جامعة الحسين التقنية
Al Hussein Technical University

Center of Professional Development and Community Outreach

Generative AI Training Program - Professional Development (Reskilling) Training Program

Generative AI - Data Preparation & Training Methods

جامعة الحسين التقنية
Al Hussein Technical University

Module 4:
Data Preparation & Training Methods

Dr. Abeer Al-Hyari
Abeer.alhyariups@htu.edu.jo

Module-4

Module	Module Name and Focus	Key Topics to Cover	PLOs Addressed
Module 4	**Data Preparation & Training Methods**	1. Data Collection & Cleaning for GenAI 2. Tokenization & Vocabulary Management 3. Fine-Tuning Strategies (Full vs. PEFT) 4. Reinforcement Learning from Human Feedback (RLHF) & Instruct Tuning 5. Data Governance & Licensing Principles	Data Prep (Primary), Model Development

Objectives

جامعة الحسين التقنية
Al Hussein Technical University

1. **Collect, clean, and prepare datasets** for GenAI applications, addressing issues such as noise, data drift, and bias.
2. **Implement tokenization and vocabulary management techniques**, including training custom tokenizers for domain-specific data.
3. **Apply and compare fine-tuning strategies**, including full fine-tuning and PEFT methods such as LoRA and QLoRA.
4. **Explain and perform foundational alignment techniques**, including instruction tuning and basic RLHF components.
5. **Follow data governance, licensing, and ethical guidelines** to ensure compliant and Responsible dataset usage.

Topics Covered

جامعة الحسين التقنية
Al Hussein Technical University

1. Data Collection & Cleaning for GenAI
2. Tokenization & Vocabulary Management
3. Fine-Tuning Strategies (Full vs. PEFT)
4. Reinforcement Learning from Human Feedback (RLHF) & Instruct Tuning
5. Data Governance & Licensing Principles

جامعة الحسين التقنية
Al Hussein Technical University

Data Collection & Cleaning for GenAI

Sources of Training Data

جامعة الحسين التقنية
Al Hussein Technical University

- Public datasets (Common Crawl, LAION, The Pile, Wikipedia)
- Curated datasets (books, academic corpora)
- Filtered code repositories (GitHub, StackOverflow dumps)

Sources of Training Data

جامعة الحسين التقنية
Al Hussein Technical University

1. Public Datasets

- **Common Crawl** – large-scale web crawl used as a core component for many LLMs.
- **LAION Image–Text Datasets** – large-scale multimodal datasets for training vision–language models.
- **The Pile** – a curated, diverse 825GB corpus designed for training language models.
- **Wikipedia** – high-quality, human-edited encyclopedic text.

Sources of Training Data

جامعة الحسين التقنية
Al Hussein Technical University

2. Curated / High-Quality Datasets

- **Digitized books** (fiction, non-fiction, textbooks, technical manuals).
- **Academic corpora** (research papers, scientific datasets, arXiv, PubMed).
- **Domain-specific corpora** (legal texts, medical texts, policy documents).

3. Filtered Code Repositories

- **GitHub repositories** (cleaned, de-duplicated, license-aware).
- **StackOverflow dumps** (Q&A paired with accepted solutions).
- **Open-source code datasets** (e.g., CodeSearchNet, The Stack).

Data Quality

جامعة الحسين التقنية
Al Hussein Technical University

- Chinchilla Scaling Laws:
  Quality > *Quantity*, optimal compute–data balance
- Impact of noisy vs. clean data on downstream performance

Chinchilla Scaling Laws

جامعة الحسين التقنية
Al Hussein Technical University

Chinchilla Scaling Laws (DeepMind, 2022)

- Proposed that **LLM performance is maximized** when **model size** and **training dataset size** are **balanced**, rather than simply increasing parameters.
- Key insight: **Undertrained large models** (too many parameters, too little data) waste compute and perform poorly.
- Optimal ratio: roughly **20 tokens per parameter** during training.

Why Quality > Quantity

جامعة الحسين التقنية
Al Hussein Technical University

- Scaling Laws show that adding **more low-quality data** is not as effective as training on **clean, diverse, and correctly labeled data**.
- High-quality datasets significantly improve:
  - Generalization
  - Reasoning ability
  - In-context learning
  - Safety and factual accuracy

Impact of Noisy vs. Clean Data

جامعة الحسين التقنية
Al Hussein Technical University

Data Type	Effect on Model	Downstream Impact
Clean, curated, diverse data	Efficient learning, stable convergence	Better accuracy, less hallucination, improved reasoning
Semi-noisy data	Model learns mixed signals	Reduced consistency, moderate error rate
Highly noisy, duplicated, or biased data	Training instability, wasted compute	Hallucinations, poor generalization, harmful biases

Dataset Bias & Data Drift

جامعة الحسين التقنية
Al Hussein Technical University

- Types of bias: sampling bias, representation bias, annotation bias
- Detecting bias in datasets
- Temporal data drift + domain drift
- Mitigation techniques (balancing, filtering, augmentation)

Types of Dataset Bias

جامعة الحسين التقنية
Al Hussein Technical University

- **Sampling Bias**
  When the collected data does not reflect the full population.
  Example: Over-collecting examples from one demographic, device type, or class.
- **Representation Bias**
  When certain groups, classes, or scenarios are under-represented even if they were sampled.
  Example: Rare events or minority classes appear too rarely for the model to learn from.
- **Annotation / Labeling Bias**
  Inconsistent or subjective labeling caused by human annotators or unclear guidelines.
  Example: Different annotators labeling borderline images inconsistently due to ambiguity.

Detecting Bias in Datasets

جامعة الحسين التقنية
Al Hussein Technical University

- **Descriptive Statistics & Visual Checks**
  Class imbalance charts, histograms, correlation matrices.
- **Feature Distribution Comparison**
  Comparing sub-groups (e.g., gender, regions) to identify systematic differences.
- **Label Agreement Metrics**
  Inter-annotator agreement (Cohen's Kappa, Fleiss' Kappa).
- **Performance Fairness Metrics**
  Accuracy per subgroup, false-negative/false-positive disparity.
- **Outlier and Coverage Analysis**
  Checking if rare or edge cases are represented adequately.

Data Drift

جامعة الحسين التقنية
Al Hussein Technical University

A. Temporal Data Drift

- Changes in the statistical distribution of input data *over time*.
  Example:
  A fraud detection model trained in 2022 underperforms in 2025 because fraud patterns evolved.

B. Domain / Covariate Drift

- Shifts in data distribution due to changes in environment, sensors, user behavior, or context.

Example:
A model trained on clinical data from one hospital performs poorly in another hospital.

Mitigation Techniques

- **Balancing**
  - Oversampling minority classes
  - Undersampling majority classes
  - Synthetic generation (SMOTE)
- **Filtering & Cleaning**
  - Removing noisy labels
  - Excluding corrupted or extreme outlier samples
- **Data Augmentation**
  - Text: paraphrasing, back-translation
  - Images: rotation, color jitter, synthetic examples
  - Tabular: noise injection, perturbation

Mitigation Techniques

جامعة الحسين التقنية
Al Hussein Technical University

- **Reweighting Samples**
  Assigning higher weights to underrepresented classes or time periods.
- **Domain Adaptation / Model Retraining**
  - Regular retraining with recent data
  - Fine-tuning with target-domain samples
  - Transfer learning techniques

جامعة الحسين التقنية
Al Hussein Technical University

Tokenization & Vocabulary Management

Why LLMs need tokenization

جامعة الحسين التقنية
Al Hussein Technical University

1. LLMs operate on numbers, not characters

Neural networks take **numeric vectors** as input.
Tokenization converts words, subwords, or characters into IDs → which are then mapped to embeddings.

Without tokenization, the model would have no consistent numerical representation of language.

Why LLMs need tokenization

جامعة الحسين التقنية
Al Hussein Technical University

2. Tokens create a standardized vocabulary

If a model tried to learn using raw characters (letters like “a”, “b”, “c”):

- sequences would be longer
- patterns would be harder to learn
- training would be slower and less accurate

Tokenization creates a vocabulary of useful pieces:

- words (“apple”)
- subwords (“read”, “ing”, “##ly”)
- symbols
- punctuation

This makes learning easier.

Why LLMs need tokenization

جامعة الحسين التقنية
Al Hussein Technical University

3. Tokenization makes language predictable

Natural text is extremely diverse.
Tokenizers (like BPE, WordPiece, SentencePiece) compress this diversity so the model can learn patterns consistently.

Example:
“unbelievable” → “un”, “believe”, “able”
This reduces sparsity and improves generalization.

Why LLMs need tokenization

جامعة الحسين التقنية
Al Hussein Technical University

4. It helps with rare words

Without tokenization:

- Every new word would be unknown.
  Tokenization breaks words into known pieces, so the model can handle:
- names
- slang
- typos
- new words
- multilingual text

Example: “hyariTech” → “hy”, “ari”, “Tech”

Why LLMs need tokenization

جامعة الحسين التقنية
Al Hussein Technical University

5. Efficiency

Tokenization reduces:

- training time
- memory usage
- model size

Instead of processing text character-by-character, a model handles **fewer, more meaningful units**.

Why LLMs need tokenization

جامعة الحسين التقنية
Al Hussein Technical University

6. Tokenization enables mathematical structure in Transformers

Transformers rely on:

- **attention** over discrete tokens
- **positional encodings** for token order

These mechanisms require the text to be split into tokens first.

Overview of tokenizers

جامعة الحسين التقنية
Al Hussein Technical University

1. BPE (Byte Pair Encoding)
2. WordPiece
3. SentencePiece
4. TikToken (OpenAI)

1. BPE — Byte Pair Encoding

جامعة الحسين التقنية
Al Hussein Technical University

Used in: GPT-2, GPT-NEO, BLOOM (variant), some early LLMs

Paper: Sennrich et al., 2015

Key idea

- Starts from characters.
- Iteratively merges the most frequent *pair* of tokens.
- Builds a vocabulary of subword units based on frequency.
- Frequent words become single tokens; rare words remain broken into pieces.

1. BPE — Byte Pair Encoding

جامعة الحسين التقنية
Al Hussein Technical University

Strengths

- Simple and efficient.
- Good coverage for unknown and rare words.
- Works well across languages.

Weaknesses

- Cannot handle out-of-vocabulary merges at inference time.
- Sensitive to text preprocessing (whitespace, casing).

2. WordPiece

Used in: BERT, DistilBERT, ALBERT, Electra
Developed at: Google (Schuster & Nakajima)

Key idea

- Similar to BPE but **selects merges that maximize likelihood** (not just most frequent pairs).
- Uses a leading marker ## to denote subwords (“##ing”, “##tion”).

2. WordPiece

جامعة الحسين التقنية
Al Hussein Technical University

Strengths

- More stable subword segmentation than BPE.
- Reduces fragmented tokens for common words.

Weaknesses

- Requires pre-tokenized (whitespace-separated) input.
- Less flexible with scripts like Chinese/Japanese since it relies on word boundaries.

3. SentencePiece

جامعة الحسين التقنية
Al Hussein Technical University

Used in: T5, ALBERT (variant), LLaMA (SPM-BPE), XLNet
Created by: Google

Key idea

- A *drop-in tokenizer* that:
  - Does **not require pre-tokenized text** (treats input as a raw stream, including spaces)
  - Supports two main algorithms:
    - **Unigram Language Model** (primary)
    - **BPE** (optional)

3. SentencePiece

جامعة الحسين التقنية
Al Hussein Technical University

Strengths

- Language-agnostic.
- Excellent for multilingual training.
- Spaces are part of vocabulary (using "▁" as a space symbol).
- Robust handling of Unicode.

Weaknesses

- Unigram LM is slower to train than BPE.
- Resulting subwords are less intuitive for humans.

4. TikToken (OpenAI)

جامعة الحسين التقنية
Al Hussein Technical University

Used in: GPT-3.5, GPT-4, GPT-4o, GPT-5
Purpose: High-performance tokenizer optimized for OpenAI's LLM architectures

Key idea

- A highly optimized Rust implementation.
- Uses a variant of **Byte-level BPE** (byte fallback).
- Fully deterministic and very fast.

4. TikToken (OpenAI)

جامعة الحسين التقنية
Al Hussein Technical University

Features

- **Byte-level encoding:**

Every sequence of bytes can be tokenized (no OOV issues).

- **Model-specific encoding:**

Different GPT models use slightly different vocabularies (e.g., cl100k_base).

- **Efficient token counting:**

Essential for prompt budgeting.

Strengths

- Very fast and memory-efficient.
- Great Unicode handling.
- Standard for OpenAI models.

When to Use Each

جامعة الحسين التقنية
Al Hussein Technical University

- **BPE:** Simple, effective for general LLMs; classic approach.
- **WordPiece:** Best for encoder-only models (BERT) focused on accuracy and stability.
- **SentencePiece:** Best for multilingual, cross-domain datasets; most flexible.
- **TikToken:** Best for OpenAI models; optimized for performance and byte-level reliability.

Comparison

جامعة الحسين التقنية
Al Hussein Technical University

Feature	BPE	WordPiece	SentencePiece	TikToken
Vocabulary rule	Merge most frequent pairs	Maximize likelihood	Unigram LM or BPE	Byte-level BPE
Pre-tokenization	Required	Required	Not needed	Not needed
Space handling	External	External	"▁" symbol	Bytes
Unicode robustness	Medium	Medium	High	Very high
Used in	GPT-2, BLOOM	BERT family	LLaMA, T5	GPT-3.5–GPT-5

جامعة الحسين التقنية
Al Hussein Technical University

- Vocabulary construction
- Merging rules
- Token efficiency

Vocabulary Construction

جامعة الحسين التقنية
Al Hussein Technical University

1. Vocabulary Construction

- Vocabulary construction is the process by which a tokenizer determines the set of tokens (subwords, characters, or units) that a language model will use. The goal is to build a vocabulary that is:
  - **Compact** (not too large)
  - **Expressive** (covers linguistic patterns well)
  - **Efficient** (minimizes the number of tokens per sentence)

Vocabulary Construction

جامعة الحسين التقنية
Al Hussein Technical University

Steps in Vocabulary Construction

1. **Collect Corpus Statistics**
   A large representative text corpus is scanned to determine the frequency of characters, subwords, and word fragments.
2. **Initialize Base Units**
   Often byte-level units (0–255) or Unicode characters.
3. **Iteratively Add Useful Tokens**
   Frequently occurring subword patterns, prefixes, suffixes, or entire common words.
4. **Stop at Target Vocabulary Size**
   Typical vocab sizes: 30k–100k (classic LLMs), 100k–200k (modern multilingual models).

Why Vocabulary Construction Matters

- Affects **model memory footprint** (embedding matrix size).
- Affects **cost** (fewer tokens = cheaper inference).
- Affects **robustness** (handling rare words, misspellings, multilingual text).

2. Merging Rules

Merging rules determine **how smaller units combine into larger tokens** during tokenization.

Different tokenization algorithms implement this differently:

2.1 Byte Pair Encoding (BPE)

- Start with single characters.
- Count the most frequent adjacent pair (e.g., “t” + “h”).
- Merge them into a new token (“th”).
- Repeat until you reach the target vocab size.

Merging Rules:

- A ranked list of merges is stored.
- Tokenization applies merges greedily from highest rank → lowest.
- Deterministic and fast.

2. Merging Rules

جامعة الحسين التقنية
Al Hussein Technical University

2.2 WordPiece (Used in BERT)

- Similar to BPE, but merges are chosen by maximizing **likelihood**, not raw frequency.
- Encourages subwords that contribute most to model performance.

Merging Rules:

- Scoring-based selection.
- Splits words using “##” prefix for continuation fragments (e.g., "play", "##ing").

2. Merging Rules

جامعة الحسين التقنية
Al Hussein Technical University

2.3 Unigram Model (SentencePiece)

- Start with a *very large* vocabulary (e.g., 1M tokens).
- Compute the loss when each token is removed.
- Iteratively prune least useful tokens.

Merging Rules:

- Not merging but **pruning**.
- Tokenization uses **probabilistic Viterbi search**.

2. Merging Rules

جامعة الحسين التقنية
Al Hussein Technical University

2.3 Unigram Model (SentencePiece)

- Start with a *very large* vocabulary (e.g., 1M tokens).
- Compute the loss when each token is removed.
- Iteratively prune least useful tokens.

Merging Rules:

- Not merging but **pruning**.
- Tokenization uses **probabilistic Viterbi search**.

2.4 TikToken (OpenAI)

- A highly optimized BPE variant.
- Works at the byte level for multilingual generality.
- Designed for speed + low fragmentation.

3. Token Efficiency

Token efficiency measures how effectively a tokenizer represents text using the **fewest possible tokens** while preserving meaning.

Key Measures of Token Efficiency

1. **Tokens per Word**
   Fewer tokens per word → more efficient.
2. **Tokens per Sentence**
   Directly affects inference speed and cost.
3. **Compression Ratio**
   Higher compression = more efficient vocabulary.
4. **Fragmentation Behavior**
   Efficient tokenizers avoid over-splitting words, especially:
   - Morphologically rich languages
   - Rare or long words
   - Technical terminology

3. Token Efficiency

جامعة الحسين التقنية
Al Hussein Technical University

3.1 What Improves Token Efficiency?

✓ Well-designed vocabulary

- Includes common words + common subwords.
- Balances between full words and sub-tokens.

✓ Effective merging rules

- High-frequency patterns get single tokens.
- Rare patterns can be decomposed safely.

✓ Byte-level fallback

- Ensures zero Out-of-Vocabulary (OOV) problems.

✓ Multilingual-aware design

- Avoids vocabulary explosion while covering many languages.

4. Why These Concepts Matter for LLMs

جامعة الحسين التقنية
Al Hussein Technical University

- Shorter sequences → faster training & inference
- Smaller vocabularies → smaller model size
- Better segmentation → better understanding & reasoning
- Fewer tokens → lower cost per prompt

Modern LLMs (GPT-4/5 class) heavily optimize tokenization because **tokenization alone affects performance, cost, and model behavior**.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

- Training cost
- Sequence length
- Model performance
- Multilingual capability

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

1. Training Cost

Tokenizer choice directly affects training compute in several ways:

a. Number of tokens per sample

- Different tokenizers segment the same text into different numbers of tokens.
- **More tokens → longer sequences → higher GPU hours → higher cost.**
- Example:
  - BPE: “internationalization” → *["international", "ization"]*
  - Character-level: 20+ tokens → far higher compute.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

1. Training Cost

b. Vocabulary size

- Larger vocabulary → larger embedding matrices → more parameters to train.
- Embedding layer and output softmax scale with vocabulary size: Parameters=$V \times d_{model}$
- Efficient tokenization reduces vocabulary while maintaining expressiveness.

c. Padding inefficiency

- Poor segmentation leads to uneven sequence lengths → more padding → wasted compute.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

2. Sequence Length

Tokenizer efficiency determines how much information fits into a model's context window.

a. Compression efficiency

- Subword tokenizers (BPE, WordPiece, SentencePiece) compress text better than:
  - character-level tokenization
  - word-only tokenization (suffers with rare words)

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

2. Sequence Length

b. Impact on effective context

- **Fewer tokens per sentence** → model can process a longer span of meaning.
- Better tokenizers increase the *semantic density* of a context window.

c. Long-context models

- With a fixed maximum length (e.g., 8k, 32k tokens), tokenizer efficiency determines:
  - how many paragraphs
  - how many code lines
  - how many multilingual glyphs can fit inside the context.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

3. Model Performance

Tokenizer affects multiple downstream performance metrics:

a. Handling rare or out-of-vocabulary tokens

- Subword-based tokenizers ensure coverage for:
  - rare words
  - morphological variants
  - new terminology

This improves downstream accuracy and reduces <unk> tokens.

b. Word-level semantics

- Good token boundaries → clearer morphemes → easier learning for the model.
- Worse segmentation → model must “work harder” to understand meaning.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

3. Model Performance

c. Domain specificity

- Code-specific tokenizers (e.g., StarCoder, DeepSeek-Coder) perform better on:
  - indentation
  - symbols
  - language keywords
- Biomedical tokenizers perform better on:
  - long chemical names
  - gene identifiers

Mismatch of tokenizer & domain → measurable performance drops.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

4. Multilingual Capability

Tokenizer choice is **critical** for multilingual models.

a. Shared subword vocabulary

- A single vocabulary may work well for Latin languages but fail for:
  - Chinese
  - Japanese
  - Arabic
  - Indic scripts
    Because words are not separated by whitespace.

b. Character coverage

- SentencePiece/BPE on UTF-8 handles multilingual text better than word-based tokenizers.

Impact of tokenizer choice on:

جامعة الحسين التقنية
Al Hussein Technical University

4. Multilingual Capability

d. Training efficiency impact

- Each language may require:
  - special handling (e.g., byte-level BPE for Chinese)
  - shared tokens for common patterns
  - decomposed characters (Unicode normalization)

Multilingual LLMs often use:

- **SentencePiece (Unigram)**
- **Byte-level BPE (ex: GPT-2, Llama-3)**
  because these are more script-agnostic.

جامعة الحسين التقنية
Al Hussein Technical University

Fine-Tuning Strategies

Full Fine-Tuning

Traditionally, fine-tuning involves updating all the parameters (weights) of the model based on the new data. This works well but the problem is that modern Large Language Models (LLMs) are huge and have billions of parameters.

- Updating **all model weights**
- Requires large GPU resources
- Maximum flexibility + highest cost

Parameter-Efficient Fine-Tuning (PEFT)

جامعة الحسين التقنية
Al Hussein Technical University

Parameter-Efficient Fine-Tuning (PEFT) is a method used to fine-tune Large Language Models (LLMs) by updating a small subset of the model's parameter while keeping the majority of the pre-trained weights frozen.

This makes fine-tuning much more efficient in terms of:

- **Computational cost**: Less computing power is required.
- **Storage**: Task-specific modules take up minimal storage.
- **Training time**: It’s faster because fewer parameters are being updated.

Parameter-Efficient Fine-Tuning (PEFT)

جامعة الحسين التقنية
Al Hussein Technical University

- **Concept**: freeze the base model, update only small trainable modules
- **Advantages**:
  - Low compute
  - Fast training
  - Modular & version-controlled

paper "[2303.15647] Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning"

PEFT Methods

جامعة الحسين التقنية
Al Hussein Technical University

- **LoRA** (Low-Rank Adaptation)
- **QLoRA** (quantized LoRA)
- Prefix-Tuning
- Adapter Modules\Layers

LoRA (Low-Rank Adaptation)

جامعة الحسين التقنية
Al Hussein Technical University

- **LoRA (Low-Rank Adaptation)** reduces the number of trainable parameters by decomposing weight updates into low-rank matrices. Instead of updating the entire weight matrix, it modifies only a small, low-rank component which approximates the changes needed for fine-tuning.
- It achieves results close to full fine-tuning but with far fewer parameters.
- It has been successfully applied to LLMs like GPT-3 and T5 making it a popular choice for parameter-efficient fine-tuning at scale.

DoRA (Weight-Decomposed Low-Rank Adaptation)

جامعة الحسين التقنية
Al Hussein Technical University

- **DoRA** builds upon the concept of LoRA but introduces a novel weight-decomposed approach to further enhance efficiency. In DoRA, the weight matrix is decomposed into two components i.e a low-rank update and a scaling factor.
- It also maintains the low computational cost of LoRA while potentially improving performance.
- It is useful in scenarios where fine-tuning must be both efficient and robust such as in cross-domain applications or when adapting models to new languages.

Prefix Tuning

جامعة الحسين التقنية
Al Hussein Technical University

- Prefix tuning works by adding a small set of learnable "prefix" tokens to the model’s input at every layer. These prefix tokens act as task-specific prompts that helps the model's behavior without changing its original parameters.
- It allows the model to retain its general knowledge while adapting to specific tasks through the learned prefixes.
- It is used for tasks like text generation where controlling the output style or content is important.

Adapter Modules\Layers

جامعة الحسين التقنية
Al Hussein Technical University

- Adapters Modules are small, trainable modules inserted between the layers of a pre-trained model. During fine-tuning, only the adapter modules are updated while the original model weights remain fixed. Once fine-tuned, it can be easily added or removed allowing for modular customization of the model.
- It allow for efficient multi-task learning where different adapters can be used for different tasks while sharing the same base model.
- For example: The Hugging Face AdapterHub provides an extensive library of pre-trained adapters for various NLP tasks.

Instruction Tuning

جامعة الحسين التقنية
Al Hussein Technical University

- **Purpose:** Adapt a pre-trained LLM to follow explicit human instructions using instruction–prompt–response pairs.
- **Training Objective:** Primarily supervised next-token prediction; may include auxiliary losses to preserve general language modeling capabilities.
- **Dataset Structure:** Each sample includes a **task description**, **input context**, and **target response**.
- **Dataset Expansion:** Automated methods (e.g., self-instruct generation, synthetic tasks, difficulty scaling) increase task diversity and robustness.
- **Model Adjustments:** Lightweight architectural additions—special tokens, attention-mask tuning, adapter layers (e.g., LoRA)—enhance instruction adherence and improve training efficiency.

جامعة الحسين التقنية
Al Hussein Technical University

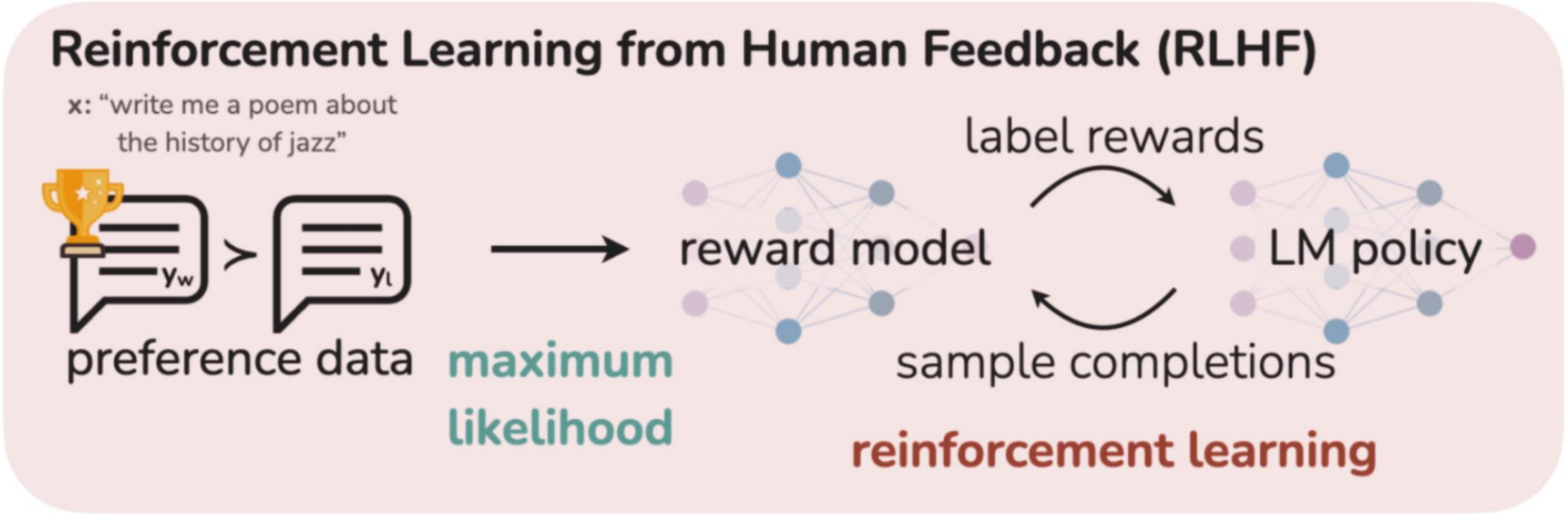
Reinforcement Learning from Human Feedback (RLHF) & Instruct Tuning

RLHF

جامعة الحسين التقنية
Al Hussein Technical University

- **Reinforcement Learning from Human Feedback (RLHF)** is a training method used to align large language models with human preferences. Instead of relying only on traditional supervised datasets, RLHF incorporates judgments from human annotators to guide the model's behavior.
- **Goal:** Align the model's behavior with human values, reduce harmful outputs, and improve helpfulness and reasoning quality.

RLHF



https://wandb.ai/byyoung3/huggingface/reports/What-is-RLHF-Reinforcement-learning-from-human-feedback-for-AI-alignment--VmlldzoxMzczMjEzMQ#:~:text=With%20RLHF%2C%20the%20model%20is,and%20considerate%20of%20user%20intent.

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

Alignment is needed to ensure that AI systems behave in ways that are safe, predictable, and aligned with human values and intentions. As models grow more capable, the consequences of misbehavior or misunderstanding become more serious. Here are the key reasons:

1. AI models don’t inherently understand human values

LLMs learn patterns from massive datasets—not ethics, safety, or intentions. Without alignment, a model may produce harmful, biased, or misleading outputs simply because those patterns existed in its training data.

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

2. Preventing harmful or unsafe behavior

Alignment reduces risks such as:

- generating toxic or offensive language
- giving dangerous instructions (e.g., harmful chemical recipes)
- providing medical/financial advice irresponsibly
- hallucinating incorrect or fabricated facts

Aligned models are trained to decline unsafe tasks and guide users responsibly.

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

3. Ensuring models follow human instructions correctly

- Without alignment, a model might:
- ignore user constraints
- misunderstand task goals
- produce irrelevant or verbose responses

Alignment improves instruction-following, making models more useful and reliable.

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

4. Reducing bias and promoting fairness

Training data contains biases from the real world.
Alignment helps identify and reduce:

- gender bias
- racial bias
- cultural or geographic bias
- stereotypes in language generation

This supports fairness and accessibility.

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

5. Making AI systems trustworthy for real-world deployment

In applications like education, healthcare, customer service, and law, AI must behave predictably and ethically.
Alignment improves:

- reliability
- transparency
- user trust
- suitability for professional environments

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

6. Balancing general creativity with boundaries

LLMs can produce creative and unrestricted outputs.
Alignment ensures this creativity remains:

- safe
- appropriate
- respectful
- aligned with user intent and context

Why alignment is needed

جامعة الحسين التقنية
Al Hussein Technical University

7. Preparing AI for long-term societal impact

As models become more autonomous and integrated into critical services, alignment ensures:

- responsible AI progress
- reduction of unintended consequences
- preservation of human oversight
- ethical growth of AI ecosystems

RLHF Pipeline

جامعة الحسين التقنية
Al Hussein Technical University

1. **Pretraining the language model**
2. **Supervised Fine-Tuning (SFT)** on instruction datasets
3. **Reward Modeling** from human preferences
4. **Policy Optimization** (PPO / DPO)

RLHF Pipeline

جامعة الحسين التقنية
Al Hussein Technical University

- **Pretraining the language model:**
  the language model is pretrained on a massive and diverse collection of human-written text
- **Supervised Fine-Tuning (SFT):**
  The model is trained on curated prompt–response pairs to learn an initial pattern of desirable outputs.
- **Reward Modeling:**
  Human evaluators rank multiple model-generated responses.
  These rankings train a *reward model* that can automatically score new outputs based on how "human-preferred" they are.
- **Reinforcement Learning (e.g., PPO):**
  The model is further optimized using reinforcement learning algorithms so that it produces outputs that maximize the reward model's score.
  This step teaches the model to behave more safely, helpfully, and consistently with human expectations.

Pretraining The LLM

جامعة الحسين التقنية
Al Hussein Technical University

- **Pretraining phase:** Model is trained on a massive, diverse corpus of human-written text.
- **Learning objective:** Predict the next word in a sentence.
- **Skills learned:** Grammar, language structure, basic reasoning patterns, and broad world knowledge.
- **Outcome:** The model gains strong general language abilities.
- **Limitation:** Pretraining does **not** teach human expectations, preferences, or values.

Supervised Fine-Tuning (SFT)

جامعة الحسين التقنية
Al Hussein Technical University

- Start with a pretrained language model.
- Train it on curated *instruction–response* pairs written by humans.
- Purpose: Give the model an initial sense of how helpful, safe answers look.

Reward Model Training (RM)

جامعة الحسين التقنية
Al Hussein Technical University

- The model generates **multiple responses** to the same prompt.
- Human evaluators **rank the responses** from best to worst.
- These rankings reflect real human preferences.
- Use the ranked comparisons to train a **reward model**.
- The RM learns to assign a numerical score to outputs:
  - Higher score = more aligned with human preference
  - Lower score = undesirable / unhelpful behavior

Reinforcement Learning Optimization (PPO)

جامعة الحسين التقنية
Al Hussein Technical University

- Use the reward model as the “environment” in RL.
- Apply a reinforcement learning algorithm (commonly **Proximal Policy Optimization – PPO**).
- Goal: Adjust the language model to **maximize the reward** provided by the reward model.

Safety Checks & Iterative Refinement

جامعة الحسين التقنية
Al Hussein Technical University

- Evaluate the updated model for:
  - safety
  - coherence
  - harmful or biased behavior
  - instruction-following quality
- Further human feedback may be collected to improve performance.

Instruct Tuning vs. Chat Tuning vs. Safety Fine-Tuning

جامعة الحسين التقنية
Al Hussein Technical University

1. Instruct Tuning (Instruction Fine-Tuning)

- Uses supervised datasets of **instruction → ideal response** pairs.
- Trains the model to **follow natural language instructions** directly.
- Focuses on general task-solving ability across many domains.
- Examples: “Explain…”, “Summarize…”, “Translate…”, “Solve…”.
- Primary goal → **Improve instruction-following and usefulness.**

Instruct Tuning vs. Chat Tuning vs. Safety Fine-Tuning

جامعة الحسين التقنية
Al Hussein Technical University

2. Chat Tuning (Conversational Fine-Tuning)

- Uses multi-turn **dialogue datasets** with human-like conversational patterns.
- Teaches the model how to behave in a chat setting (role-taking, turn-taking).
- Emphasizes:
  - polite and natural interaction
  - maintaining context across turns
  - clarifying user intent
  - adopting roles (teacher, assistant, coder…)
- Primary goal → **Create a smooth, helpful conversational assistant.**

Instruct Tuning vs. Chat Tuning vs. Safety Fine-Tuning

جامعة الحسين التقنية
Al Hussein Technical University

3. Safety Fine-Tuning (Safety Alignment / Constitutional Fine-Tuning)

- Focuses on teaching the model **safe, ethical, and policy-aligned behavior**.
- Often includes:
  - refusal training (e.g., declining harmful requests)
  - de-biasing
  - reducing toxic, misleading, or unsafe outputs
  - enforcing content policies or safety rules
- Uses human feedback, rules, or "constitutions" to guide safe responses.
- Primary goal → **Prevent harmful, biased, or unsafe behaviors.**

"Helpful, Harmless, Honest" framework

جامعة الحسين التقنية
Al Hussein Technical University

- The **HHH framework** is a foundational guideline for aligning large language models to behave in ways that serve users effectively while minimizing risks. It defines three essential qualities that an aligned AI system should maintain:

"Helpful, Harmless, Honest" framework

جامعة الحسين التقنية
Al Hussein Technical University

- The **HHH framework** is a foundational guideline for aligning large language models to behave in ways that serve users effectively while minimizing risks. It defines three essential qualities that an aligned AI system should maintain:

"Helpful, Harmless, Honest" framework

جامعة الحسين التقنية
Al Hussein Technical University

1. Helpful

A helpful model provides responses that are:

- **Relevant** to the user's request
- **Accurate and informative**
- **Clear and easy to understand**
- **Effective at completing the intended task**
- **Adaptive** to the user's level and context

Goal: Maximize usefulness and task completion.

"Helpful, Harmless, Honest" framework

جامعة الحسين التقنية
Al Hussein Technical University

2. Harmless

A harmless model avoids causing harm by:

- Declining **dangerous**, **illegal**, or **unethical** requests
- Avoiding toxic, biased, or discriminatory language
- Preventing misinformation or harmful advice
- Respecting sensitive contexts (health, finance, legal, minors)

Goal: Protect users, society, and the system from harmful outcomes.

"Helpful, Harmless, Honest" framework

جامعة الحسين التقنية
Al Hussein Technical University

3. Honest

An honest model:

- Avoids making up facts ("no hallucinations")
- Admits uncertainty when information is unknown
- Clearly differentiates between **facts, opinions, and assumptions**
- Avoids misleading content or fabricated instructions
- Provides transparent reasoning when appropriate

Goal: Build trust through truthful and reliable communication.

Why HHH Matters

جامعة الحسين التقنية
Al Hussein Technical University

This framework is used in RLHF, alignment training, and safety fine-tuning because it ensures that models:

- Provide **high-quality help**
- Avoid **harmful behaviors**
- Stay **truthful and transparent**

Together, HHH supports the development of AI systems that are safe, trustworthy, and aligned with human expectations.

Relationship to Responsible AI

جامعة الحسين التقنية
Al Hussein Technical University

How They Fit Together (Simple Pipeline)

1. **Responsible AI**
   → Defines ethical principles and system-level requirements.
2. **HHH Framework**
   → Specifies desired model behaviors aligned with those principles.
3. **RLHF**
   → Implements those behaviors through data, reward modeling, and training.

Relationship to Responsible AI

جامعة الحسين التقنية
Al Hussein Technical University

- **Responsible AI** = the ethical and governance framework
- **HHH** = the behavioral targets for aligned model responses
- **RLHF** = the technical method used to make models behave according to HHH and Responsible AI principles

They form a **top-down alignment hierarchy**:

Responsible AI → HHH Principles → RLHF Training → Aligned AI Behavior

جامعة الحسين التقنية
Al Hussein Technical University

Data Governance & Licensing Principles

Legal Foundations for GenAI Datasets

جامعة الحسين التقنية
Al Hussein Technical University

- Copyright
- Fair Use
- Creative Commons licenses
- Terms of Service compliance

Legal Foundations for GenAI Datasets

جامعة الحسين التقنية
Al Hussein Technical University

1. Copyright

- Copyright protects **original creative works**, including text, images, code, and audio.
- Using copyrighted material for training GenAI models requires understanding:
  - Who owns the rights
  - Whether the dataset includes copyrighted content
  - Whether permission or licensing is required
- Unauthorized use may result in infringement unless an exception (e.g., fair use) applies.

Legal Foundations for GenAI Datasets

جامعة الحسين التقنية
Al Hussein Technical University

2. Fair Use (or Similar Exceptions)

- A legal doctrine (mainly in the U.S.) that allows certain uses of copyrighted material without permission.
- Courts consider four factors:
  1. **Purpose** (transformative use is favored)
  2. **Nature** of the work
  3. **Amount** used
  4. **Effect** on the market
- AI training may be considered transformative, but legal interpretations are evolving and vary across jurisdictions.

Legal Foundations for GenAI Datasets

جامعة الحسين التقنية
Al Hussein Technical University

3. Creative Commons (CC) Licenses

- A set of standardized licenses that enable creators to share their work with specific permissions.
- CC licenses specify:
  - Attribution requirement (BY)
  - Non-commercial use (NC)
  - No derivatives (ND)
  - Share-alike (SA)
- For GenAI datasets:
  - **CC BY, CC BY-SA** → generally compatible with model training
  - **CC BY-NC, CC BY-ND** → restrictions may prevent training or downstream use
- Always verify the exact license before including data in a training corpus.

Legal Foundations for GenAI Datasets

جامعة الحسين التقنية
Al Hussein Technical University

4. Terms of Service (ToS) Compliance

- Platforms (e.g., websites, APIs, digital libraries) have ToS that regulate:
  - Data scraping
  - Automated collection
  - Reuse and redistribution
  - Commercial use
- Violating ToS can lead to:
  - Contractual breach
  - Service account termination
  - Legal action
- Responsible dataset creation requires respecting the ToS of data sources.

Dataset Provenance Tracking

جامعة الحسين التقنية
Al Hussein Technical University

Dataset provenance tracking refers to the systematic process of recording, documenting, and managing the **origin, history, and transformations** of all data used in developing GenAI systems. It ensures that every component of the dataset is **traceable, lawful, ethical, and reproducible**.

Dataset Provenance Tracking

جامعة الحسين التقنية
Al Hussein Technical University

Why It Matters

- **Legal compliance:** Ensures data sources respect copyright, licenses, and Terms of Service.
- **Ethical safety:** Prevents the use of harmful, biased, or low-quality data.
- **Transparency:** Helps explain model behavior and training data lineage.
- **Accountability:** Enables audits and regulatory reporting for Responsible AI.
- **Reproducibility:** Allows teams to rebuild datasets and models consistently.

Avoiding “contaminated data”

جامعة الحسين التقنية
Al Hussein Technical University

Contaminated data refers to any data in a training set that *should not* be included because it creates unfair advantages, undermines evaluation validity, or violates ethical/legal constraints. Avoiding contamination ensures that models are trained responsibly and evaluated accurately.

Ethical Sourcing and Dataset Documentation

جامعة الحسين التقنية
Al Hussein Technical University

1. Ethical Sourcing

Ethical sourcing ensures that all data collected for training GenAI models is obtained **legally, responsibly, and with respect for human rights and societal norms**.
It goes beyond legality to include fairness, transparency, and minimizing harm.

Ethical Sourcing and Dataset Documentation

جامعة الحسين التقنية
Al Hussein Technical University

Key Principles of Ethical Sourcing

- **Respect for creators' rights:**
  Use data that complies with copyright, licensing, and Terms of Service.
- **Consent and privacy protection:**
  Avoid personal data unless explicit, informed consent is provided.
- **Fair representation:**
  Avoid datasets that overrepresent or underrepresent specific groups (bias risk).
- **Safety and harm reduction:**
  Remove harmful, abusive, extremist, or discriminatory content.
- **Transparency in collection methods:**
  Document how the data was collected (crawl, API, manual, purchase).
- **Cultural and regional respect:**
  Avoid appropriating sensitive cultural materials without permission.
- **Avoid exploitation:**
  Ensure annotators and data labelers are paid fairly and not exposed to psychological harm.

Ethical Sourcing and Dataset Documentation

جامعة الحسين التقنية
Al Hussein Technical University

2. Dataset Documentation

Dataset documentation provides detailed records of the dataset's **origin, content, structure, risks, and intended use**.
It enhances transparency, reproducibility, and responsible governance.

Common Documentation Standards

- Datasheets for Datasets (Gebru et al.)
- Data Statements for NLP
- Model Cards / Data Cards (Google, Meta)
- OpenML metadata schemas
- Provenance metadata (JSON-LD / DCAT)

Module 4 Summary & Transition

جامعة الحسين التقنية
Al Hussein Technical University

- Module 4 established the full pipeline required to build and align GenAI systems.
- It began with **data collection and cleaning**, ensuring that training data is high-quality, diverse, and ethically sourced. This led into **tokenization and vocabulary management**, where raw text is transformed into efficient units for model consumption. Next, we explored **fine-tuning strategies**, comparing full-model updates with lighter **PEFT** approaches that reduce computational cost. The module concluded with **RLHF and instruct tuning**, which shape model behavior toward human preferences, and with **data governance and licensing principles**, ensuring legal, transparent, and responsible model development.

Module 4 Summary & Transition

جامعة الحسين التقنية
Al Hussein Technical University

- With the foundation for building and aligning models in place, **Module 5 transitions from creation to evaluation and optimization**. Here, the focus shifts to measuring how well GenAI models actually perform in practice.
- Students learn key **LLM metrics** (perplexity, BLEU, ROUGE, human evaluation) and **image/media metrics** such as FID. The module further examines **model limitations**—hallucinations, knowledge gaps, and contradictions—before introducing **optimization techniques** including quantization, distillation, and compression for efficient deployment. Finally, **red-teaming and adversarial testing** equip learners with methods to assess safety, robustness, and real-world reliability.