AI Link Collection Report - 2025-07-08

Generated on July 08, 2025 at 11:54 AM

Total Items: 1

Report Statistics	
Total Links Processed	1
Generation Date	2025-07-08
Generation Time	11:54:45

■ 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.

- The bottleneck is not just having experience but collecting the **right kind of experience** that benefits learning.

- Future AI progress will focus on:

- **Exploration**: acquiring new and informative experiences.

- **Cost of Experience Collection**

- Scaling is fundamentally a question of resources:

- Compute cycles

- Synthetic-data generation

- Data curation pipelines

- Human oversight

- Any expenditure that creates learning signals.

- Introduced a bookkeeping unit called **flops**:

- Represents one floating-point operation.

- Used as a common currency for measuring effort consumed by systems.

- Discussion focuses on relative spend, not specific resources.

- **Exploration in Data-Driven Systems**

- Exploration is often associated with **Reinforcement Learning (RL)** but is broader:

- Every data-driven system must decide which experiences to collect before learning.

- 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 allocate 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 desired objectives.

- Without large-scale pretraining, RL struggles to progress.

- Observations:

- Smaller models show improved reasoning when distilled using chain-of-thought from larger models.

- Some interpret this as evidence that large scale isn't necessary for effective reasoning, but this view is misguided.

- Key question:

- If model capacity isn't the bottleneck, why do smaller models need to distill from larger ones?

- Explanation:

- The cost of pretraining is an **upfront exploration tax**.

- Models without pretraining or smaller pretrained models find it harder to explore the solution space.

- Pretraining pays this tax by investing compute in diverse data to learn a rich sampling distribution.

- Distillation allows smaller models to inherit this exploration capability from larger models.

The Era of Exploration

Importance of Pre-Paid Exploration

- **Pretrained models** vs. smaller models:

- Smaller models struggle to explore the solution space effectively.

- Pretraining involves significant compute resources to learn a rich sampling distribution.

- **Distillation**:

- Allows smaller models to inherit exploration capabilities from larger models.

- Bootstraps exploration from the investment made in larger models.

Reinforcement Learning (RL) Loop

- General form of the RL loop:

- **Exploration**: Agent generates randomized exploration trajectories.

- **Reinforce**: Good trajectories are up-weighted; bad ones are down-weighted.

- **Coverage in RL**:

- Essential for the agent to generate a minimal number of "good" trajectories during exploration.

- In 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.

Challenges of Exploration

- Exploration without prior information is difficult:
- Even in simple tabular RL, extensive trials are needed.
- Sample complexity lower-bound: $\Omega(\frac{SAH^2}{\epsilon^2})$ (Dann & Brunskill, 2015)
- **Variables**:
- S: Size of the state space
- A: Size of the action space
- H: Horizon
- ϵ: Distance to the best possible solution
- Minimum episodes grow linearly with state-action pairs and quadratically with the horizon.
- For LLMs:
- State space includes every possible text prefix.
- Action space is any next token, both very large.
- Without prior information, RL becomes practically impossible.

Role of Pretraining

- Pretraining has done the hard work of exploration:
- Learns a better prior for sampling trajectories.
- Types of trajectories sampled are constrained by the prior.
- Need to move beyond the prior for further progress.

Exploration and Generalization

- Historical focus in RL:
- Solving a single environment (e.g., Atari, MuJoCo).
- Equivalent to training and testing on the same datapoint.
- Importance of generalization:
- Success in unseen or unanticipated problems is crucial.
- Generalization performance is critical for language models (LLMs).
- LLM training vs. deployment:
- Trained on a finite set of prompts.
- Must handle arbitrary user queries at deployment.
- Current LLMs excel in tasks with verifiable rewards (e.g., coding puzzles, formal proofs).
- Challenges in fuzzier domains (e.g., generating research reports, writing novels):
- Feedback is sparse or ambiguous.
- Large-scale training and data collection are difficult.

Options for Training Generalizable Models

- Data diversity drives robust generalization:

- Exploration directly controls data diversity.
- In supervised learning:
- A labeled example reveals all details in a single forward pass.
- Increasing data diversity requires collecting more data.
- In RL:
- Each interaction exposes a narrow slice of the environment.
- Agents must gather varied trajectories to build a representative picture.
- Lack of diversity in collected trajectories can lead to overfitting.

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.

Current Approaches and Limitations

- Many existing methods treat the problem as a **representation learning** issue:
- Apply regularization techniques from supervised learning (e.g., dropout, data augmentation).
- These methods overlook **exploration**, a crucial component of RL.

- **Exploration Strategies**:
- Agents can improve generalization by changing exploration methods.
- Previous work showed that pairing RL algorithms with stronger exploration strategies can:
- Double generalization performance on Procgen without explicit regularization.
- Allow models to leverage more expressive architectures and resources.

Comparison with LLMs

- While Procgen is less complex than problems faced by **Large Language Models (LLMs)**:
- The problem structure is similar: trained on finite problems, tested on new ones without further training.
- Current exploration in LLMs is simple, often limited to sampling from autoregressive distributions.
- There is significant potential for better exploration approaches.

- **Challenges in Exploration**:

- Few successful examples of improved exploration strategies.

- Possible reasons for limited success:

- Difficulty of the problem.

- Inefficiency in terms of computational resources.

- Lack of rigorous attempts to explore this area.

- If gains from Procgen-style exploration translate to LLMs:

- 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**:

- Deciding where to learn (specific problems to solve).

- 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).

- World sampling limits the information any agent can learn.

2. **Path Sampling**:

- Deciding how to gather data within a chosen world (unique to RL).

- Agents choose trajectories to collect (e.g., random walks, curiosity-driven policies, tree search).

- Different path-sampling strategies can vary in computational cost and training distributions.

- In supervised learning or unsupervised pretraining:

- The second axis incurs a constant cost due to access to all information in a data point.

- Exploration cost primarily resides on the first axis (world sampling).

- Computational resources 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 using a trade-off curve between world sampling and 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 ones.

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:

- 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 objective; open-ended learning requires defining the universe of environments or subjective judgment of interesting outcomes.

- Infinite space of environments vs. finite resources necessitates expressing preferences over environments.

Designing Environment Specs

- Challenges in optimizing world sampling objectives:

- Designing environments may resemble selecting pretraining data.

- Difficulty in determining why one environment aids another.

- Likely scenario:

- Designing specs within individual expertise or domain.

- Learning common principles from sufficient “human-approved” and “useful” specs.

Conclusion

- Preliminary evidence suggests that fewer environments may suffice for achieving generality in decision-making, contrasting with the need for extensive pretraining data.

The Era of Exploration

- **Objective**: Train an agent for general exploration and decision-making in out-of-distribution environments.

- **Design Process Acceleration**:

- Utilizing existing **Large Language Models (LLMs)** can significantly speed up the design process.

- Anticipated trend: Individuals will design specifications within their own expertise or domain.

- **Learning from Specifications**:

- Accumulating enough "human-approved" and "useful" specifications may allow for the identification of common principles.

- Potential to automate the 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 decision-making generality.

- Preliminary evidence suggests that a small number of environments can suffice for training an agent in out-of-distribution scenarios.

- **Scaling Challenges**:

- Scaling exploration and decision-making is less straightforward than scaling pretraining.

- Reliable methods for world sampling and intelligent path sampling are needed.

- Expected outcome: Isoperformance curves that bend inwards towards the origin, indicating efficient resource allocation between environments and agents.

Final Thoughts

- **Exploration Focus**:

- Exploration (world sampling and path sampling) is a promising direction for future research.

- Current scaling paradigms are effective but may eventually reach saturation.

- The challenge lies in determining where to allocate additional computational resources.

- **Future Considerations**:

- Unknowns include the right scaling laws, environment generators, and exploration objectives.

- The next few years will reveal if exploration can enhance computational efficiency beyond existing paradigms.

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 Insights

- **RL Optimization Objective**:

- A valid alternative possibility is that the RL optimization objective may not perform well with smaller models, but this is likely not the case as previous successful RL applications involved small models.

- **Information Exploitation**:

- Models 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, a new state-of-the-art performance was achieved on the "25M easy" benchmark of ProcGen.

- **Random Sampling Effectiveness**:

- Random sampling works reasonably well for many problems, such as Atari, indicating more about the environments than the exploration method itself.

- **Exploration Algorithms**:

- A variety of RL algorithms, such as posterior sampling or information-directed sampling, aim to reduce model uncertainty during exploration.

- These methods are generally too costly to implement at the scale of LLMs, and existing approximations have not been widely adopted.

- **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 11:54 AM