Proceedings of CONF-SEML 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA33097

The Reasoning Capability of LLMs on Scientific Tasks: A Survey

Dingjun Zhao

School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
dingjun.zhao@unswalumni.com

Abstract. Recent advances in large language models (LLMs) have substantially improved their ability to perform complex reasoning tasks, with some models reaching or exceeding human expert performance in specific domains. Despite this progress, a systematic and comprehensive synthesis of existing research on LLMs' scientific reasoning capabilities remains limited. Existing studies either focus on narrow subdomains or single aspects of the field, lacking integration of benchmarks, models, and evaluation methodologies. To address this gap, this survey provides a structured review of three core aspects of scientific reasoning in LLMs: (1) benchmark datasets used to evaluate scientific reasoning performance, (2) representative LLMs exhibiting differing reasoning capabilities, and (3) evaluation methodologies designed to assess both reasoning outcomes and reasoning processes. In addition, this work analyses key challenges that constrain current progress, including hallucinations, domain-specific data scarcity, and inefficient overthinking in multi-step reasoning. This survey aims to provide a coherent reference framework for researchers and practitioners engaged in the development, evaluation, and application of LLMs for scientific reasoning.

Keywords: Large Language Models, Reasoning, Science

1. Introduction

A review of the evolution of large language models (LLMs) in recent years reveals substantial acceleration, with model advancements exemplified by architectures grounded in Transformer [1]. Among those preeminent LLMs, GPT-5 [2], Gemini-2.5-Pro [3], Qwen-3 [4], and DeepSeek-R1 [5], competencies have been displayed not only across canonical natural language processing (NLP) tasks including question-answering and summary generation but also within specialized scientific domains; instances may be drawn from mathematical reasoning [6,7] combined with programming tasks [8,9].

Within the purview of cognition, situates reasoning itself as an elemental faculty inherent to human thought processes, facilitating an inference-drawing capability derived from available information [10]. The domain of scientific inquiry is demonstrably undergirded by this same aptitude: hypothesis construction, experiment design for empirical substantiation, iterative development of theoretical frameworks, and emergent technological methodologies all rest upon nuanced deployments of logical deduction [11,12].

DOI: 10.54254/2755-2721/2026.BA33097

Research on LLMs' reasoning abilities in scientific contexts has grown rapidly. However, most existing studies concentrate on narrow subdomains, such as specific types of mathematical problems or introductory physics tasks [13,14].

To address this gap, this survey provides a comprehensive review of LLMs' current capabilities in scientific reasoning. Our goal is to offer researchers and practitioners a clear overview of the field's state. We also aim to highlight persistent limitations and suggest targeted directions for future work.

2. Background

Success is observable among LLMs within the domain of NLP tasks extensively. Upon these foundational accomplishments, recent scholarly attention toward harnessing LLM for tackling complex scientific problems has markedly intensified. Manifest differences demarcating scientific reasoning from NLP paradigms, differences reflecting not solely the augmented intricacy inhering in scientific data itself, but also the logical rigour peculiar thereto, standing alongside obligatory engagement with esoteric bodies of disciplinary knowledge and the pronounced fluidity governing mechanisms intrinsic to scientific inferencing activities.

Constitutive of scientific cognition are two salient properties: multimodality and multi-scalarity. Many undertakings assigned to this sphere dictate amalgamation of textual materials, formal sign-systems, regularised datatypes, together with graphical instantiations [15-17]. Within scientific inference,requiblue is alignment holding semantic material in coherence across modalities and gradations alien to common NLP tasks. A mathematical resolution developed through simultaneous interpretation of expository narratives, formally expressed equations, and spatial depictions exemplifies the phenomenon [18]. Properties registered at the macroscale in materials science emerge only via inferential traversals bridging atomic configuration states to corresponding system-wide outcomes [19].

The fundamental to the epistemicity at issue remain elaborated inferential continua entailing both deductions and causalities whose unbroken propagation constitutes criteria for analytic validity, any interruption therewithin negating the conclusions proffered [20-22]. Domain-specific conceptual frameworks necessitate integrativisation, a synthesis requisite between induction over empirical forms, canonical deductivism, and evaluation under hypotheticals contrary-to-fact [23]; the requirement emerges thus that the functional underpinnings of ordinary language model machinery become subject to qualitative realignment when directed towards scientific inference.

Barriers, structural in their origin, admit differentiation into dual typologies. Preeminent stands the hierarchical stratification ordering scientific informational corpora, dependence relations of which best find representation within a highly ramified, specialist schema designed to capture entwined verticality and causative connectivity alike [24]. Acceleration marking epistemic advancement within disciplines illustrates temporal discordance: static inventories maintained by extant LLM configurations encounter increasing inadequacy as knowledge currency decays, the lag endemic to such repositories assuming notable saliency [25,26].

These stringent requirements make scientific reasoning a rigorous test of LLMs' advanced cognitive capabilities. Notably, state-of-the-art models now excel in specific scientific domains: on GPQA, top models attain 83.3%-87.5% accuracy, surpassing the 74% corrected accuracy of human specialists [27]. This demonstrates that reasoning-enhanced LLMs can surpass human performance in targeted scientific reasoning.

DOI: 10.54254/2755-2721/2026.BA33097

3. Scientific reasoning benchmarks

For systematic evaluation of LLMs' scientific reasoning, standardized benchmarks enable quantitative, cross-model comparison and are categorized by two core dimensions: domain coverage and evaluation focus into three types: general-purpose, domain-specific, and logical-symbolic.

Table 1. Summarizes representative benchmarks in each category, including their open-source status, release year, domain coverage, and dataset size

Benchmark	Open-Source	Year	Domain	Size
GSM8K [7]	Yes	2021	Mathematics	8.5k
Livecodebench [8]	Yes	2024	Mathematics	511
GPQA [16]	Yes	2023	Physics & Chemistry & Biology	448
NuminaMath [21]	Yes	2024	Mathematics	8.6k
ARB [28]	No	2023	Math & Physics & Chemistry & Biology & Law	1.1k
UGPhysics [29]	Yes	2025	Physics	5.5k
ChemBench [30]	Yes	2025	Chemistry	2.8k
BioASQ [31]	Yes	2023	Biology	4.7k
MATH [32]	Yes	2021	Mathematics	12.5k
SVAMP [33]	Yes	2021	Mathematics	1.0k
Putnambench [34]	Yes	2024	Mathematics	776

3.1. General-purpose benchmarks

Across the terrain of cross-disciplinary scientific evaluation, benchmarks designated as general-purpose have been advanced, among which can be enumerated ARB [28] and GPQA. Contained within the boundaries demarcated by GPQA are 448 queries in multiple-choice format, wherein an expansive disciplinary span is observed: not confined to biology but equally incorporating domains coextensive with both chemistry and physics. Subdivisions pertinent thereto include thematic regions consonant with quantum mechanics, organic analytical chemistry, as well as the intricacies characterising molecular biology systems. A construction purposefully resistant to simple retrieval through external web resources being postulated for this dataset, it emerges that trained non-specialists' performance manifests in accuracy rates merely reaching 34%, while domain-specific experts demonstrate a raised ceiling, approximating 65%.

3.2. Domain-specific benchmarks

Over breadth, the primacy of analytical depth finds endorsement within domain-specific benchmarks, wherein a concentration upon mono-disciplinary reasoning furnishes evaluative measures attuned to models' grasp upon specialised domains. Instances typified by UGPhysics [29] in physics, ChemBench [30] situating itself within chemistry, BioASQ [31] engaging biomedicine, along with MATH [32-34] concerning mathematics, collectively illustrate such frameworks: each circumscribes its focus to discipline-bounded content. It is through alignment with empirical patterns discerned in authentic research that these instruments permit identification of lacunae resident within the field-specific interpretive capacities possessed by respective systems.

DOI: 10.54254/2755-2721/2026.BA33097

3.3. Logical-symbolic benchmarks

Logical-symbolic benchmarks target abstract reasoning capabilities that are essential for mathematically rigorous disciplines. In contexts distanced from stockpiles of discrete factual knowledge, models become evaluated instead upon criteria related fundamentally to their means of performing formal deduction and sustaining coherent symbolic inferential sequences unbroken by extraneous context. Usually manifesting as tasks employing prescriptive rule sets or exclusively symbolic representation, reliance weakens on peripheral corpus-driven information, foregrounding elementary cognitive mechanisms underlying structured thought. Of note stands PutnamBench [34], which is explicitly built to separate pure logical competence from domain-specific factual memory, providing a direct metric for formal scientific reasoning capacity.

4. Large language models

Early representative LLMs include the GPT (Generative Pre-trained Transformer) series developed by OpenAI [35-37], which established the pre-training and prompting paradigm underlying modern LLMs. Building on subsequent advances in model architecture, training efficiency, and reasoning mechanisms, recent LLMs can be broadly categorized into three types based on their design objectives and implications for scientific reasoning: general-purpose LLMs, domain-specific LLMs, and reasoning-enabled LLMs.

Table 2. Summarizes representative models in each category, including their open-source status, release year, domain coverage, and parameter size

LLM	Open-Source	Year	Domain	Size
GPT-5 [2]	No	2025	General LLM	N/A
Gemini-2.5-Pro [3]	No	2025	Reasoning LLM	N/A
Qwen-3 [4]	Yes	2025	General LLM	235B
DeepSeek-R1 [5]	Yes	2025	Reasoning LLM	671B
Llama-4-Scout [40]	Yes	2025	General LLM	17B
Gemma-3 [41]	Yes	2025	General LLM	27B
DeepSeek-V3 [42]	Yes	2024	General LLM	671B
ChemDual [44]	No	2025	Chemical LLM	N/A
Med-PaLM-2 [45]	No	2025	Medical LLM	N/A
ChemDFM-R [48]	No	2025	Chemical LLM	14B
DeepSeekMath [49]	No	2024	Math LLM	7B
GLM-4.5 [53]	Yes	2025	Reasoning LLM	355B
OpenAI-o3-Pro [54]	No	2025	Reasoning LLM	N/A

4.1. General-purpose LLMs

LLMs whose orientation is toward general-purpose applicability disclose an infrastructural inclination that privileges adaptability above circumscribed specialisation. Central to these system architectures emerges a foundational pursuit: attenuation of procedural dependencies on individualised task recalibrations, which can be discerned through their operational employment of methodologies such as pre-training executed over corpora of monumental scale in consonance with

DOI: 10.54254/2755-2721/2026.BA33097

sophisticated prompting regimens; thereby, burdens imposed by narrowly tailored fine-tuning data requisitioned for singular deployments are significantly alleviated. It is from such methodological underpinnings that phenomena exemplified by GPT-4o's documented performance become manifest, a manifestation observable in its attainment of an 88.1% accuracy metric when subjected to MMLU benchmarking protocols within constraints defined by zero-shot chain-of-thought (CoT) inferential modalities, domains comprising fifty-seven rigorously demarcated academic territories [38,39].

Arising from careful examination of application contexts centred on scientific reasoning is a dialectical dynamic wherein notable dualities persist: appreciable abilities pertaining to generalised domain transfer coexist alongside palpable limitations regarding depth necessary for advanced epistemological engagement within specialised fields [40-42]. Primary origins of such observed inadequacies may be situated in both the absence of thoroughly integrated disciplinary knowledge bases and a partiality or lack at the level of bespoke logical schemas requisite for discipline-specific inference-making. Within empirical evaluative frameworks, the trace of this limitation becomes evident. For instance, even where apex-level universal models are concerned, demonstrable outcomes seldom exceed a 43.22% threshold across measures like the SciBench suite [43].

4.2. Domain-specific LLMs

To the elevation of professional reasoning capabilities within discrete scientific disciplines, specific developmental trajectories for domain-focused LLMs may be ascribed. Reliance upon high-fidelity data, restricted to disciplinary specificity, such data encompassing academic treatises, technical expositaries, experimental data compendia, and systematised repositories for formal knowledge, is evidenced in these models [44,45]. Distilled from such selective datascapes, specialisation is effected not only through precise extractional mechanisms by which specialised knowledges are retrieved, but equally via semantic comprehension at profundities unique to professional contexts, coupled with structural rigour manifested in the formalisation processes assigned to discipline-contained reasoning operations.

Discernible within this domain-specific LLMs' fabrication exists an overarching workflow, whose staged progression can thus be outlined: Initially, curation under expert arbitration results in a corpus composed of high-veracity,domain-relevant materials; subsequently, procedural application of Domain-Adaptive Pre-training in conjunction with Parameter-Efficient Fine-Tuning recalibrates foundational models toward context-congruence proper to the selected field; finally, refinement centres on augmentation of professional inference abilities, operationalised by protocols such as knowledge distillation, reinforcement learning, a process aligned strictly to discipline and customised fine-tunings that privilege advanced inferential behaviours over broad generalisations [46-49]. Among exemplars emergent, Med-PaLM-2 constitutes a model identified for its preeminence in medical inquiry response capacity; credit for such proficiency may be located in both tailoblue fine-tuning directed to clinical discourses and strategic employment of reasoning-adapted prompt configurations. ChemDual, bearing derivation from the LLaMA design space, demonstrates applicability to tasks like chemical reaction prognostication and retrosynthetic analysis, drawing scaffolding from expansive instructional datasets and tokenisations distributed across scales, multi-dimensionally conceived.

4.3. Reasoning-enabled LLMs

Reasoning-enhanced LLMs prioritize depth and complexity handling in reasoning, rather than broad generalization or single-domain specialization. Into pre-existing architectures, advanced reasoning

DOI: 10.54254/2755-2721/2026.BA33097

schemas are incorporated by such models; integrated are methodologies exemplified through augmentation via external tool deployment [50], CoT prompting [51], adoption of divergent inferential stratagems [52], alongside progressions from preliminary nascent inference to sophisticated strong-reasoning modalities. As one instance illustrates, a dual-modal reasonative framework is actualised within Qwen-3, wherein speedy heuristic deduction finds combination with meticulous stepwise analytical progression for the resolution of problems deemed complex.

Strong performances by these reasoning-augmented LLMs can be discerned on benchmarks designed to probe high-level complexity across scenarios markedly varied in character. Yet several conspicuous inadequacies persist. Observable in their operational metrics, reduced competence emerges when faced with routine linguistic tasks encountered daily, a tendency arising as a result of optimisation efforts oriented scantily toward objectives outside the scope of rigorous reasoning [53,54]. Furthermore, observation reveals that, confronted by highly specialised scientific subject matter, deficiencies remain present within reasoning capacity, professional knowledge of structural form, together with rigorously consistent domain-based constraint systems, having not found sufficient integration. From this can be inferblue impediments limiting both performativity within general-use communicative settings and attainment of expert-level analytical penetration, particular to select disciplinary fields.

5. Evaluations for reasoning LLMs

Characterisation of scientific reasoning, observed in the context of multi-step inferential processes, extends beyond reliance upon empirical evidence alone; involvement occurs with meticulous decomposition of complex problems and employment of rigorously justified logical deduction at every constituent juncture. Dominant assessment paradigms, the evaluation of cognitive faculties in both human subjects and LLMs included, have customarily rested on verification of final response accuracy, selection of terminal correctness frequently regarded as an immediate index for generalised reasoning proficiency. Instances arising from emergent disciplinary deployments of LLMs dedicated to nontrivial scientific inquiries demonstrate that recent scholarly investigations have enacted a methodological shift: focal points are increasingly established on scrutinising comprehensive reasoning trajectories,outcome-adjacent aspects receiving less exclusive attention. From this,higher-fidelity appraisals become attainable concerning coherence and formal soundness characterising underlying inference mechanisms within such AI entities.

5.1. Result-based methods

Situated within the evaluative locus dominant among inquiries into LLM modalities, preponderance is evident in frameworks constructed with orientational alignment to result-driven protocols. These protocols, formulated such that the outputs rendered by computational models undergo juxtaposition against referential responses curated antecedently to empirical undertakings, may be distinguished through their structural emphasis on judgment criteria wherein alignment between predictive artefacts and ground truths is foregrounded most saliently. Apparent from textual analyses prevails a systematised interrelation among these paradigms; harboured not exclusively upon operational streamlining or replicative feasibility, it stands but rather suffused also with an implementative directness, one persistently sought throughout analytical and pragmatically technical spheres.

Instituted among contemporary evaluative instruments, BLEU [55] configures its modus operandi around intersectivity quantification of n-gram fragments, calibrating resultant valuations by imposing brevity-related penalties so as to generate representations proportionate to linguistic

DOI: 10.54254/2755-2721/2026.BA33097

concordance. Further advanced is ROUGE [56], whose focus resides in subsequence co-occurrence recurrent across generated and referential texts, thereby designating analytic centrality to contiguous segment overlap in its estimative functions. METEOR [57], developed subsequent to aforementioned schemes, introduces accommodation for lexical variant substitution and permissive rearrangement restructuring, in this allowance for functional analogy, expansion unfolds beyond precedent singular pattern-recognition approaches deemed erstwhile adequate. An assemblage thus emerges whose methodological intricacies reflect typological distinctness intrinsic to differential computation-linguistic ventures, discernibly highlighted amid specificities particularised by each protocol within result-centric assessment territories.

5.2. Process-based methods

A discernible paradigm shift towards methodologies privileging processual progression has emerged, observable amid increasing discontent with evaluativity grounded solely in result-finality. Singular attention to binary verdicts at completion is, within these frameworks, not the analytic endpoint chosen; instead, a gaze diffused longitudinally along multilayered structures of inference phases one finds emphasised. Particularised scrutiny extends toward logical warrant permeating discrete procedural instances. By this mechanism, augmentation occurs within the diagnostic granularity so crucial for scaffolding reason-appraisal metrics, and through such stratified analyses, one can trace the formative stages convergent with principles endemic to anthropic logic-processing [58-60].

Manifestation of an approach oriented thus is detectable in Receval's conception [61], which operates without reliance on external determinants of correctness, given its intention to construct taxonomizations addressing the spectrum inherent in complex deductive chains. There, structuring arises via axes permitting decomposition: each operational unit is subjected initially to atomization, subsequent examination targeting connective webs linking these primitives. Contrasted against this, ReasonEval [62] institutes an appraisal regime arranged upon dual coordinate perspectives, on one hand, veracity attributed to atomic movements, and on the other, their instrumental efficacy in propelling conversion toward holistic resolution. Surfaces from this evaluative scaffold are three annotation typologies demarcated by operational import: those precipitating both accuracy and advancement along the solution continuum; steps adherent to local validity but fictionally inert regarding teleologic momentum; and finally, annotations denoting conceptual derailments emerging where logical fissures, fallacious mappings, or arithmetical deviations arise as disruptive elements within reasoning progressions.

6. Challenges

The application of LLMs to scientific reasoning is constrained by three core challenges. These challenges are hallucinations, data deficiency, and overthinking. They are not independent. Instead, they are mutually reinforcing. Data deficiency leads to insufficient domain knowledge. This lays the foundation for hallucinations. Overthinking emerges as a side effect. It results from using CoT prompting to mitigate hallucinations and enhance reasoning depth. Addressing these interrelated issues requires a systematic and integrated approach.

DOI: 10.54254/2755-2721/2026.BA33097

6.1. Hallucinations

As pertains to the phenomenon of hallucination within generative artificial intelligence, it is denoted as content produced by models that diverges from factual accuracy or exhibits inconsistency with the supplied input context [63]. Conventional taxonomization bifurcates such outputs into the categories recognised as factual hallucinations and faithfulness hallucinations, distinctions whose prevalence reflects not chance occurrence but rather an entrenchment across systemic deficiencies embedded deeply throughout processes foundational to LLM construction. Chronologically distributed, these limitations are: data curation stages, training methodologies employed, and inferential procedures executed during deployment phases [64].

It is apparent from curated datasets deployed at the initial phase of LLM development that much of the hallucinatory output encountered can be ascribed to imperfections therein. As exemplified by pretraining corpora and alignment-list datasets, one frequently encounters inaccuracies regarding factual matters, the manifestation of societal biases, subtle and overt, as well as terminal absences in domain completeness [65]. Under circumstances governed by such incomplete representation, models become compelled, upon facing queries situated outside knowledge boundaries inscribed by their training, toward fabrication, with evidence visible in tasks necessitating extrapolation [66]. Manifest too does this issue remain where protocols for supervised fine-tuning (SFT), insufficiently discriminating in filtration rigour, perpetuate patterns wherein unreliable responses receive inadvertent reinforcement.

SFT paradigms, institutionalising a normativity around answering irrespective of uncertainty, systematically marginalise expressions of epistemic humility, thus incentivising answers plausible only by surface criteria. When RLHF mechanisms dominate, preference induction orients models toward utterances optimised for user fluency alignment, utility perceptions superseding fidelity to facticity; recent literature suggests the emergence of strategic behaviours comprising reward manipulation and excessive deference to anticipated user stances, consequences robustly observable [67,68].

Inferential architecture and prompt-level decoding further exacerbate the tendency toward hallucinated outputs. Stochastic sample-based strategies, integrating diversity objectives, invariably trade off veracity for novelty via augmentation of low-probability selection spaces. In tandem, constraints on reasoning depth inherent to prevailing frameworks and the absence of integrated consistency verification modules result repeatedly in episodic productions devoid of empirical grounding or logical accountability [69].

6.2. Data deficiency

Sufficiency and quality of accessible domain-specific datasets for science remain inadequate; core dependencies exist between the acquisition of high-calibre empirical corpora and both effective parameter training paradigms and operational benchmark calibration [59]. Despite the unremitting proliferation characteristic of general-purpose data corpora within contemporary machine learning research, a pronounced lag persists regarding curated scientific repositories. Constituting a particularity distinct from ordinary textual data are such scientific materials: embedded within them lies a requirement not only for stringent experimental substantiation but also systematic peer evaluation and domain-specialist labelling. From this, it emerges that deficit phenomena in dataset availability have established themselves as a persistent impediment whose influence pervades LLM-mediated scientific cognition, and thorough problem segmentation or resilient inferential synthesis becomes necessary [70].

DOI: 10.54254/2755-2721/2026.BA33097

Evidence exists that highlights initiatives intent on ameliorating these limitations by introducing models constructed upon frameworks characterised by either inclusive or frugal data utilisation. By means of such alternative algorithmic structures, pressures imposed by authentic dataset scarcity may experience partial alleviation. Instances include methodologies akin to Language Self-Play (LSP) [71,72]. Embedded within the LSP architecture is an internal role division: A solitary pre-trained LLM unit alternately assumes dual identities, a proposer devising successively intricate tasks unique to scientific deduction, and, contrarily, a reasoner charged with articulating sequentially derived resolutions. Substantiative results disseminated through assessments place the LLaMA-3.2-3B-Instruct variant, trained via the aforementioned method, at performance parity with contingently sourced baselines for metrics aligned with ScienceQA and MATH experimental platforms. Further augmentation in output has been ascertainvisd where reinforcement strategies coalesce with LSP procedures [73].

Running parallel, recent years have witnessed an intensification in approaches centred around the synthetic generation of scientific data. To exemplify, frameworks typified by SYNTHLLM [74] employ stratified procedural mechanisms: Initial stages encompass thesaural extraction from authoritative academic aggregations, including arXiv and PubMed, alongside canonical didactic treatises. This stage is subsequently followed by structural reshuffling of elementary domain idiolects to instigate broader deductive latitude. Culmination ensues with fine-tuned open-source networks facilitating automated production of logic-sequenced inferential cascades, thereby ensuring synthesis outputs replicate, at scale, the compositional regularities integral to authentic scientific exegesis.

6.3. Overthinking

Frequently observed in applications of large language models, the utilisation of CoT prompting for complex multi-step reasoning problems has become a paradigm attracting considerable attention; embedded within this framework is a discernible tension. Enhanced task performance may be reached through extended processes of logical decomposition, by disaggregating formidable inferential leaps into tractable segments, yet with notable computational cost unavoidably incurred. When increments in problem-solving efficacy align minimally with lengthening chains, alongside a surge in computation demands, an overthinking phenomenon is posited within these systems [75,76].

Restrictive strategies developed at the model level have assumed significance where modulation of internal properties occurs to curtail blueundancies latent in generation pathways; Self-Braking Tuning exemplifies such approaches, by inducing compactness in LLMs' reasoning sequences, an approximate 60% decrease measublue in token expenditure encounteblue on mathematical task sets emerging, with calibration allowing output parity initial performance referents being rendered manifest [77].

A distinct orientation, compressive techniques directed towards reasoning outputs target the structural distillation of protracted cognitive traces: blueuctionist representations formed thereby. It is illustrated by frameworks such as CODI, encoding CoT expressions found in natural-language outputs into latent continuities via a process of self-distillation, the resultant effects being compression characterised by heightened efficiency, yet accompanied by competitive outcomes that surpass those implicit methodologies registered on standard benchmarks, GSM8K among them [78], as attested by measurements.

On another operational front, the regulation at the prompt-level leverages explicit instructionality to delimit complexity native to inference tasks. By example, Chain of Draft confines intermediary

DOI: 10.54254/2755-2721/2026.BA33097

deductions to succinct textual drafts and induces an optimisation concerning both token usage and user-facing latency during model deployment, although only nuanced depressions in accuracy are evidenced across principal LLM platforms examined [79].

7. Conclusion

This paper focuses on the scientific reasoning capabilities of LLMs. It synthesizes findings from mainstream research. Based on this synthesis, it provides a structured analysis of three core aspects. The first aspect is benchmarks designed to evaluate scientific reasoning. The second is LLM performance on such tasks. The third is evaluation methodologies specifically adapted to this setting. Additionally, the paper identifies key challenges. These challenges currently limit the effectiveness and reliability of LLMs in scientific reasoning

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need, " Advances in neural information processing systems, vol. 30, 2017.
[2] OpenAI, "Gpt-5, " https: //openai.com/gpt-5, 2025, accessed: 2025-09-27. Commercial LLM; no open-source technical report available.
[3] G. DeepMind, "Gemini 2.5 pro, " https: //deepmind.com/product/gemini/gemini-2-5-pro, 2024, accessed: 2025-09-27. Proprietary model; technical details not publicly disclosed.
[4] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv et al., "Qwen3 technical report, " arXiv preprint arXiv: 2505.09388, 2025.
[5] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, " arXiv preprint arXiv: 2501.12948, 2025.
[6] K. Huang, J. Guo, Z. Li, X. Ji, J. Ge, W. Li, Y. Guo, T. Cai, H. Yuan, R. Wang et al., "Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations, " arXiv preprint arXiv: 2502.06453, 2025.
[7] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, " arXiv preprintarXiv: 2410.05229, 2024.
[8] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica, "Livecodebench: Holistic and contamination free evaluation of large language models for code, " arXiv preprint arXiv: 2403.07974, 2024.
[9] M. Tian, L. Gao, S. Zhang, X. Chen, C. Fan, X. Guo, R. Haas, P. Ji, K. Krongchon, Y. Li et al., "Scicode: A research coding benchmark curated by scientists, " Advances in Neural Information Processing Systems, vol. 37, pp. 30 624–30 650, 2024.
[10] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with language model prompting: A survey, " arXiv preprint arXiv: 2212.09597, 2022.
[11] D. Shapere, "The structure of scientific revolutions, " The Philosophical Review, vol. 73, no. 3, pp. 383–394, 1964.
[12] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, "Large language models for mathematical reasoning: Progresses and challenges, " arXiv preprint arXiv: 2402.00157, 2024.
[13] Wang, Peng-Yuan, et al. "A survey on large language models for mathematical reasoning." ACM Computing Surveys (2025).
[14] F. Meng, W. Shao, L. Luo, Y. Wang, Y. Chen, Q. Lu, Y. Yang, T. Yang, K. Zhang, Y. Qiao et al., "Phybench: A physical commonsense benchmark for evaluating text-to-image models, " arXiv preprint arXiv: 2406.11802, 2024.
[15] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering, " Advances in Neural Information Processing Systems, vol. 35, pp. 2507–2521, 2022.
[16] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q& a benchmark, " in First Conference on Language Modeling, 2024.
[17] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "Pmc-vqa: Visual instruction tuning for medical visual question answering, " arXiv preprint arXiv: 2305.10415, 2023.
[18] J. Wei, C. Jia, Q. Chen, H. He, L. Sun, C. He, L. Wu, B. Yu, and C. Tan, "Geoint-r1: Formalizing multimodal geometric reasoning with dynamic auxiliary constructions, " arXiv preprint arXiv: 2508.03173, 2025.

DOI: 10.54254/2755-2721/2026.BA33097

[19] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm, " npj Computational Materials, vol. 6, no. 1, p. 138, 2020.
[20] Q. Chen, L. Qin, J. Liu, D. Peng, J. Guan, P. Wang, M. Hu, Y. Zhou, T. Gao, and W. Che, "Towards reasoning era: A survey of long chain-of-thought for reasoning large language models, " arXiv preprint arXiv: 2503.09567, 2025.
[21] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, K. Rasul, L. Yu, A. Q. Jiang, Z. Shen et al., "Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions, " Hugging Face repository, vol. 13, no. 9, p. 9, 2024.
[22] A. Plaat, A. Wong, S. Verberne, J. Broekens, N. van Stein, and T. Back, "Reasoning with large language models, a survey, " arXiv preprint arXiv: 2407.11511, 2024.
[23] A. Patil and A. Jadon, "Advancing reasoning in large language models: Promising methods and approaches, " arXiv preprint arXiv: 2502.03671, 2025.
[24] H. Liu, Z. Fu, M. Ding, R. Ning, C. Zhang, X. Liu, and Y. Zhang, "Logical reasoning in large language models: A survey, " arXiv preprint arXiv: 2502.09100, Feb 2025, arXiv preprint arXiv: 2502.09100v1 [cs.AI] 13 Feb 2025.
[25] N. L. of Medicine, "Medline/pubmed baseline statistics, " https: //www.nlm.nih.gov/bsd/medline pubmed production stats.html, 2023, accessed: 2025-10-17.
[26] A. Merchant and E. D. Cubuk, "Millions of new materials discovered with deep learning, " DeepMind Blog, Nov. 2023, accessed: 2025-10-17. Available at: https: //deepmind.google/blog/millions-of-new-materials-discovered-with-deep-learning/. [Online]. Available: https: //deepmind.google/blog/millions-of-new-materials-discovered-with-deep-learning/
[27] BRACAI, "Gpqa benchmark leaderboard: Testing llms on graduate-level questions, " Web Page, 7 2024, accessed: 2025-10-05. [Online].Available: https: //www.bracai.eu/post/gpqa-benchmark
[28] T. Sawada, D. Paleka, A. Havrilla, P. Tadepalli, P. Vidas, A. Kranias, J. J. Nay, K. Gupta, and A. Komatsuzaki, "Arb: Advanced reasoning benchmark for large language models, " arXiv preprint arXiv: 2307.13692, 2023.
[29] X. Xu, Q. Xu, T. Xiao, T. Chen, Y. Yan, J. Zhang, S. Diao, C. Yang, and Y. Wang, "Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models, " arXiv preprint arXiv: 2502.00334, 2025.
[30] A. Mirza, N. Alampara, S. Kunchapu, M. R´ıos-Garc´ıa, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh et al., "A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists, "Nature Chemistry, pp. 1–8, 2025.
[31] A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras, "Bioasqqa: A manually curated corpus for biomedical question answering, " Scientific Data, vol. 10, no. 1, p. 170, 2023.
[32] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset, " arXiv preprint arXiv: 2103.03874, 2021.
[33] A. Patel, S. Bhattamishra, and N. Goyal, "Are nlp models really able to solve simple math word problems?" arXiv preprint arXiv: 2103.07191, 2021.
[34] G. Tsoukalas, J. Lee, J. Jennings, J. Xin, M. Ding, M. Jennings, A. Thakur, and S. Chaudhuri, "Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition, " Advances in Neural Information Processing Systems, vol. 37, pp. 11 545–11 569, 2024.
[35] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., "Improving language understanding by generative pre-training, " 2018.
[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language models are unsupervised multitask learners, " OpenAI blog, vol. 1, no. 8, p. 9, 2019.
[37] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners, " Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.
[38] Kaggle, "Mmlu (massive multitask language understanding) - open benchmarks, " https: //www.kaggle.com/benchmarks/open-benchmarks/mmlu, 2024, accessed: 2025-12-21.
[39] D. Hendrycks, C. Zhang, S. Basart, and et al., "Measuring massive multitask language understanding, " arXiv preprint arXiv: 2009.03300, 2020.
[40] M. AI, "The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, " https: //ai.meta.com/blog/llama-4-multimodal-intelligence/, Meta Platforms, Inc., Apr 2025, ac-cessed: 2025-10-05; Published on Meta AI Blog.
[41] DeepMind, "Gemma 3: Lightweight, state-of-the-art open mul-timodal models, " https: //deepmind.google/models/gemma/gemma-3/, Google LLC, 2025, accessed: 2025-11-05; Published by DeepMind (Google); Includes model family, benchmarks, and deployment tools.

DOI: 10.54254/2755-2721/2026.BA33097

[42] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan et al., "Deepseek-v3 technical report, " arXiv preprint arXiv: 2412.19437, 2024.
[43] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang, "Scibench: Evaluating college-level scientific problem-solving abilities of large language models, " arXiv preprint arXiv: 2307.10635, 2023.
[44] X. Lin, Q. Liu, H. Xiang, D. Zeng, and X. Zeng, "Enhancing chemical reaction and retrosynthesis prediction with large language model and dual-task learning, " arXiv preprint arXiv: 2505.02639, 2025.
[45] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis et al., "Toward expert-level medical question answering with large language models, " Nature Medicine, vol. 31, no. 3, pp. 943–950, 2025.
[46] C. Yang, R. Zhao, Y. Liu, and L. Jiang, "Survey of specialized large language model, " arXiv preprint arXiv: 2508.19667, 2025.
[47] V. Prabhakar, M. A. Islam, A. Atanas, Y.-T. Wang, J. Han, A. Jhun-jhunwala, R. Apte, R. Clark, K. Xu, Z. Wang et al., "Omniscience: A domain-specialized llm for scientific reasoning and discovery, " arXiv preprint arXiv: 2503.17604, 2025.
[48] Z. Zhao, B. Chen, Z. Wan, L. Chen, X. Lin, S. Yu, S. Zhang, D. Ma, Z. Zhu, D. Zhang et al., "Chemdfm-r: An chemical reasoner llm enhanced with atomized chemical knowledge, " arXiv preprint arXiv: 2507.21990, 2025.
[49] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models, " arXiv preprint arXiv: 2402.03300, 2024.
[50] C. Li, M. Xue, Z. Zhang, J. Yang, B. Zhang, B. Yu, B. Hui, J. Lin, X. Wang, and D. Liu, "Start: Self-taught reasoner with tools, " in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 13 523–13 564.
[51] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models, " Advances in neural information processing systems, vol. 35, pp. 24 824–24 837, 2022.
[52] H. Puerto, T. Chubakov, X. Zhu, H. T. Madabushi, and I. Gurevych, "Fine-tuning with divergent chains of thought boosts reasoning through self-correction in language models, " arXiv preprint arXiv: 2407.03181, 2024.
[53] A. Zeng, X. Lv, Q. Zheng, Z. Hou, B. Chen, C. Xie, C. Wang, D. Yin, H. Zeng, J. Zhang et al., "Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, " arXiv preprint arXiv: 2508.06471, 2025.
[54] OpenAI, "o3-pro model documentation: Technical specifications and api reference, " https: //platform.openai.com/docs/models/o3-pro, OpenAI, Inc., 2025, accessed: 2025-10-05; Published by OpenAI; Includes model reasoning capabilities, 200k context window, API endpoints, and pricing details.
[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation, " in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[56] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries, " in Text summarization branches out, 2004, pp. 74–81.
[57] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, " in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[58] L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng et al., "Improve mathematical reasoning in language models by automated process supervision, " arXiv preprint arXiv: 2406.06592, 2024.
[59] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step, " in The Twelfth International Conference on Learning Representations, 2023.
[60] W. Xiong, W. Zhao, W. Yuan, O. Golovneva, T. Zhang, J. Weston, and S. Sukhbaatar, "Stepwiser: Stepwise generative judges for wiser reasoning, " arXiv preprint arXiv: 2508.19229, 2025.
[61] P. Mondorf and B. Plank, "Beyond accuracy: evaluating the reasoning behavior of large language models–a survey, " arXiv preprint arXiv: 2404.01869, 2024.
[62] S. Xia, X. Li, Y. Liu, T. Wu, and P. Liu, "Evaluating mathematical reasoning beyond accuracy, " in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 26, 2025, pp. 27 723–27 730.
[63] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, " ACM Transactions on Information Systems, vol. 43, no. 2, pp. 1–55, 2025.
[64] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models, " arXiv preprint arXiv: 2401.01313, vol. 6, 2024.
[65] C. Si, Z. Gan, Z. Yang, S. Wang, J. Wang, J. Boyd-Graber, and L. Wang, "Prompting gpt-3 to be reliable, " arXiv preprint arXiv: 2210.09150, 2022.

DOI: 10.54254/2755-2721/2026.BA33097

[66] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y. H. Sung, D. Zhou, Q. Le et al., "Freshllms: Refreshing large language models with search engine augmentation, " in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 13 697–13 720.
[67] K. Li, O. Patel, F. Vi´egas, H. Pfister, and M. Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model, " Advances in Neural Information Processing Systems, vol. 36, pp. 41 451–41 530, 2023.
[68] H. Zhang, S. Diao, Y. Lin, Y. R. Fung, Q. Lian, X. Wang, Y. Chen, H. Ji, and T. Zhang, "R-tuning: Teaching large language models to refuse unknown questions, " arXiv preprint arXiv: 2311.09677, vol. 63, p. 67, 2023.
[69] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models, " in Findings of the association for computational linguistics: ACL 2024, 2024, pp. 3563–3578.
[70] K. Wang, J. Zhu, M. Ren, Z. Liu, S. Li, Z. Zhang, C. Zhang, X. Wu, Q. Zhan, Q. Liu et al., "A survey on data synthesis and augmentation for large language models, " arXiv preprint arXiv: 2410.12896, 2024.
[71] A. Zhao, Y. Wu, Y. Yue, T. Wu, Q. Xu, M. Lin, S. Wang, Q. Wu, Z. Zheng, and G. Huang, "Absolute zero: Reinforced self-play reasoning with zero data, " arXiv preprint arXiv: 2505.03335, 2025.
[72] X. Zhao, Z. Kang, A. Feng, S. Levine, and D. Song, "Learning to reason without external rewards, " arXiv preprint arXiv: 2505.19590, 2025.
[73] J. G. Kuba, M. Gu, Q. Ma, Y. Tian, and V. Mohan, "Language self-play for data-free training, " arXiv preprint arXiv: 2509.07414, 2025.
[74] Z. Qin, Q. Dong, X. Zhang, L. Dong, X. Huang, Z. Yang, M. Khademi, D. Zhang, H. H. Awadalla, Y. R. Fung et al., "Scaling laws of synthetic data for language models, " arXiv preprint arXiv: 2503.19551, 2025.
[75] Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, H. Chen, and X. Hu, "Stop overthinking: A survey on efficient reasoning for large language models, " 2025. [Online]. Available: https: //arxiv.org/abs/2503.16419
[76] M. Hassid, G. Synnaeve, Y. Adi, and R. Schwartz, "Don't overthink it. preferring shorter thinking chains for improved llm reasoning, " arXiv preprint arXiv: 2505.17813, 2025.
[77] H. Zhao, Y. Yan, Y. Shen, H. Xu, W. Zhang, K. Song, J. Shao, W. Lu, J. Xiao, and Y. Zhuang, "Let llms break free from overthinking via self-braking tuning, " arXiv preprint arXiv: 2505.14604, 2025.
[78] Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, and Y. He, "Codi: Compressing chain-of-thought into continuous space via self-distillation, " arXiv preprint arXiv: 2502.21074, 2025.
[79] S. Xu, W. Xie, L. Zhao, and P. He, "Chain of draft: Thinking faster by writing less, " arXiv preprint arXiv: 2502.18600, 2025.