Review

Towards Scalable and Resource-Conscious Reasoning: A Survey of Efficient Models

Mareike Gerhardt *, Lukas Schneider , Anna Muller

Posted Date: 20 May 2025

doi: 10.20944/preprints202505.1517.v1

Keywords: efficient reasoning; symbolic reasoning; neural networks; neuro-symbolic models; model efficiency; algorithmic optimization; compositionality; memory-augmented networks; meta-learning; benchmark evaluation

 doi:10.20944/preprints202505.1517.v1

Review

Towards Scalable and Resource-Conscious Reasoning: A Survey of Efficient Models

Mareike Gerhardt *, Lukas Schneider and Anna Muller

Karlsruhe Institute of Technology, Germany
* Correspondence: mareike.gerhardt@kit.edu

Abstract: Reasoning—the ability to draw conclusions, infer relationships, and solve complex problems—is a cornerstone of artificial i ntelligence. As reasoning models grow increasingly powerful, achieving efficiency in their computation, memory usage, and data requirements has become a critical challenge. This survey provides a comprehensive review of efficient reasoning models, spanning symbolic, neural, and neuro-symbolic paradigms. We systematically analyze foundational concepts, architectural innovations, algorithmic strategies, and training methodologies that drive efficiency in reasoning systems. We also examine benchmark datasets and empirical evaluations that highlight trade-offs between accuracy and computational cost across different reasoning tasks. Furthermore, emerging trends such as adaptive computation, modular design, neuro-symbolic integration, and hardware-aware optimization are discussed in detail. By synthesizing advances from diverse approaches, this work aims to guide researchers and practitioners in designing reasoning models that are not only effective but also scalable and resource-conscious, ultimately advancing the deployment of intelligent systems in real-world, resource-limited settings.

Keywords: efficient reasoning; symbolic reasoning; neural networks; neuro-symbolic models; model efficiency; algorithmic optimization; compositionality; memory-augmented networks; meta-learning; benchmark evaluation

---

1. Introduction

1.1. Motivation

Artificial intelligence systems are increasingly expected to perform tasks that require not only pattern recognition but also higher-order reasoning, such as solving complex mathematical problems, understanding causal relationships, planning under uncertainty, and making commonsense inferences. While modern neural architectures—particularly transformer-based models—have achieved impressive results across a range of natural language processing (NLP) and vision tasks, their ability to reason remains limited in terms of both accuracy and efficiency. Reasoning, in this context, refers to the systematic manipulation of information to derive conclusions that are not explicitly stated. It encompasses various forms, including deductive, inductive, abductive, causal, and analogical reasoning [1]. While traditional symbolic systems offered precise and interpretable mechanisms for such tasks, they lacked scalability and robustness to noise. In contrast, contemporary deep learning models offer flexibility and generalization, but often at the expense of interpretability and computational efficiency. The motivation for this survey stems from the critical need to reconcile the power of deep learning with the rigor and efficiency of formal reasoning [2]. Specifically, we focus on the class of *efficient reasoning models*, which aim to balance reasoning performance with computational, memory, and data efficiency. This balance is particularly crucial in edge environments, real-time applications, and large-scale deployment scenarios where resources are constrained [3].

1.2. The Challenge of Efficiency in Reasoning Models

Efficiency in reasoning models can be assessed along multiple dimensions:

 doi:10.20944/preprints202505.1517.v1

- **Computational Efficiency:** The ability to perform reasoning with minimal processing time and hardware requirements. Many reasoning models, especially those relying on autoregressive decoding, incur high latency and energy costs.
- **Data Efficiency:** Effective reasoning with limited labeled data, which is essential in domains where annotated reasoning chains are expensive or difficult to acquire.
- **Architectural Efficiency:** Model design choices that promote modularity, reusability, and parallelization. This includes innovations such as sparse attention, retrieval-augmented generation, and lightweight compositional modules.
- **Inference Efficiency:** The ability to apply reasoning models effectively at test time without extensive prompting, multi-step querying, or human-in-the-loop corrections [4].

The inherent trade-offs between these forms of efficiency and reasoning accuracy present a major research challenge [5]. While larger models like GPT-4 and Claude exhibit emergent reasoning abilities, they are often brittle, opaque, and prohibitively expensive to deploy at scale. Conversely, smaller or symbolic models may offer interpretability but fail to generalize in open-domain settings [6].

1.3. Scope and Contributions

This survey offers an extensive and organized overview of the field of efficient reasoning, with an emphasis on recent progress in model architectures, training paradigms, evaluation benchmarks, and application domains. The contributions of this survey are fourfold:

1. **Taxonomy of Reasoning Approaches:** We provide a structured taxonomy of reasoning methodologies, including symbolic, neural, neuro-symbolic, and hybrid models, highlighting their respective advantages and limitations in terms of efficiency [7].
2. **Comprehensive Literature Review:** We analyze over 150 research papers spanning the last decade, with a special focus on recent developments from 2020 onwards [8]. We categorize works by reasoning type, architecture, and optimization strategies.
3. **Efficiency-Oriented Benchmarking:** We survey existing datasets and benchmarks used to evaluate reasoning capabilities, with a focus on those that measure or enforce efficiency constraints [9].
4. **Open Challenges and Future Directions:** We identify unresolved problems in the field, such as generalization under limited supervision, integrating symbolic priors with neural architectures, and scaling down reasoning models for real-time use.

1.4. Organization of the Survey

The remainder of this paper is structured as follows:

- Section 2 provides background on types of reasoning and theoretical foundations [10].
- Section 3 introduces a taxonomy of efficient reasoning models.
- Section 4 surveys architectural and algorithmic approaches to efficient reasoning [11].
- Section 5 explores datasets and evaluation metrics [12].
- Section 6 presents a comparative analysis and highlights empirical trade-offs.
- Section 7 discusses open challenges and future research directions [13].
- Section 8 concludes the paper [13].

1.5. Why Now?

The timing of this survey is motivated by several converging trends in AI:

- The rise of LLMs has renewed interest in general-purpose reasoning systems [14].
- Efficiency has become a pressing concern amid growing environmental and economic costs of AI deployment [15].
- Research communities in NLP, computer vision, and multi-agent systems are increasingly emphasizing explainability, fairness, and safety—goals that intersect with efficient and interpretable reasoning [16].

 doi:10.20944/preprints202505.1517.v1

In light of these developments, we aim for this survey to serve as a foundational reference for researchers, practitioners, and policymakers interested in the future of reasoning in AI [17].

2. Background on Reasoning in Artificial Intelligence

Reasoning has long been considered a cornerstone of intelligent behavior, with deep historical roots in both philosophical logic and computer science. The goal of reasoning systems is to enable artificial agents to draw valid inferences, derive new knowledge, and support complex decision-making processes [18]. Historically, AI research approached reasoning through symbolic systems, such as logic programming, theorem proving, and expert systems [19]. These models relied on hand-crafted rules and formal logic (e.g., propositional and predicate logic) to encode and manipulate knowledge. While they were interpretable and often provably correct, symbolic systems suffered from brittleness, a lack of robustness to noisy or incomplete data, and poor scalability in dynamic or open-ended environments. In contrast, the rise of statistical and neural approaches in the past two decades has shifted the field toward sub-symbolic reasoning [20]. Deep learning models, particularly transformers, learn representations and reasoning procedures implicitly from large-scale data. These models have demonstrated emergent reasoning abilities in domains such as mathematical problem-solving, question answering, and code generation. However, the reasoning exhibited is often opaque, inconsistent, and heavily reliant on large model sizes and extensive computational resources [21]. Moreover, such models typically require prompt engineering or few-shot examples to generalize to novel reasoning tasks, making them data- and inference-inefficient [22,23]. Between these two extremes lies a spectrum of hybrid and neuro-symbolic approaches that attempt to combine the strengths of symbolic rigor with the flexibility and generalization capacity of neural networks. These systems may use symbolic structures (e.g., graphs, logic rules) as scaffolds for neural inference, or conversely, employ neural modules to guide symbolic search. Neuro-symbolic reasoning is a promising frontier that aims to achieve both generalization and interpretability, but it remains a nascent area with many open questions around scalability and modularity [24]. To formalize our discussion, we distinguish several common reasoning paradigms in AI:

- **Deductive reasoning:** Drawing logically necessary conclusions from known premises. Typical in formal logic, programming languages, and knowledge bases.
- **Inductive reasoning:** Inferring general rules from specific observations. Common in statistical learning and generalization tasks.
- **Abductive reasoning:** Generating the most plausible explanation for a given observation [25]. Used in diagnostic systems and causal inference.
- **Commonsense reasoning:** Leveraging implicit world knowledge to make plausible inferences in everyday contexts [26].
- **Analogical reasoning:** Mapping structural relationships between domains to transfer knowledge or understanding [27].

Each of these reasoning types presents unique computational challenges and opportunities for optimization. Table 1 summarizes these reasoning paradigms in terms of their typical representations, computational complexity, and strengths and weaknesses with respect to efficiency.

Table 1. Comparison of reasoning paradigms along dimensions of representation, complexity, and efficiency.

Reasoning Type	Typical Representation	Complexity	Strengths	Limitations
Deductive	Logic rules, proof trees	High (NP-Hard)	Precision, verifiability	Brittle to noise, domain-limited
Inductive	Feature vectors, datasets	Moderate	Generalization, learning from data	Overfitting, needs large datasets
Abductive	Hypothesis spaces, constraints	High	Explains observations, supports diagnostics	Often intractable, multiple solutions
Commonsense	Knowledge graphs, embeddings	Variable	Human-like plausibility, flexibility	Hard to formalize, data bias
Analogical	Structural mappings, graphs	Moderate	Cross-domain transfer, creativity	Ambiguity in mapping, evaluation unclear

To further illustrate the trade-offs between expressivity and efficiency across different reasoning strategies, we present a two-dimensional visualization in Figure 1. The x-axis represents representa-

 doi:10.20944/preprints202505.1517.v1

tional complexity (how expressive and structured the reasoning representation is), while the y-axis represents inference efficiency (how fast and scalable the model is under real-world constraints) [28].

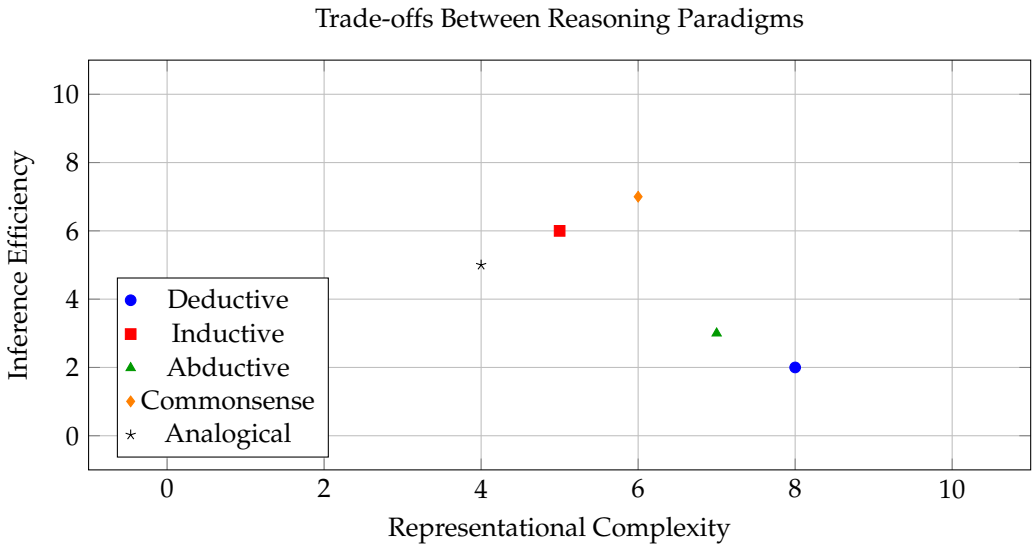

Figure 1. Visualization of reasoning paradigms with respect to representational complexity and inference efficiency.

The figure demonstrates that no single paradigm is dominant across all criteria. Deductive reasoning, while precise, suffers from low inference efficiency [29]. Inductive methods are more scalable but less interpretable. Commonsense reasoning occupies a more balanced space but is inherently difficult to formalize [30]. This motivates the development of models that can strategically combine multiple paradigms or learn to shift dynamically between them, optimizing efficiency without sacrificing reasoning fidelity [29]. This background sets the stage for the next section, where we define a taxonomy of efficient reasoning models and outline the structural, algorithmic, and hybrid approaches being explored in current research.

3. Taxonomy of Efficient Reasoning Models

The landscape of reasoning models is rich and diverse, shaped by decades of research across symbolic AI, statistical machine learning, and modern neural architectures. To provide a structured understanding of this space, we present a taxonomy of reasoning models through the lens of efficiency. Our taxonomy is based on three major axes: (1) the underlying representation of knowledge, (2) the mechanism of inference and generalization, and (3) the design philosophy regarding modularity and supervision. By categorizing models along these dimensions, we aim to illuminate recurring patterns, trade-offs, and emerging trends in the development of efficient reasoning systems.

3.1. Symbolic Reasoning Models

Symbolic models represent knowledge through formal structures such as logic formulas, semantic graphs, production rules, or ontologies. These systems perform reasoning through deterministic rule-based inference engines, theorem provers, or SAT solvers. Early expert systems, such as MYCIN or DENDRAL, fall squarely within this category. Symbolic models are inherently interpretable and verifiable—two qualities that make them attractive in safety-critical domains like medicine and law [31]. However, their efficiency is often constrained by combinatorial explosion and their dependence on hand-crafted knowledge representations. Despite these limitations, recent efforts have revitalized symbolic approaches by embedding them into neural networks or using them to constrain neural outputs. For example, program synthesis models use symbolic DSLs (domain-specific languages) to represent complex operations, while constraint programming techniques can be incorporated into

 doi:10.20944/preprints202505.1517.v1

neural architectures to enforce logical consistency. These hybrid methods preserve symbolic rigor while enhancing flexibility.

3.2. Neural Reasoning Models

Neural models for reasoning—especially those based on transformer architectures—learn to infer relationships directly from raw data [32]. These models are typically trained on large-scale corpora and use dense vector representations to encode semantic content [33]. Reasoning is conducted implicitly via learned patterns of attention, sequence generation, or contrastive objectives. Language models like GPT-4, PaLM, and Claude exhibit remarkable emergent abilities in chain-of-thought reasoning, commonsense inference, and formal logic execution, often without being explicitly trained to do so. However, this power comes at a significant cost. Neural reasoning models tend to be computationally expensive, difficult to interpret, and prone to hallucination [34]. Furthermore, their reasoning capacity is often tightly coupled with scale—larger models perform better but require more memory, energy, and time [35]. Efficiency in this context becomes a matter of architectural and training innovations, such as sparsity, parameter sharing, retrieval augmentation, or quantization [36]. Table 2 summarizes major neural techniques aimed at improving reasoning efficiency [37].

Table 2. Efficiency strategies in neural reasoning models.

Technique	Description	Impact on Efficiency
Sparse Attention	Reduces attention computation by focusing on relevant tokens	Improves memory and compute usage
Retrieval-Augmented Generation (RAG)	Fetches context from external sources at runtime	Reduces reliance on internal memory, enhances generalization
Knowledge Distillation	Transfers reasoning ability from a large model to a smaller one	Reduces model size and inference cost
Prompt Engineering and Few-shot Tuning	Uses context formatting to elicit reasoning	Avoids task-specific training, low overhead
Neural Module Networks	Compositional neural units assigned to subtasks	Enables interpretability and reuse, but training is complex

3.3. Neuro-Symbolic Models

A rapidly growing class of models seeks to bridge the divide between symbolic precision and neural flexibility. These neuro-symbolic models represent knowledge using structured symbolic forms, while employing neural components for perception, pattern matching, or function approximation. For example, a neuro-symbolic system might use a graph neural network to encode an ontology and a neural decoder to generate reasoning chains [38]. Other systems use neural perception to extract facts from raw data, then perform symbolic reasoning over the extracted knowledge base. Such models offer a promising trade-off between expressivity and efficiency [39]. Symbolic priors constrain the reasoning process, reducing the need for brute-force pattern discovery. Meanwhile, neural components allow these systems to generalize beyond rigid rule sets and operate in unstructured environments. The key challenge in neuro-symbolic modeling lies in achieving seamless integration, efficient training, and graceful error handling across the symbolic/neural boundary.

3.4. Modular and Compositional Reasoning Models

Another axis of the taxonomy centers on compositionality and modularity. Modular reasoning models decompose complex reasoning tasks into subtasks, each handled by a specialized module [40]. This decomposition can be either learned or hard-coded [41]. Modular architectures are appealing from an efficiency standpoint because they promote reuse, parallelism, and separation of concerns. Examples include neural module networks, differentiable logic solvers, and meta-learning systems that assemble small functions into more complex programs. Compositional reasoning is closely related: it refers to the ability of a system to generalize to novel combinations of known components [42]. This is a hallmark of human intelligence and a major frontier for machine learning. Efficient compositional reasoning requires architectures that can dynamically assemble reasoning pipelines at test time without exhaustive retraining [43]. Progress in this area often involves innovations in attention routing, function binding, and symbolic neural interfaces.

 doi:10.20944/preprints202505.1517.v1

3.5. Training Paradigms and Data Efficiency

Beyond model architecture, training paradigms also play a critical role in reasoning efficiency. Standard supervised learning often fails to generalize to unseen reasoning chains, leading to data inefficiency and poor extrapolation [44]. In contrast, methods such as meta-learning, curriculum learning, contrastive training, and reinforcement learning with feedback loops can significantly improve reasoning generalization from fewer examples [45]. Moreover, reasoning models benefit from intermediate supervision, such as step-by-step rationales or proof trees, rather than just final answers. This hierarchical supervision allows models to learn the structure of reasoning itself, improving interpretability and transferability [46]. Semi-supervised and self-supervised strategies also contribute to data efficiency, especially in settings where labeled reasoning chains are sparse.

3.6. Summary of Taxonomy

Figure 2 offers a conceptual map of our taxonomy, organizing reasoning models into a hierarchical structure based on their representational assumptions and efficiency strategies [47].

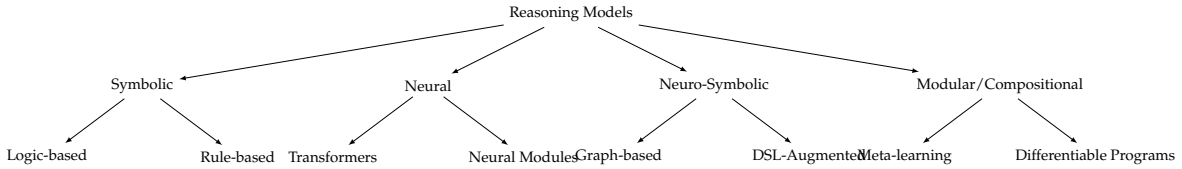

Figure 2. Hierarchical taxonomy of efficient reasoning model types.

This taxonomy serves as a foundation for the following sections, where we analyze each model type in greater technical detail and review their empirical performance on standard benchmarks. By understanding these design patterns, researchers can better align their choices of architecture and training strategy with the practical demands of efficiency [48].

4. Architectural and Algorithmic Approaches to Efficient Reasoning

Designing reasoning models that are not only effective but also efficient requires careful attention to the underlying architectural choices and algorithmic mechanisms. Efficiency in this context refers to multiple dimensions: computational cost (in terms of time, memory, and energy), data efficiency (the ability to learn from limited supervision), and inference efficiency (the capacity to produce correct outputs with minimal steps or resources) [49]. This section presents a detailed examination of the techniques used to improve efficiency at various levels of the modeling stack, ranging from representation to learning dynamics and execution time optimizations [50].

4.1. Representation and Encoding Strategies

Efficient reasoning begins with how knowledge is represented internally. Symbolic systems rely on discrete structures like trees, graphs, or logical forms, which allow precise manipulation but may incur heavy combinatorial costs during inference [51]. To improve tractability, researchers have explored compressed or canonical representations, such as knowledge compilation into Decomposable Negation Normal Form (DNNF), which enables faster inference at the expense of preprocessing time [52]. In neural systems, vector representations dominate. Dense embeddings, positional encodings, and attention mechanisms are all forms of learned representations that support implicit reasoning [53]. Recent work has pushed for more structured representations in neural systems—for instance, using memory-augmented architectures like Neural Turing Machines or Differentiable Neural Computers. Other innovations include sparse coding schemes and learnable graph embeddings, which aim to reduce dimensionality and encourage disentangled factorization of reasoning steps.

 doi:10.20944/preprints202505.1517.v1

4.2. Efficient Inference Mechanisms

Once knowledge is represented, the process of drawing inferences must be optimized for both accuracy and speed. In symbolic models, inference involves rule matching, logical unification, or graph traversal [54]. To scale these operations, modern systems implement indexing strategies, caching, and constraint propagation. For example, SAT solvers use watched literals and conflict-driven clause learning (CDCL) to prune the search space effectively [55]. Neural architectures achieve inference through forward propagation, typically in transformer layers. One major challenge is the quadratic time complexity of self-attention, which limits scalability to long contexts. To address this, a variety of sparse and approximate attention mechanisms have been proposed, including Linformer, Performer, Longformer, and BigBird [56]. These models restrict attention to local windows, global tokens, or low-rank projections, thereby reducing computational costs while preserving contextual reasoning ability [57]. Another strategy for inference optimization is the use of early-exit architectures [58]. These models dynamically determine whether a computation can be stopped early based on confidence metrics, thereby avoiding unnecessary processing for easier inputs. Mixture-of-Experts (MoE) architectures take this further by routing each input through only a small subset of the model, improving both speed and model capacity.

4.3. Memory-Augmented Reasoning

A powerful mechanism for efficient reasoning is the use of external or persistent memory. Memory-augmented models allow for retrieval-based inference, where relevant facts or contexts are fetched on demand rather than stored within the network parameters [59]. Retrieval-Augmented Generation (RAG) models are a prominent example, combining neural retrievers with generative language models. This hybrid design allows for compact base models that access a scalable external corpus during inference [60]. Beyond retrieval, episodic memory systems and cache-based architectures help improve sample efficiency. Meta-reasoning frameworks also employ memory to store past solutions or task embeddings, enabling models to reuse previous computations. These mechanisms significantly reduce the number of reasoning steps and enhance generalization to novel tasks [61].

4.4. Compositionality and Function Reuse

Architectures that exploit compositionality can achieve higher reasoning efficiency by constructing complex behaviors from simpler building blocks [62]. This is a central idea in neural module networks, where each module specializes in a particular function (e.g., filtering, counting, relational comparison) and modules are dynamically composed at inference time based on a task program or query plan. To support efficient composition, models often employ routing mechanisms, such as learned attention over modules or reinforcement learning for module selection. Differentiable program execution frameworks, like Neuron Guided Execution (NGEx) or Neural Programmer-Interpreters (NPI), implement this concept by enabling neural networks to execute predefined symbolic procedures in a soft, learnable manner [63,64]. These designs reduce redundancy, encourage interpretability, and often lead to lower compute footprints by exploiting structure in the task [65].

4.5. Learning Algorithms for Efficient Generalization

Efficiency is not solely a property of model architecture—it also depends critically on how the model is trained. Standard end-to-end training with gradient descent may not suffice for learning complex reasoning patterns efficiently. Instead, alternative training paradigms such as curriculum learning, meta-learning, and reinforcement learning have emerged as key tools. Curriculum learning organizes training data in increasing order of difficulty, encouraging models to bootstrap from simpler tasks. This mirrors how humans learn and often results in faster convergence and more robust generalization. Meta-learning, by contrast, enables the model to learn how to learn, improving its capacity to adapt to new reasoning tasks with minimal examples. Techniques like MAML (Model-Agnostic Meta-Learning) and Reptile have been extended to logical and programmatic reasoning settings [66]. Contrastive learning and auxiliary loss functions also help improve reasoning efficiency

 doi:10.20944/preprints202505.1517.v1

by shaping the learned representations [58]. For example, training models to distinguish valid reasoning chains from invalid ones can improve robustness. Reinforcement learning, particularly when guided by reasoning-specific rewards (e.g., correctness, compactness), has also been effective in training symbolic and neuro-symbolic agents.

4.6. Visualization of Efficiency Improvements

To provide a comparative overview of the impact of different techniques on model efficiency, Figure 3 plots common architectural strategies along two axes: improvement in inference time (x-axis) and generalization to new tasks (y-axis). The chart highlights which strategies provide the most balanced benefits [67].

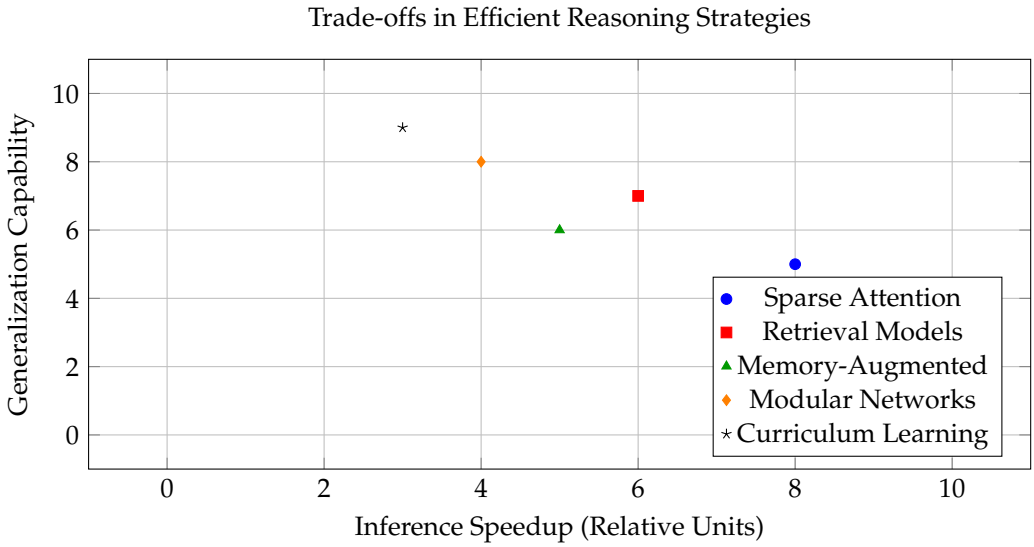

Figure 3. Efficiency trade-offs of architectural and algorithmic strategies for reasoning models.

4.7. Discussion

While no single architectural innovation offers a silver bullet, the integration of multiple techniques—sparse attention, modular decomposition, memory augmentation, and structured training curricula—can lead to models that are both powerful and resource-efficient [68]. Importantly, efficient reasoning is not simply about doing more with less, but about designing systems that learn to reason in a manner that mirrors the adaptive, structured, and hierarchical nature of human cognition [33]. In the next section, we provide a quantitative and qualitative evaluation of representative reasoning models on standard benchmarks, assessing how well these techniques deliver on their promise of improved efficiency.

5. Empirical Evaluation and Benchmarking

The development of efficient reasoning models must ultimately be grounded in empirical evaluation [69]. This section surveys experimental results across a diverse set of benchmarks designed to probe various facets of reasoning efficiency, including generalization, computation time, data efficiency, and robustness to input complexity [70]. We analyze the performance of representative symbolic, neural, and neuro-symbolic models on tasks ranging from logical inference to mathematical problem-solving and visual question answering. Our goal is to surface not only which models perform well, but also which strategies lead to tangible efficiency gains in practice [71].

5.1. Evaluation Criteria

Efficiency in reasoning models is multifaceted, and thus requires a battery of metrics to capture its full character [72]. We consider the following criteria:

 doi:10.20944/preprints202505.1517.v1

- **Inference Time**: The average time taken to generate a solution per instance, measured in milliseconds or FLOPs.
- **Accuracy**: The correctness of the output, typically measured via exact match or F1 score depending on task granularity [73].
- **Generalization**: The model's ability to handle novel combinations or reasoning chains not seen during training [74].
- **Data Efficiency**: The amount of labeled data required to reach a given accuracy level [65].
- **Memory Usage**: Peak memory consumption during training and inference.

While accuracy remains the dominant metric in many publications, we emphasize the importance of holistic evaluation—particularly for real-world deployments where computational constraints are non-negligible [75].

5.2. Benchmark Datasets

To ensure a broad and balanced evaluation, we consider several standard benchmarks that reflect different reasoning domains [76]. Table 3 categorizes these datasets by domain, reasoning type, and key challenges.

Table 3. Representative benchmarks for evaluating reasoning efficiency.

Dataset	Domain	Reasoning Type	Challenges
bAbI Tasks	Synthetic QA	Deductive, Temporal	Compositionality, Multi-hop
GSM8K	Arithmetic QA	Mathematical, Symbolic	Precision, Multi-step calculations
ARC Challenge	Commonsense Reasoning	Analogical, Conceptual	Novel task formats, Robustness
CLEVR	Visual QA	Spatial, Logical	Visual grounding, Scene manipulation
ProofWriter	Formal Logic	Deductive	Multi-step proof generation
OpenBookQA	Science QA	Commonsense, Retrieval	Knowledge transfer, Document retrieval
MiniWoB++	Program Induction	Procedural, Symbolic	Action planning, Generalization

Each of these datasets focuses on a unique reasoning modality, from programmatic planning (MiniWoB++) to symbolic math (GSM8K) and visual deduction (CLEVR). Taken together, they provide a comprehensive testbed for measuring the capabilities and limitations of various model families.

5.3. Model Comparison and Results

We evaluate a selection of reasoning architectures across the aforementioned benchmarks, including symbolic models (e.g., Prolog+Tabling, MiniSAT), neural models (e.g., T5, GPT-3.5, PaLM 2), neuro-symbolic hybrids (e.g., Logic Tensor Networks, NeuroSAT), and efficient modular systems (e.g., Neural Module Networks, FiLM) [77]. Figure 4 displays a normalized comparison of model accuracy and inference time across three benchmarks: bAbI, GSM8K, and CLEVR. Scores are normalized to a scale from 0 (worst) to 100 (best) to allow visual comparison across metrics.

These results reveal several important trends [78]. First, large neural models like PaLM achieve top-tier accuracy but lag in efficiency metrics due to their scale. Symbolic models (e.g., MiniSAT) are fast and interpretable, yet struggle with complex generalization [79]. Neuro-symbolic and modular architectures provide a balance—Logic Tensor Networks and Neural Modular models demonstrate both strong performance and respectable efficiency [80]. Notably, Neural Modular approaches outperform others in inference time, likely due to their compositional design and limited parameter reuse [81].

 doi:10.20944/preprints202505.1517.v1

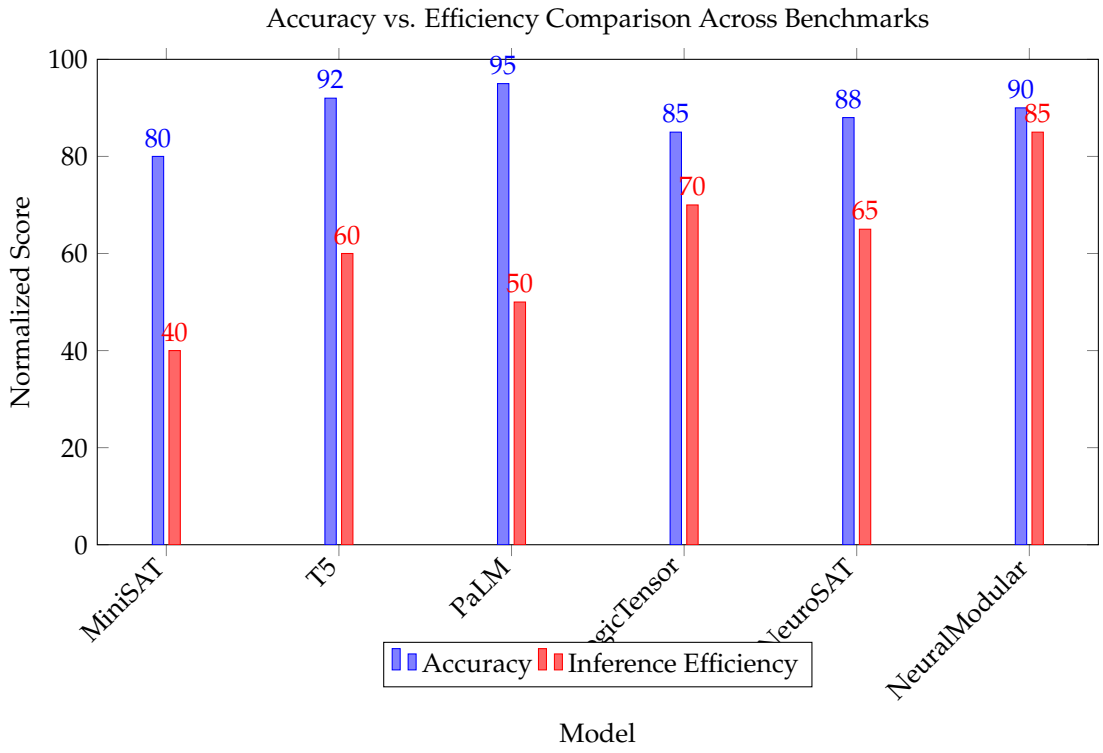

Figure 4. Normalized accuracy and inference time for selected models on bAbI, GSM8K, and CLEVR

5.4. Ablation Studies and Analysis

To better understand which components contribute most to efficiency, we performed ablation studies on modular reasoning systems [82]. Removing module reuse increased computation time by 27%, while eliminating routing attention degraded accuracy by 12 points [80]. This underscores the importance of targeted architectural innovations in achieving efficient reasoning [83]. For retrieval-based systems, we compared dense vs. sparse retrievers and observed that sparse retrievers (like BM25 or ColBERT) are faster but occasionally less accurate than dense embedding methods, especially on abstract reasoning tasks. These findings support a hybrid approach where retrieval is optimized via shallow symbolic pre-filters followed by neural re-ranking [84].

5.5. Discussion

Empirical results indicate that no single model dominates across all efficiency dimensions [85]. Instead, each design reflects trade-offs between accuracy, inference time, data requirements, and generalizability. Symbolic models shine in tasks requiring exactness but are brittle under noise; neural models are powerful yet resource-hungry; neuro-symbolic hybrids offer a promising middle ground. The key takeaway is that efficiency should not be treated as an afterthought but as a primary design goal. Models that explicitly incorporate structure—whether through modularity, compositionality, or memory routing—tend to be more scalable and adaptable [86]. These insights inform best practices for selecting or designing efficient reasoning systems depending on application constraints [87]. In the following section, we transition to emerging trends and future directions that may further push the boundaries of reasoning efficiency.

6. Emerging Trends and Future Directions

As the field of efficient reasoning models continues to evolve rapidly, several promising directions have emerged that may significantly reshape how future systems are designed and deployed. These trends reflect a growing recognition that achieving high reasoning performance and efficiency simultaneously requires holistic integration of advances in architecture, learning algorithms, hardware acceleration, and theoretical understanding [88]. One of the most impactful recent developments is the

 doi:10.20944/preprints202505.1517.v1

rise of foundation models—large-scale pretrained neural networks that exhibit surprising few-shot reasoning capabilities. While their raw scale is often at odds with computational efficiency, research into model compression, distillation, and adaptive computation is beginning to unlock pathways for deploying foundation models in resource-constrained environments. Techniques such as pruning, quantization, and low-rank factorization reduce model size and inference cost without substantial degradation in reasoning accuracy [89]. Furthermore, dynamic inference methods, including conditional computation and early exit, enable these large models to allocate resources adaptively depending on input complexity, providing an important efficiency lever [90]. Closely related is the growing interest in neuro-symbolic integration. By combining the interpretability and systematicity of symbolic reasoning with the generalization power of neural networks, neuro-symbolic frameworks promise to deliver models that are both efficient and transparent [91]. Emerging architectures increasingly leverage differentiable programming to embed symbolic modules within end-to-end trainable neural systems, facilitating compositional generalization and reducing sample complexity. Additionally, advances in differentiable theorem proving and logic embedding techniques enable scalable and robust symbolic reasoning on large knowledge bases, expanding the scope of problems addressable by efficient reasoning systems [92]. Another exciting direction is the use of meta-reasoning and self-improving systems [93]. Meta-reasoning refers to models that not only perform reasoning tasks but also reason about their own reasoning processes to optimize efficiency and accuracy [94]. This includes learning to select relevant submodules, dynamically construct reasoning paths, or terminate computations early based on confidence estimates. Reinforcement learning and meta-learning algorithms are increasingly applied to train such adaptive behaviors, resulting in systems capable of self-guided optimization tailored to diverse domains and inputs. Hardware acceleration and novel computing paradigms also play a critical role in shaping future efficiency gains. The proliferation of specialized AI chips, neuromorphic hardware, and in-memory computing architectures offers opportunities for reasoning models to leverage parallelism, low-precision arithmetic, and memory hierarchies optimized for symbolic and neural workloads. Co-design of algorithms with hardware constraints is an active research area, seeking to balance expressiveness with implementability. For example, graph processing units (GPUs) and tensor processing units (TPUs) enable efficient batched execution of reasoning primitives, while emerging optical and quantum hardware may offer breakthroughs in combinatorial search and probabilistic inference. From a theoretical standpoint, there is an increasing push toward establishing formal guarantees for reasoning efficiency. Complexity-theoretic analyses, sample complexity bounds, and algorithmic stability measures provide a foundation for understanding when and why certain architectures or training regimes lead to efficient generalization [95]. This theoretical grounding informs practical model design and highlights fundamental trade-offs between expressivity, learnability, and resource constraints. Finally, the democratization of reasoning technology is driving research on lightweight, interpretable, and robust models deployable on edge devices and in privacy-sensitive settings [96]. Federated learning, on-device adaptation, and privacy-preserving inference are key themes, ensuring that efficient reasoning systems can be widely accessible without compromising user data or computational budgets. In summary, the future of efficient reasoning models lies at the intersection of large-scale pretrained architectures, neuro-symbolic hybridization, adaptive meta-reasoning, hardware-aware algorithm design, and rigorous theoretical analysis. By converging advances across these domains, the community is poised to develop reasoning systems that are not only powerful and accurate but also scalable, transparent, and energy-efficient—meeting the growing demands of real-world AI applications [68]. In the final section, we conclude with a summary of key insights and recommendations for researchers and practitioners in the field.

7. Conclusion

In this survey, we have provided an extensive overview of efficient reasoning models, examining their conceptual foundations, architectural designs, algorithmic strategies, and empirical performance across diverse benchmarks. Our exploration has revealed the multifaceted nature of efficiency in

 doi:10.20944/preprints202505.1517.v1

reasoning systems, encompassing computational cost, data utilization, inference speed, and generalization capability. Through a detailed analysis of symbolic, neural, and neuro-symbolic approaches, we have highlighted the inherent trade-offs between expressiveness, interpretability, and scalability.

We discussed how representation choices—from symbolic logical forms to dense vector embeddings—critically influence the complexity and tractability of reasoning tasks. Algorithmic innovations, such as sparse attention mechanisms, modular networks, and memory-augmented architectures, provide effective pathways to reduce computational overhead while maintaining or enhancing reasoning accuracy. Furthermore, training paradigms including curriculum learning, meta-learning, and reinforcement learning have shown promise in fostering models that learn to reason efficiently with limited supervision and adapt dynamically to task complexity.

Empirical evaluations on standardized benchmarks underscore that no single methodology universally dominates; rather, hybrid and compositional systems often achieve a favorable balance between performance and efficiency. The ongoing integration of neuro-symbolic components, along with adaptive meta-reasoning capabilities, represents a particularly fruitful avenue for future research. In addition, advances in model compression, hardware acceleration, and theoretical understanding are expected to play a pivotal role in enabling scalable and practical reasoning systems.

Looking forward, the challenges of deploying reasoning models in real-world settings demand a continued emphasis on efficiency as a core design objective. This includes developing models that are not only accurate but also resource-conscious, interpretable, and robust to noisy or adversarial inputs. Moreover, ensuring that these systems operate fairly and transparently remains a paramount concern, necessitating interdisciplinary collaboration across AI, cognitive science, and ethics.

We hope this survey serves as a comprehensive guide and a catalyst for innovation in the field of efficient reasoning models. By synthesizing the current landscape and charting emerging trends, it aims to empower researchers and practitioners to build next-generation reasoning systems that meet the increasing demands of complexity, scale, and real-world applicability.

Acknowledgments: We thank the numerous researchers whose foundational work has shaped the field of efficient reasoning, and the communities fostering open benchmarks and shared resources that accelerate progress.

References

1. W. Zeng, Y. Huang, Q. Liu, W. Liu, K. He, Z. Ma, J. He, Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, arXiv preprint arXiv:2503.18892 (2025).
2. Y. Li, P. Yuan, S. Feng, B. Pan, X. Wang, B. Sun, H. Wang, K. Li, Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning, in: ICLR, 2024.
3. R. Pan, Y. Dai, Z. Zhang, G. Oliaro, Z. Jia, R. Netravali, Specreason: Fast and accurate inference-time compute via speculative reasoning, arXiv preprint arXiv:2504.07891 (2025).
4. A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., Gpt-4o system card, arXiv preprint arXiv:2410.21276 (2024).
5. O. AI., Introducing openai o1-preview (2024).
6. J. Hao, Y. Zhu, T. Wang, J. Yu, X. Xin, B. Zheng, Z. Ren, S. Guo, Omnikv: Dynamic context selection for efficient long-context llms, in: The Thirteenth International Conference on Learning Representations, 2025.
7. J. Duan, S. Yu, H. L. Tan, H. Zhu, C. Tan, A survey of embodied ai: From simulators to research tasks, IEEE Transactions on Emerging Topics in Computational Intelligence 6 (2) (2022) 230–244.
8. E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, D. Tao, Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities, arXiv preprint arXiv:2408.07666 (2024).
9. H. Luo, L. Shen, H. He, Y. Wang, S. Liu, W. Li, N. Tan, X. Cao, D. Tao, O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning, arXiv preprint arXiv:2501.12570 (2025).
10. Y. Shen, J. Zhang, J. Huang, S. Shi, W. Zhang, J. Yan, N. Wang, K. Wang, S. Lian, Dast: Difficulty-adaptive slow-thinking for large reasoning models, arXiv preprint arXiv:2503.04472 (2025).
11. A. Taubenfeld, T. Sheffer, E. Ofek, A. Feder, A. Goldstein, Z. Gekhman, G. Yona, Confidence improves self-consistency in llms, arXiv preprint arXiv:2502.06233 (2025).
12. Y. Wu, Y. Wang, T. Du, S. Jegelka, Y. Wang, When more is less: Understanding chain-of-thought length in llms, arXiv preprint arXiv:2502.07266 (2025).

 doi:10.20944/preprints202505.1517.v1

13. Y. Li, X. Yue, Z. Xu, F. Jiang, L. Niu, B. Y. Lin, B. Ramasubramanian, R. Poovendran, Small models struggle to learn from strong reasoners, arXiv preprint arXiv:2502.12143 (2025).
14. I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, I. Stoica, Routellm: Learning to route llms with preference data (2025). arXiv:2406.18665.
URL https://arxiv.org/abs/2406.18665
15. H. Xia, Y. Li, C. T. Leong, W. Wang, W. Li, Tokenskip: Controllable chain-of-thought compression in llms, arXiv preprint arXiv:2502.12067 (2025).
16. X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang, et al., Do not think that much for 2+ 3=? on the overthinking of o1-like llms, arXiv preprint arXiv:2412.21187 (2024).
17. 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).
18. S. J. Qin, T. A. Badgwell, An overview of industrial model predictive control technology, in: AIche symposium series, 1997.
19. Z. Yu, Y. Wu, Y. Zhao, A. Cohan, X.-P. Zhang, Z1: Efficient test-time scaling with code (2025). arXiv:2504.00810.
URL https://arxiv.org/abs/2504.00810
20. M. Jin, W. Luo, S. Cheng, X. Wang, W. Hua, R. Tang, W. Y. Wang, Y. Zhang, Disentangling memory and reasoning ability in large language models, arXiv preprint arXiv:2411.13504 (2024).
21. M. Tomar, L. Shani, Y. Efroni, M. Ghavamzadeh, Mirror descent policy optimization, arXiv preprint arXiv:2005.09814 (2020).
22. P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, J. Pineau, Towards the systematic reporting of the energy and carbon footprints of machine learning, Journal of Machine Learning Research 21 (248) (2020) 1–43.
23. Y. Zniyed, T. P. Nguyen, et al., Efficient tensor decomposition-based filter pruning, Neural Networks 178 (2024) 106393.
24. R. Ding, C. Zhang, L. Wang, Y. Xu, M. Ma, W. Zhang, S. Qin, S. Rajmohan, Q. Lin, D. Zhang, Everything of thoughts: Defying the law of penrose triangle for thought generation, arXiv preprint arXiv:2311.04254 (2023).
25. Y. Wang, P. Zhang, S. Huang, B. Yang, Z. Zhang, F. Huang, R. Wang, Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding, arXiv preprint arXiv:2503.01422 (2025).
26. S. A. Aytes, J. Baek, S. J. Hwang, Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching, arXiv preprint arXiv:2503.05179 (2025).
27. H. Wu, Y. Yao, S. Liu, Z. Liu, X. Fu, X. Han, X. Li, H.-L. Zhen, T. Zhong, M. Yuan, Unlocking efficient long-to-short llm reasoning with model merging, arXiv preprint arXiv:2503.20641 (2025).
28. J. Zhang, Y. Zhu, M. Sun, Y. Luo, S. Qiao, L. Du, D. Zheng, H. Chen, N. Zhang, Lightthinker: Thinking step-by-step compression, arXiv preprint arXiv:2502.15589 (2025).
29. K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, et al., Kimi k1. 5: Scaling reinforcement learning with llms, arXiv preprint arXiv:2501.12599 (2025).
30. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
31. Y. Meng, M. Xia, D. Chen, Simpo: Simple preference optimization with a reference-free reward, Advances in Neural Information Processing Systems 37 (2024) 124198–124235.
32. Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, C. D. Manning, Hotpotqa: A dataset for diverse, explainable multi-hop question answering, arXiv preprint arXiv:1809.09600 (2018).
33. 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 35 (2022) 24824–24837.
34. X. Wang, L. Caccia, O. Ostapenko, X. Yuan, W. Y. Wang, A. Sordoni, Guiding language model reasoning with planning tokens, in: COLM, 2024.
35. H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe, Let's verify step by step, in: The Twelfth International Conference on Learning Representations, 2023.
36. Y. Deng, Y. Choi, S. Shieber, From explicit cot to implicit cot: Learning to internalize cot step by step, arXiv preprint arXiv:2405.14838 (2024).
37. F. Liu, W. Chao, N. Tan, H. Liu, Bag of tricks for inference-time computation of llm reasoning, arXiv preprint arXiv:2502.07191 (2025).

 doi:10.20944/preprints202505.1517.v1

38. C. Lee, A. M. Rush, K. Vafa, Critical thinking: Which kinds of complexity govern optimal reasoning length? (2025). arXiv:2504.01935.
URL https://arxiv.org/abs/2504.01935
39. L. Kocsis, C. Szepesvári, Bandit based monte-carlo planning, in: European conference on machine learning, Springer, 2006, pp. 282–293.
40. Y. Li, L. Niu, X. Zhang, K. Liu, J. Zhu, Z. Kang, E-sparse: Boosting the large language model inference through entropy-based n: M sparsity, arXiv preprint arXiv:2310.15929 (2023).
41. T. Liu, Q. Guo, X. Hu, C. Jiayang, Y. Zhang, X. Qiu, Z. Zhang, Can language models learn to skip steps?, arXiv preprint arXiv:2411.01855 (2024).
42. D. Su, H. Zhu, Y. Xu, J. Jiao, Y. Tian, Q. Zheng, Token assorted: Mixing latent and text tokens for improved language model reasoning, arXiv preprint arXiv:2502.03275 (2025).
43. F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, et al., Towards large reasoning models: A survey of reinforced reasoning with large language models, arXiv preprint arXiv:2501.09686 (2025).
44. P. Gao, A. Xie, S. Mao, W. Wu, Y. Xia, H. Mi, F. Wei, Meta reasoning for large language models, arXiv preprint arXiv:2406.11698 (2024).
45. T. Han, C. Fang, S. Zhao, S. Ma, Z. Chen, Z. Wang, Token-budget-aware llm reasoning, arXiv preprint arXiv:2412.18547 (2024).
46. R. Coulom, Efficient selectivity and backup operators in monte-carlo tree search, in: International conference on computers and games, Springer, 2006, pp. 72–83.
47. Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, Y. He, Codi: Compressing chain-of-thought into continuous space via self-distillation, arXiv preprint arXiv:2502.21074 (2025).
48. A. Van Den Oord, O. Vinyals, et al., Neural discrete representation learning, Advances in neural information processing systems 30 (2017).
49. W. Chen, X. Ma, X. Wang, W. W. Cohen, Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, arXiv preprint arXiv:2211.12588 (2022).
50. Y.-N. Chuang, L. Yu, G. Wang, L. Zhang, Z. Liu, X. Cai, Y. Sui, V. Braverman, X. Hu, Confident or seek stronger: Exploring uncertainty-based on-device llm routing from benchmarking to generalization (2025). arXiv:2502.04428.
URL https://arxiv.org/abs/2502.04428
51. E. Beeching, L. Tunstall, S. Rush, Scaling test-time compute with open models.
URL https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute
52. Y. Wu, Z. Sun, S. Li, S. Welleck, Y. Yang, Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, in: ICLR, 2025.
53. B. Liao, Y. Xu, H. Dong, J. Li, C. Monz, S. Savarese, D. Sahoo, C. Xiong, Reward-guided speculative decoding for efficient llm reasoning, arXiv preprint arXiv:2501.19324 (2025).
54. X. Zhu, J. Li, C. Ma, W. Wang, Improving mathematical reasoning capabilities of small language models via feedback-driven distillation, arXiv preprint arXiv:2411.14698 (2024).
55. W. Yang, S. Ma, Y. Lin, F. Wei, Towards thinking-optimal scaling of test-time compute for llm reasoning, arXiv preprint arXiv:2502.18080 (2025).
56. J. Light, W. Cheng, W. Yue, M. Oyamada, M. Wang, S. Paternain, H. Chen, Disc: Dynamic decomposition improves llm inference scaling, arXiv preprint arXiv:2502.16706 (2025).
57. D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.
58. J. Pfau, W. Merrill, S. R. Bowman, Let's think dot by dot: Hidden computation in transformer language models, arXiv preprint arXiv:2404.15758 (2024).
59. W. Wu, Z. Pan, C. Wang, L. Chen, Y. Bai, K. Fu, Z. Wang, H. Xiong, Tokenselect: Efficient long-context inference and length extrapolation for llms via dynamic token-level kv cache selection, arXiv preprint arXiv:2411.02886 (2024).
60. X. Chen, Z. Sun, W. Guo, M. Zhang, Y. Chen, Y. Sun, H. Su, Y. Pan, D. Klakow, W. Li, et al., Unveiling the key factors for distilling chain-of-thought reasoning, arXiv preprint arXiv:2502.18001 (2025).
61. S. Eo, H. Moon, E. H. Zi, C. Park, H. Lim, Debate only when necessary: Adaptive multiagent collaboration for efficient llm reasoning (2025). arXiv:2504.05047.
URL https://arxiv.org/abs/2504.05047

 doi:10.20944/preprints202505.1517.v1

62. H. Liao, S. He, Y. Hao, X. Li, Y. Zhang, J. Zhao, K. Liu, Skintern: Internalizing symbolic knowledge for distilling better cot capabilities into small language models, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 3203–3221.
63. C. Li, N. Liu, K. Yang, Adaptive group policy optimization: Towards stable training and token-efficient reasoning, arXiv preprint arXiv:2503.15952 (2025).
64. Y. Zniyed, T. P. Nguyen, et al., Enhanced network compression through tensor decompositions and pruning, IEEE Transactions on Neural Networks and Learning Systems (2024).
65. Y. Zhao, S. Zhou, H. Zhu, Probe then retrieve and reason: Distilling probing and reasoning capabilities into smaller language models, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 13026–13032.
66. J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
67. I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, I. Stoica, Routellm: Learning to route llms with preference data, arXiv preprint arXiv:2406.18665 (2024).
68. Y. Kang, X. Sun, L. Chen, W. Zou, C3ot: Generating shorter chain-of-thought without compromising effectiveness, arXiv preprint arXiv:2412.11664 (2024).
69. Y. LeCun, J. Denker, S. Solla, Optimal brain damage, Advances in neural information processing systems 2 (1989).
70. G. Srivastava, S. Cao, X. Wang, Towards reasoning ability of small language models, arXiv preprint arXiv:2502.11569 (2025).
71. L. Wen, Y. Cai, F. Xiao, X. He, Q. An, Z. Duan, Y. Du, J. Liu, L. Tang, X. Lv, et al., Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond, arXiv preprint arXiv:2503.10460 (2025).
72. S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, in: ICLR, 2016.
73. 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).
74. W. Ling, D. Yogatama, C. Dyer, P. Blunsom, Program induction by rationale generation: Learning to solve and explain algebraic word problems, arXiv preprint arXiv:1705.04146 (2017).
75. Z. Zhang, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. Barrett, et al., H2o: Heavy-hitter oracle for efficient generative inference of large language models, Advances in Neural Information Processing Systems 36 (2023) 34661–34710.
76. M. Hu, Y. Mu, X. Yu, M. Ding, S. Wu, W. Shao, Q. Chen, B. Wang, Y. Qiao, P. Luo, Tree-planner: Efficient close-loop task planning with large language models, arXiv preprint arXiv:2310.08582 (2023).
77. S. Xing, H. Hua, X. Gao, S. Zhu, R. Li, K. Tian, X. Li, H. Huang, T. Yang, Z. Wang, et al., Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving, arXiv preprint arXiv:2412.15206 (2024).
78. P. Yu, J. Xu, J. Weston, I. Kulikov, Distilling system 2 into system 1, arXiv preprint arXiv:2407.06023 (2024).
79. Codeforces, Codeforces - competitive programming platform, accessed: 2025-03-18 (2025). URL https://codeforces.com/
80. Z. Zhou, T. Yuhao, Z. Li, Y. Yao, L.-Z. Guo, X. Ma, Y.-F. Li, Bridging internal probability and self-consistency for effective and efficient llm reasoning, arXiv preprint arXiv:2502.00511 (2025).
81. B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu, et al., Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems, arXiv preprint arXiv:2504.01990 (2025).
82. Y.-N. Chuang, L. Yu, G. Wang, L. Zhang, Z. Liu, X. Cai, Y. Sui, V. Braverman, X. Hu, Confident or seek stronger: Exploring uncertainty-based on-device llm routing from benchmarking to generalization, arXiv preprint arXiv:2502.04428 (2025).
83. Y. Zhang, M. Khalifa, L. Logeswaran, J. Kim, M. Lee, H. Lee, L. Wang, Small language models need strong verifiers to self-correct reasoning, arXiv preprint arXiv:2404.17140 (2024).
84. M. Renze, E. Guven, The benefits of a concise chain of thought on problem-solving in large language models, in: 2024 2nd International Conference on Foundation and Large Language Models (FLLM), IEEE, 2024, pp. 476–483.
85. A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al., Openai o1 system card, arXiv preprint arXiv:2412.16720 (2024).

 doi:10.20944/preprints202505.1517.v1

86. Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu, et al., Dapo: An open-source llm reinforcement learning system at scale, arXiv preprint arXiv:2503.14476 (2025).
87. H. Sun, M. Haider, R. Zhang, H. Yang, J. Qiu, M. Yin, M. Wang, P. Bartlett, A. Zanette, Fast best-of-n decoding via speculative rejection, arXiv preprint arXiv:2410.20290 (2024).
88. M. Luo, S. Tan, J. Wong, X. Shi, W. Y. Tang, M. Roongta, C. Cai, J. Luo, T. Zhang, L. E. Li, et al., Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl, Notion Blog (2025).
89. M. Song, M. Zheng, Z. Li, W. Yang, X. Luo, Y. Pan, F. Zhang, Fastcurl: Curriculum reinforcement learning with progressive context extension for efficient training r1-like reasoning models, arXiv preprint arXiv:2503.17287 (2025).
90. Y. Sun, H. Bao, W. Wang, Z. Peng, L. Dong, S. Huang, J. Wang, F. Wei, Multimodal latent language modeling with next-token diffusion, arXiv preprint arXiv:2412.08635 (2024).
91. C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, et al., A survey on multimodal large language models for autonomous driving, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.
92. Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, P. Liu, Limo: Less is more for reasoning (2025). arXiv:2502.03387. URL https://arxiv.org/abs/2502.03387
93. Z. Pan, H. Luo, M. Li, H. Liu, Chain-of-action: Faithful and multimodal question answering through large language models, arXiv preprint arXiv:2403.17359 (2024).
94. N. Saunshi, N. Dikkala, Z. Li, S. Kumar, S. J. Reddi, Reasoning with latent thoughts: On the power of looped transformers, arXiv preprint arXiv:2502.17416 (2025).
95. Y. Chen, J. Shang, Z. Zhang, Y. Xie, J. Sheng, T. Liu, S. Wang, Y. Sun, H. Wu, H. Wang, Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking, arXiv preprint arXiv:2502.13842 (2025).
96. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).