AI Link Collection Report - 2025-07-08

Generated on July 08, 2025 at 11:21 AM

Total Items: 1

Report Statistics	
Total Links Processed	1
Generation Date	2025-07-08
Generation Time	11:21:48

■ Link Collection Details

1. The Era of Exploration

■ https://yidingjiang.github.io/blog/post/exploration/

■ Full Article Content:

The Era of Exploration - **Large Language Models (LLMs)** - Unintended byproduct of three decades of freely accessible human text online. - Ilya Sutskever compares this information reservoir to **fossil fuel**: abundant but finite. - Studies suggest: - At current token-consumption rates, frontier labs could exhaust high-quality English web text before the decade ends. - Today's models consume data faster than humans can produce it. - **Era of Experience** - Coined by David Silver and Richard Sutton. - Meaningful progress will depend on data generated by learning agents themselves. - Key point: The bottleneck is not just any experience but the **right kind of experience** that benefits learning. - Future AI progress will focus on **exploration** rather than merely stacking parameters. - **Cost of Experience Collection** - Scaling involves resource considerations: - Compute cycles - Synthetic-data generation - Data curation pipelines - Human oversight - Any expenditure that creates learning signals. - Introduced term: **flops** - Represents one floating-point operation. - Used as a common abstract currency for measuring effort consumed by systems. - Discussion focuses on relative spend, not specific resources. - **Exploration in Data-Driven Systems** - Exploration is crucial for every data-driven system to decide which experiences to collect. - Broader definition of exploration beyond reinforcement learning (RL). - Inspired by Minqi’s article: **General intelligence requires rethinking exploration**. - **Post Organization** - The following sections will cover: 1. How pre-training inadvertently solved part of the exploration problem. 2. Why better exploration translates into better generalization. 3. Where to invest the next hundred thousand GPU-years. - **Pretraining as Exploration** - Standard LLM pipeline: 1. Pretrain a large model on next-token prediction using extensive text. 2. Fine-tune the model with RL for specific objectives. - Without large-scale pretraining, RL struggles to progress. - Observations: - Smaller models show improved reasoning when distilled from larger models. - Misinterpretation: Large scale is not a prerequisite for effective reasoning. - Key question: If model capacity isn't the bottleneck, why do smaller models need distillation from larger ones? - Explanation: - Pretraining incurs a significant **exploration tax**. - Models without pretraining or smaller pretrained models struggle to explore the solution space effectively. - Pretraining invests vast compute resources to learn a rich sampling distribution for likely correct continuations. - Distillation allows smaller models to inherit exploration capabilities from larger models' investments. ## The Era of Exploration ### Importance of Pre-Paid Exploration - **Smaller pretrained models** struggle to explore the solution space effectively. - **Pretraining** involves significant computational resources to learn a rich sampling distribution for likely correct continuations. - **Distillation** allows smaller models to inherit the exploration capabilities from larger models, leveraging prior investments. ### Reinforcement Learning (RL) Loop - The general RL loop consists of: - **Exploration**: The agent generates randomized exploration trajectories. - **Reinforce**: Good trajectories are up-weighted, while bad ones are down-weighted. - For effective learning: - The agent must generate a minimal number of **“good” trajectories** during exploration. - This concept is known as **coverage** in RL. - In **Large Language Models (LLMs)**: - Exploration is achieved through sampling from the model’s autoregressive output distribution. - Correct solutions must be likely in the naive sampling distribution. - Lower-capacity models may struggle to find valid solutions through random sampling, leading to ineffective reinforcement. ### Challenges of Exploration - **Exploration without prior information** is difficult: - Even in simple tabular RL, extensive trials are needed for learning. - A known lower-bound on sample complexity is \(\Omega(\frac{SAH^2}{\epsilon^2})\) (Dann & Brunskill, 2015): - **S**: Size of the state space - **A**: Size of the action space - **H**: Horizon - **ε**: Distance to the best solution

- Minimum episodes grow linearly with state-action pairs and quadratically with the horizon. - For LLMs: - The state space includes every possible text prefix. - The action space consists of any next token, both of which are large. - Without prior information, RL becomes nearly impossible. ### Role of Pretraining - Pretraining has done the heavy lifting for exploration by learning a better prior for trajectory sampling. - However, this constrains the types of trajectories that can be sampled naively. - To advance, we need strategies to move beyond the prior. ### Exploration and Generalization - Historically, RL research focused on single environments (e.g., Atari, MuJoCo): - This is akin to training and testing on the same data point. - Performance in a single environment does not indicate how well a model handles novel situations. - **Generalization** is crucial in machine learning: - Success in unseen problems is more important than solving known ones. - For LLMs: - Training involves a finite set of prompts, but deployment requires handling arbitrary user queries. - Current LLMs excel in tasks with verifiable rewards (e.g., coding puzzles) due to easily checkable correctness. - The challenge lies in generalizing to ambiguous domains (e.g., generating reports) where feedback is sparse. ### Options for Training Generalizable Models - **Data diversity** is key for robust generalization in deep learning. - Exploration directly influences data diversity: - In supervised learning, each labeled example reveals all details in one pass, necessitating more data for diversity. - In RL, each interaction reveals a narrow slice of the environment. - Agents must collect varied trajectories to build a representative picture. - Lack of diversity in collected trajectories can lead to overfitting, even within the same environment. ## The Era of Exploration ### Supervised Learning vs. Reinforcement Learning (RL) - **Supervised Learning**: - Labeled examples reveal all details in a single forward pass. - Data diversity can only be increased by collecting more data. - **Reinforcement Learning (RL)**: - Each interaction exposes a narrow slice of the environment. - Agents must gather varied trajectories to build a representative picture. - Lack of diversity in trajectories (e.g., naive random sampling) can lead to overfitting. ### Challenges in Multiple Environments - **Procgen Benchmark**: - A collection of Atari-like games with procedurally generated environments. - Each game theoretically contains “infinitely” many environments. - Objective: Train on a fixed number of environments and generalize to unseen ones. ### Existing Approaches and Limitations - Many approaches treat the problem as **representation learning** and apply regularization techniques from supervised learning (e.g., dropout, data augmentation). - These techniques help but overlook **exploration**, a crucial component of RL. - Agents can improve generalization by changing exploration strategies. ### Research Findings - Previous work showed that pairing an RL algorithm with a stronger exploration strategy can: - **Double generalization performance** on Procgen without explicit regularization. - Recent findings indicate that better exploration allows models to: - Leverage more expressive architectures and computational resources. - Generalize better on Procgen. ### Exploration in LLMs - While Procgen is simpler than current LLM challenges, the problem structure is similar: - RL agents are trained on a finite set of problems and tested on new problems without further training. - Current exploration methods in LLMs are basic: - Typically involve sampling from the model’s autoregressive distribution with tweaks (e.g., temperature, entropy bonus). - There is potential for better exploration approaches, but few successful examples exist. ### Potential Issues with Exploration - Challenges in improving exploration could stem from: - Difficulty of the problem. - Inefficiency in terms of computational resources. - Lack of effort in exploring new strategies. - If Procgen-style exploration gains translate, we may be missing out on efficiency and new capabilities. ## Two Axes of Scaling Exploration - **Exploration** involves deciding what data the learner will see, occurring on two axes: ### 1. World Sampling - Refers to deciding where to learn: - In supervised learning, this includes data collection, synthetic generation, and curation. - In RL, it involves designing or generating environments (e.g., math puzzles, coding problems). - Can arrange worlds into curricula. - Determines the limit on information any agent can learn. ### 2. Path Sampling - Refers to deciding how to gather data within a world (unique to RL): - After choosing a world, the agent selects trajectories to collect (e.g., random walks, curiosity-driven policies, tree search, tool-use). - Different strategies incur varying computational costs and produce different training distributions. - Path sampling is about what the learner “wants” to see. ### Cost Considerations - In supervised learning or unsupervised pretraining: - The second axis incurs a constant cost due to access to all information in each data point. - In RL, exploration costs primarily reside on the first axis (world sampling): - Flops can be allocated to acquiring new worlds or processing existing ones. ## The Era of Exploration

Key Concepts - **Supervised Learning** and **Unsupervised Pretraining**: - Constant cost on the second axis due to access to all information in data points (e.g., cross-entropy loss). - Exploration cost primarily on the first axis – **world sampling**. - **Reinforcement Learning (RL)**: - Greater flexibility in both axes (world sampling and path sampling). - Random trajectories often reveal little about ideal behavior, leading to lower information density (useful bits per flop). - Naïve trajectory sampling risks wasting flops on noise. ### Spending Flops Wisely - Options for exploring within each world: - Sample more trajectories from a single environment. - Spend flops on strategizing the next trajectory to discover high-value states and actions. ### Maximizing Information per Flop - High-level goal in machine learning: - **Maximize information per flop**. - Trade-off curve between: - Resources spent on **world sampling**. - Resources spent on **path sampling**. - Risks: - Too much focus on world sampling may lead to meaningless experiences. - Overfitting to a small set of worlds may hinder generalizable behavior. - Ideal scenario: - Balanced resource allocation between sampling new worlds and extracting information from existing worlds. ### Scaling Laws and Performance Curves - Similarity to **Chinchilla scaling laws**: - Two axes correspond to compute used for different types of sampling rather than parameters and data. - Isoperformance curve can be traced at each performance level: - X-axis: Compute for interacting with environments. - Y-axis: Compute for generating or running environments (e.g., generative verifier with CoT). ### Path Sampling vs. World Sampling - **Path Sampling**: - Well-defined problem with a clear objective: reduce model uncertainty. - Existing approaches have strong sample complexity but can be expensive. - **World Sampling**: - Less clear objectives; open-ended learning requires defining the universe of environments or subjective judgments on interesting outcomes. - Infinite space of environments vs. finite resources necessitates expressing preferences over environments. ### Designing Environment Specs - Challenges in designing environments: - Similar to selecting pretraining data; hard to determine why one environment aids another. - Likely scenario: designing specs within individual expertise or domain. - Future possibilities: - Learning common principles from sufficient “human-approved” and “useful” specs. - Potential for automation in the process, akin to current pretraining data selection. - Preliminary evidence suggests that fewer environments may suffice for achieving generality in decision-making. ## The Era of Exploration - **Objective**: Train an agent capable of **general exploration** and **decision making** in entirely out-of-distribution environments. - **Design Process Acceleration**: - Utilizing existing **Large Language Models (LLMs)** can significantly speed up the design process. - Likely scenario: Individuals will design specifications within their own **expertise** or **domain of interest**. - **Learning from Specifications**: - Accumulating enough “**human-approved**” and “**useful**” specifications may allow for the identification of common principles. - This could lead to automation in the design process, similar to current **pretraining data selection**. - **Generalization Concerns**: - It would be inconvenient if the same number of environments as pretraining data is needed for equivalent decision-making generality. - Preliminary evidence suggests that a **small number of environments** can suffice for training an agent in out-of-distribution settings. - **Scaling Challenges**: - Scaling the two axes of world sampling and path sampling is less straightforward than scaling pretraining. - A reliable method for introducing scale into world sampling and a more intelligent approach to path sampling could yield **isoperformance curves** that bend inward towards the origin. - This would inform optimal allocation of computational resources between environments and agents. ## Final Thoughts - **Exploration Focus**: - While there are many potential avenues (e.g., better **curiosity objectives**, **open-endedness**, **meta-exploration**), the key point is the importance of exploration. - Existing scaling paradigms have been effective but will eventually reach saturation. - The critical question is where to invest the next significant computational resources. - Exploration (world sampling and path sampling) is proposed as a promising direction. - **Future Considerations**: - The right scaling laws, environment generators, and exploration objectives are still unknown but should be achievable. - The upcoming years will determine if exploration can extend computational capabilities beyond existing paradigms. - The investment in exploration is deemed worthwhile. ## Acknowledgements - Special thanks to: - Allan Zhou - Sam Sokota - Minqi Jiang - Ellie Haber - Alex Robey - Swaminathan Gurumurthy - Kevin Li - Calvin Luo - Abitha Thankaraj - Zico Kolter - For their feedback and discussions on the draft. ## Additional Notes - **RL Optimization Objective**: - A valid alternative possibility is that the **Reinforcement Learning (RL)** optimization objective may not perform well with smaller models, though this is likely not

the case as successful RL applications prior to LLMs often involved small models. - **Model Limitations**: - The model may not fully exploit available information due to computational limitations, but the information remains accessible if desired. - **Generalization Assumption**: - For generalization to be feasible, it is assumed that a “**good enough**” policy exists for all environments, similar to the assumption of minimal label noise in supervised learning. - **Performance Benchmark**: - At the time of writing, this work sets a new **state-of-the-art** performance on the “**25M easy**” benchmark of **ProcGen**. - **Random Sampling**: - For many problems, such as **Atari**, random sampling performs reasonably well, indicating more about the environments than the exploration method itself. - **Exploration Algorithms**: - A variety of RL algorithms, known as **posterior sampling** or **information-directed sampling**, aim to guide exploration to reduce model uncertainty but are generally too costly for LLMs at scale. Various approximations exist but are not widely utilized for LLMs.

■ **Shared by:** Sai
■ **Shared on:** Jul 07, 2025 at 11:21 AM
■ **Domain:** yidingjiang.github.io
■ **Length:** 6,325 words
■ **Processed:** Jul 08, 2025 at 11:21 AM