Deploying AI

Data Engineering

```
$ echo "Data Sciences Institute"
```

Introduction

Agenda

Agenda

- Data curation
- Data augmentation and synthesis
- Data processing

Dataset Engineering

- The quality of a model depends heavily on the quality of its training data.
- Even with infinite compute, a model cannot perform well without appropriate data.
- Dataset engineering aims to create datasets that allow you to train the best model within budget.
- As models demand more data, dataset handling requires more talent and infrastructure investments.

Data-Centric AI

- Data-centric AI focuses on improving AI performance through better datasets.
- Model-centric AI instead emphasizes architectures, model sizes, and training techniques.
- Data-centric practices are increasingly necessary as data becomes the main differentiator of performance.
- High-quality datasets can often matter more than slight architectural improvements.

Training Phases and Data Needs

- Different training phases require datasets with different attributes.
- Pre-training data quantity is measured in tokens.
- Supervised finetuning data quantity is measured in the number of examples.
- Post-training data requires curation tailored to application-specific needs.

Key Aspects of Dataset Engineering

- **Data curation**: deciding what data is needed and how much.
- **Data synthesis**: generating or augmenting data to fill gaps.
- **Data processing**: cleaning, filtering, and preparing data.
- These steps are iterative and often require going back and forth.

Data Curation

- Data should have essential characteristics: compliant, diverse, high quality, and sufficient.
- Compliance means adhering to laws, regulations, and internal policies.
- Coverage ensures training data represents the variety of real-world problems.
- High-quality data should exhibit the behaviors you want your model to learn.

Compliance in Data

- Training data must comply with relevant policies, including privacy and PII restrictions.
- Non-compliant data can cause legal, ethical, and operational risks.
- Thinking through compliance early prevents costly rework later.

Data Coverage

- Data should capture the diversity of usage patterns expected in your application.
- Examples should include long and short instructions, formal and informal queries, and multiple languages if relevant.
- Diversity dimensions include domain, topic, style, and length.
- Llama 3's performance gains were largely driven by improvements in data quality and diversity.

Data Quality Challenges

- High-quality annotations are always difficult to obtain.
- Chain-of-thought annotations require detailed step-by-step reasoning, which is labor-intensive.
- Tool-use data requires careful design or simulation of task execution.
- Human annotations may not always align with efficient model usage.

Experimenting with Small Datasets

- Finetuning with small datasets (50–100 examples) can still show improvements.
- Plotting performance against dataset size helps estimate how much data is needed.
- Gains usually diminish as dataset size grows.
- Diversity of tasks can matter as much as the number of examples.

Data Acquisition

- Sources of data include public datasets, purchased datasets, in-house annotation, and synthetic generation.
- Data acquisition strategies balance quality, diversity, budget, and compliance.
- User-generated application data often provides the most valuable feedback loop.
- Building a “data flywheel” can continually improve product performance.

Annotation Guidelines

- Annotation requires clear, detailed guidelines to ensure consistency.
- Guidelines define what makes a response “good” or “bad.”
- Without strong guidelines, annotation quality may drift over time.
- Evaluation data guidelines can be reused for annotation purposes.

Data Augmentation and Synthesis

- Data augmentation modifies real data to create new examples (e.g., flipping an image).
- Data synthesis generates artificial data mimicking real properties.
- Both aim to increase dataset size, diversity, and quality.
- Synthetic data is increasingly common, especially for post-training.

Synthetic Data Use Cases

- Synthetic data can improve **quantity** by producing large-scale datasets.
- It can expand **coverage** by creating examples across domains and tasks.
- It can enhance **quality** by balancing distributions and reducing bias.
- It also helps mitigate privacy risks by replacing sensitive data.

Instruction Data Synthesis

- Instruction finetuning requires (instruction, response) pairs.
- AI can generate instructions, responses, or both.
- Example: Alpaca used 175 human-written seed pairs and expanded to 52,000 pairs with GPT-3.
- Multi-turn datasets like UltraChat rely heavily on AI synthesis.

Synthetic Data for Preference Training

- AI judges can be used to decide which response is better.
- Biases like first-position bias must be mitigated with careful design.
- Synthetic preference data reduces reliance on costly human feedback.
- Validation of synthetic outputs is crucial to prevent error propagation.

Model Bootstrapping

- Models can generate data to train smaller or newer models.
- Example: Nemotron-4 was finetuned using synthetic data generated by a teacher model.
- Verified synthetic data improves performance, while unverified data may degrade it.
- Bootstrapping requires mechanisms for quality control.

Data Processing

- Processing steps include cleaning, deduplication, normalization, and formatting.
- The order of operations should optimize efficiency.
- Always validate scripts on small samples before scaling to full datasets.
- Keeping original data copies helps protect against script errors.

Inspecting Data

- Inspecting data provides insights into quality beyond automated checks.
- Statistical analysis can reveal distribution patterns, lengths, and token usage.
- Manual inspection often uncovers issues quickly, saving downstream effort.
- Deduplication prevents biases and test contamination.

Deduplication Issues

- Duplicated data skews distributions and inflates model confidence in false correlations.
- It can also cause contamination between training and test sets.
- Even small duplication rates can significantly degrade performance.
- Deduplication improves both efficiency and fairness of training.

Limitations of Synthetic Data

- Synthetic data cannot fully replace human-generated data.
- Mixing human and AI data often yields the best results.
- Poor-quality synthetic data may propagate errors and biases.
- Careful evaluation is required to ensure synthetic contributions are beneficial.

Summary of Dataset Engineering

- Data quality, diversity, and compliance are the foundation of effective AI systems.
- Data-centric AI emphasizes curating and synthesizing datasets over model tweaks.
- Synthetic and augmented data offer scalable solutions but require quality controls.
- Dataset engineering is iterative, requiring constant inspection, evaluation, and improvement.

References

References

- Huyen, Chip. Designing machine learning systems. O'Reilly Media, Inc., 2022