

Reasoning in Large Language Models

Aleksei Medvedev

Abstract. Large language models (LLMs) perform reasoning statistically, identifying patterns in data rather than applying strict logical rules. This leads to unreliability in complex tasks like logic, math, and causal inference, compounded by a severe shortage of high-quality training data. Initial approaches rely on human-crafted prompts and reinforcement learning with human feedback (RLHF), but these remain fundamentally limited by their dependence on human input. The critical shift comes with Reinforcement Learning from Verifier Feedback (RLVR), which eliminates human judges entirely. RLVR employs automated verifiers to evaluate reasoning steps internally, enabling self-sustaining improvement and resolving both reasoning deficiencies and data scarcity without human oversight.

The current stage of AI and large language models (LLMs) development faces a critical constraint: the exhaustion of readily available, high-quality raw textual data suitable for training autoregressive LLMs. As experts often wryly note, “re-learning the entire internet seems increasingly impractical.” This scarcity necessitates the discovery of novel data sources and training methodologies.

Simultaneously, the reasoning capabilities of pure LLMs — encompassing logical deduction, mathematical problem-solving, commonsense reasoning, and causal inference — consistently fall short of expectations when tackling complex tasks.

These dual challenges — data limitations and reasoning deficiencies — are driving intensive research into innovative techniques specifically designed to qualitatively advance the reasoning prowess of language models.

Classical AI systems (e.g., expert systems, theorem provers) rely on explicit symbolic rules and deductive logic. Reasoning is deterministic, interpretable, and rule-bound. In contrast, LLMs perform statistical reasoning: they infer patterns from vast data, generating responses probabilistically based on learned correlations. This enables flexibility but lacks inherent guarantees of correctness, interpretability, or systematicity.

One of the most promising modern approaches to improving LLM’s reasoning capabilities leverages inference-time computation scaling. This technique strategically allocates greater computational resources during model inference to enhance

output quality—analogous to providing sufficient 'thinking time' for complex problems. Rather than merely prompting for a direct answer, this paradigm employs structured reasoning frameworks like Tree of Thoughts (ToT). Here, problems are decomposed into branching sequences of intermediate steps, followed by systematic path exploration and selection.

Building on these prompting strategies, LLMs can be further enhanced through supervised fine-tuning (SFT) on curated datasets of verifiable reasoning tasks. These datasets — spanning mathematical proofs, algorithmic problems, causal inference challenges, and domain-specific questions — provide explicit demonstrations of structured reasoning. By training on such high-quality, step-by-step solutions, SFT enables models to internalize robust reasoning patterns, substantially improving their baseline capabilities beyond prompt-dependent approaches.

The next evolutionary phase employs reinforcement learning (RL) to further refine reasoning capabilities. The established approach, Reinforcement Learning from Human Feedback (RLHF), operates through the following stages: collecting human rankings of LLM responses to reasoning tasks, training a reward model to predict these preferences, and fine-tuning the LLM to maximize predicted rewards. While valuable for basic alignment, RLHF faces inherent scalability limitations.

This is where Reinforcement Learning from Verifier Feedback (RLVR) delivers a paradigm shift. By replacing human verification with automated assessment of reasoning validity, RLVR creates a self-sustaining training environment. Crucially, it introduces stepwise reward signals - at each reasoning step, an automated verifier evaluates logical coherence, mathematical correctness, or causal validity, providing granular feedback for continuous optimization.

This trajectory culminates in a new era of human-independent reasoning optimization, defined by a decisive pivot toward fully automated reasoning pipelines. Central to this paradigm are three breakthroughs: the remarkable efficacy of 1-Shot RLVR, where verifier feedback applied to a single canonical example yields dramatic quality improvements previously requiring massive datasets; the emergence of Absolute Zero Reasoners — autonomous systems where LLMs simultaneously act as proposer of novel problems and solver of complex challenges; and synthetic multi-environment training frameworks that develop cross-domain reasoning through verifiable simulations. Together, these advances establish the foundation for next-generation AI research, conclusively demonstrating how we can achieve orders-of-magnitude reductions in human verification while permanently resolving the data scarcity crisis.

Aleksei Medvedev
Saint Petersburg Electrotechnical University "LETI"
Saint Petersburg, Russia
e-mail: alkomedvedev@gmail.com