Aurojit Ghosh

9/4/2025

**Week 4 Report**

**Video Summary 7: Live Session 21(Combined)- Data preparation techniques for fine-tuning**

**Key Points**

* **Data preparation techniques**
* **Different types of datasets**
* **Impact on models**
* **Steps in creating SFT Datasets**

Data preparation techniques involve modules, tools, processors, and architecture. At end of day, you know what your data is, what your goals are for fine-tuning. Lot of it aligns with what data you have.

For data preparation for LLMS

1. Start with pretrained model. Pretrained models are large language models which is trained in general data. It specializes in specific tasks through fine tuning.
2. Fine-tuned teaches LLM to understand new patterns in data. It adapts the pretrained model to specific domains or tasks. It allows the model to learn task specific vocabulary, context and platforms. Main purpose of fine tuning is to enhance model’s performance on targeted applications.
3. Fine tuning has data importance that high quality, task-specific data is essential for effective fine tuning. Data should represent intended use cases and cover many edge cases.

For data importance in fine-tuning, high-quality task specific data is essential for effective fine tuning. The data should represent the intended use case and cover edge cases as much as possible.

There are different types of datasets. Instruction-following Datasets, Chat/ Dialogue Datasets, Domain Specific Text Datasets, and Reasoning/Chain of Thought Datasets.

Domain-specific text datasets are collections of text related to a particular field such as legal, medical, and financial. Because adapting LLMs to specialized vocabularies. One example is feeding your own company’s knowledge base to the model.

Reasoning Chain of Thought Datasets are tasks requiring the model to show intermediate reasoning. It improves logical accuracy for tasks like diagnostics or decision support.

Dataset types are 4 things. Instruction, Chat/Dialogue, Domain Specific, and Reasoning.

Instructions are summarization and action items.

Chat/Dialogue is context aware extractions and clarifications during multipurpose chart

Domain specifically is correcting use of internal or industry specific jargon.

The challenge in creating data is noisy or low-quality data, duplicates and redundant examples, and data decontamination.

Noisy or low-quality data has transcripts which often contain filler words or irrelevant chats.

It’s impact is models trained on this will hallucinate and start providing incorrect information.

Solution is preprocessing steps such as cleaning fillers, punctuation, and speaker tags.

Duplicates and Redundant Examples has repeated conversations or save meetings transcribed multiple times.

Its direct impact would be skewing models towards frequent examples which leads to biased outputs

Solution is deduplication strategy with fuzzy matching or embeddings

Data Decontamination is training data which may contain company secret names, emails, phone numbers

Its direct impact is privacy violations, data leaks, regulatory and legal issues

Solution is automated PII redaction tools

Good datasets have accuracy, complexity, and diversity. Models should have free of typos.

Dataset has wide range of topics, styles, and meeting types. Dataset requires complex tasks to challenge model.

To create SFT Datasets, there are 6 steps.

1. Collect Raw Data by gathering text from open datasets, internal documents, and transcripts.
2. Clean the Data by normalizing text
3. Deduplicate and Decontaminate by detecting and removing near duplicates
4. Define tasks and Annotate Examples by writing clear instructions
5. Add Diversity and Complexity by including varied topics and speaking styles
6. Evaluate Diversity by inspecting random samples and calculating basic dataset stats

**Video Summary 8: Live Session 20 (Combined) - Introduction to fine-tuning models**

**Key Points Covered:**

* Fine tuning and whether prompt engineering is necessary
* LLM Training

To use fine tuning, start with prompt engineering if possible. Fine tuning needs development of robust evaluation framework with various metrics. It’s used for non-technical reasons. Fine tuning should be considered if prompt engineering isn’t sufficient. Purpose of fine tuning is specializing pre-trained model for specific tasks, optimizes performance on narrower task specific datasets, and builds upon general language knowledge from pre-training. Enterprises customize their fine-tuned models about 72%, create custom models about 6%, and their only retrieval-augmented generation about 22%. But there are potential downsides of fine tuning. Fine-tuning has been viewed as less necessary as model capabilities improve. Companies instead found out prompt engineering gave better results, often at much lower cost. Moving away from fine tuning helped companies avoid model lock in as fine-tuning models required high costs. LLM Training Lifecycle has three stages pre-training, supervised fine-tuning, and preference alignment. Pretraining uses raw text for next token prediction, resulting in base model. Supervised fine tuning uses question answer pairs to teach models to follow instructions. Preference alignment aligns the model with human preferences.

**Video Summary 9: Lab Session - Intro to GTM with Sridev**

**Key Points Covered in this video:**

* **Go to market strategy**
* **Determine pain points for products**

The Go to Market Strategy is a strategy to bring new product to market and drive demand for it. It aligns sales, marketing, and other stakeholders. You always need go to market strategy when launching a new product in a new market, new product in an existing market, and existing product in a new market. The prerequisites are understanding product market fit and credibility. Product market fit has a journey of formation, validation, and growth. It has 3 steps which is validating, scaling, and establishing.

Product has pain points. Paint points are reason why products exist. Entire messaging and positioning needs to be centered around customer’s pain points. How big is pain point really depends on how effectively does product solve for paint point?

To determine pain points here are steps.

1. Identify the problem and list out all the pain points
2. Define your target group by creating an ideal customer profile which requires primary demographic such as age and industry, geography, preferred channels, size/function, pain points, price points, and decision making factors
3. Determine competitor analysis such as who are your competitors, who are they targeting and what geographies, what is USP of product of product compared to competitors, and is market oversaturated
4. Focus on key messaging using value matrix such as pain points, product value, and message.
5. Map out buyer journey
6. Determine your best marketing channels such as social media or podcast
7. Determine inbound/outbound sales funnels and self serve models
8. Determine metrics to track for success
9. Break it down into team functions and create project plan and align team
10. Supervise, Visit, and Iterate