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, 22 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

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

2.1 Evolution of Reasoning in AI

The evolution of artificial reasoning systems traces a path from the early dominance of explicit symbolic logic frameworks to the contemporary prevalence of neural and, most notably, transformer-based paradigms. The earliest AI systems were anchored in symbolic representations, rule-based inference mechanisms, and logic programming, valued for their interpretability and the transparency of their algorithmic operations [69, 91, 95, 105]. Although such approaches enabled rigorous deductive reasoning, they were frequently limited by brittleness in open or ambiguous domains and demanded extensive manual construction of knowledge bases [91].

The subsequent maturation of connectionist models, particularly deep neural networks, marked a paradigm shift towards data-driven representations. These architectures achieved remarkable success in synthesizing hierarchical abstractions, empowering AI systems to handle an expansive array of reasoning tasks without the need for hand-crafted logic [69]. Nonetheless, conventional neural networks exhibited notable deficiencies in generalization, particularly on tasks demanding compositionality, recursion, or algorithmic processing—challenges that highlighted the enduring value of symbolic methods in fields such as arithmetic, logic, combinatorics, and structured stepwise reasoning [95, 105]. Addressing these shortcomings, hybrid neural-symbolic (NeSy) models emerged, aiming to harmonize the perceptual strengths of neural networks with the explicit, interpretable inference afforded by symbolic modules [69, 95]. These frameworks have demonstrated improved performance in domains such as mathematical problem solving, retrosynthetic analysis, and other multi-step reasoning tasks [95]. However, key challenges remain, notably in achieving robust compositional generalization, scalability to extensive knowledge bases, and the seamless integration of symbolic and neural reasoning. This persistent challenge positions the unification of these paradigms as a central research frontier [91, 95].

In tandem, the field has witnessed the unprecedented impact of large-scale transformer-based language models (LLMs), including GPT, T5, PaLM, LLaMA, and Flan, each underpinned by extensive pre-training on massive corpora [10, 25, 38, 45, 54, 55, 64, 76, 91, 99, 105]. These models exhibit emergent reasoning capabilities across tasks in arithmetic, logic, and algorithmic problem solving, especially when advanced prompting strategies—such as chain-of-thought (CoT) techniques—are applied [38, 105]. For instance, CoT prompting has been shown to elicit intermediate reasoning steps, significantly enhancing the models' performance on multi-step problems compared to standard zero- or few-shot settings [38, 105]. Despite these advances, systematic evaluation across benchmarks consistently reveals persistent gaps between the reasoning abilities of current LLMs and those of human experts, particularly on tasks that demand systematic abstraction, logical compositionality, or integration of expansive world knowledge [10, 45, 64]. Even state-of-the-art models like GPT-4, while capable of generating cogent rationales for complex clinical or scientific problems, frequently display errors traceable to superficial correlation learning and remain susceptible to logical fallacies or hallucinations—especially when required to extrapolate beyond their training data [25, 55, 76, 105].

Empirical studies further demonstrate that LLM performance is highly sensitive to the design of prompts, the selection of exemplars, and the mechanisms of knowledge retrieval. Critical reasoning failures persist in multi-step logical inference, combinatorial puzzles, and causal reasoning [10, 38, 76]. Retrieval-augmented CoT strategies have delivered substantive gains for multi-modal scientific and mathematical domains by dynamically identifying and leveraging relevant knowledge sources [99, 105]; however, these gains are often incremental. Persistent brittleness is observed on tasks engineered to probe compositional generalization or causal inference [45, 55, 64, 91]. Collectively, the evidence underscores that while transformer-based LLMs have advanced automated reasoning, their abilities remain largely emergent and stochastic rather than grounded in explicit abstraction or robust causal modeling, thus motivating the continued exploration of hybrid, neurosymbolic, and biologically inspired approaches [91, 95, 105].

2.2 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. 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, have further extended semantic alignment and resilience to contextual variation [23, 79].

Transfer learning—particularly via pre-trained checkpoints from models such as BERT, GPT-2, and RoBERTa—has become a standard modality for adapting large-scale models to downstream tasks with minimal additional training [75]. This paradigm shift has substantially improved access to high-performing models, reducing resource requirements and enabling widespread success across tasks such as translation, summarization, and machine reasoning [75]. 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].

Notably, encoder-decoder architectures now explicitly model tabular structure or document salience to support long-context reasoning and accurate summarization, while retrieval-augmented systems incorporate external information to improve reasoning fidelity [75, 79, 93].

Despite these advances, significant architectural limitations endure:

- Models frequently underperform when processing extended input contexts, with accuracy declining as relevant information is dispersed across longer sequences [55].
- Standard embedding mechanisms, adept at capturing local and semantic dependencies, are less effective for structured data (e.g., tables, knowledge graphs) absent specialized architectural enhancements [75, 93].
- Innovations such as structured attention, field-content selective encoders, and advanced pooling strategies are actively being explored to address these challenges [23, 75, 79, 93].

These ongoing research directions aim to bridge the gap between flexible, general-purpose architectures and the demands of explicitly structured or long-context reasoning tasks.

2.3 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 complex wiring and dynamics of biological connectomes are widely hypothesized to underpin the cognitive flexibility and generalization observed in human reasoning. This perspective has fostered the design of neuromorphic systems and reservoir computing models that emulate canonical features of brain networks, such as modularity and criticality [83]. Recent empirical results indicate that reservoir computing architectures incorporating brain-inspired topologies outperform random networks on tasks requiring flexible generalization, underscoring the computational benefits of structural segregation and integrative dynamics [83].

The advantages of biologically inspired architectures are manifold:

- They provide explanatory models for the emergence of cognitive flexibility and compositionality in biological reasoning [83].
- They offer practical strategies for enhancing the efficiency, adaptability, and robustness of artificial reasoning systems, particularly in dynamic and ambiguous environments.
- They motivate integrative approaches that synthesize cognitive, neural, and symbolic methods, with the aspiration of achieving the recursive and adaptive reasoning capacities characteristic of biological intelligence [83, 95].

In summary, the historical and foundational landscape of AI reasoning comprises a dynamic interplay between symbolic, neural, and hybrid paradigms; innovations in representation and architecture; and the emerging influence of neuroscience-inspired methodologies. Each trajectory contributes unique strengths and faces distinct limitations (see Table 1), collectively informing the ongoing evolution and future direction of reasoning-enabled AI research [69, 75, 83, 91, 95, 105].

3 Benchmarking Speech and Language Models

3.1 Standardized Frameworks and Leaderboards

The evaluation of speech and language models has evolved significantly due to the emergence of standardized benchmarking frameworks and public leaderboards, which facilitate systematic assessments of generalization, robustness, and task coverage. In the domain of speech processing, the Speech processing Universal PERformance Benchmark (SUPERB) serves as a comprehensive, extensible, and reproducible platform designed to evaluate foundation models across a diverse set of 15 tasks, including phoneme

Table 1: 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

recognition, keyword spotting, speaker identification, and automatic speech recognition. Through unified evaluation protocols and methodically constructed multi-task procedures—such as fixed feature encoders paired with task-specific prediction heads, and statistically robust metric aggregation—SUPERB enables rigorous comparison across 33 models encompassing both self-supervised and conventional paradigms. Importantly, SUPERB's insistence on reproducibility, robust statistical testing, and open-source benchmarking resources has accelerated the community's ability to reach consensus on model performance and limitations, while also revealing persistent vulnerabilities in generative and low-resource scenarios [68, 103].

Analogously, natural language processing (NLP) relies on frameworks such as HELM and DIoR to offer methodologically robust, scenario-based leaderboards that encompass a broad range of multi-domain tasks spanning Wikipedia and news text to biomedical corpora. These frameworks transcend surface-level metrics by incorporating evaluations centered on societal impact, reliability, and efficiency [47, 68]. The deliberate inclusion of well-curated, domain-diverse datasets is vital: contemporary large-scale studies show that benchmark composition can appreciably influence model rankings and perceived performance, particularly as the range of covered domains and task types expands [47].

A noteworthy innovation is the introduction of continual learning benchmarks, such as CL-MASR for multilingual automatic speech recognition. These benchmarks systematically arrange sequences of tasks and languages specifically to reveal deficiencies in models' capacity to acquire new skills without succumbing to catastrophic forgetting. The CL-MASR benchmark supplies standardized, reproducible task sequences and a comprehensive suite of metrics—including Word Error Rate, measures of forgetting, backward transfer, and intransigence—to facilitate systematic evaluation of catastrophic forgetting, cross-lingual interference, and data/resource imbalance issues, especially in low-resource or highly typologically diverse environments. Furthermore, the open-source nature of CL-MASR advances direct reproducibility and collaborative method development within the community [53]. Collectively, these trends illustrate the rising expectations for multi-domain, resource-robust, and reproducible benchmarking in both speech and language modeling research.

3.2 Evaluation Metrics and Best Practices

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]. Recent evaluation approaches have shifted toward semantically grounded metrics that more accurately reflect human preferences and understanding—for example, leveraging cross-encoder or bi-encoder models fine-tuned for semantic similarity, which consistently outperform traditional n-gram overlap measures in both general and specialized contexts [16, 81]. Notably, metrics informed by question-answering (QAEval) or natural language inference consistently exhibit stronger alignment with human assessments of comprehension than do surface lexical overlap metrics, a trend particularly pronounced in specialized or layperson-facing applications [16, 28].

Despite such advancements, challenges in metric selection persist. Many widely adopted metrics, especially in generative or low-data settings, still fail to deliver robust, statistically reliable distinctions between models, and some exhibit instability under variations in evaluation setup—such as changes in scenario grouping, example sampling, or aggregation strategy, as evidenced by analyses of DIoR within the HELM benchmark [16, 47, 68]. The resultant volatility of metric-based leaderboards in response to moderate alterations in benchmark composition or evaluation protocol underlines the necessity for transparency in metric definitions, comprehensive statistical reporting, and precise articulation of scenario aggregation methods. Furthermore, there is growing recognition of the need for composite or scenario-weighted evaluation methodologies [39, 47, 68]. 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

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 2.

Table 2: 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

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 and weaknesses.
- The evolution of evaluation metrics now emphasizes meaningful alignment with human judgment, particularly in complex and layperson-facing tasks.
- Comparative studies underscore the trade-offs between LLMs and fine-tuned domain-specific models, spotlighting ongoing requirements for task adaptation and careful resource allocation.
- Diversity, extensibility, and open scientific practices are foundational to future-proofing benchmarks and maximizing their impact across domains.

4 Probing, Reasoning, and Linguistic Competence Benchmarks

4.1 Linguistic and Reasoning Probing

The evaluation of large language models (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.

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]. 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, 97]. 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 a 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 [15, 81, 100]. Ensemble approaches, which combine these methods, achieve broader coverage, yet no single paradigm independently solves a majority of cases, illustrating the persistent difficulty of abstract reasoning and broad generalization [15, 100].

Specialized domains have further spurred the development of targeted benchmarks. For example, biomedical and clinical reasoning datasets such as MedS-Bench and arckit extend probing

into domain-specific abstraction and reasoning. Results show that even the most advanced LLMs, including GPT-4 and Claude-3.5, exhibit divergent abilities between real-world and multiple-choice scenarios; these models excel at the latter but consistently underperform on tasks requiring complex clinical information extraction or summarization [9, 15]. Such outcomes spotlight the ongoing disconnect between benchmark performance and deployable, real-world reasoning competence.

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].

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, the limitations inherent to single-view and unisource evaluation frameworks have become even more pronounced. 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].

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].

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].

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

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 multimodal 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]. For example, models that exhibit high scores on reasoning benchmarks through chain-of-thought or instruction-based prompting are, upon closer scrutiny, sometimes found to yield improvements that lack statistical significance under rigorous experimental replication. This reality exposes concerns over the replicability and meaningfulness of current evaluation pipelines [14, 23, 27, 50].

Furthermore, 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 [42, 97].

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]. Yet, 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 3, each major benchmarking theme offers crucial diagnostic capabilities while simultaneously exposing core limitations that remain unresolved.

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

4.4.1 Prompt-based Evaluation and Knowledge Probing. Prompt-based evaluation has emerged as a cornerstone for assessing the knowledge and reasoning capacities of large language models (LLMs). Notably, benchmarks such as the LAMA probe utilize cloze-style prompts to estimate factual recall. However, evidence indicates that such prompts systematically underestimate model knowledge due to their rigid syntactic format and lack of paraphrastic diversity [42]. Recent developments, including paraphrasing-based and mining-based approaches as implemented in the LPAQA suite, reveal that a more diverse set of high-quality prompts can extract substantially greater knowledge—achieving up to an 8.5% absolute improvement on LAMA. This demonstrates a considerably tighter lower bound on ascertainable model knowledge [42].

Yet, these advances introduce new challenges. Chief among them is prompt sensitivity: minor variations in prompt phrasing can yield large fluctuations in answer accuracy, resulting in instability across experimental runs and impeding robust inter-study comparisons. Furthermore, limitations in prompt-based benchmarks, particularly those focused on simple factual recall or compositionality (such as the Two Word Test, TWT), reveal that LLMs continue to struggle with distinguishing meaningful linguistic compositions from nonsensical ones. Even state-of-the-art models frequently rely on superficial lexical or vector similarities rather than genuine compositional semantics, a pattern not observed in human language understanding [73]. Such findings underscore the need for caution when equating high performance on surface-level tasks with true language comprehension.

Despite ongoing progress in benchmark development, prompt-based knowledge measurement exhibits several persistent limitations:

- Susceptibility to surface-level artifacts and syntactic cues;
- High variability and unpredictability arising from prompt paraphrasing;
- Inadequate robustness and reproducibility of results, particularly across diverse experimental conditions.

These challenges are magnified in specialized domains such as biomedical and clinical contexts, where domain-specific terminologies and schemas can further exacerbate unpredictable generalization patterns [16, 100]. Achieving transparency and comparability requires open access to probing datasets (e.g., TWT and LPAQA) and rigorous, systematic reporting of prompt construction methodologies [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]. 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]. 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, 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, 96].

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.

Table 3: 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

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]. 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 4, 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—especially in biomedical and clinical tasks—highlight the paramount importance of robust evaluation methodologies and transparent reporting. Comparative assessments between general-purpose LLMs and specialized, fine-tuned models (e.g., BioBERT, PubMedBERT, BART) reveal that while general models often outperform on tasks requiring open-domain reasoning or complex question answering (such as medical licensure examinations), they typically lag behind specialized systems in extraction or classification benchmarks. Moreover, large LLMs exhibit elevated rates of hallucination, missingness, and output inconsistency, particularly in zero- or few-shot settings [16, 100].

A notable trend involves closed-source models (such as GPT-4) achieving state-of-the-art reasoning performance, albeit at higher computational cost, while open-source models often derive greater benefit from broad, instruction-optimized data rather than domain-specific pretraining [16]. Dynamic prompting methods (including few-shot CoT and instruction-based tuning) help mitigate inconsistency and hallucination, but do not fully close the performance gap. Even advanced instruction-tuned models (e.g., MMedIns-Llama 3) face ongoing challenges regarding comprehensive scenario coverage, multilingual capabilities, and real-world clinical applicability [100].

Transparency in reporting—encompassing benchmark releases, dataset availability (such as TWT and LPAQA), model code dissemination, and standardized evaluation protocols—has become essential for scientific progress and reproducibility in this space [42, 73, 100]. In high-stakes domains, such transparency is both a methodological and ethical imperative, ensuring that errors, limitations, and failure modes are openly recognized and addressed [16, 42, 73, 100].

In summary, the convergence of rigorous prompt engineering, adaptive training and self-correction methods, and transparent, domain-sensitive evaluation practices defines the present boundary of robust knowledge measurement and reasoning in LLMs. Nonetheless, ongoing research must confront the intertwined challenges of prompt sensitivity, adaptation robustness, domain complexity, and reproducibility if it is to realize genuinely reliable, generalizable, and interpretable language models.

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

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

Table 4: 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

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

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 combination of graph neural networks (GNNs) and LLMs has yielded significant benefits for tasks requiring the synthesis of unstructured and structured knowledge, such as 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]. These architectures encode structured information (e.g., knowledge graphs or tabular data) as graph representations, facilitating fine-grained reasoning by way of message passing, aggregation, and selective propagation, while leveraging the extensive contextual knowledge inherent in LLMs. As a prominent example, LBR-GNN fuses contextualized linguistic and graph representations, utilizing edge aggregation and targeted message passing to enhance common-sense question answering beyond the capacity of individual paradigms [102]. Additionally, frameworks that align multi-modal and textual data through chain-of-thought demonstrations have enabled complex scientific reasoning by jointly leveraging neural and structured elements [56].

In scientific, mathematical, and biomedical domains, these methods provide powerful mechanisms for encoding domain constraints, probabilistic relations, and hierarchically organized knowledge—attributes critical for reliable inference and interpretability [18, 23, 26, 27, 30, 44, 52, 56, 79, 87, 89, 95, 102, 107]. For combinatorially challenging problems, such as mathematical theorem proving, molecular property prediction, or scientific discovery, the fusion of neural networks with symbolic and probabilistic reasoning confers considerable performance enhancements paired with interpretable modeling [18, 23, 26, 27, 30, 44, 79, 87, 89].

In the biomedical field, in particular, the application of graph-based, symbolic, and hybrid reasoning methods has yielded tangible real-world impact. Approaches such as LLMs augmented with domain-specific symbolic and graph-based modules have proven superior to generic LLMs for tasks including extraction of social determinants from electronic health records (EHRs), clinical text classification, diagnosis assignment, and information extraction [10, 14, 16, 25, 27, 32, 34, 35, 43, 52, 54, 60, 64, 74, 76, 87, 94, 95, 100, 107, 108]. For example, enhancements through structured knowledge codes lead to improved detection of adverse social determinants and show reductions in demographic bias [34, 35, 60, 76]. Diagnostic frameworks leveraging these methods excel in DRG classification, rare disease recognition, and clinical narrative interpretation, providing interpretable rationales that facilitate actionable clinical insights [16, 32, 34, 52, 54, 64, 74, 87, 100, 107, 108].

Nevertheless, integrating graph-based, symbolic, and neural methodologies presents significant challenges:

- Scaling GNNs to handle massive, evolving knowledge graphs remains non-trivial.
- Managing compounded errors or hallucinations at the neural-symbolic interface is difficult.
- Automated construction of high-fidelity graph structures from noisy or heterogeneous data sources is an ongoing obstacle.
- Biomedical and scientific fields are further challenged by limited annotated data, incomplete or inconsistent ontologies, and bias within domain corpora, all of which impair generalizability and trustworthiness [16, 27, 32, 34, 64, 102, 107].

Progress has been made through advancements such as standardized benchmarks for knowledge graphs, robust data augmentation, and instruction-tuned LLMs adapted to clinical and scientific content. However, achieving scalable, reliable, and fully explainable graph-based reasoning in practical applications remains contingent on continued methodological and theoretical innovations [10, 14, 16, 18, 25, 27, 35, 44, 52, 56, 60, 74, 87, 94, 95, 100, 102, 107].

Table 5: 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

6 Evaluation Methodologies, Interpretability, and Transparency

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]. Evaluation of LLM-generated plain language summaries illustrates that, whereas automated and even subjective metrics may indicate close resemblance to references, outputs frequently yield lower actual comprehension when subjected to rigorous objective assessments. This discrepancy underscores the necessity for metrics that transcend lexical or stylistic similarity and instead emphasize downstream impacts, such as actionable understanding or decision support [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 advances advocate for the adoption of multi-faceted evaluation strategies, integrating question-answering-based metrics, semantic similarity scoring, and rigorous human-in-the-loop assessments that prioritize comprehension and trust calibration over mere surface agreement [16, 68, 103]. At the same time, reproducibility has emerged as a core methodological concern. The prevalence of heterogeneous experimental designs, insufficient transparency regarding code and data, and environment-specific dependencies have contributed to widespread reproducibility challenges in LLM and deep learning research [28, 47, 100]. To address these issues, current guidelines urge:

- Replicating computational environments,
- Providing comprehensive documentation of model architectures and pipelines,
- Sharing data and code through open repositories, and
- Conducting systematic sensitivity analyses

to bolster trustworthiness and promote scientific progress [28, 39, 47].

Benchmark design is itself subject to increasing scrutiny in pursuit of both efficiency and rigor. Studies indicate that efficient benchmarking—implemented by reducing redundant evaluation without compromising reliability—can notably decrease computational costs and environmental impacts, provided aggregation strategies and scenario diversity are judiciously selected [39, 103]. Importantly, limitations inherent to widely-used benchmarking approaches have become apparent: static benchmarks often fail to capture the interactive, causal, or real-world reasoning capacities of contemporary models. This observation underscores the need for more dynamic, robust, and reproducible benchmarks in future evaluation protocols [28, 68, 81].

As outlined in Table 6, 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 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. However, the logical coherence of their rationales often correlates with answer correctness—logical errors serve as possible signals for human oversight [12, 38]. 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].

Table 6: 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

To enable interpretability beyond post-hoc justification, contemporary methods have begun to embed explanation mechanisms directly within model training and input representations. Techniques including hierarchical clustering and feature learning frameworks facilitate attribution of outputs to specific input features or groups. This enables both:

- Local interpretability (instance- or case-specific explanations),
- Global interpretability (class- or cluster-level insights)

thus enhancing transparency across scales [46, 67, 107]. Neural symbolic computing (NeSy) further attempts to integrate deep learning's representational capability with symbolic AI's logical structure and auditability, exhibiting promising outcomes in mathematical, scientific, and decision-making applications. Nonetheless, challenges persist concerning compositional generalization and performance scaling [10, 17, 33, 82].

Interpretability in unsupervised tasks—such as clustering or feature extraction—poses unique obstacles due to the lack of ground-truth labels. The advent of neuralized clustering models and mutual information-based hierarchical clustering offers solutions for efficient attribution, enabling explanations of why data points form particular groups and supporting both interpretability and model quality assessment [32, 33, 67]. 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, potentially fostering miscalibrated trust [36, 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—risking the exacerbation of 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 employing model prediction analysis, confidence calibration, and demographic impact assessments have documented failures in both traditional and novel architectures, including increased sensitivity to demographic descriptor variables and uneven accuracy across groups [34, 45, 90, 95]. For instance, fine-tuned models addressing social determinants of health attenuated (but did not eliminate) bias compared to zero- or few-shot LLMs, indicating the need for both data- and architecture-driven mitigation strategies [8, 90].

Transparency throughout the modeling pipeline—including dataset composition, model objective specification, and parameter sharing—remains a prerequisite for detecting and mitigating such risks [3, 17, 72, 89, 107]. Contemporary literature increasingly calls for:

- Open and representative datasets,
- Public code and evaluation resources, and
- Transparent evaluation protocols,

to facilitate robust, community-driven audits and reproducibility [17, 45, 72, 76, 95, 107]. In parallel, transparency within modeling workflows—including visibility into intermediate representations, decision rationales, and potential failure points—is essential for regulatory oversight and informed engagement by diverse stakeholders [3, 32, 67, 72, 89, 95].

Mitigating hallucination and misinformation necessitates coupled strategies: technical interventions (such as factual verification modules or knowledge-grounded models) and organizational safeguards (including red-teaming, continual post-deployment monitoring, and unambiguous user communication) [38, 48, 64, 72, 94, 105]. Furthermore, privacy and security considerations accentuate the importance of open, auditable, and securely managed data practices—especially in high-impact environments like medicine and law [3, 34, 72, 89, 105, 107]. Despite progress, ongoing gaps demand further attention, including the development of truly representative training corpora, robust adversarial testing procedures, and longitudinal audits to monitor emergent risks and behaviors throughout the model lifecycle [8, 17, 48, 72, 105].

In summary, the convergence of advanced assessment methodologies, interpretability frameworks, and bias/fairness auditing is transforming evaluation protocols for LLMs. The field is moving decisively away from narrow, surface-based metrics in favor of comprehensive, reproducible, and ethically attuned approaches that:

- Integrate diverse stakeholder perspectives,
- Foster open scientific practices, and
- Directly confront the central risks and opportunities inherent in contemporary 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

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. 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]. 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.

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;
- Detailed reporting of all hyperparameters, code versions, and environmental dependencies [39].

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]. The integration of standardized provenance mechanisms such as PROV ensures that workflows are not only repeatable but also interpretable across diverse contexts. 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].

These distinctions are captured in Table 7, 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]. 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]. Other progressive frameworks emphasize:

- Protocol registration and systematic appendices;
- Adherence to FAIR (Findable, Accessible, Interoperable, Reusable) principles;
- Explicit empirical validation of methods across diverse settings [29, 71, 78].

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, MedS-Bench, CL-MASR) highlights the need for unified, scalable, and statistically robust evaluation protocols that balance efficiency with breadth and scenario coverage [16, 47, 100].

7.3 Policy Recommendations and Incentives

Addressing the reproducibility crisis requires a dual approach that targets both procedural reform and incentive structures. Foremost is the need for explicit disambiguation of improvement sources in all published research, achieved through mandatory ablation studies, clearly reported experimental conditions, and rigorous benchmarking against well-tuned baselines [39, 65]. Such criteria should be embedded in journal and conference submission standards and underpinned by specialist review focused on statistics and experimental rigor.

Structural incentives are indispensable. Openness in benchmarking, code, and artifact sharing not only enables community evaluation but also fosters scientific accountability—an effect reflected in elevated citation rates and research impact for transparent publications [16, 39, 47, 100]. To this end, policy mechanisms including checklist-mandated artifact submission, embargoed yet verifiable

Table 7: Comparative features of widely used workflow management systems supporting reproducible research.

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

code and dataset releases, and post-publication discussion platforms are recommended to support the systemic shift toward open scientific practice. Moreover, institutionalizing workflow-based repeatability—leveraging tools such as Snakemake and PROV—should become standard for all empirical studies, particularly those of significant societal consequence [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].

Ultimately, a durable solution to reproducibility in AI and NLP research necessitates not only sophisticated computational infrastructure but also robust cultural and procedural transformations. Aligning incentives, rigorously upholding open and transparent standards, and cultivating a research environment that rewards both meticulousness and innovation together constitute the pathway toward resolving the current reproducibility crisis and ensuring continued scientific progress.

8 Safety, Robustness, Scalability, and Automated Pipelines

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. Such findings highlight intrinsic limitations in both proactive training regimes and post-hoc defense strategies. The proliferation of unaligned—at times intentionally adversarial—models underscores an escalating motivation for adversarial usage, a risk that intensifies as access and model training become increasingly democratized [106].

To counteract these evolving adversarial threats, the community has widely investigated out-of-distribution (OOD) detection methods, emphasizing frameworks based on generative adversarial networks (GANs) and autoencoders. These strategies pinpoint anomalous or untrusted inputs by learning granular characteristics of the expected data distribution. Notably, approaches 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, all without necessitating exhaustive manual annotation of unsafe queries [29]. Nevertheless, current robustness remains hampered by the expressiveness constraints of generative models and limited representation of OOD scenarios sampled during training. This shortfall underscores the necessity for more systematic, scalable methodologies capable of dynamically updating detection protocols as adversarial tactics evolve.

A further, interrelated dimension of the safety discourse encompasses privacy, security, and fairness—each exerting critical influence over both open-source and proprietary LLM deployments [34, 35, 38, 45, 64, 67, 72, 76, 90, 95]. Privacy concerns are multifaceted, spanning inadvertent leakage of sensitive data in model outputs, vulnerability to inversion attacks, and risks of re-identification, particularly when LLMs are tasked with processing confidential health or financial information [38, 45, 76]. Security challenges—prompt injection, model extraction, and exploitation of entrenched biases—further complicate institutional adoption and public trust in systems underpinned by LLMs [34, 67, 95]. Compounding these issues is the persistent challenge of fairness: the potential for LLMs to encode and perpetuate societal, racial, or gender biases, thereby propagating or amplifying inequities, notably in domains such as healthcare, law, and finance [35, 64, 72, 90]. Comparative experimental studies have indicated that domain-specific fine-tuning, as well as the integration of synthetic multi-demographic datasets, may decrease model susceptibility to demographic biases. However, such progress remains incremental, necessitating ongoing audits and rigorous benchmarking [45, 72].

In summary, the safety and robustness of LLMs are contingent not on model scale alone, but rather on a systemic integration of adversarial evaluation, OOD detection, privacy-preserving mechanisms, and fairness-aware design—each undergirded by iterative external audit and transparent reporting. Despite sustained research initiatives, LLM safety and robustness remain locked in an adversarial dynamic, wherein defensive techniques must constantly adapt to match the pace and ingenuity of emergent threats [29, 64, 76, 106].

- **Key challenges** addressed in recent literature include:
  - Robust OOD detection under diverse threat models
  - Privacy preservation during sensitive data handling
  - Security against injection, extraction, and misuse
  - Fairness in mitigating demographic and societal biases
- **Mitigation strategies** increasingly emphasize:
  - Model audits and transparent reporting
  - Continuous updating of defense frameworks
  - Domain-specific and synthetic data augmentation

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. This is especially valuable for multi-modal tasks, where stratified retrieval and advanced reranking can elevate both task accuracy and resource efficiency, even under stringent computational constraints [14, 44, 101]. 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, 27, 30, 50, 62, 70, 91, 98]. Notably, the convergence of modular RL and LLM architectures with outcome-driven reward modeling streamlines deployment, particularly in cloud and distributed environments.

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]. Recent benchmarking initiatives, facilitated by efficient evaluation suites and adaptive model compression tools, demonstrate that meticulous pipeline optimization—including minimization of redundant computation, document signal refinement, and aggregation strategy tuning—can materially lower operational costs and carbon emissions with negligible detriment to performance [27, 56, 64, 68]. The widespread adoption of open-source, modular orchestration libraries further accelerates research reproducibility and expedites technology transfer into industrial and public-sector applications [62, 64, 91, 101].

Despite these advancements, important challenges persist. End-to-end automated pipelines remain prone to error propagation, OOD failures, and emergent behaviors as system complexity increases. Achieving a balance between efficiency, accessibility, and rigorous safety or fairness constraints thus demands systematic trade-off analyses and the standardization of auditing protocols across both research and production environments [44, 64, 68, 98]. As the adoption of LLMs accelerates, the continued development of scalable, automated, and cost-conscious orchestration frameworks represents a crucial determinant in unlocking—safely and equitably—the transformative societal potential of advanced AI.

- **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

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

9.1 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

Table 8: Representative paradigms for LLM workflow orchestration

Paradigm	Core Methodology	Notable Advantages
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]
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, 30, 50, 62, 70, 91, 98]
Automated Hierarchical Pipelines	Integration of operator modules and schedulers to choreograph complex, heterogeneous workflows	Facilitates modularity, scalability, and reliability; supports reproducibility [7, 50, 91, 101]

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 9.

9.2 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. As a notable example, structured electronic health record (EHR) codes are often inadequate for reporting social determinants of health, whereas advanced text-mining methods leveraging language models demonstrate improved recall of disparate factors, especially those relating to marginalized groups [34, 45]. The application of synthetic data augmentation and targeted fine-tuning for underrepresented classes has further reduced vulnerability to demographic bias, thus reinforcing the necessity for systematically balanced data pipelines in AI model development [35, 48, 72, 90].

Nevertheless, entrenched and emergent challenges remain:

- Algorithmic audits and benchmarking continue to reveal systematic disparities in model outputs along axes such as race, gender, and socio-economic status, exposing neglected failure modes and driving calls for more nuanced, intersectional evaluation protocols [34, 38, 64, 67, 76].
- The lack of unified standards for evaluating LLMs, combined with a proliferation of ad hoc prompt engineering approaches, has impeded replicability and undermined confidence in observed advances in fairness [8, 48, 90, 105].
- This replication crisis underscores an urgent need for robust experimental design, open data/code sharing, and reproducibility standards to accurately assess and rectify demographic risks.

In parallel, significant regulatory and ethical developments—including GDPR, the EU AI Act, and growing mandates for explainable AI—are shaping both technical design and evaluation practices [72, 95, 105]. Leading research advocates for the integration of fairness constraints, causal inference, and interpretability objectives directly into training and inference workflows, so that regulatory compliance is embedded as a foundational design principle rather than as a post hoc consideration [38, 45, 67, 72, 95]. Legal-theoretic formalisms and hybrid neuro-symbolic systems facilitate the encoding of precedential knowledge, offering promising directions for transparent and auditable AI in sensitive domains [34, 52, 89, 95].

In summary, advancing inclusion, ethics, and demographic representation in multi-modal, multi-view AI necessitates continuous cross-disciplinary engagement, methodological transparency, and a willingness to rigorously confront both the technical and socio-ethical complexities intrinsic to scalable real-world deployment.

10 Societal, Ethical, and Policy Considerations

10.1 Oversight and Accountability

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

Table 9: 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

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 10.

10.2 Human-Centric and Transparent AI Systems

A shift toward trustworthy, human-centered AI necessitates technical excellence embedded within systems explicitly designed for transparency, auditability, and collaborative engagement. Current LLMs exhibit remarkable emergent abilities—engaging in decision-support, delivering recommendations, and mediating high-stakes interactions. Nevertheless, their efficacy is undermined by persistent challenges, including hallucination, systemic bias, and poor calibration of uncertainty, all of which threaten the societal value of these technologies if left unresolved [63, 82, 89, 105]. There exists a recurrent gap between model confidence levels and user perception: users routinely misjudge system certainty, especially when interpretability mechanisms inadvertently inflate apparent confidence. This demonstrates the urgent need for design strategies that:

- Transparently communicate model uncertainty,
- Align explanation style with true model confidence,
- Calibrate user trust to actual system reliability [27, 89].

Diverse strategies to enhance transparency and explainability have emerged. Precedent-based interpretability frameworks, inspired by legal reasoning, allow model decisions to be explicitly traced to underlying training instances or logical deductions, thereby elevating both auditability and contestability of black-box models [67]. Neural-symbolic (NeSy) systems further bridge connectionist learning and symbolic reasoning, yielding semantic explanations that traverse the boundaries between statistical inference and formal logic—improving both user trust and the systems' corrective capacity [45, 67]. While such hybrid interpretability solutions are not yet universally scalable, their conceptual promise delineates a priority research direction for explainable AI, particularly in domains implicating legal, healthcare, and policy decision-making [45, 67].

In aggregate, ecosystem-level transparency includes the adoption of:

- Open, standardized benchmarks,
- Comprehensive evaluation protocols,
- Proactive and reproducible release practices,
- Infrastructure that supports transparent, standardized reporting [45, 63, 67].

Table 10: 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

Persistent reliance on superficial metrics is increasingly recognized as insufficient; broad confidence intervals, infrequent statistically significant improvements, and substantial unexplained variance collectively underscore the necessity of refined methodologies and confidence-calibrated reporting practices [105]. The intricate dynamics of human-LLM interaction introduce fresh error modes and biases—such as automation bias and overreliance—that can only be meaningfully addressed by centering system design and evaluation on human factors, complementarity, and inter-disciplinary collaboration [27].

Ultimately, the realization of human-centric AI is contingent not only upon technical interventions but also on systemic changes in research culture and policy. Key pillars include:

- Comprehensive pre-registration of studies,
- Specialist ethics review at both organizational and publication levels,
- Automated, transparent data reporting,
- Sustained discourse and post-publication monitoring [34, 67, 84].

By embedding these protocols into academic norms and industrial practices alike, the AI community advances toward systems that are not merely powerful, but demonstrably fair, accountable, and aligned with the public good [8, 27, 34, 67, 69].

11 Persistent Gaps, Open Challenges, and Strategic Recommendations

11.1 Identification of Persistent Gaps

Despite substantial advances in large language models (LLMs) and their integration into diverse natural language processing (NLP) 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 suggests that LLMs remain brittle 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, 10, 25, 26, 42, 49, 101]. 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, 65, 67, 87]. Though the field has witnessed a proliferation of new datasets and evaluation frameworks, these do not fully capture the intricacies of human linguistic reasoning, which can result in the overestimation of LLMs' true capabilities [42, 45, 53, 93, 96, 97, 101].

There remain pronounced disparities between human and model performance, especially on tasks demanding true compositional semantics or abstraction [26, 42, 45, 49, 97]. Even when language proficiency is high, LLMs typically fail to exhibit the flexible abstraction and robust common sense shown by humans [45, 97]. Many models obtain seemingly high scores by exploiting dataset artifacts or superficial correlations, but their performance degrades sharply under adversarial, out-of-distribution (OOD), or compositionally challenging conditions [27, 42, 93].

Challenges in fairness, auditability, and demographic robustness remain unresolved. Although data augmentation and the use of synthetic data offer partial mitigation, significant risks of demographic or social bias persist, exacerbated by both the composition of training data and model architectures. This is particularly problematic in sensitive sectors such as healthcare and law [10, 25, 27, 35, 43, 52, 81, 83, 95]. Calls for comprehensive, multi-level auditing and beyond-token bias mitigation strategies are widespread but have not reached widespread adoption or implementation [14, 25, 43, 83, 95].

Interpretability also presents 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]. Although advances in neurosymbolic reasoning and explainable AI have shown promise, issues with scalability and practical integration into LLM pipelines persist [9, 11, 14, 16, 18, 40, 69]. Innovations such as neural-symbolic hybrids, structural concept extraction, and probabilistic explanations, though valuable, have yet to achieve accessible and efficient scaling for broad applications [9, 11, 16, 18, 64, 69].

Robustness to input perturbation and adversarial attacks is a further area of concern. Recent adversarial testing and deployment experience 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, 30, 63, 81, 93, 100, 106]. Such vulnerabilities highlight the continuous need for advanced robustness evaluation.

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, 55]. The situation is exacerbated by the incomplete adoption of open tools and standardized reporting protocols, which complicate efforts toward reproducibility and replicability [39, 47, 65].

Finally, the lack of universally adopted definitions and quantitative measures of replicability and reproducibility undermines

comparability and reliability in the field. Despite progress via open-source initiatives and reporting checklists, the community's fragmented practices continue to impede fair, transparent, and effective scientific progress [4, 5, 8, 27, 32, 34, 39, 47, 60, 61, 65, 75, 80, 83, 90, 100, 103].

11.2 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 protocols that go beyond conventional accuracy, incorporating semantic and structural faithfulness, robustness to adversarial and noisy inputs, fairness across demographics, and evaluation for multilingual/multimodal competencies [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:** Systematically increase scenario and data diversity in benchmarking, ensuring coverage for compositional, OOD, multilingual, and real-world tasks. Human-in-the-loop evaluation and transparent, objective comprehension metrics should be embedded in model assessment, especially for systems intended for general users [4, 6, 15, 16, 29, 33, 36, 38, 44–46, 51, 55, 62, 68, 93, 101, 106]. Existing benchmarks should be redesigned to eliminate superficial artifacts and better reflect genuine reasoning and semantic competency [42, 45, 53, 93, 97, 106].
- **Hybrid Reasoning Architectures:** Promote the integration of symbolic, neurosymbolic, probabilistic, and neural approaches to address current bottlenecks in compositionality, interpretability, and generalization [18, 27, 40, 50, 56, 60, 62, 64, 69, 72, 74, 78, 86, 107]. Community-driven, open-source initiatives and algorithmic transparency should be incentivized to support research, rapid prototyping, and education [9, 11, 18, 40, 69, 89, 91, 102]. Process-level annotation, trace-based supervision, and outcome-oriented reward mechanisms—especially within reinforcement and hybrid learning contexts—should be given priority [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) paradigm. This includes publishing code, data, models, and complete workflow specifications—preferably containerized and version-controlled to maximize 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 rigorous validation standards, such as comprehensive ablation studies, controlled comparison of pre-training and fine-tuning factors, and 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 discourse are essential to counteract reporting biases and ensure that claimed advances reflect true progress [27, 32, 39, 47, 56, 60, 65, 90, 103].

These technical recommendations can be abstracted into a structured overview for clarity. For this purpose, key persistent gaps and targeted strategies are summarized in Table 11.

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

- 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, also considering environmental sustainability [4, 5, 27, 39, 47, 55, 56, 65, 66, 68, 78, 87–89, 103].

Embedding these strategic priorities into the foundational practices of NLP and AI research is imperative. Only through such coordinated and community-driven efforts can the field ensure trustworthy, equitable, and sustainable innovation in language technologies.

12 Conclusion

12.1 Synthesis of Key Findings

This survey has systematically mapped the swiftly evolving landscape of large language models (LLMs) and foundation models, foregrounding their notable advances while critically examining persistent and emergent challenges in reasoning, benchmarking, interpretability, fairness, robustness, and reproducibility.

Substantial progress has been achieved in enhancing the reasoning capacities of LLMs through novel prompting strategies such as chain-of-thought (CoT) and retrieval-augmented demonstration selection. These techniques have led to significant performance breakthroughs in complex domains, including clinical diagnostics, scientific discovery, and multimodal inference [8, 37, 40, 60, 87, 91, 106]. Such advances are supported by innovations in modular architectures, scalable training paradigms, and the integration of external reasoning modules—including neuro-symbolic and reinforcement learning-based frameworks [14, 20, 67, 91, 95, 102]. Despite these gains, a critical evaluation reveals a persistent gap between current LLMs' linguistic and reasoning abilities and true human-like abstraction; models continue to rely heavily on statistical patterning rather than genuine causal inference or semantic compositionality [8, 20, 95].

Benchmarking efforts have also become more rigorous and diversified, addressing tasks such as biomedical information extraction, negotiation, tabular reasoning, and resilient multi-agent coordination. Nevertheless, contemporary studies consistently demonstrate that even state-of-the-art models maintain vulnerabilities

Table 11: 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

regarding semantic understanding, factual robustness, and cross-modal integration. These findings underscore the imperative to develop new benchmarks and evaluation protocols, specifically designed to reveal failure modes not captured by conventional metrics [2, 5, 40, 71, 78, 87, 103].

The domains of interpretability, fairness, and transparency have similarly attracted focused attention. The deployment of probing classifiers, explainability tools suitable for both unsupervised and supervised models, and rationale-generating architectures has opened new avenues for model introspection and for calibrating user trust [13, 18, 26, 27, 49, 52, 69, 107]. However, significant challenges remain, including the documented risks of end-user over-reliance on persuasive yet potentially misleading explanations, as well as the perpetuation of demographic and algorithmic biases. These issues are particularly acute in high-stakes contexts such as healthcare and law [26, 32, 34, 58, 63, 107]. Contemporary discourse on fairness now attends not only to algorithmic debiasing but also to the centrality of inclusive data practices and ongoing empirical audits.

Despite the proliferation of open-sourcing initiatives, reproducibility persists as a central and unresolved concern. Although the availability of open datasets and libraries—comprising model checkpoints, annotated corpora, and workflow tools—has improved standardization, systemic challenges remain. These include inconsistency in code sharing, undocumented computational environments, artifacts arising from stochastic training, and frequent shifts in hardware or software platforms [23, 36, 44, 46, 51, 79, 81, 91, 102]. Recent attempts to formally define and quantify reproducibility at multiple levels have consistently revealed substantial gaps between nominal claims and practical replicability, a situation further exacerbated by academic incentives that privilege positive results and benchmark overfitting [29, 51, 81, 101, 102]. Although there has been encouraging progress in the form of checklists, community-driven reporting protocols, and the refinement of transparency standards at leading conferences [33, 36, 46], these measures have not yet fully mitigated the threat to scientific trust or facilitated cross-team collaboration.

In summary, it is evident that future advances in LLM research will depend not only on technical innovation but also on structural changes that promote openness and transparency. Adopting modular, standardized workflows—including transparent data management, well-documented codebases, and communal evaluation platforms—remains crucial for fostering robust, trustworthy, and reproducible LLM research and practical deployment [29, 36, 46, 81, 91].

12.2 Future Outlook

The synthesis of recent developments points decisively to an imperative: the advancement of AI systems characterized by modularity, explainability, reproducibility, and responsibility by design [9, 11, 15, 16, 39, 42, 47, 53, 65, 68, 73, 75, 81, 92, 96, 97, 100, 103]. Achieving this vision will require methodological innovation at multiple levels:

- **Modularization in Models and Workflows:** Prioritizing modular design—both in model architectures and experimental processes—will allow for agile, composable testing of new models, datasets, and evaluative techniques. The recent development of high-level architectural blueprints and standardized operator libraries is already fostering the democratization of AI research, accelerating adaptive experimentation across research domains [42, 75, 91, 102, 106].
- **Integrated Explainability:** Making explainability a foundational (rather than optional) component of system design is critical. Advances such as rationale generation, formal causal inference, human-interpretable representation learning, and neuro-symbolic integration hold promise for transitioning from superficial interpretability to actionable and trustworthy model transparency. This is especially vital in sensitive applications—such as clinical, legal, and scientific settings—where the need for error correction and auditability is paramount [13, 26, 27, 52, 70, 73, 95, 103, 107]. Ultimately, robust interpretability will demand integrated approaches: algorithmic explanations, user-centered interfaces, and rigorous empirical studies assessing explanation trustworthiness and impact.
- **Reproducibility and Open Science:** Sustained progress depends on building robust, community-driven infrastructure for research and evaluation. This includes universal adoption of workflow management systems, public repositories for datasets and code, version-controlled software and data, and transparent, standardized reporting practices—all of which have proven effective in fields such as bioinformatics [9, 36, 39, 46, 47, 65, 81, 92, 97, 100]. Broader initiatives are also necessary: incentivizing the sharing of negative results, promoting comprehensive ablation studies, benchmarking against strong baselines, and tracking both code and computational environments to ensure replicability.
- **Responsibility and Ethical Integration:** Every stage of the research and deployment lifecycle should be informed

by a commitment to responsibility. This requires open, collaborative benchmarks and evaluation frameworks that explicitly address inclusion, ethical alignment, and real-world societal contexts [11, 16, 39, 42, 65, 68, 73]. As LLMs and foundation models increasingly underpin key decision-making processes in high-impact sectors, the field is progressively accountable for developing systems that prioritize fairness, accountability, and societal welfare.

Consequently, the trajectory of LLM and foundation model research is intrinsically linked to the ongoing cultivation of a transparent, inclusive, and modular research culture. Progress toward this ideal will be grounded in the following foundational pillars:

Future work must actively reinforce these pillars. As emphasized by recent literature [9, 11, 15, 16, 39, 42, 47, 53, 65, 68, 73, 75, 81, 92, 96, 97, 100, 103], only through a collective commitment to transparency, inclusiveness, and modularity in both technical and cultural dimensions can the field fulfil the promise of next-generation LLMs—for scientific advancement, societal integration, and the broader 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

[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*

Table 12: 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.

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