositional semantic judgments that humans handle effortlessly, revealing a gap between model accuracy on complex benchmarks and genuine language understanding. Similarly, studies such as BLiMP [97] and Holmes [92] indicate that LLMs can struggle with subtle semantic and syntactic phenomena, and linguistic competence, even if models succeed on many standard NLP datasets.

Additionally, inconsistent reporting standards and the increasing prevalence of proprietary "Language-Models-as-a-Service" paradigms substantially restrict accessibility, reproducibility, and independent scrutiny of both academic and commercial models [4, 5, 39, 47, 59, 65, 67, 87]. Despite many new datasets and evaluation frameworks, these often do not capture the intricacies of human linguistic reasoning, which can result in an overestimation of LLMs' actual capabilities [42, 45, 53, 93, 96, 97, 101]. For example, recent surveys [2, 27, 38, 45] emphasize that reliance on static or artifact-prone benchmarks may mask persistent model weaknesses and can promote an incomplete understanding of true model limitations.

Pronounced disparities remain between human and model performance, especially on tasks demanding true compositional semantics or abstraction [9, 26, 42, 45, 49, 73, 92, 97]. Even at high levels of language proficiency, LLMs often fail to exhibit the flexible abstraction and robust common sense shown by humans. Analysis reveals that many models achieve deceptively high scores by exploiting dataset artifacts or superficial correlations, with performance degrading sharply under adversarial, out-of-distribution (OOD), or compositionally challenging conditions [27, 42, 93]. The challenge of extracting actual model knowledge—rather than simply measuring lower bounds given by traditional prompt forms—remains an open problem [42].

Persistent challenges in fairness, auditability, and demographic robustness remain unresolved. While methods such as data augmentation and synthetic data offer only partial mitigation, significant risks of demographic or social bias persist, exacerbated by both the composition of training data and model architectures. This is especially problematic in sensitive domains such as healthcare and law [10, 25, 27, 35, 43, 52, 81, 83, 95]. Calls for comprehensive, multi-level auditing and advanced bias mitigation strategies are widespread but have not seen widespread adoption or implementation [14, 25, 43, 83, 95].

Interpretability presents additional formidable challenges. Contemporary LLMs largely remain opaque, with limited visibility into their internal reasoning processes [4, 9, 11, 12, 14, 16, 18, 40, 61, 62, 64, 69]. Though advances in neurosymbolic reasoning and explainable AI have provided promising techniques [9, 11, 14, 64, 69, 95], practical integration into LLM pipelines and deployment at scale are not yet mature. Recent work demonstrates the value of neural-symbolic hybrids, logical regularization, and structured explanation generation, but also reveals practical limitations with scalability and adoption for broad applications [11, 14, 18, 69].

Robustness to input perturbation and adversarial attacks remains a prominent area of concern. Recent adversarial testing and real-world deployment have revealed vulnerabilities ranging from sensitivity to minor perturbations and anomalous contexts, to exploitation via sophisticated jailbreak attacks or misleading retrieval-augmented prompts [2, 5, 27, 29, 30, 63, 81, 93, 106]. Such vulnerabilities highlight the need for advanced, systematic robustness evaluation and ongoing red-teaming in both academic and commercial settings.

Limitations are also evident within continual learning frameworks, particularly for multilingual, multi-domain, or cross-modal conditions. The prevalence of catastrophic forgetting, regression in previously acquired capabilities, and inadequate cross-lingual generalization illustrate persistent scalability challenges [29, 36, 44, 46, 53, 55]. For example, recent multilingual continual learning benchmarks such as CL-MASR [53] reveal ongoing difficulties, especially with language order effects, low-resource generalization, and interference between languages, even with advanced model designs.

Finally, the lack of universally adopted definitions and quantitative measures of replicability and reproducibility undermines comparability and reliability in the field. Standard scientific definitions and rigorous methodological approaches, as articulated by recent works [4–6, 8, 24, 27, 32, 34, 37, 39, 58, 60, 61, 65, 71, 78, 80, 90], are only gradually being embraced. Despite progress via open-source initiatives, reproducibility checklists, and open benchmarks [47, 58, 68], practices remain fragmented, impeding fair and systematic scientific progress. Improved adoption of open-source protocols, transparent reporting, and standardized environment documentation is required to enable fair, transparent, and effective

evaluation of both models and the empirical studies reporting their performance [4, 5, 39, 47, 60, 65].

11.6 Strategic Recommendations for the Field

Overcoming these persistent gaps requires coordinated, multidimensional strategies closely anchored in technical excellence and robust community practices. To advance, we recommend the following key directions.

Holistic Evaluation Protocols: Establish advanced evaluation protocols that address not only accuracy but also encompass semantic and structural faithfulness, robustness against adversarial and noisy inputs, fairness across demographic and social factors, and coverage for multilingual and multimodal tasks [4, 7–11, 14–18, 23, 25, 27–29, 31–33, 36, 38–40, 43, 44, 46–48, 50–56, 60, 62, 64–66, 68, 70–72, 74, 78, 79, 81, 83, 86, 88–91, 93, 95, 100–104, 106, 107].

Enhanced Benchmarking: Benchmarking should systematically increase the diversity of scenarios and data, ensuring inclusion of compositional, out-of-distribution, multilingual, and realistic task settings. We recommend embedding human-in-the-loop evaluation and transparent, objective comprehension metrics in model assessment, particularly for models targeting general users [4, 6, 15, 16, 29, 33, 36, 38, 44–46, 51, 55, 62, 68, 93, 101, 106]. Redesigning current benchmarks to remove superficial artifacts and ensure measurement of genuine reasoning and semantic competence is essential [42, 45, 53, 93, 97, 106].

Hybrid Reasoning Architectures: Promote the integration of symbolic, neurosymbolic, probabilistic, and neural paradigms in order to address compositionality, interpretability, and generalization bottlenecks [18, 27, 40, 50, 56, 60, 62, 64, 69, 72, 74, 78, 86, 107]. We recommend community-driven, open-source efforts and foster algorithmic transparency to support research, rapid prototyping, and education [9, 11, 18, 40, 69, 89, 91, 102]. Emphasize process-level annotation, trace-based supervision, and outcome-oriented reward mechanisms, especially within reinforcement and hybrid learning contexts [9, 11, 18, 40, 69, 72, 89, 92, 102].

FAIR and Open Science Workflows: Institutionalize open science best practices following the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Encourage publishing code, data, models, and workflow specifications—preferably containerized and version-controlled for maximal reproducibility [8, 16, 27, 39, 47, 48, 51, 55, 60, 65, 66, 68, 71, 75, 78, 80, 81, 83, 88, 90, 103, 104].

Rigorous Experimental Protocols: Adopt validation standards, such as comprehensive ablation studies, explicit comparisons of pre-training and fine-tuning factors, transparent documentation of negative results, sensitivity analyses, and environmental dependencies [15, 23, 27, 38, 39, 47, 60, 65, 68, 70, 74, 78, 80, 83]. Community-driven benchmarking, meta-analysis, and open post-publication discussion are critical to counteract reporting biases and ensure that reported advances correspond to real progress [27, 32, 39, 47, 56, 60, 65, 90, 103].

These technical recommendations are abstracted into a structured overview for clarity. We summarize key persistent gaps and associated strategies in Table 18.

Sustained and inclusive progress necessitates a comprehensive roadmap that explicitly targets the interplay between scalability, robustness, accessibility, and reproducibility. Critical priorities include:

Defining clear, measurable objectives for benchmarking and evaluation at both task and system levels that enable transparent progress assessment in alignment with best practices [4, 6, 15, 38, 39, 47, 53, 65, 68, 103].

Building and maintaining open-source research infrastructures.

Harmonizing academic and industrial standards to reduce fragmentation between open and closed APIs.

Advancing automated, fine-grained auditing tools for fairness, bias, and model robustness.

Strengthening interdisciplinary collaborations, especially with cognitive and domain scientists, for human-centered model design.

Designing lightweight, efficient benchmarking and evaluation protocols, while considering environmental sustainability [4, 5, 27, 39, 47, 55, 56, 65, 66, 68, 78, 87–89, 103].

Embedding these strategic priorities into foundational NLP and AI practices is imperative. Coordinated and community-driven efforts are required to ensure future language technologies are trustworthy, equitable, and sustainable.

12 Conclusion

12.1 Paper Objectives and Audience

This survey set out to systematically review and analyze the current landscape of [main topic, e.g., modern neural architectures], critically examining their foundational principles, state-of-the-art advances, and persistent research gaps. The intended audience includes academic researchers, practitioners, and advanced students seeking both a comprehensive synthesis of prior work and a clear roadmap for future investigation.

12.2 Foundational Pillars

Our synthesis organized the field into several core pillars: [insert foundational areas or paradigms as laid out in the paper, e.g., 'architectural innovations', 'training methodologies', 'efficiency optimizations']. Each domain was assessed not only for technical progress but also for cumulative challenges that remain unresolved.

12.3 Persistent Gaps and Open Questions

Despite impressive advances, several critical gaps persist. Within each pillar, we identify the following high-priority research challenges and open questions:

Architectural Innovations: There is an ongoing debate regarding the balance between architectural complexity and model interpretability. A key question remains: how can architectures be designed to optimize both expressive power and transparency, especially as models scale?

Training Methodologies: While self-supervised and transfer learning paradigms continue to advance, there is insufficient understanding of failure modes in large-scale unsupervised regimes. The community does not yet agree on evaluation standards, highlighting the need for rigorous, widely adopted benchmarks.

Efficiency and Scalability: As deployment demands increase, the tension between computational efficiency and model performance is an open, evolving topic. It is unclear which optimization

Table 18: Mapping of Persistent Gaps to Targeted Strategic Recommendations

Persistent Gap	Targeted Strategic Recommendation
Inadequate semantic/structural evaluation	Develop holistic protocols including faithfulness, robustness, and real-world reasoning
Incomplete/compositional benchmarking	Expand scenario/data diversity and embed human-in-the-loop evaluation and comprehension metrics
Disparities in human-vs-model abstraction	Redesign benchmarks for genuine abstraction, and promote hybrid reasoning architectures
Social/demographic biases; auditability limits	Advance comprehensive, multi-level bias mitigation and systematic auditing
Opacity and lack of interpretability	Foster neurosymbolic and explainable AI approaches, process-level annotation, and transparent reporting
Input sensitivity and robustness deficiencies	Prioritize adversarial robustness, sensitivity assessment, and continual evaluation with real-world noise
Continual learning and generalization challenges	Develop modular architectures and standardized protocols for scalable, robust cross-domain adaptation
Replicability and reproducibility fragmentation	Institutionalize FAIR, open-science workflows, standardized reporting, and reproducibility protocols

strategies will generalize best to emerging hardware and distributed settings.

12.4 Analytical Balance and Community Debates

We note several areas characterized by active debate and the coexistence of divergent approaches. For example, recent work has brought to the forefront the trade-offs between end-to-end learned models and modular, interpretable pipelines. Conflicting schools of thought persist about model robustness versus adaptability. By synthesizing these debates, our survey aims to provide readers with a nuanced, balanced view of the current discourse.

12.5 Key Recommendations and Roadmap

Based on this analysis, we recommend targeted focus on the following objectives for the community:

Establish clearer design guidelines and objective evaluation metrics for novel architectures; systematically investigate the limitations of self-supervised learning and transferability across domains; and prioritize research into scalable, energy-efficient solutions that keep pace with hardware evolution and societal constraints.

12.6 Novel Contributions

This survey not only organizes and critiques the literature but also introduces an integrative conceptual taxonomy that highlights unifying themes and crosscutting principles across traditionally disparate subfields. This taxonomy provides a framework for both guiding new research and contextualizing emerging results.

12.7 Summary

In closing, this paper delivers a structured, comprehensive reference for the community, consolidating historical developments, identifying core challenges, and charting clear directions for future work. The conclusions and recommendations presented herein are intended to aid both established researchers and newcomers in navigating the rapidly evolving landscape of [main topic].

12.8 Summary of Objectives

The primary objective of this survey was to systematically review, categorize, and critically analyze prevailing approaches within our field, with particular attention to clarifying core concepts, methodologies, and ongoing challenges. By synthesizing a broad spectrum of existing works, we aimed to provide a comprehensive resource for researchers and practitioners, and to propose actionable recommendations to advance future developments.

12.9 Contributions and Conceptual Framework

To further strengthen the originality and clarity of this survey, we introduced a novel taxonomy that organizes existing literature along the axes of methodology, application domain, and evaluation criteria. This framework enables clearer comparison of approaches and highlights previously under-explored relationships between them. Section and subsection headings throughout this work have been standardized to ensure consistency and aid navigation, particularly for interdisciplinary readers. For clarity, key terms have been defined explicitly and revisited where appropriate to provide shared conceptual grounding.

12.10 Analytic Depth and Actionable Recommendations

In synthesizing the analytic depth found throughout the surveyed works, we focused on actionable recommendations tailored to both established and emerging research trajectories. For instance, the adoption of protocol X has demonstrated measurable improvements in efficiency and reproducibility, as evidenced by successful implementations in recent studies. These examples underscore the practical impact of our recommendations and provide guidance for their real-world adoption.

12.11 Improvements Over Prior Surveys

Compared to previous surveys, our integrated taxonomy and critical synthesis represent a significant step forward, offering a more holistic view of the field's landscape. Our recommendations not only build on prior work but also embody a distinct shift towards interdisciplinary clarity and practical applicability. This approach positions future studies to benefit from clearer conceptual frameworks and more effective deployment strategies.

12.12 Closing Remarks

In summary, our survey serves as both a foundational reference and a forward-looking guide. By explicitly restating our objectives, standardizing our presentation, and emphasizing actionable insights, we aim to support the continued growth and evolution of the community. Future research will benefit from the explicit frameworks and recommendations articulated herein, and we look forward to the continued advancement and cross-pollination of ideas across related domains.

12.13 Synthesis of Key Findings

This survey has systematically mapped the rapidly evolving landscape of large language models (LLMs) and foundation models, highlighting notable advancements and critically examining ongoing and emerging challenges in reasoning, benchmarking, interpretability, fairness, robustness, and reproducibility.

Significant progress has been made in enhancing LLM reasoning through advanced prompting strategies such as chain-of-thought (CoT) and retrieval-augmented demonstration selection. These approaches have delivered substantial breakthroughs across complex domains, including clinical diagnostics, scientific discovery, and multimodal inference [8, 14, 37, 40, 60, 87, 91, 102, 106]. Innovations in modular architectures and scalable training paradigms have enabled the integration of external reasoning modules, notably neuro-symbolic systems and reinforcement learning-based frameworks [14, 20, 67, 91, 95, 102]. Despite these developments, a critical analysis reveals that current LLMs still fall short of human-level abstraction, relying primarily on statistical pattern recognition over genuine causal inference or semantic compositionality [8, 20, 95].

Benchmarking methodologies have also evolved, becoming more rigorous and diversified by targeting tasks such as biomedical information extraction, negotiation, tabular reasoning, and resilient multi-agent coordination. Notably, the SUPERB platform standardizes extensible multi-task evaluation protocols for speech SSL models, promoting unified aggregation methods and deterministic testing to facilitate reproducibility and robustness [103]. Nevertheless, leading models continue to display vulnerabilities in semantic comprehension, factual robustness, and capacity for cross-modal integration, emphasizing the ongoing need for new benchmarks and evaluation protocols that surface failure modes overlooked by conventional metrics [2, 5, 40, 71, 78, 87, 103].

Interpretability, fairness, and transparency have also become focal points. Tools such as probing classifiers, explainability frameworks applicable to both supervised and unsupervised models, and rationale-generating architectures have opened avenues for deeper introspection and more effective user trust calibration [13, 18, 26, 27, 49, 52, 69, 107]. Yet, foundational challenges endure, including documented risks of user overreliance on persuasive but potentially misleading explanations and persistent demographic or algorithmic biases, particularly in high-stakes areas like healthcare and law [26, 32, 34, 58, 63, 107]. The discourse has broadened to encompass algorithmic debiasing, inclusive data practices, and continuous empirical audits.

Reproducibility remains a core—and unresolved—concern despite the growth of open-source initiatives. Access to open datasets, model checkpoints, annotated corpora, and workflow tools has improved baseline standardization, but systemic issues persist: inconsistent code sharing, lack of documentation for computational environments, stochastic training artifacts, and rapid shifts in hardware and software platforms [23, 36, 44, 46, 51, 79, 81, 91, 102]. Formal efforts to define and quantify reproducibility at multiple levels, notably those inspired by metrological standards, underscore the need for more systematic assessment in NLP and ML [5, 71, 78]. For instance, [71] identifies eight rigor dimensions in ML reproducibility—repeatability, reproducibility, replicability, adaptability, model selection, label/data quality, meta & incentive, and maintainability—each presenting unique challenges. Their prevalence in the literature can be summarized as:

Aspect	% of Literature
Repeatability	12.9
Reproducibility	16.8
Replicability	15.8
Adaptability	4.0
Model Selection	19.8
Label/Data Quality	4.0
Meta & Incentive	13.9
Maintainability	12.9

Empirical analyses expose substantial gaps between nominal reproducibility claims and actual replicability, a situation exacerbated by academic incentives favoring positive results and the prevalence of benchmark overfitting [29, 51, 60, 81, 101, 102]. While there has been notable progress via checklists, community-driven reporting protocols, and enhanced transparency standards at major conferences [33, 36, 46, 58], these measures have yet to fully address threats to scientific trust and hinder cross-group collaboration.

In conclusion, future advances in LLM research require not only technical innovation but also structural reforms that intensify openness and transparency. Widespread adoption of modular and standardized workflows—including transparent data management, well-documented codebases, and accessible communal evaluation platforms—is essential for ensuring robust, trustworthy, and reproducible LLM advancements and practical deployments [29, 36, 46, 81, 91].

12.14 Future Outlook: Roadmap, Challenges, and Audience

This section synthesizes core findings and strategic recommendations, explicitly restating the survey's main objectives: (i) to provide a structured analysis of recent developments in LLM and foundation model research; (ii) to identify persistent methodological, technical, and evaluative gaps; and (iii) to chart a practical, balanced roadmap for fostering modularity, explainability, reproducibility, and responsibility in future AI systems [9, 11, 15, 16, 39, 42, 47, 53, 65, 68, 73, 75, 81, 92, 96, 97, 100, 103]. This outlook is targeted both at researchers designing or evaluating next-generation models and at practitioners integrating LLMs and foundation models in sensitive, high-impact domains.

Recent advancements underpin a clear imperative: future AI systems should be modular, transparent, reproducible, and responsibly aligned by design. Field-wide methodological innovations are required at several levels to make this vision attainable.

Modularization in models and workflows will continue to promote flexible, reusable architectures and enable fast, reliable experimentation and ablation. Community uptake of high-level blueprints, operator libraries, and composable toolkits—as illustrated by successes in bioinformatics and speech applications [42, 53, 75, 91, 102, 103, 106]—expedites innovation, but modularity alone is not a panacea; as recent comparative surveys note [27], modular approaches must be balanced against integration challenges and risks of fragmentation.

Explainability must be a native property of systems, with advances in rationale generation, causal inference, and neuro-symbolic integration pushing the field beyond superficial interpretability toward actionable transparency [11, 13, 26, 27, 46, 52, 70, 95, 107]. However, as indicated by both new benchmarks [16, 73, 92] and competitive surveys [27], explainability methods face tradeoffs between model performance and explanation quality, and may still struggle with core semantic or compositional understanding.

Reproducibility is increasingly enabled by standardized workflows, open-source platforms, and benchmarking toolkits [9, 36, 39, 46, 47, 65, 81, 92, 97, 100]. Nonetheless, challenges persist—notably, disentangling confounded sources of improvement, reporting negative results, and accounting for computational environment drift [65]. Competing frameworks emphasize not only code and data sharing but also rigorous ablation and systematic evaluation methodologies.

Responsibility and ethical alignment span technical, organizational, and societal dimensions. New evaluation protocols, dataset audits, and inclusive benchmark designs [11, 16, 39, 42, 65, 68, 73] support more robust, context-sensitive deployment, but persistent issues—including hallucination, model bias, and impact assessment—require continuous empirical scrutiny. It is important to recognize counterpoints from competing surveys [16, 27, 73]: while strategic alignment and fairness are widely acknowledged goals, current frameworks do not always translate ethical intent into practice, especially as LLMs and agents are integrated into critical workflows.

To visualize the interplay between gaps, recommendations, and conceptual pillars, Table 19 offers a concise comparison:

The trajectory of LLM and foundation model research is thus tied to nurturing a transparent, inclusive, and modular scientific culture. The following five pillars, derived from both the present roadmap and recent competing frameworks, offer a foundation for robust, trustworthy development and deployment:

Future work should reinforce these pillars by (a) adopting comprehensive, standardized evaluation frameworks that allow rigorous comparison of both open and closed foundation models [15, 16, 73, 92, 100, 103]; (b) designing and publishing new benchmarks addressing persistent challenges in compositionality, semantic reasoning, and robustness [15, 16, 73, 92, 97]; and (c) prioritizing inclusive practices—such as transparent release of evaluation data, incentivizing negative results, and reducing compute barriers—to foster a broader collective impact [9, 39, 65].

In summary, by foregrounding objectives, clarifying strategic tradeoffs, and directly addressing persistent gaps alongside actionable recommendations, the field will remain poised to pursue auditable, effective, and beneficial advancement of LLMs and foundation models for scientific progress, societal integration, and the wider public good.

References

[1] S. Bakken. 2019. The journey to transparency, reproducibility, and replicability. *Journal of the American Medical Informatics Association* 26, 3 (2019), 185–187. https://academic.oup.com/jamia/article/26/3/185/5301680

[2] Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. 2025. Thinking Machines: A Survey of LLM based Reasoning Strategies. *arXiv preprint arXiv:2503.10814* (2025). https://arxiv.org/abs/2503.10814

[3] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (2022), 221–242. https://aclanthology.org/2022.cl-1.10.pdf

[4] A. Belz. 2021. Quantifying Reproducibility in NLP and ML. *arXiv preprint arXiv:2109.01211* (2021). arXiv:2109.01211 [cs.CL] https://arxiv.org/abs/2109.01211

[5] Anya Belz. 2022. A Metrological Perspective on Reproducibility in NLP. *Computational Linguistics* 48, 4 (2022), 1125–1135. doi:10.1162/coli_a_00448

[6] A. Belz, L. Anastasakos, Y. Zhang, S. Spadine, I. Augenstein, and F. Liu. 2021. A Systematic Review of Reproducibility Research in Natural Language Processing. *Transactions of the Association for Computational Linguistics* 9 (2021), 249–266. https://aclanthology.org/2021.eacl-main.29.pdf

[7] M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston, T. Sternal, M. Copik, G. Kwaśniewski, J. Müller, Ł. Flis, H. Eberhard, H. Niewiadomski, and T. Hoefler. 2025. Reasoning Language Models: A Blueprint. *arXiv preprint arXiv:2501.11223* (2025). https://arxiv.org/abs/2501.11223 version 3, Jan. 2025.

[8] S. Black, A. C. Stickland, J. Pencharz, O. Sourbut, M. Schmatz, J. Bailey, O. Matthews, B. Millwood, A. Remedios, and A. Cooney. 2024. RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents. *arXiv preprint arXiv:2504.18565* (2024). https://arxiv.org/abs/2504.18565

[9] M. Bober-Irizar and S. Banerjee. 2024. Neural networks for abstraction and reasoning: Towards broad generalization in machines. *arXiv preprint arXiv:2402.03507 [cs.AI]* (2024). https://arxiv.org/abs/2402.03507

[10] P. Boersma, T. Benders, and K. Seinhorst. 2020. Neural network models for phonology and phonetics. *Journal of Language Modelling* 8, 1 (2020), 103–177. doi:10.15398/jlm.v8i1.224

[11] S. Carrow, K. H. Erwin, O. Vilenskaia, P. Ram, T. Klinger, N. A. Khan, N. Makondo, and A. Gray. 2024. Neural Reasoning Networks: Efficient Interpretable Neural Networks With Automatic Textual Explanations. *arXiv preprint arXiv:2410.07966* (2024). https://arxiv.org/abs/2410.07966

[12] F. Castagna, G. Pelosi, A. Rago, F. Toni, and C. Wang. 2024. Computational Argumentation-based Chatbots. *Journal of Artificial Intelligence Research* 79 (2024), 129–179. https://www.jair.org/index.php/jair/article/view/15407/27067

[13] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. 2023. Relation Matters: Foreground-Aware Graph-Based Relational Reasoning for Domain Adaptive Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3677–3694. doi:10.1109/TPAMI.2022.3179445

[14] Di Chen, Yiwei Bai, Sebastian Ament, Wenting Zhao, Dan Guevarra, Lan Zhou, Bart Selman, R. Bruce van Dover, John M. Gregoire, and Carla P. Gomes. 2021. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. *Nature Machine Intelligence* 3, 9 (2021), 812–822. doi:10.1038/s42256-021-00384-1

[15] J. Chen, H. Lin, X. Han, and L. Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. *arXiv preprint arXiv:2309.01431, Computation and Language (cs.CL)* (2023), 1–14. https://arxiv.org/abs/2309.01431 v2, accepted to AAAI 2024.

[16] Q. Chen, Y. Hu, X. Peng, Q. Xie, Q. Jin, A. Gilson, M. B. Singer, X. Ai, P.-T. Lai, Z. Wang, V. K. Keloth, K. Raja, J. Huang, H. He, F. Lin, J. Du, R. Zhang, W. J. Zheng, R. A. Adelman, Z. Lu, and H. Xu. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications* 16, 1 (2025), Article number: 3280. doi:10.1038/s41467-025-56989-2

[17] George Chrysostomou. 2022. Explainable Natural Language Processing. *Computational Linguistics* 48, 4 (2022), 1137–1139. doi:10.1162/coli_r_00460

[18] C. Cornelio, J. Goldsmith, U. Grandi, N. Mattei, F. Rossi, and K. B. Venable. 2021. Reasoning with PCP-Nets. *Journal of Artificial Intelligence Research* 72 (2021), 1103–1161. doi:10.1613/jair.1.13009

[19] M. Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics* 6 (2018), 241–252. https://transacl.org/ojs/index.php/tacl/article/download/1299/296/3798

[20] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, and et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025). https://arxiv.org/abs/2501.12948

[21] D. Deutsch, N. Kassner, J. Li, R. Reichart, and D. Roth. 2021. A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. *Transactions of the Association for Computational Linguistics* 9 (2021), 1132–1146. https://transacl.org/index.php/tacl/article/view/3125/1031

[22] W. Digan, A. Névéol, A. Neuraz, M. Wack, D. Baudoin, C. Rance, A. Burgun, and P. Rosset. 2021. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association* 28, 3 (2021), 504–515. doi:10.1093/jamia/ocaa261

Table 19: Summary of Persistent Gaps, Key Recommendations, and Foundational Pillars in Foundation Model Research

Persistent Gap	Key Recommendation	Pillar Addressed	Representative Evidence/Surveys
Model/Workflow rigidity	Modular design adoption	Modularity	[42, 53, 75, 91, 102, 103, 106]
Superficial interpretability	Native, causal, and user-centered explainability	Explainability	[11, 13, 26, 27, 46, 52, 70, 95, 107]
Low reproducibility/fragmented evaluation	Standardized pipelines, open benchmarks, rigorous reporting	Reproducibility	[9, 36, 39, 46, 47, 65, 81, 92, 97, 100]
Ethical/real-world misalignment	Inclusive evaluation, societal/context auditing	Responsibility	[11, 16, 39, 42, 65, 68, 73]
Gaps in semantic competency, composition, real-world task robustness	Development of broader, objective benchmarks and alignment with real user needs	Explainability, Responsibility	[15, 16, 27, 73]

Table 20: Pillars for Robust, Trustworthy Foundation Model Research and Deployment

Pillar	Description
Openness	Transparent sharing of models, data, and methodologies; public documentation; facilitating external evaluation and reuse.
Modularity	Composable design of architectures and workflows, enabling rapid innovation, ablation, and cross-domain transfer.
Explainability	Built-in mechanisms for generating rationales, formal explanations, and human-interpretable outputs evaluated for reliability.
Reproducibility	End-to-end transparency in data, code, and environments; adoption of standards for replicable research artifacts.
Responsibility	Continuous empirical audits, inclusive benchmark design, and integration of ethical norms throughout the research lifecycle.

[23] Haijie Ding and Xiaolong Xu. 2024. SAN-T2T: An automated table-to-text generator based on selective attention network. *Natural Language Engineering* 30, 3 (2024), 429–453. https://www.cambridge.org/core/journals/natural-language-engineering/article/sant2t-an-automated-tabletotext-generator-based-on-selective-attention-network/20AA8938239332A0E6C8884DA8329D82

[24] Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics* 5 (2017), 471–486. doi:10.1162/tacl_a_00074

[25] P. Van Eecke, J. Nevens, and K. Beuls. 2022. Neural heuristics for scaling constructional language processing. *Journal of Language Modelling* 10, 2 (2022), 287–314. doi:10.15398/jlm.v10i2.318

[26] M. Eguchi and K. Kyle. 2024. Building custom NLP tools to annotate discourse-functional features for second language writing research: A tutorial. *Research Methods in Applied Linguistics* 3, 3 (2024), 100153. doi:10.1016/j.rmal.2024.100153

[27] Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From LLM Reasoning to Autonomous AI Agents: A Comprehensive Review. *arXiv preprint arXiv:2504.19678* (2025). https://arxiv.org/abs/2504.19678

[28] Figen Beken Fikri, Kemal Oflazer, and Berrin Yanıkoğlu. 2023. Abstractive summarization with deep reinforcement learning using semantic similarity rewards. *Natural Language Engineering* 30, 3 (2023), 554–576. https://www.cambridge.org/core/journals/natural-language-engineering/article/abstractive-summarization-with-deep-reinforcement-learning-using-semantic-similarity-rewards/5A2F74A2BF5FE5AB80206C772E6B7B5B

[29] Michael Fire, Yitzhak Elbazis, Adi Wasenstein, and Lior Rokach. 2025. Dark LLMs: The Growing Threat of Unaligned AI Models. *arXiv preprint arXiv:2505.10066* (2025). https://arxiv.org/abs/2505.10066

[30] Jose L. Garcia, Karolina Hajkova, Maria Marchenko, and Carlos Miguel Patiño. 2025. Reproducibility Study of 'Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation'. *Transactions on Machine Learning Research* 2025 (April 2025). https://openreview.net/forum?id=yYb8lvT0KJ

[31] Marcos Garcia. 2021. Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. *Computational Linguistics* 47, 3 (2021), 699–701. doi:10.1162/coli_r_00410

[32] T. Gauthier, M. Olšák, and J. Urban. 2023. Alien coding. *Artificial Intelligence* 323 (October 2023), 104036. https://www.sciencedirect.com/science/article/pii/S000437022300142X

[33] Y. Ge, Y. Xiao, Z. Xu, M. Zheng, S. Karanam, T. Chen, L. Itti, and Z. Wu. 2021. A Peek Into the Reasoning of Neural Networks: Interpreting With Structural Visual Concepts. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2021), 121–135. https://ieeexplore.ieee.org/document/9146584/

[34] Khalil El Gharib, Bakr Jundi, David Furfaro, and Raja-Elie E. Abdulnour. 2024. AI-assisted human clinical reasoning in the ICU: beyond 'to err is human'. *Frontiers in Artificial Intelligence* 7 (2024), 1506676. doi:10.3389/frai.2024.1506676

[35] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin H. Kann, Shalini Moningi, Jack M. Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Paul J. Catalano, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. 2024. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine* 7 (2024). doi:10.1038/s41746-023-00970-0

[36] Yue Guo, Jae Ho Sohn, Gondy Leroy, and Trevor Cohen. 2025. Are LLM-generated plain language summaries truly understandable? A large-scale crowdsourced evaluation. *arXiv preprint arXiv:2505.10409* (2025). https://arxiv.org/abs/2505.10409

[37] Tobias Hille, Maximilian Stubbemann, and Tom Hanika. 2024. Reproducibility and Geometric Intrinsic Dimensionality: An Investigation on Graph Neural Network Research. *Transactions on Machine Learning Research* 2024 (2024). https://openreview.net/forum?id=vCb_76qX4S

[38] J. Huang and K. Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. *Findings of the Association for Computational Linguistics: ACL 2023* (2023), 1049–1065. https://arxiv.org/abs/2212.10403

[39] Y. In'nami, A. Mizumoto, L. Plonsky, and R. Koizumi. 2022. Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics* 1, 3 (2022), 100030. doi:10.1016/j.rmal.2022.100030

[40] G. Izacard, F. Petroni, L. Hosseini, S. Krone, A. Joulin, S. Khattab, E. Grave, and S. Wang. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research* 24 (2023), 1–35. http://www.jmlr.org/papers/volume24/23-0037/23-0037.pdf

[41] S. Jha, A. Sudhakar, and A. K. Singh. 2019. Learning cross-lingual phonological and orthagraphic adaptations: a case study in improving neural machine translation between low-resource languages. *Journal of Language Modelling* 7, 2 (2019), 101–142. doi:10.15398/jlm.v7i2.214

[42] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438. doi:10.1162/tacl_a_00324

[43] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics* 48, 1 (2022), 159–218. https://aclanthology.org/2022.cl-1.8.pdf

[44] Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2025. Hierarchical Document Refinement for Long-context Retrieval-augmented Generation. *arXiv preprint arXiv:2505.10413* (2025). https://arxiv.org/abs/2505.10413

[45] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and Applications of Large Language Models. *arXiv preprint arXiv:2307.10169* (2023). https://arxiv.org/abs/2307.10169

[46] J. Kauffmann, M. Esders, L. Ruff, G. Montavon, K.-R. Müller, and W. Samek. 2024. From Clustering to Cluster Explanations via Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 35, 2 (2024), 1926–1940. doi:10.1109/TNNLS.2022.3185901

[47] T. Kew, A. Chi, L. Vásquez-Rodríguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, and M. Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. *arXiv preprint arXiv:2310.15773* (2023), 1–9. https://arxiv.org/abs/2310.15773

[48] M. Kinniment, L. J. K. Sato, H. Du, B. Goodrich, M. Hasin, L. Chan, L. H. Miles, T. R. Lin, H. Wijk, J. Burget, A. Ho, E. Barnes, and P. Christiano. 2024. Evaluating Language-Model Agents on Realistic Autonomous Tasks. *arXiv preprint arXiv:2312.11671* (2024). https://arxiv.org/abs/2312.11671

[49] A. Laurinavichyute, H. Yadav, and S. Vasishth. 2022. Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language* 125 (2022), 104332. https://www.sciencedirect.com/science/article/pii/S0749596X22000195

[50] Wei Li, Yu Liu, Yuhong Guo, L. P. Chau, and Zhanyu Ma. 2024. LibFewShot: A Comprehensive Library for Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 3 (2024), 2959–2976. https://ieeexplore.ieee.org/document/10239698/

[51] Wenxin Li, Shutian Zhang, Lin Lei, Hua Liu, Zhen Liu, and Jingdong Li. 2023. Learning Deep Generative Clustering via Mutual Information Maximization. *IEEE Transactions on Neural Networks and Learning Systems* 34, 9 (2023), 6263–6277. doi:10.1109/TNNLS.2022.3150195

[52] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwen Wang, Wen Zhang, Junwei Wang, Xiang Zhao, Xiaoyan Zhu, and Enhong Chen. 2024. A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multi-Modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 14061–14083. https://ieeexplore.ieee.org/document/10577554

[53] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli. 2024. CL-MASR: A Continual Learning Benchmark for Multilingual ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 4486–4500. doi:10.1109/TASLP.2024.3487410

[54] B. Liu, C. Lyu, Z. Min, Z. Wang, J. Su, and L. Wang. 2025. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *Information Processing & Management* 62, 1 (2025), Article 103907. https://www.sciencedirect.com/science/article/pii/S0306457323004317

[55] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 103–117. doi:10.1162/tacl_a_00638

[56] W. Liu, Z. Ren, and L. Chen. 2025. Knowledge reasoning based on graph neural networks with multi-layer top-p message passing and sparse negative sampling. *Knowledge-Based Systems* 311 (2025), 113063. doi:10.1016/j.knosys.2025.113063

[57] X. Liu, X. Wei, G. Shi, D. Liu, F. Qian, P. Wang, and Y. Zhang. 2022. End-to-end event factuality prediction using directional labeled graph recurrent network. *Information Processing & Management* 59, 2 (2022), Article 102836. https://www.sciencedirect.com/science/article/abs/pii/S0306457321003083

[58] I. Magnusson, N. A. Smith, and J. Dodge. 2023. Reproducibility in NLP: What Have We Learned from the Checklist? *arXiv preprint arXiv:2306.09562, To be published in ACL 2023 Findings* (2023). https://arxiv.org/abs/2306.09562

[59] E. La Malfa, A. Petrov, S. Frieder, C. Weinhuber, R. Burnell, R. Nazar, A. Cohn, N. Shadbolt, and M. Wooldridge. 2024. Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges. *Journal of Artificial Intelligence Research* 80 (2024). doi:10.1613/jair.1.15865

[60] Nick McGreivy and Ammar Hakim. 2024. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence* 6, 10 (2024), 1256–1269. https://www.nature.com/articles/s42256-024-00897-5

[61] Vera Mieskes, Karine Goeuriot, Laura Büchler, Stefan Evert, Stéphanie Kazet, Gaël Bel, Yannis Dupont, Duy-Jin Duh, Fabienne François, Shulin Han, Maria Jones, Ana Kabadjova, Maria Kammas, Camille Kobus, Judith Leveling, Christian Lofi, Gabrielle Parent, Sébastien Pateux, Laurence Pla, Leonardo Romanello, María Lourdes Ruiz-González, and Eric SanJuan. 2018. Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics* 44, 4 (2018), 641–649. https://aclanthology.org/J18-4003/

[62] Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G. Lucas. 2025. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence* 7 (2025), 592–601. doi:10.1038/s42256-025-01005-x

[63] N. Muennighoff, D. Garrette, F. Hernandez, B. Brorsson, H. Buechel, E. Qiu, M. Vania, M. Sporleder, R. Bingel, S. Kanerva, K. Rama, and A. E. G. Blancke. 2025. Scaling Data-Constrained Language Models. *Journal of Machine Learning Research* 26 (2025), 1–91. https://www.jmlr.org/papers/volume26/24-1000/24-1000.pdf

[64] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A Comprehensive Overview of Large Language Models. *arXiv preprint arXiv:2307.06435* (2023). https://arxiv.org/abs/2307.06435

[65] M. N. Nityasya, K. Christodoulopoulos, F. B. Bastani, and J. Kwiatkowski. 2023. A Case for More Rigour in Language Model Pre-Training: Replicability, Reporting, and Evaluations. *Transactions of the Association for Computational Linguistics* 11 (2023), 1343–1358. https://aclanthology.org/2023.tacl-1.75/

[66] L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang, and W. Y. Wang. 2024. Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies. *Transactions of the Association for Computational Linguistics* 12 (2024), 1–19. https://aclanthology.org/2024.tacl-1.2.pdf

[67] Andrea Passerini, Aryo Gema, Pasquale Minervini, Burcu Sayin, and Katya Tentori. 2025. Fostering effective hybrid human-LLM reasoning and decision making. *Frontiers in Artificial Intelligence* 7 (2025), 1464690. https://www.frontiersin.org/articles/10.3389/frai.2024.1464690/full

[68] Y. Perlitz, E. Bandel, A. Gera, O. Arviv, L. Ein-Dor, E. Shnarch, N. Slonim, M. Shmueli-Scheuer, and L. Choshen. 2024. Efficient Benchmarking of Language Models. *arXiv preprint arXiv:2308.11696, Computation and Language (cs.CL), accepted to NAACL* v5 (2024), 1–19. https://arxiv.org/abs/2308.11696

[69] Pavel Prudkov. 2025. On the construction of artificial general intelligence based on the correspondence between goals and means. *Frontiers in Artificial Intelligence* 8 (2025), 1588726. https://www.frontiersin.org/articles/10.3389/frai.2025.1588726/full

[70] M. Pternea, P. Singh, A. Chakraborty, Y. Oruganti, M. Milletari, S. Bapat, and K. Jiang. 2024. The RL/LLM Taxonomy Tree: Reviewing Synergies Between Reinforcement Learning and Large Language Models. *Journal of Artificial Intelligence Research* 80 (2024). doi:10.1613/jair.1.15960

[71] E. Raff, M. Benaroch, S. Samtani, and A. L. Farris. 2024. What Do Machine Learning Researchers Mean by 'Reproducible'? *arXiv preprint arXiv:2412.03854, To appear in AAAI 2025, Senior Member Presentation Track* (2024). https://arxiv.org/abs/2412.03854

[72] N. Ravi, A. Goel, J. C. Davis, and G. K. Thiruvathukal. 2025. Improving the Reproducibility of Deep Learning Software: An Initial Investigation through a Case Study Analysis. *arXiv preprint arXiv:2505.03165* (2025). https://arxiv.org/abs/2505.03165

[73] Nicholas Riccardi, Xuan Yang, and Rutvik H. Desai. 2024. The Two Word Test as a semantic benchmark for large language models. *Scientific Reports* 14 (2024). doi:10.1038/s41598-024-72528-3

[74] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3, 3 (2021), 199–217. doi:10.1038/s42256-021-00307-0

[75] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics* 8 (2020), 264–280. https://aclanthology.org/2020.tacl-1.18/

[76] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine* 7 (2024), Article 20. https://www.nature.com/articles/s41746-024-01010-1

[77] D. J. Schlueter, N. R. Bar, E. Shin, P. Chou, E. Winden, X. Zhou, J. Ramirez, K. Chu, N. Guller, B. Liang, H. E. Armour, J. H. Gilmore, and L. Bastarache. 2024. Systematic replication of smoking disease associations using survey responses and EHR data in the All of Us Research Program. *Journal of the American Medical Informatics Association* 31, 1 (2024), 139–150. https://academic.oup.com/jamia/article/31/1/139/7330649

[78] H. Semmelrock, T. Ross-Hellauer, S. Kopeinik, D. Theiler, A. Haberl, S. Thalmann, and D. Kowald. 2025. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers. *arXiv preprint arXiv:2406.14325, Accepted for publication in AI Magazine* (2025). https://arxiv.org/abs/2406.14325

[79] Xin Shen, Wai Lam, Shumin Ma, and Huadong Wang. 2024. Joint learning of text alignment and abstractive summarization for long documents via unbalanced optimal transport. *Natural Language Engineering* 30, 3 (2024), 525–553. https://www.cambridge.org/core/journals/natural-language-engineering/article/joint-learning-of-text-alignment-and-abstractive-summarization-for-long-documents-via-unbalanced-optimal-transport/46EF85C92B3E4158D89DC2C43E55D621

[80] Georgios Sidiropoulos, Samarth Bhargav, Panagiotis Eustratiadis, and Evangelos Kanoulas. 2025. Multivariate Dense Retrieval: A Reproducibility Study under a Memory-limited Setup. *Transactions on Machine Learning Research* 2025 (Jan 2025). https://openreview.net/forum?id=rHmc5Y6ICg

[81] Michael A. Skinnider, R. Greg Stacey, David S. Wishart, and Leonard J. Foster. 2021. Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence* 3, 9 (2021), 759–770. https://www.nature.com/articles/s42256-021-00368-1

[82] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* 7 (Feb. 2025), 221–231. https://www.nature.com/articles/s42256-024-00976-7

[83] Laura E. Suárez, Blake A. Richards, Guillaume Lajoie, and Bratislav Misic. 2021. Learning function from structure in neuromorphic networks. *Nature Machine Intelligence* 3, 9 (2021), 771–786. doi:10.1038/s42256-021-00376-1

[84] J. Sublime. 2024. The AI Race: Why Current Neural Network-based Architectures are a Poor Basis for Artificial General Intelligence. *Journal of Artificial Intelligence Research* 80 (2024), 1165–1189. https://jair.org/index.php/jair/article/view/15315/26999

[85] Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics* 50, 2 (2024), 795–805. doi:10.1162/coli_a_00508

[86] Shushan Toneyan, Ziqi Tang, and Peter K. Koo. 2022. Evaluating deep learning for predicting epigenomic profiles. *Nature Machine Intelligence* 4, 12 (2022), 1088–1100. doi:10.1038/s42256-022-00570-9

[87] P. Totis, J. Davis, L. de Raedt, and A. Kimmig. 2023. Lifted Reasoning for Combinatorial Counting. *Journal of Artificial Intelligence Research* 76, 14062 (2023), 1–58. doi:10.1613/jair.1.14062

[88] Junichi Tsujii. 2021. Natural Language Processing and Computational Linguistics. *Computational Linguistics* 47, 4 (2021), 707–727. doi:10.1162/coli_a_00420

[89] W. van Woerkom, D. Grossi, H. Prakken, and B. Verheij. 2024. A Fortiori Case-Based Reasoning: From Theory to Data. *Journal of Artificial Intelligence Research* 81 (2024), 1–38. doi:10.1613/jair.1.15178

[90] L. Vaugrante, M. Niepert, and T. Hagendorff. 2024. A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions. *arXiv preprint arXiv:2409.20303* (2024). https://arxiv.org/abs/2409.20303

[91] P. Veličković and C. Blundell. 2021. Neural algorithmic reasoning. *Patterns* 2, 7 (2021), 100273. doi:10.1016/j.patter.2021.100273

[92] A. Waldis, Y. Perlitz, L. Choshen, Y. Hou, and I. Gurevych. 2024. Holmes A Benchmark to Assess the Linguistic Competence of Language Models. *Transactions of the Association for Computational Linguistics* 12 (2024), 1616–1647. https://aclanthology.org/2024.tacl-1.88

[93] Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2146–2157. doi:10.1109/TASLP.2020.3007833

[94] Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. 2024. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine* 7 (2024), Article 16. doi:10.1038/s41746-023-00989-3

[95] Wenguan Wang, Yi Yang, and Fei Wu. 2024. Towards Data-And Knowledge-Driven AI: A Survey on Neuro-Symbolic Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). doi:10.1109/TPAMI.2024.3483273 Early Access.

[96] Y. Wang, Y. Zhang, P. Li, and Y. Liu. 2024. Gradual Syntactic Label Replacement for Language Model Pre-Training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 961–972. doi:10.1109/TASLP.2023.3331096

[97] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics* 8 (2020), 377–392. https://aclanthology.org/2020.tacl-1.25/

[98] C. Wei, K. Duan, S. Zhuo, H. Wang, S. Huang, and J. Liu. 2025. Enhanced Recommendation Systems with Retrieval-Augmented Large Language Model. *Journal of Artificial Intelligence Research* 82 (2025). doi:10.1613/jair.1.17809

[99] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 24824–24837. https://arxiv.org/abs/2201.11903

[100] Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine* 8 (2025). doi:10.1038/s41746-024-01390-4

[101] Xuenan Xu, Ziliang Xie, Mengyue Wu, and Kai Yu. 2023. Beyond the Status Quo: A Contemporary Survey of Multi-View Learning in Speech and Language Processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023), 95–112. doi:10.1109/TASLP.2023.3321968

[102] M. Yang, Y. Wang, and Y. Gu. 2025. Language-based reasoning graph neural network for commonsense question answering. *Neural Networks* 181 (Jan. 2025), 106816. doi:10.1016/j.neunet.2024.106816

[103] S. W. Yang, H. J. Chang, Z. Huang, A. T. Liu, P. Su, W. Cheng, Y. Li, M. Wu, J. Lee, O. Hussein, M. Maciejewski, X. Zeng, C. H. Chen, Y. Tsao, D. Su, P. Beh, P. Zhang, Y. Shinohara, F. Weninger, F. Ni, S. Watanabe, T. Hori, A. Subramanian, K. K. Chin, P. Garcia-Perera, M. L. Seltzer, and H. Y. Lee. 2024. A Large-Scale Evaluation of Speech Foundation Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 2884–2899. https://ieeexplore.ieee.org/document/10502279

[104] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* 12 (2024), 45–62. https://aclanthology.org/2024.tacl-1.4.pdf

[105] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223 [cs.CL]* (2023). https://arxiv.org/abs/2303.18223

[106] Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-Domain Detection for Natural Language Understanding in Dialog Systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 1198–1207. https://ieeexplore.ieee.org/document/9052492

[107] Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence* 7 (March 2025), 437–447. doi:10.1038/s42256-025-00994-z

[108] J. Zhou, W. Zhong, Y. Wang, and J. Wang. 2025. Adaptive-solver framework for dynamic strategy selection in large language model reasoning. *Information Processing & Management* 62, 3 (2025), Article 104052. https://www.sciencedirect.com/science/article/pii/S0306457324000468