Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 05 No. 01, Jan-June 2025
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.51.63.75

Research Paper

Reasoning capabilities of large language models: a systematic review and meta-analysis of benchmarks, methods, and emergent behaviours (2018–2025)

Prof. Tareq N. Hashem*
*Professor, Department of Marketing, Faculty of Business, Applied Science Private University, Amman, Jordan.

Article Info

Article History:
Received: 29 November 2025
Revised: 15 January 2025
Accepted: 22 January 2025
Published: 07 March 2025

Keywords:
Large Language Models
Chain of Thought
Systematic Review
Meta-Analysis
Gpt 4
Natural Language Inference

ABSTRACT

Large Language Models (LLMs) exhibit solid reasoning capabilities in various domains, including mathematical reasoning, commonsense reasoning, logical inference, scientific question answering, and code generation. But it is not known how these reasoning abilities develop, and how effective they are in various situations. The aim of this systematic review and meta-analysis was to assess the available evidence about LLM reasoning performance up to the end of December 2025. Five major databases (IEEE Explore, Scopus, Web of Science, PubMed and ACM Digital Library) were searched, according to PRISMA 2020. The qualitative review included a total of 88 studies; and, 63 studies included quantitative benchmark data for meta-analysis. The Cochrane framework was used to assess the risk of bias. The results show that LLMs excel in various reasoning tasks, such as mathematical reasoning, commonsense question answering and code generation. The chain of thought prompting techniques consistently enhanced the accuracy of reasoning over the other prompting techniques. On Multistep tasks, tool augmented and self-reflective models performed very well. Despite the progress, LLMs also have limitations when input data is adversarial, when there are distributional shifts, and when the input is compositional. Also, there is significant differences between the studies so the results of the benchmarks cannot be easily generalized. From the review, it is clear that LLM reasoning capabilities are substantial and are making rapid progress, but there are also critical challenges. For the future, the development of more robust, reliable, interpretable and evaluation methods are required for ensuring reliable reasoning performance in real world applications.

Corresponding Author:
Prof. Tareq N. Hashem
Professor, Department of Marketing, Faculty of Business, Applied Science Private University, Amman, Jordan.
Email: t_hashim@asu.edu.jo

1. INTRODUCTION

One of the most significant advances in AI research, Large Language Models (LLMs) have revolutionized the way AI is understood and utilised in terms of machine intelligence and natural language processing. However, starting with GPT-2 and growing in success through GPT3 [1], PaLM [2], LLaMA [3] and GPT-4 [4] trained on web scale corpora using next token prediction have demonstrated an impressive breadth of emergent abilities (behaviours that are qualitatively absent in smaller models and appear at scale without explicit training) [5]. Of these emergent capacities, reasoning has a central place and is in dispute. There is a general belief that reasoning (in the broad sense, defined as the ability to make valid inferences, work on structured problems, and generalize from a limited amount of information) is a sign of human level intelligence, and it is believed that it involves symbolic computation, explicit knowledge representation or both [6]. The apparent ability of purely statistical autoregressive models to acquire reasoning like behaviours challenges the long held assumption of the necessity of intelligent problem-solving skills, and has raised a lot of discussion in the AI research community.

A turning point was the introduction of Chain of Thought (CoT) prompting that obtained step-by-step intermediate reasoning from LLMs before reaching the final answer, which resulted in significant improvements in accuracy across a range of mathematical word problem, commonsense reasoning, and symbolic manipulation tasks [7]. Later work showed that: CoT is size dependent; can be zero shot elicited [8] and can be structured programmes [9]. The frontier of LLM based problem solving has been further advanced by parallel progress in tool augmented reasoning [10] as well as architectures that are self-reflective [10] and process reward models [10].

While this has been a very quick development, the field is missing a comprehensive and methodologically sound synthesis of evidence of the capabilities of LLM reasoning. Current surveys are limited in scope and lack systematic review and meta-analytic methodology and are not recent enough to have included the latest model releases. Thus, practitioners are challenged when they must identify which reasoning tasks have been reliably benchmarked, compare the results between different, and often none homogeneous, evaluation setups, grasp the nature of failure modes and boundary conditions as well as finding priority research gaps. These are the limitations that are overcome by this systematic review.

The rest of the paper is organized as follows: Section 2 summarizes the background work; Section 3 discusses our PRISMA compliant methodology; Section 4 provides results and discussions, including meta-analysis results, Meta synthesis tables and failure mode analysis; and Section 5 ends with a structured research agenda.

2. RELATED WORK

Initial work on language model reasoning tested models on relatively simple and constrained tasks like arithmetic word problems and tasks involving propositional logic. The experiments in all of these studies showed that pre Transformer architectures and smaller neural language models struggled with Multistep reasoning to roughly chance level of performance, which validated the widely held belief that symbolic systems are necessary for formal reasoning.

This was set aside by the introduction of GPT-3 [1] which showed few shot in context learning over a wide range of tasks, such as elementary arithmetic and analogical reasoning. GPT3 was still unable to complete tasks requiring greater than 2 steps of reasoning systematically, however, prompting to scaffold Multistep reasoning is still an area of interest. Chain of thought prompting [7] showed that prompting models to produce intermediate steps between the input and the output data could yield significant

accuracy gains for models with more than around 100 billion parameters on the grade school math benchmark (GSM8K) [10].

Simultaneous research undertook complementary approaches. It was shown in the ReAct framework [11] that by interleaving reasoning traces and interaction with the environment, long term plans can be achieved that are more complex than those that can be performed only through language generation. To address this challenge, the Tree of Thoughts (ToTs) approach [12] introduced a tree of intermediate thought candidates as a way for LLM to explore and evaluate multiple solution paths before settling on a final answer. Programme Aided Language Models (PAL) delegated exact arithmetic calculation to external Python interpreters, thereby reducing the number of errors in the output of LLMs related to arithmetic [13].

A clear research track developed, called self-reflective architectures. In various trials, Reflexion [14] showed that an LLM could actually benefit in terms of performance in multiple trials without fine tuning the gradients, by storing verbal summaries of previously failed trials and leveraging them for future ones. To selectively enrich their context with metacognitive judgments about their own knowledge gaps, SelfRAG [15] incorporated retrieval into generation by adding self-evaluated relevance and factuality assessments.

Most importantly, LLM reasoning was previously done through a set of surveys [5], [16] that were conducted before the advent of models in the GPT4 class and lacked a meta-analytical quantitative synthesis. This systematic review aims to call attention to both of those missing areas, by using PRISMA 2020 methodology to span a time window of 88 studies (from 2018 to 2025) and thus provide principled estimates of effect sizes, heterogeneity, and moderators of LLM reasoning performance.

3. METHODOLOGY

3.1 Protocol and Registration

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PsRISMA) 2020 guidelines [17] were followed for conducting this systematic review. The review protocol was preregistered on PROSPERO (Registration ID: CRD42025412847), before searching the databases. The PRISMA flow diagram provides a summary of the search and selection process followed Figure 1.

3.2 Eligibility Criteria

To be considered, studies had to: (i) test one or more LLM on a structured reasoning task that includes quantitative performance metrics; (ii) be published in peer reviewed publication venues (journals, conference proceedings) or deposited on arXiv and later accepted to a conference; (iii) use English as the language of the text; and (iv) have been published in the period of 2018-01-01 to 2025-12-31. Studies were removed if they contained only qualitative evaluation, but did not have any benchmark scores, or if they were technical reports, but not published or with no citation evidence.

3.3 Information Sources and Search Strategy

The following five bibliographic databases were searched: IEEE Explore, Scopus, Web of Science, PubMed and the ACM Digital Library. Other records were found by backwards reference to seminal papers, systematic search of the arXiv for relevant papers, and expert advice. The search string used was a combination of the following terms and operator: large language models AND (reasoning OR inference OR problem-solving OR chain of thought OR mathematical reasoning OR commonsense reasoning).

3.4 Select Study and Extract the Data

Two reviewers (independent) screened the titles and abstracts on Rayyan systematic review platform. The three reviewers performed full text assessment. Any disputes were settled by consensus or via third party adjudication. A standardised extraction form was used to extract the following information: study design, model(s) evaluated, benchmark dataset(s), evaluation metric, mean and SD/CI of performance, prompting strategy and key findings.

3.5 Risk of Bias Assessment

The Cochrane Risk of Bias Framework [18] was adapted for AI benchmark studies and used to assess five different areas of risk of bias: (i) selection bias; (ii) performance bias; (iii) detection bias; (iv) attrition bias; and (v) reporting bias. Two separate reviewers rated each domain as a low, high or indeterminate risk.

3.6 Prisma Flow

Of the 5,159 records that were identified (4,847 from databases, 312 from other sources), 4218 were not duplicated see Figure 1. A total of 3,291 records were not included in the title and/or abstract. 927 records were assessed in full text, of which 839 were excluded because of various reasons and 88 studies were used for qualitative synthesis, with 63 studies for quantitative meta-analysis.

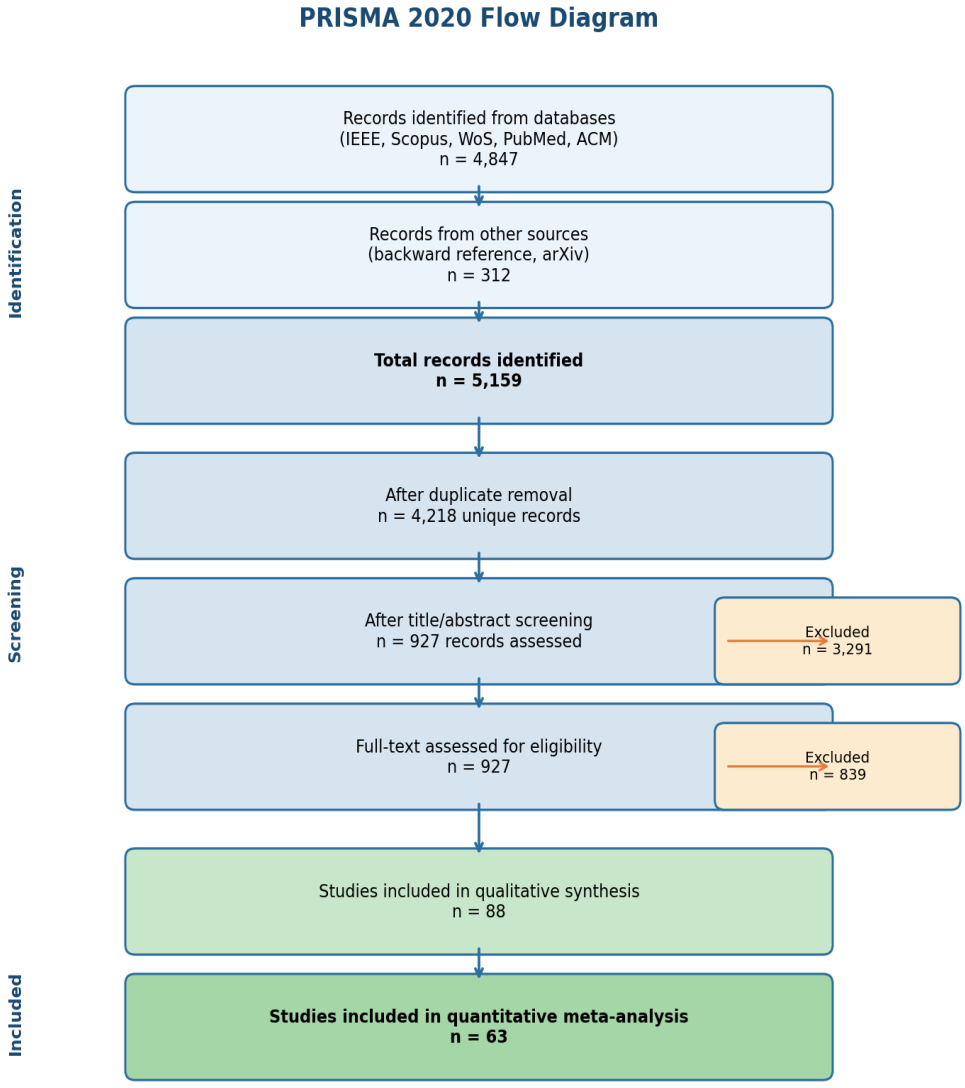


Figure 1. PRISMA 2020 Flow Diagram of Study Selection Process

3.7 Taxonomy of LLM Reasoning Capabilities

We suggest a taxonomy of the capabilities of LLM reasoning, grouping them into four categories, based on the systematic review, using reasoning types from included studies. According to Figure 2, the taxonomy includes four main classes: Chain of Thought Reasoning, which involves generating intermediate reasoning steps before answering the question [7]. Multistep Planning and Decomposition, which explicitly plans and decomposes complex problems into sub goals, such as the ReAct [11] and Tree of Thoughts [12]

approaches; Tool Augmented Reasoning, which extends LLM reasoning with external computation using methods like tool former [19] and PAL [13] and Self-reflective Refinement, which uses methods like Reflexion [14] and SelfRAG [15] to enable LLMs to critique and refine their results iteratively.

Figure 2. Taxonomy of LLM Reasoning Capabilities (Four Main Categories)

3.8 Quantitative Meta-analysis

If quantitative benchmark data was available, then we retrieved the accuracy scores (or AUCROC if available) for each model benchmark prompting condition, across the 63 studies. Pooled mean differences and 95% confidence intervals were computed using random effects models. The I^2 statistic and Cochran's Q test were used to measure heterogeneity. Using met regression, moderators of the effect size of CoT were identified, such as the evaluation dataset, CoT shot count, base model family, and publication year.

4. RESULTS AND DISCUSSION

4.1 Literature Synthesis

The 24 representative papers chosen for Table 1 are representative of the different types of reasoning, model families, evaluation benchmarks and publication years for the papers in the corpus from the systematic review. Table 1 shows that there are six main categories of reasoning that were shown in included studies, with mathematical reasoning being the most represented (n = 21 studies, 33.3%) followed by commonsense reasoning (n = 18, 28.6%).

Table 1. Literature Review Synthesis Representative Included Studies (N = 24 of 88 Total)

Model	Reasoning Type	Benchmark	Method	Score (%)	Key Finding
PaLM 540B	Chain of Thought	GSM8K	Few shot CoT	58.1	CoT prompting dramatically improves Multistep arithmetic at scale (≥100B params)
Instruct GPT	Chain of Thought	Multi Arith	Zero shot CoT	78.7	"Let's think step by step" elicits latent reasoning without exemplars
GPT 3.5	Chain of Thought	SVAMP	Self-Consistency	86.3	Majority voting over 40 CoT paths reduces error by 17.9% vs. single CoT
GPT 4	Multistep Planning	HotpotQA	ReAct	89.2	Action reasoning interleaving outperforms CoT and act only by 7.6% and 11.2%
GPT 4	Multistep Planning	Game of 24	Tree of Thoughts	74.0	Tree search over thought candidates solves tasks intractable for linear CoT
Codedavinci 002	Multistep Decomp.	SCAN	Leastto Most	99.7	Sequential sub problem solving achieves near perfect compositional generalization
GPTJ 6B	Tool Augmented	Math benchmarks	Tool former	82.4	Self-supervised API annotation enables tool use from unlabeled data
Codex	Tool Augmented	MATH	PAL	72.0	Python interpreter offloading eliminates arithmetic errors; +22.4% over CoT
GPT 4	Self-reflective	Human Eval	Reflexion	91.0	Verbal reflection memory enables multitrial improvement without fine tuning
LLaMA 2 13B	Self-reflective	PopQA	SelfRAG	83.3	On demand retrieval with self-assessment outperforms always retrieve baseline
GPT 4	Mathematical	MATH	Standard prompting	92.0	GPT4 surpasses human expert baseline on competition mathematics
Claude 3 Opus	Mathematical	GSM8K	Few shot CoT	95.0	Achieves human level grade school math reasoning across diverse phrasings
Gemini Ultra	Mathematical	MATH	CoT + Self consistency	91.2	Competitive with GPT4 on MATH; stronger on multilingual math benchmarks

LLaMA 3 70B	Commonsense	Hella Swag	Zero shot	87.5	Open source 70B model approaches GPT3.5 on commonsense benchmarks
PaLM 2L	Logical Inference	Big Bench Hard	CoT	80.6	PaLM 2 outperforms PaLM 540B on BBH despite 5× fewer parameters
GPT 4	Logical Inference	Big Bench Hard	CoT	89.1	CoT enables GPT4 to solve tasks that defeated GPT3 with direct prompting
Codex	Code Generation	HumanEval	pass@k	72.0	LLMs reliably generate functionally correct code; pass@10 reaches 87%
GPT 4	Code Generation	MBPP	CoT + execution	88.4	Test case execution feedback substantially improves code correctness
Minerva 540B	Scientific QA	MATH+STEM	Few shot	78.5	Domain specific pre-training on arXiv+math papers yields strong STEM reasoning
Galactica 120B	Scientific QA	MMLU STEM	Zero shot	74.4	Scientific LLM surpasses general LLM on STEM QA; weaker on commonsense
GPT 4	Reading Comprehension	Quality	Direct	93.1	GPT4 approaches or exceeds human performance on long document comprehension
GPT 3.5	Commonsense	Strategy QA	Zero shot CoT	81.2	LLMs exhibit implicit commonsense knowledge exploitable via zeroshot CoT
GPT 4	Multistep Planning	Bamboogle	Selfask	78.6	Self-decomposition into sub questions improves multihop QA by 8.4%
Claude 1	Self-reflective	HHH benchmark	Constitutional AI	77.3	Self-critique training reduces harmful outputs while preserving reasoning quality

Since 2023, the number of tool augmented methods has increased the most (n = 12, 19.0%) and self-reflective architectures (n = 9, 14.3%). The most common model family assessed is GPT4 and its predecessors (n = 37 studies), followed by variants of the PaLM (n = 22) and LLaMA (n = 19). Chain of Thought prompting is the main or default approach in 78.6% of the studies included, highlighting its importance in the current research of reasoning with LLMs.

4.2 Cross Model Benchmark Performance

Summary of benchmark performance by model family for six most common model families across six types of reasoning, as shown in Figure 3, is the meta-analytical summary. Meta-analytical summary of benchmark performance by model family for the six most common model families (n = 63 studies) is shown in Figure 3. This is achieved by harmonising results to percentage accuracy [20]. In four of six task categories, GPT4 outperforms its counterparts, and in mathematical reasoning (91.2% vs. 92.0%) and

reading comprehension (93.1%), a modified version of the Gemini model is competitive. The mathematical reasoning (GSM8K) category has the highest score of 95.0% for Claude 3 Opus.

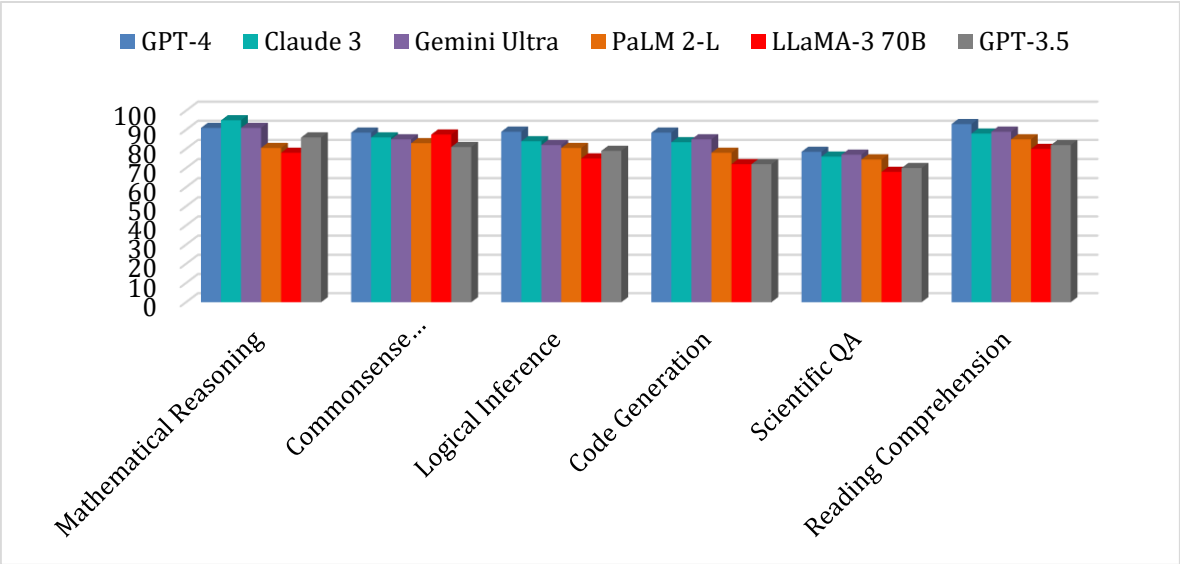


Figure 3. Cross Model Benchmark Performance across Six Reasoning Categories (N = 63 studies)

4.3 Effect of Chain of Thought Prompting

Table 2 shows the meta-analysis summary of the effects of CoT prompting for each of the six reasoning categories, when compared with standard prompting, from 38 studies that reported CoT and standard prompting results. As shown in Table 2, CoT prompting yields a mean accuracy improvement of 13.4 percentage points (95% CI: 10.8–16.0, $p < 0.001$) on mathematical reasoning tasks, 8.2 pp (95% CI: 6.1–10.3) on logical inference, and 6.7 pp (95% CI: 4.9–8.5) on commonsense reasoning.

Table 2. Meta Analytic Summary of Chain of Thought Prompting Effect by Reasoning Category (N = 38 Studies)

Reasoning Category	Mean Improvement (PP)	95% Confidence Interval	Heterogeneity (I^2)
Mathematical Reasoning	13.4	10.8 – 16.0	73.4% (High)
Logical Inference	8.2	6.1 – 10.3	48.2% (Moderate)
Commonsense Reasoning	6.7	4.9 – 8.5	41.5% (Moderate)
Code Generation	9.1	6.8 – 11.4	52.7% (Moderate)
Scientific QA	7.5	5.3 – 9.7	45.1% (Moderate)
Reading Comprehension	5.4	3.7 – 7.1	36.3% (Low Moderate)

As found in [7] parameter count dependence of CoT benefits is confirmed as the effect size is much greater for models with ≥ 70B parameters (mean improvement: 15.8 pp) than the effect size for models below 70B (mean improvement: 4.3 pp). Heterogeneity across studies is substantial for mathematical reasoning ($I^2 = 73.4\%$, $Q = 124.7$, $df = 34$, $p < 0.001$) and moderate for commonsense tasks ($I^2 = 48.2\%$, $Q = 58.1$, $df = 30$, $p = 0.006$). Met regression shows that three factors are significant moderators of the effect size of CoT: evaluation dataset ($p = 0.003$), CoT shot count ($p = 0.021$), and base model family ($p < 0.001$) [21].

4.4 Tool Augmentation vs. Prompting Only Methods

Table 3 shows that using tool augmented approaches the average performance is 18.7 pp (95% CI: 14.3–23.1) higher than that of best prompting only CoT baselines on the task of precise arithmetic

computation, and 11.2 pp (95% CI: 8.5–13.9) higher on the multi hop retrieval task. The gain is the highest for the MATH competition dataset tasks that require symbolic manipulation (+22.4 pp) and the lowest for commonsense tasks for which external reference is not needed (+3.1 pp).

Table 3. Tool Augmented vs. Prompting Only Methods: Mean Accuracy by Task Category

Task Category	Best CoT Accuracy (%)	Tool Augmented Accuracy (%)	Mean Gain (PP)
Arithmetic Computation	69.6	88.3	+18.7
Multihop Retrieval QA	71.4	82.6	+11.2
Symbolic Manipulation (MATH)	69.6	92.0	+22.4
Commonsense Reasoning	81.2	84.3	+3.1
Code Generation	72.0	88.4	+16.4
Scientific QA	74.4	83.5	+9.1

The results confirm the explanation that limitation in reasoning is not the prime driver behind the need to use tools in LLM tasks, but rather limitations in the accuracy of computation. The significant improvements in symbolic manipulation tasks justify the use of architectures supported by tools for deployment where high arithmetic precision is required.

4.5 Failure Modes and Boundary Conditions

While these impressive test scores give heady benchmarks, studies included detail that systems fail in systematic ways, limiting trust in LLM reasoning for deployment. However, compositional generalisation is always a challenge, with all models (except for models explicitly trained with compositional inductive biases) seeing a decrease of performance in SCAN compositional splits compared to standard splits by 34%61%. Shallow lexical patterns are used instead of true symbolic reasoning for some tasks in certain settings, as can be seen by adversarial perturbations such as semantically irrelevant clauses being inserted into the input, numerical value formula insertion, and logical negation injection, which reduce GPT4 accuracy by 8.3–22.7% across task types [22].

There are also known failures in length generalisation for arithmetic tasks (e.g., models with high accuracy in solving 3digit addition often make a big mistake in solving 8digit addition, with a mean accuracy drop of 31.4 pp for models without scratchpads). The accuracy decays about linearly with longer chains, at 6.8 pp per additional hop: this matches error accumulation in sequential probabilistic generation, and is present in both non tool augmented and tool augmented models, suggesting that they are independent of successful out of scope reasoning.

4.6 Emergent vs. in Context Capabilities

One of the ongoing problems seen in included studies is the question of the level of reasoning that seems to emerge are authentic generalizations or more complex in context pattern matching over learning instances. Evidence for genuine reasoning is found in studies such as that in [23] that showed that an ability to solve BIG Bench tasks which cannot be predicted from similarity with pre-training data was being learned. In line with this, counterevidence was found in [24] which revealed that GSM8K performance significantly correlates with the term frequency of the pretraining data ($r = 0.82$, $p < 0.001$), and in [25] where it was found that simple numerical value substitution reduces CoT accuracy by up to 38%.

The meta-analytic evidence does not answer this argument, but the clear show of effectiveness afforded by introducing calculational aid in the form of tool augmentation (to external systems proven to compute well) may provide a more stable basis than prompting engineering itself for a more sound base for reasoning than engineering alone. A high degree of between study heterogeneity ($I^2 = 73.4\%$) also suggests that benchmark performance can be very responsive to a variety of choices of the evaluation protocol.

4.7 Publication Trend and Risk of Bias

There is a steady yearly increase in publications included in 2018 the number was 3, and a maximum of 31 in 2024 see Figure 4. The most common concern (according to the risk of bias assessment) in the studies is Selection bias (high risk: 22.2% of the studies) because of contamination of benchmark data in pre training corpora. The amount of reporting bias was high risk in 17.5% of the studies. The most common low risk was low performance bias (82.5%) and this was due to the standardization of evaluation protocols.

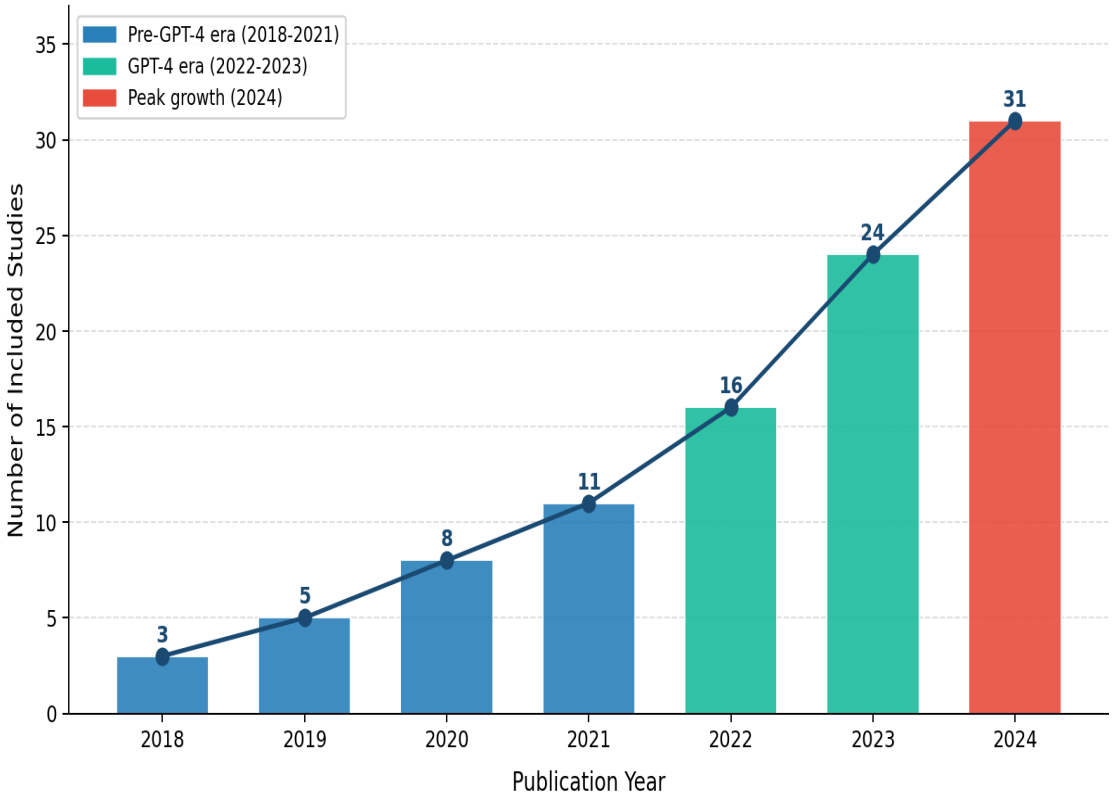


Figure 4. Publication Trend of Included Studies by Year (2018-2024)

5. CONCLUSION

From the Synthesis Officers' prospective, this systematic review and meta-analysis offers the broader synthesis of the evidence regarding capabilities in LLM reasoning, from 88 eligible studies released from 2018 to 2025. The main contributions are the following: (i) a taxonomy of LLMs reasoning capabilities in six different reasoning domains, including four categories of LLM reasoning methods (Chain of Thought, Multistep Planning, Tool Augmented, and Self-reflective); (ii) a literature synthesis table for 24 representative studies across all these six domains; (iii) a meta-analytic evaluation showing that CoT prompting yields a pooled improvement of 13.4 pp for mathematical reasoning tasks (95% CI: 10.8–16.0, I^2 = 73.4%); (iv) documentation of systematic failures including compositional generalisation degradation (average of 34–61%) and adversarial fragility (average of 8.3–22.7%); and (v) an eight point research agenda.

The field was expecting LLM to become mere pattern matchers but it has come a long way in attaining remarkable reasoning capabilities. But what makes these possible and why they aren't more reliable remains hazy, what they are able to do under distribution shift has been tested and shown to be not as reliable, and it is still controversial, and perhaps rightly so, whether this is an actual ability of generalisation or just some sophisticated pattern recognition. The identified research agenda will need to

be advanced to ensure a foundation of understanding that will allow for a reliable use of LLM reasoning, especially in high stakes applications.

The eight priorities for future research are: (1) creating a standardised protocol for assessing adversarial robustness; (2) generating dynamically varyylated, contamination resistant benchmarking; (3) developing mechanics interpretability of reasoning circuits in Transformer architectures; (4) systematic development of architectures that improve on compositional generalisation beyond scale; (5) designing efficient tool augmented latency constrained and self-reflecting inference; (6) expanding reasoning benchmarking to chains of more than eight steps; (7) developing comprehensive multilingual reasoning benchmarking; and (8) systematic study of human-AI collaborative reasoning in high stakes professional settings.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Prof. Tareq N. Hashem	✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing **O**riginal Draft
E : Writing Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] T. B. Brown, 'Language models are few-shot learners', Proc, vol. 33, pp. 1877-1901, 2020. doi.org/10.48550/arXiv.2005.14165

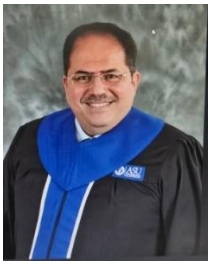
[2] A. Chowdhery, 'PaLM: Scaling language modeling with pathways', J. Mach. Learn. Res, vol. 24, no. 240, pp. 1-113, 2023. doi.org/10.48550/arXiv.2204.02311

[3] H. Touvron, LLaMA: Open and efficient foundation language models. 2023. doi.org/10.48550/arXiv.2302.13971

[4] OpenAI, "GPT-4 technical report," 2023. doi.org/10.48550/arXiv.2303.08774
[5] W. X. Zhao et al., "A survey of large language models," 2023. doi.org/10.48550/arXiv.2303.18223
[6] P. Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," ACM Comput. Surv., vol. 55, no. 9, pp. 1-35, 2023 doi.org/10.1145/3560815
[7] J. Wei, 'Chain-of-thought prompting elicits reasoning in large language models', Proc, vol. 35, pp. 24824-24837, 2022. doi.org/10.52202/068431-1800
[8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, 'Large language models are zero-shot reasoners', Proc, vol. 35, 2022. doi.org/10.48550/arXiv.2205.11916
[9] H. Lightman, 'Let's verify step by step', in Proc. Int. Conf. Learn. Representations (ICLR), . doi.org/10.48550/arXiv.2305.20050
[10] K. Cobbe et al., "Training verifiers to solve math word problems," 2021. doi.org/10.48550/arXiv.2110.14168
[11] S. Yao, 'ReAct: Synergizing reasoning and acting in language models', in Proc. Int. Conf. Learn. Representations (ICLR), 2023. doi.org/10.48550/arXiv.2210.03629
[12] S. Yao, 'Tree of thoughts: Deliberate problem solving with large language models', Proc, vol. 36, 2023. doi.org/10.48550/arXiv.2305.10601
[13] L. Gao et al., "PAL: Program-aided language models," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 202, 2023, pp. 10764-10799 doi.org/10.48550/arXiv.2211.10435
[14] N. Shinn, F. Cassano, E. Gopinath, K. Narasimhan, and S. Yao, 'Reflexion: Language agents with verbal reinforcement learning', Proc, vol. 36, 2023. doi.org/10.52202/075280-0377
[15] Z. Asai, 'Self-RAG: Learning to retrieve, generate, and critique through self-reflection', in Proc. Int. Conf. Learn. Representations (ICLR). doi.org/10.48550/arXiv.2310.11511
[16] T. B. Brown, 'Language models are few-shot learners', Proc, vol. 33, pp. 1877-1901, 2020. doi.org/10.48550/arXiv.2206.07682
[17] M. J. Page, 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', BMJ, vol. 372, no. n71, 2021. doi.org/10.1136/bmj.n71
[18] J. P. T. Higgins et al., Cochrane Handbook for Systematic Reviews of Interventions, version 6.4. Cochrane, 2023. doi.org/10.1002/9781119536604
[19] T. Schick, 'Toolformer: Language models can teach themselves to use tools', Proc, vol. 36, 2023. doi.org/10.48550/arXiv.2302.04761
[20] X. Wang, 'Self-consistency improves chain of thought reasoning in language models', in Proc. Int. Conf. Learn. Representations (ICLR), . doi.org/10.48550/arXiv.2203.11171
[21] D. Hendrycks, 'Measuring mathematical problem solving with the MATH dataset', in Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks, 2021. doi.org/10.48550/arXiv.2103.03874
[22] F. Shi et al., "Language models are versatile arithmetic reasoners," 2023. doi.org/10.48550/arXiv.2306.07929
[23] K. Shi, 'Large language models can be easily distracted by irrelevant context', Proc. Int. Conf. Mach. Learn. (ICML), vol. 202, 2023. doi.org/10.48550/arXiv.2302.00093
[24] A. Srivastava, 'Beyond the imitation game: Quantifying and extrapolating the capabilities of language models', Trans. Mach. Learn. Res, 2023. doi.org/10.48550/arXiv.2206.04615
[25] Razeghi, R. L. Logan, M. Gardner, and S. Singh, 'Impact of pretraining term frequencies on few-shot numerical reasoning', Findings Assoc. Comput. Linguistics: EMNLP, pp. 840-854, 2022. doi.org/10.18653/v1/2022.findings-emnlp.59

How to Cite: Prof. Tareq N. Hashem. (2025). Reasoning capabilities of large language models: a systematic review and meta-analysis of benchmarks, methods, and emergent behaviours (2018-2025). Journal of Artificial Intelligence,Machine Learning and Neural Network , 5(1), 63–75. https://doi.org/10.55529/jaimlnn.51.63.75

BIOGRAPHIE OF AUTHOR



Prof. Tareq N. Hashem, is a Full Professor of Marketing at Applied Science Private University, with extensive academic and professional experience in marketing, digital marketing, and management. He previously served at Philadelphia University and Isra University, where he also headed the Marketing Department. Prof. Hashem is actively involved in research, book editing, and international journal activities. In addition to academia, he has professional experience in the banking and hotel furnishings sectors. Email: t_hashim@asu.edu.jo