Evolutionary System 2 Reasoning: An Empirical Proof

Zeyuan Ma[14], **Wenqi Huang**[1], **Guo-Huan Song**[23], **Hongshu Guo**[14],
Sijie Ma[1], **Zhiguang Cao**[5], **Yue-Jiao Gong**[1]*

[1]South China University of Technology, [2]Zhejiang Normal University, [3]Northern Computility,
[4]Panorama Optimization, [5]Singapore Management University

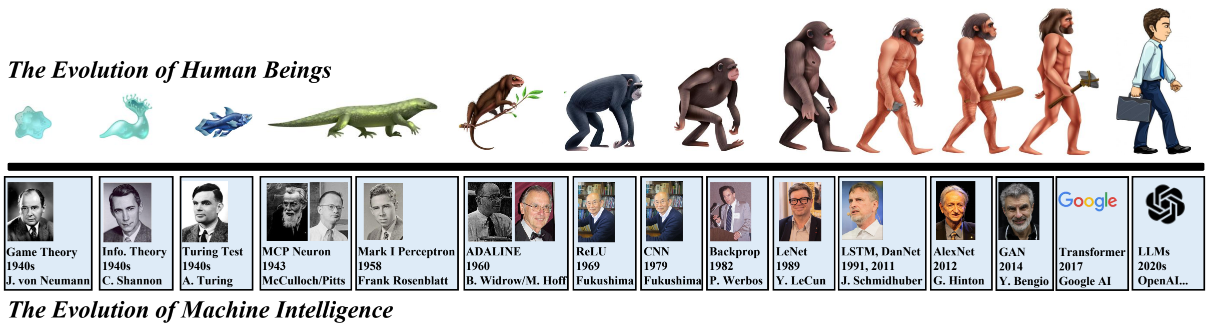

The Evolution of Human Beings
Game Theory 1940s J. von Neumann
Info. Theory 1940s C. Shannon
Turing Test 1940s A. Turing
MCP Neuron 1943 McCulloch/Pitts
Mark I Perceptron 1958 Frank Rosenblatt
ADALINE 1960 B. Widrow/M. Hoff
ReLU 1969 Fukushima
CNN 1979 Fukushima
Backprop 1982 P. Werbos
LeNet 1989 Y. LeCun
LSTM, DanNet 1991, 2011 J. Schmidhuber
AlexNet 2012 G. Hinton
GAN 2014 Y. Bengio
Google
Transformer 2017 Google AI
LLMs 2020s OpenAI...
The Evolution of Machine Intelligence

Figure 1: An intuitive comparison between the evolution paths of human beings and machine intelligence.

Abstract

Machine intelligence marks the ultimate dream of making machines' intelligence comparable to human beings. While recent progress in Large Language Models (LLMs) show substantial *specific skills* for a wide array of downstream tasks, they more or less fall shorts in *general intelligence*. Following correlation between intelligence and system 2 reasoning (slow thinking), in this paper, we aim to answering a worthwhile research question: could machine intelligence such as LLMs be evolved to acquire reasoning ability (not specific skill) just like our human beings? To this end, we propose evolutionary reasoning optimization (ERO) framework which performs *survival of the fittest* over a population of LLMs to search for individual with strong reasoning ability. Given a reasoning task, ERO first initializes multiple LLMs as a population, after which an evolutionary strategy evolves the population to maximize quantified reasoning score of the best individual. Based on experiments on representative testsuites, we claim two surprising empirical discoveries: i) the latest LLMs such as GPT-5 still show limited system 2 reasoning ability; ii) with simple evolution-loop of ERO, a relatively weak model (Qwen-7B) could be enhanced to emerge powerful reasoning ability. Our project can be accessed at https://github.com/MetaEvo/ERO for reproduction needs.

This paper does not advertise for LLMs, but explores more possibilities. — *The authors*

1 Introduction

Machine intelligence (often interchangeably used with AI) has experienced ups and downs within a long river of history (Legg and Hutter 2007; Minsky 2007; LeCun, Bengio, and Hinton 2015). Since the initial proposal of AI at 1950s (McCarthy et al. 2006), an evolution path has been observed: from basic theories (Shannon 1948; Turing 1950) to concrete architectures (Rosenblatt 1958; Fukushima 1980; Hochreiter and Schmidhuber 1997; Vaswani et al. 2017; Gu and Dao 2024) and algorithms (Robbins and Monro 1951; Werbos 1994; Graves 2013; Loshchilov and Hutter 2017). Today, the application of AI has spread to every corner of the world. Domains such as image processing (Gonzalez 2009), nature language processing (Bengio et al. 2003) and scientific discovery (Jumper et al. 2021) benefit from its learning power and corresponding human-competitive performance.

However, we should not overlook the dark side of advanced machine intelligence (i.e., LLMs) simply due to its twinkling academic and engineering achievements (Zhou et al. 2024; Li et al. 2024; Novikov et al. 2025a). In other words, we have to realize that LLMs, though pre-trained with massive human knowledge prior, may still operate at the pattern recognition (fast thinking, System 1 reasoning) level, and hence lacks long-chain, deep, logical reasoning ability (slow thinking, System 2 reasoning), as testified in recent competitions[1].

As illustrated in Figure 1, such System 2 reasoning inability potentially roots from the essential difference be-

*Corresponding author (gongyuejiao@gmail.com)

[1] https://arcprize.org/leaderboard

tween the evolution of machine intelligence and that of our human beings (Cosmides and Tooby 1994; Pinker 2003). For human beings, we are continually involved in evolutionary process under open-ended environmental selection pressure, which follows the *survival of the fittest* principle proposed by Darwin (Darwin, Burrow, and Burrow 1958). The "open-ended" term is used to reference extreme generalization scenario where environmental uncertainty is naturally unknown by human beings (Wolpert 2024). In contrast, almost all machine intelligence instances are trained for specific application scopes explicitly restricted by their developers (human beings). The feedback or learning signal in their learning loops may inherently restricts them from general intelligence with logic reasoning (Wolpert and Macready 2002). To make this point clearer, we borrow the valuable perspective from developmental psychology (Spelke and Kinzler 2007), which holds the position that: human-level intelligence shows generalization and open-endedness and is capable of expanding far beyond its evolution path. More importantly, human is born with innate and evolution-driven knowledge priors such as elementary physics, goal-directness, arithmetic and geometry. These priors enable us to acquire certain skills efficiently (Chollet 2019), by System 2 slow thinking.

The gap between existing LLMs and general System 2 reasoning ability motivates us to explore possible solutions. An intuitive yet under-explored thought would be: Given that existing advanced LLMs have absorbed massive knowledge priors through pre-training with internet-scale corpus, can we further evolve them (e.g., Neuroevolution (Stanley et al. 2019a)) to attain System 2 reasoning ability? To answer this research question empirically, we in this paper propose **Evolutionary Reasoning Optimization (ERO)** framework that enables human-like evolution process for LLMs to adapt themselves in complex tasks that require System 2 reasoning. In our framework, the neural network parameters of a LLM are regarded as a holistic genotype space. At the beginning, given a complex reasoning task, a population of LLMs are randomly born via sampling from the genotype space. Then a (μ+λ) Evolutionary Strategy (ES) (Rechenberg 1978; Beyer and Schwefel 2002) is applied to guide the LLM population toward more powerful System 2 reasoning performance on the target task. The evolution rule in our framework is purely objective-oriented: the LLM individual with higher reasoning ability survives and contributes to the reproduction of offspring, which is closely analogous to evolution of human beings. We provide an intuitive illustration in Figure 4 to showcase how ERO evolves a weak Qwen-7B model to surpass powerful GPT-5 model on reasoning tasks. We next provide a brief review of related works in Sec. 2, elaborate the technical details of **ERO** in Sec. 3 and discuss empirical results in Sec. 4 respectively.

2 Related Works

2.1 Reasoning in LLMs

Reasoning ability is regarded as a key for achieving human-level machine intelligence (Li et al. 2025b). In particular, it relies on logical reasoning and systematic step-by-step thinking to ensure solving effectiveness on complex tasks, which is typically termed as System 2 reasoning. Compared to System 1 reasoning, which features fast, pattern recognition-based decision mapping, System 2 reasoning presents deliberate slow thinking, resulting in concise and rational problem solving via mitigating cognitive biases in System 1 reasoning. While the swift development of LLMs (e.g., DeepSeek-v3 (Liu et al. 2024a), GPT-5 (OpenAI 2025)) shows promising results on understanding and performing human-competitive tasks, they may still lack matched cognitive abilities with human beings in complex reasoning tasks (Chollet et al. 2025).

To improve the capability of reasoning LLMs, initial exploration includes Chain-of-Thought (CoT) (Wei et al. 2022; Kim et al. 2023) and Tree-of-Thought (ToT) (Yao et al. 2023), which focus on preparing high-quality, step-by-step and fine-grained supervision data through decomposing the complex reasoning process into chain or tree structure. Given the data scaling difficulty and single-pass reasoning pattern in CoT and ToT, subsequent works further apply Monte Carlo Tree Search (MCTS) to allow LLMs revisit, reflect and refine their reasoning paths dynamically (Li et al. 2025a; Cheng et al. 2024; Zhao et al. 2024), or self-improvement strategies (Huang et al. 2023) that bootstrap training data from either iterative self-reflection (Zelikman et al. 2022) or rule-based reasoning path augmentation (Guan et al. 2025). Beside these data curation designs, the training paradigm itself also plays crucial role in attaining robust reasoning LLMs. Common practice in up-to-date literature leans to reinforcement fine-tuning (RFT) with output reward modeling (ORM) (Cobbe et al. 2021) or process reward modeling (ORM) (Lightman et al. 2023). The former emphasizes scoring for final answer correctness and the latter pays efforts on fine-grained step-by-step reward labeling. Test time training (TTT) (Yang et al. 2025) is also adopted as effective post-training strategy to mitigate reasoning hallucination. For further reading, we suggest these surveys (Li et al. 2025b; Huang and Chang 2023; Chen et al. 2025).

2.2 Reasoning LLMs Benchmarks

While recent advance of LLMs demonstrates that, with large-scale pre-training on massive and diverse corpus, these novel machine intelligence models rival or even surpass human's performance at specific testbeds (Wang et al. 2019; Adiwardana et al. 2020; Huang, Cheng, and Tan 2025), more evidences argue that they lack compositional System 2 reasoning ability in solving complex tasks as general intelligence (Lee et al. 2024). To this end, a large body of related benchmarks have been curated to provide objective and challenging reasoning tasks for evaluating reasoning LLMs. According to their concrete task types, we could generally document them as: 1) Olympic-level mathematical reasoning benchmarks (He et al. 2024; Yao, Cheng, and Tan 2025); 2) Real world programming challenges summarized from Github (Jimenez et al. 2024); 3) Scientific discovery process (Wang et al. 2024) in physics, chemistry, etc.; 4) Agentic automation workflow tests, e.g., constructing web application from zero (Zhou et al. 2023); 5) Human-level cognitive ability tests (Chollet 2019; Chollet et al. 2025) that

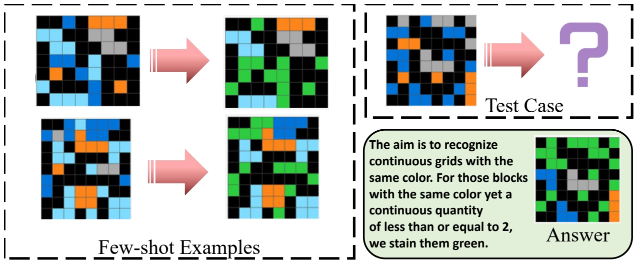

Test Case
The aim is to recognize continuous grids with the same color. For those blocks with the same color yet a continuous quantity of less than or equal to 2, we stain them green.
Answer
Few-shot Examples

Figure 2: A reasoning task example in ARC benchmark.

analog IQ examination.

In this paper, we focus on the last benchmark type, of which a representative benchmark is Abstraction and Reasoning Corpus (ARC) benchmark (Chollet 2019). As illustrated in Figure 2, the testing task instance in ARC benchmark includes multiple few-shot examples and a test case for machine intelligence to solve, which stays close in format of psychometric intelligence test (Carpenter, Just, and Shell 1990). To figure out each puzzle, an intelligence must coherently enable its innate prior on object persistence and contact influence, goal-directedness, numbers and counting, etc., just like our human beings. According to the latest leader board (Chollet et al. 2025), even the most powerful reasoning-reinforced LLMs (GPT 5 and Gemini 3) could only achieve scores no more than 50%, with an evident reasoning gap against human panel (100%). This benchmark provides us a desirable testbed.

2.3 Evolutionary LLMs Enhancement

Evolutionary Algorithms (EAs) (Jin and Branke 2005) are meta-heuristics that follow evolutionary principle in nature to optimize given problems through reproduction and selective pressure. Given EAs' high-level alignment with the evolution process of human beings and robust global optimization capability, they have been validated as powerful optimization techniques for many applications (Slowik and Kwasnicka 2020), except LLMs. Recently, initial attempts have been made to explore the possibility of leveraging EAs to enhance LLMs' performances. While limited, these efforts have seen delightful effects such as prompt optimization through textual evolution (Guo et al. 2023), program evolution through LLM-level genetic programming (Liu et al. 2024b; Novikov et al. 2025b), novel ability composition through model merge recipes (Akiba et al. 2025; Abrantes, Lange, and Tang 2025), incremental and dynamic prompting through evolutionary context engineering (Zhang et al. 2025). However, to the best our knowledge, none of prior works focus on the core vision of LLMs: reasoning like human beings. This highlights the motivation of our paper.

3 Evolutionary Reasoning Optimization

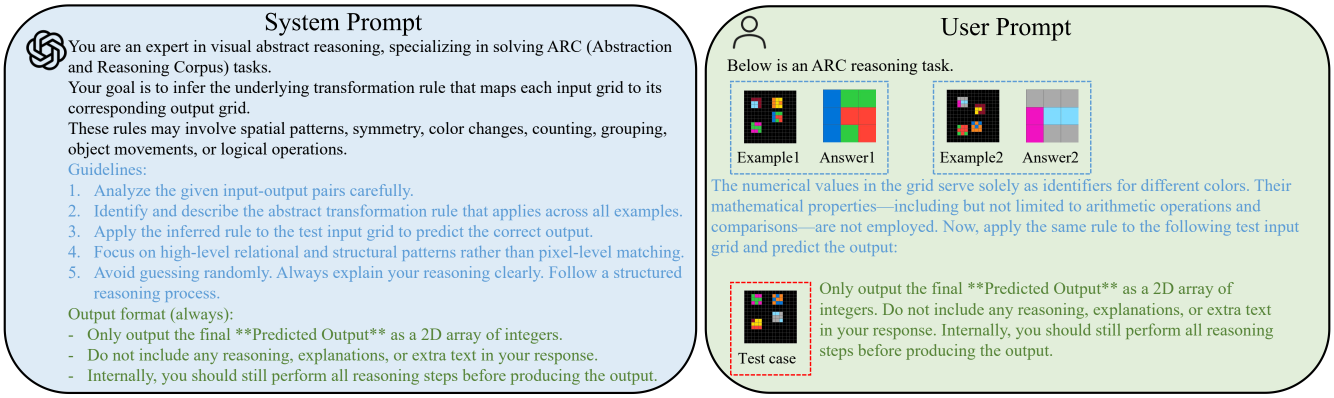
In this section, we elaborate both the general picture and specific designs of our ERO framework to clarify how we address reasoning enhancement for LLMs via EAs perspective. Generally speaking, ERO operates as a neuroevolution (Stanley et al. 2019a) approach, which is under the umbrella of evolutionary strategy (ES) (Rechenberg 1978; Beyer and Schwefel 2002) framework. We present the overall workflow of ERO in Alg. 1, where starting from an existing LLM, an iterative searching process is deployed to evolve the parameters of the LLM toward high reasoning performance on the given reasoning task. However, we must note that it is neither practical nor efficient to run ERO in a standard ES procedure. We next detail the key challenges and corresponding tailored designs in ERO.

Sampling Strategy: In ERO, we have to first determine sampling strategy (i.e., mean and covariance parameters) to serve as initialization module and hence kick out subsequent evolution process. For the mean parameter μ, we could simply set it as the weights of the LLM (denoted as θ^0). The real challenge is how to determine covariance parameter Σ. Although ES has been previously used in evolving relatively smaller neural networks (Stanley et al. 2019b; Wierstra et al. 2014), where the value ranges of parameters are controllable and hence we could set identical entries for Σ matrix, it is absolutely not the case in LLMs. This is backed up by our preliminary experiment, where we conducted a statistical summary on different LLMs and found out that the value ranges of LLMs' parameters vary a lot. However, we also found out that the value ranges of layer-wise parameters are more stable. Based on such observation, we determine the entries of variance matrix Σ by the principle below:

$$\Sigma[k,k] = \epsilon \times \frac{1}{|L_k|} \sum_{n=1}^{|L_k|} \theta^{(0)}[L_k][n] \tag{1}$$

where k denotes k-th neural network parameters of the selected LLM, L_k is the network layer where k-th parameter locate at, ϵ is value between $0 \sim 1$ to control the variance strength, $[\cdot]$ is the indexing operation. We set Σ once leave it fixed until the end. A population of LLMs with the same architecture with $\theta^{(0)}$ are then sampled by the constructed gaussian distribution (line 4 in Alg. 1).

Scoring Function: Given a population of λ sampled LLMs at g-th generation: $\{\theta^{(g),i}\}_{i=1}^{\lambda}$, the underlying ES process in ERO needs proper evaluation metric (scoring function) to measure the reasoning performances of these LLM individuals on the given task τ. A general form of such scoring function can be formulated as $\mathbb{S}(\theta|\tau)$, where θ denotes a tested LLM. We would like to clarify that our ERO does not restrict concrete implementation of the scoring function, instead, it can be quite flexible to tailor appropriate scoring schemes for different reasoning tasks. One can surely use generic schemes such as process reward model (Lightman et al. 2023) that regards any reasoning task as standard reasoning chain and credits those matched reasoning steps. On the other hand, one can also customize special $\mathbb{S}$ function for specific task. Since our ERO is a purely objective-oriented optimization system, all it need is a scalar objective to minimize or maximize. We take the testbed we select for this paper (ARC benchmark) as an example. In ARC, the answer of a reasoning task instance is typically a 1-D or 2-D array indicating the colors of grids. By representing them as



You are an expert in visual abstract reasoning, specializing in solving ARC (Abstraction and Reasoning Corpus) tasks.
Your goal is to infer the underlying transformation rule that maps each input grid to its corresponding output grid.
These rules may involve spatial patterns, symmetry, color changes, counting, grouping, object movements, or logical operations.
Guidelines:

1. Analyze the given input-output pairs carefully.
2. Identify and describe the abstract transformation rule that applies across all examples.
3. Apply the inferred rule to the test input grid to predict the correct output.
4. Focus on high-level relational and structural patterns rather than pixel-level matching.
5. Avoid guessing randomly. Always explain your reasoning clearly. Follow a structured reasoning process.

Output format (always):

- Only output the final **Predicted Output** as a 2D array of integers.
- Do not include any reasoning, explanations, or extra text in your response.
- Internally, you should still perform all reasoning steps before producing the output.

Below is an ARC reasoning task.

The numerical values in the grid serve solely as identifiers for different colors. Their mathematical properties—including but not limited to arithmetic operations and comparisons—are not employed. Now, apply the same rule to the following test input grid and predict the output:

Only output the final **Predicted Output** as a 2D array of integers. Do not include any reasoning, explanations, or extra text in your response. Internally, you should still perform all reasoning steps before producing the output.

Figure 3: System prompt and User prompt we used across all baselines.

strings, one can simply compute the score as:

$$\mathbb{S}(\theta|\tau) = 1 - \frac{lev(\hat{A}(\tau|\theta), A(\tau))}{maxLen(A(\tau), \hat{A}(\tau|\theta))} \tag{2}$$

where $lev(\cdot,\cdot)$ is the Levenshtein distance (Lcvenshtcin 1966) between two strings, A and $\hat{A}$ is the ground truth and predicted answer respectively. In this paper, ERO aims at maximizing the LLM's performance on ARC tasks through maximizing corresponding scoring function values.

Island Architecture: Given the massive searching space of LLM's neural network parameters, island-based population architecture could be a useful strategy to enhance the searching diversity of underlying ES process, which may further improve the final optimization performance (Gong et al. 2015). Besides, since LLMs are inherently aligned with multi-card computational resources and distributed computational methods, island architecture is a coherent choice in LLM-based evolution frameworks (Romera-Paredes et al. 2024; Lee et al. 2025). To this end, our ERO instantiates multiple LLM populations as independent islands, which sample and evaluate LLM individuals (lines $4 \sim 5$ of Alg. 1) in parallel. The communication (fitness aggregation) across different islands occurs when we have to aggregate elite LLM individuals and accordingly update the mean and variance parameters of ES process (lines $6 \sim 7$ of Alg. 1). Unlike vanilla ES, the μ elite LLM individuals are selected as the $\lfloor \frac{\mu}{Z} \rfloor$ best individuals per island, where Z is the number of islands deployed. Once the elite individuals are voted out, we update the mean parameters used for next-generation sampling by averaged aggregation. Note that we keep a fixed variance matrix Σ to maintain continuous exploration strength along the evolution process.

Ray Acceleration: As we mentioned above, the island architecture allows us incorporate advanced distributed ML techniques to reduce the running complexity of LLM-based evolution frameworks. This is particular useful in our ERO, since the scoring evaluation is actually time-consuming, where each LLM individual is fed with reasoning questions and prompted to output reasoning steps and answers. We hence introduce Ray[2], a large-scale ML-enabled computational framework, to distribute each island in ERO onto a separate GPU of a multi-GPU computer/cluster. The Ray parallelism not only enables distributed island-based evolution, but also further facilitates fine-grained parallel evaluation within each island, reducing the running time of ERO from days to hours. In practice, the concrete parallel degree varies due to different hardware conditions.

Algorithm 1: Evolutionary Reasoning Optimization

Input: LLM $\theta^{(0)}$; reasoning task τ; population size λ; elite group size μ; optimization budget G.
Output: best LLM individual θ^* found ever.

```
1: Attain layer-wise covariance $\Sigma$ from $\theta^{(0)}$
2: Let $g = 1$
3: while $g \leq G$ do
4:     Sample $\lambda$ LLMs: $\{\theta^{(g),i}\}_{i=1}^{\lambda} \sim \mathcal{N}(\theta^{(g-1)}, \Sigma)$
5:     Evaluate their reasoning scores: $\{\mathbb{S}(\theta^{(g),i}|\tau)\}_{i=1}^{\lambda}$
6:     Select $\mu$ top-scoring LLMs: $\{\hat{\theta}^{(g),j}\}_{j=1}^{\mu}$
7:     Update $\theta^{(g)} = \frac{1}{\mu}\sum_{j=1}^{\mu}\hat{\theta}^{(g),j}$
8:     $g = g + 1$
9: end while
10: return the LLM individual with the best score
```

Cache Optimization: One may question about how could a large population of LLMs be loaded within a single 4-GPU or 8-GPU computer/cluster, since a single LLM may require at least $10 \sim 20$ GB GPU memory. The solution we propose is to subtly and flexibly leverage limited cache memory. In specific, we only maintain necessary LLM information in an *on-the-fly* fashion for each island (i.e., each GPU node). The necessary LLM information includes the mean parameters at current optimization generation ($\theta^{(g)}$), the layer-wise variance matrix Σ, the elite pool used for maintaining $\lfloor \frac{\mu}{Z} \rfloor$ elite LLM individuals. The elite pool is dynamically updated when a newly sampled LLM individual gets better reasoning score than those in the pool, where the older elite is in-place replaced by the newly sampled one. With such cache memory optimization, ERO could evolve hundreds of LLM individuals simultaneously on GPU memory-limited platform.

[2] https://github.com/ray-project/ray

Tasks	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
Properties															
Object cohesion	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓
Object persistence	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Object influence via contact	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗
Goal-directedness	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Numbers and Counting	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗
Basic Geometry and Topology	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓	✗	✓	✓	✓	✓
Performance															
Ours (ERO+Qwen2.5-7B)	0.7765	**0.7820**	**0.9845**	0.7828	**0.9016**	**1.0000**	**0.8100**	**1.0000**	0.9461	**0.8182**	**1.0000**	0.9627	0.7193	**0.7315**	0.7073
Qwen2.5-7B	0.7059	0.1132	0.9380	0.4253	0.2623	0.9344	0.3584	0.6400	0.8491	0.3333	0.6400	0.9379	0.6486	0.5370	0.6098
Qwen2.5-32B	0.6118	0.3831	0.9535	0.4143	0.5164	0.6393	0.0924	**1.0000**	0.4315	0.4848	0.5200	0.4752	**0.7235**	0.6205	0.4268
GPT-4o-mini	0.1401	0.1200	0.9380	0.3875	0.3115	0.9344	0.3122	0.6400	0.9212	0.4545	0.6400	0.9379	0.5489	0.5507	0.6341
GPT-4o	0.6195	0.2699	0.9767	0.4434	0.2295	0.7541	0.6380	**1.0000**	0.8880	0.6061	0.6400	**0.9689**	0.6861	0.7205	0.6341
GPT-5	**0.8647**	0.6505	0.4961	**1.0000**	0.8934	**1.0000**	0.4027	**1.0000**	**0.9647**	0.5455	0.8200	0.9472	0.4376	0.3260	**0.7195**

Table 1: Pass@1 scores of LLMs baselines across 15 ARC tasks, with their task properties attached at the top of the table.

4 Empirical Validation and Discussion

4.1 Experimental Setup

We list detailed settings of each parts in ERO here, which could be generally divided into three categories:

ERO's Settings: We select *Qwen-7B*[3] as the initial LLM $\theta^{(0)}$ to be evolved. The reason behind such selection is that this relatively poor-reasoning model could facilitate validation on effectiveness of our ERO. For the hyper-parameters of the underlying island-based ES, we set its population size $\lambda = 1000$ which are evenly distributed to $Z = 4$ islands, elite pool size $\mu = 4$ and the optimization budget $G = 12$ generations. All experiments are run on a high-performance instance of a GPU cluster, which comprises an Intel Xeon 8558P CPU, 128 GB RAM and 4×64 GB virtual GPU nodes based on Nvidia H20 GPU.

Testbed: As a preliminary study and due to limited computational resources, in this paper, we have randomly sampled 15 reasoning task instances from hundreds of instances in ARC-1 benchmark (Chollet 2019). We mark these 15 tested instances as T1∼T15. We present at upper half of Table 1 the fine-grained properties of these instances in terms of their correspondence to innate cognitive abilities of human beings. Refer to our project for their correspondence to ARC-1 indices and concrete task descriptions and visualizations.

Baselines: We include 6 baselines in the comparison experiments: 1) *Ours*: the *Qwen-7B* model evolved by our ERO on the given ARC-1 reasoning task instance; 2) *Qwen-7B*: the same *Qwen-7B* pre-trained checkpoint, without ERO's evolution; 3) *Qwen-32B*[4]: a much larger *Qwen* model with stronger general task solving ability than the 7B model; 4) *GPT-4o-mini*[5], 5) *GPT-4o*[6] and 6) *GPT-5*[7], which are three GPT-series models enhanced with multi-modal processing ability and chain-based reasoning capability. For *Ours* and *Qwen-7B*, we deploy their checkpoints at our local GPU server. For the rest of baselines, we call their corresponding APIs. Their key hyper-parameters such as temperature and top-p sampling rate follow default values. For GPT-5, we use its default reasoning efforts level ("minimal").

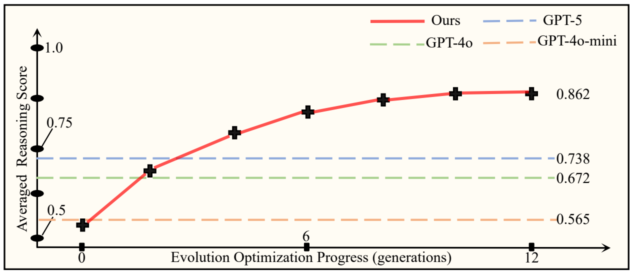


Figure 4: Evolution curve of ERO on ARC benchmark.

4.2 Major Results

For all of the selected baselines, we use a pre-designed standard prompt template to ensure fair evaluation, as illustrated in Figure 3. By using this standard template, we could test selected baselines on the 15 reasoning tasks sampled from ARC-1 benchmark, and then compute their per-instance pass@1 reasoning scores (as we defined in Eq. (2)). We next present these results and corresponding discussion.

Evolutionary Convergence: We first demonstrate the effectiveness of our ERO by illustrating its evolutionary convergence curve as shown in Figure 4, where each scalar point in the red line is the average reasoning score of ERO across 15 tested reasoning tasks. We also attach the average scores of three advanced GPT-series baselines with dashed lines. The results in Figure 4 demonstrate that while a simple pre-trained Qwen-7B model underperforms the GPT models due to its limited capacity and pre-training data scale, it could be evolved by our ERO to surpass these advanced baselines on reasoning tasks. This finding may also indicates the knowledge prior redundancy of existing LLMs. We may not need continually scale both the model capacity and training data size to enable LLM's human-level reasoning ability. On the contrary, such ability may conceal itself within the LLM's parameters, and could be adapted to specific reasoning task through evolution. The results above at least demonstrate potential of evolutionary algorithms on LLM's post-tuning.

Performance Comparison: We further present the per-instance performance comparison between our ERO and other baselines in the lower half of Table 1. Where the best and second-best are labeled in bold and underlined respectively. We also specifically mark the results of our ERO and

[3] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

[4] https://dashscope.aliyuncs.com/compatible-mode/v1

[5] https://platform.openai.com/docs/models/gpt-4o-mini

[6] https://platform.openai.com/docs/models/gpt-4o

[7] https://platform.openai.com/docs/models/gpt-5

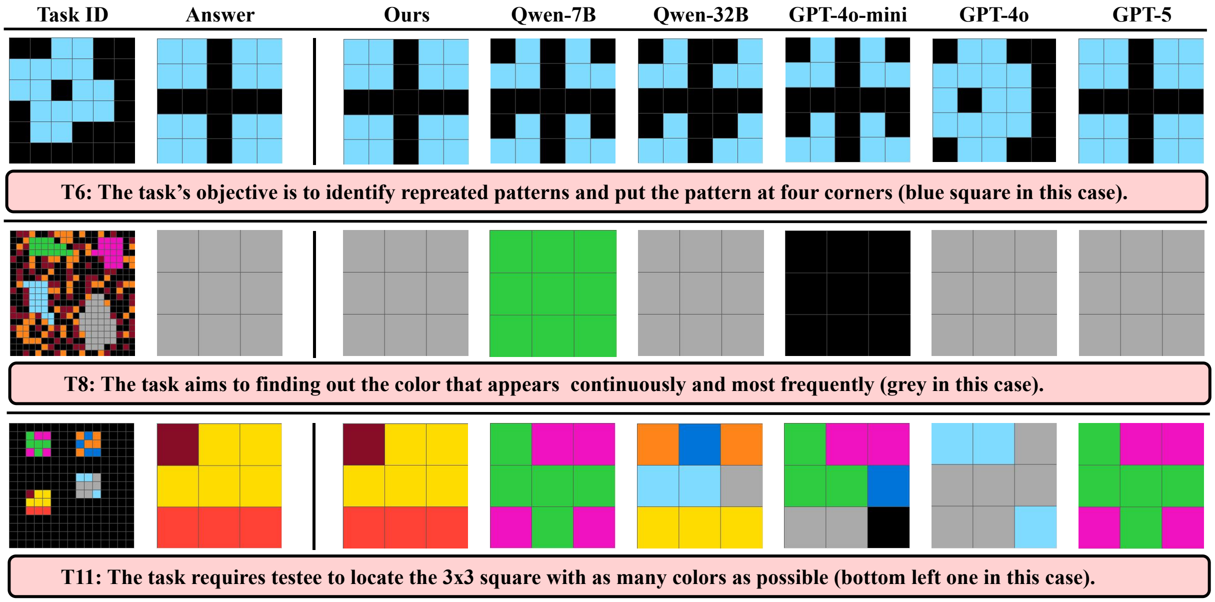

Task ID
Answer
Ours
Qwen-7B
Qwen-32B
GPT-4o-mini
GPT-4o
GPT-5
T6: The task's objective is to identify repreated patterns and put the pattern at four corners (blue square in this case).
T8: The task aims to finding out the color that appears continuously and most frequently (grey in this case).
T11: The task requires testee to locate the 3x3 square with as many colors as possible (bottom left one in this case).

Figure 5: Showcases on the effectiveness our ERO for boosting the understanding and reasoning ability of LLMs.

Qwen-7B in light blue to highlight the relative improvement. From the results, we can observe that: 1) ERO could significantly improve the reasoning capability of Qwen-7B through 12 evolution generations, which cross-validates that intelligence (whether organic or machine-based) obeys evolution principle (*survival-of-the-fittest*); 2) With our ERO, a relatively weak Qwen-7B LLM could be evolved to perform competitively with one of the most advanced LLMs: GPT-5. On 8 of the 15 tested task instances, ERO presents significant performance advantage; 3) The reasoning capability of LLMs may not root from existing scaling law in training these LLMs. A direct evidence lies in the comparison between Qwen-7B and Qwen-32B models. On 8 of the 15 reasoning tasks, a smaller Qwen-7B model presents better logical reasoning and understanding level than its "improved version". This might indicate that we should pay more attention on multi-dimensional solutions for reasoning enhancement of LLMs, not only the scale of LLM pre-training.

We also showcase in Figure 5 three task instances (T6, T8 and T11) where our ERO successfully evolves the initial Qwen-7B model from completely wrong reasoning to crystal correct answer. As their descriptions and visualizations presented in the figure, these ARC-1 reasoning tasks challenge the innate abilities of intelligence of human beings, let alone the LLMs never being trained on such tasks.

4.3 An Important Future Work

In this paper, we mainly focus on the evolution of LLM's reasoning capability under a give reasoning task. While the results mentioned in previous sections have clearly demonstrated that introducing evolutionary perspective into LLM's intelligence enhancement could result in surprising and promising effects, we have to note that the evolution of human beings may not be such simple, i.e., in an *adaption-per-task* fashion. On the contrary, the subtle evolution of human beings emerges in the remix of complex environmental dynamics and concurrent multitasking. This outlines an important and promising future work of our ERO, which is the meta-evolution across reasoning task distribution:

$$\mathbb{S}_{meta} = E_{\tau\sim\Omega}[\mathbb{S}(\theta|\tau)], \tag{3}$$

which is the expectation of reasoning scores over a reasoning task distribution Ω. As computing power and evolution paradigm (e.g., (Sarkar et al. 2025)) continue to iterate and update, we may witness in the near future the emergence of machine intelligence species with diverse behavior and characteristics (e.g., "*The Matrix*" movie), purely by evolution.

5 Conclusion

The position of this paper bridges the evolutionary computation community and LLMs community by proposing the ERO framework, which iteratively evolves LLM's parameters to maximize its System 2 reasoning scores on given reasoning tasks. At the core of ERO, we introduce island architecture-based evolutionary strategy to ensure searching diversity and quality, which attains reasoning performance gain effectively. Combined with specially designed cache optimization and ray acceleration mechanisms, ERO is capable of evolving a large population of LLMs on relatively limited computational resources. We validate ERO's potential by comparing it to existing representative LLMs on ARC benchmark. The promising results not only demonstrate evolution of LLMs is useful for intelligence enhancement, but may also reveal implicit connections between organic human beings and connectionism-based machine intelligence. We hope this work could appeal for more research efforts on evolutionary machine intelligence, and more importantly, exploration on more possibilities.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62276100), in part by the Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (Grant No. 2024B1515040010), in part by Guangzhou Science and Technology Elite Talent Leading Program for Basic and Applied Basic Research (Grant No. SL2024A04J01361), and in part by the Fundamental Research Funds for the Central Universities (Grant No. 2025ZYGXZR027). We also acknowledge that the source material we used in Figure 1 are partly designed by Freepic and Wikipedia.

References

Abrantes, J.; Lange, R.; and Tang, Y. 2025. Competition and Attraction Improve Model Fusion. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 1217–1225.

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Akiba, T.; Shing, M.; Tang, Y.; Sun, Q.; and Ha, D. 2025. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2): 195–204.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb): 1137–1155.

Beyer, H.-G.; and Schwefel, H.-P. 2002. Evolution strategies–a comprehensive introduction. *Natural computing*, 1(1): 3–52.

Carpenter, P. A.; Just, M. A.; and Shell, P. 1990. What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological review*, 97(3): 404.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Cheng, J.; Liu, X.; Wang, C.; Gu, X.; Lu, Y.; Zhang, D.; Dong, Y.; Tang, J.; Wang, H.; and Huang, M. 2024. Spar: Self-play with tree-search refinement to improve instruction-following in large language models. *arXiv preprint arXiv:2412.11605*.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Chollet, F.; Knoop, M.; Kamradt, G.; Landers, B.; and Pinkard, H. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cosmides, L.; and Tooby, J. 1994. Origins of domain specificity: The evolution of functional organization. *Mapping the mind: Domain specificity in cognition and culture*, 853116.

Darwin, C.; Burrow, J. W.; and Burrow, J. W. 1958. *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt New York.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4): 193–202.

Gong, Y.-J.; Chen, W.-N.; Zhan, Z.-H.; Zhang, J.; Li, Y.; Zhang, Q.; and Li, J.-J. 2015. Distributed evolutionary algorithms and their models: A survey of the state-of-the-art. *Applied Soft Computing*, 34: 286–300.

Gonzalez, R. C. 2009. *Digital image processing*. Pearson education india.

Graves, A. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Gu, A.; and Dao, T. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. In *Forty-second International Conference on Machine Learning*.

Guo, Q.; Wang, R.; Guo, J.; Li, B.; Song, K.; Tan, X.; Liu, G.; Bian, J.; and Yang, Y. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3828–3850.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Huang, B.; Cheng, R.; and Tan, K. C. 2025. EvoGit: Decentralized Code Evolution via Git-Based Multi-Agent Collaboration. *arXiv preprint arXiv:2506.02049*.

Huang, J.; and Chang, K. C.-C. 2023. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, 1049–1065.

Huang, J.; Gu, S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2023. Large language models can self-improve. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 1051–1068.

Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. R. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations*.

Jin, Y.; and Branke, J. 2005. Evolutionary optimization in uncertain environments-a survey. *IEEE Transactions on evolutionary computation*, 9(3): 303–317.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Kim, S.; Joo, S.; Kim, D.; Jang, J.; Ye, S.; Shin, J.; and Seo, M. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12685–12708.

Lcvenshtcin, V. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Lee, K.-H.; Fischer, I.; Wu, Y.-H.; Marwood, D.; Baluja, S.; Schuurmans, D.; and Chen, X. 2025. Evolving deeper llm thinking. *arXiv preprint arXiv:2501.09891*.

Lee, S.; Sim, W.; Shin, D.; Seo, W.; Park, J.; Lee, S.; Hwang, S.; Kim, S.; and Kim, S. 2024. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Transactions on Intelligent Systems and Technology*.

Legg, S.; and Hutter, M. 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4): 391–444.

Li, Q.; Xia, W.; Dai, X.; Du, K.; Liu, W.; Wang, Y.; Tang, R.; Yu, Y.; and Zhang, W. 2025a. Rethinkmcts: Refining erroneous thoughts in monte carlo tree search for code generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 8103–8121.

Li, Y.; Wen, H.; Wang, W.; Li, X.; Yuan, Y.; Liu, G.; Liu, J.; Xu, W.; Wang, X.; Sun, Y.; et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025b. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, F.; Tong, X.; Yuan, M.; Lin, X.; Luo, F.; Wang, Z.; Lu, Z.; and Zhang, Q. 2024b. Evolution of Heuristics: Towards Efficient Automatic Algorithm Design Using Large Language Model. In *ICML*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

McCarthy, J.; Minsky, M. L.; Rochester, N.; and Shannon, C. E. 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4): 12–12.

Minsky, M. 2007. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1): 8–30.

Novikov, A.; Vũ, N.; Eisenberger, M.; Dupont, E.; Huang, P.-S.; Wagner, A. Z.; Shirobokov, S.; Kozlovskii, B.; Ruiz, F. J.; Mehrabian, A.; et al. 2025a. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

Novikov, A.; Vũ, N.; Eisenberger, M.; Dupont, E.; Huang, P.-S.; Wagner, A. Z.; Shirobokov, S.; Kozlovskii, B.; Ruiz, F. J.; Mehrabian, A.; et al. 2025b. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

OpenAI. 2025. GPT-5 System Card. Technical report, OpenAI.

Pinker, S. 2003. *The blank slate: The modern denial of human nature*. Penguin.

Rechenberg, I. 1978. Evolutionsstrategien. In *Simulationsmethoden in der Medizin und Biologie: Workshop, Hannover, 29. Sept.–1. Okt. 1977*, 83–114. Springer.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Balog, M.; Kumar, M. P.; Dupont, E.; Ruiz, F. J.; Ellenberg, J. S.; Wang, P.; Fawzi, O.; et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995): 468–475.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6): 386.

Sarkar, B.; Fellows, M.; Duque, J. A.; Letcher, A.; Villares, A. L.; Sims, A.; Cope, D.; Liesen, J.; Seier, L.; Wolf, T.; et al. 2025. Evolution Strategies at the Hyperscale. *arXiv preprint arXiv:2511.16652*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Slowik, A.; and Kwasnicka, H. 2020. Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications*, 32(16): 12363–12379.

Spelke, E. S.; and Kinzler, K. D. 2007. Core knowledge. *Developmental science*, 10(1): 89–96.

Stanley, K. O.; Clune, J.; Lehman, J.; and Miikkulainen, R. 2019a. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1): 24–35.

Stanley, K. O.; Clune, J.; Lehman, J.; and Miikkulainen, R. 2019b. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1): 24–35.

Turing, A. M. 1950. Computing machinery and intelligence. In *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer*, 23–65. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Werbos, P. J. 1994. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. John Wiley & Sons.

Wierstra, D.; Schaul, T.; Glasmachers, T.; Sun, Y.; Peters, J.; and Schmidhuber, J. 2014. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1): 949–980.

Wolpert, D. H. 2024. What can we know about that which we cannot even imagine? In *New Frontiers in Science in the Era of AI*, 301–331. Springer.

Wolpert, D. H.; and Macready, W. G. 2002. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1): 67–82.

Yang, L.; Yu, Z.; Cui, B.; and Wang, M. 2025. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *arXiv preprint arXiv:2502.06772*.

Yao, J.; Cheng, R.; and Tan, K. C. 2025. VAR-MATH: Probing True Mathematical Reasoning in LLMS via Symbolic Multi-Instance Benchmarks. *arXiv preprint arXiv:2507.12885*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35: 15476–15488.

Zhang, Q.; Hu, C.; Upasani, S.; Ma, B.; Hong, F.; Kamanuru, V.; Rainton, J.; Wu, C.; Ji, M.; Li, H.; et al. 2025. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*.

Zhao, Y.; Yin, H.; Zeng, B.; Wang, H.; Shi, T.; Lyu, C.; Wang, L.; Luo, W.; and Zhang, K. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Zhou, H.; Hu, C.; Yuan, Y.; Cui, Y.; Jin, Y.; Chen, C.; Wu, H.; Yuan, D.; Jiang, L.; Wu, D.; et al. 2024. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*, 27(3): 1955–2005.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.