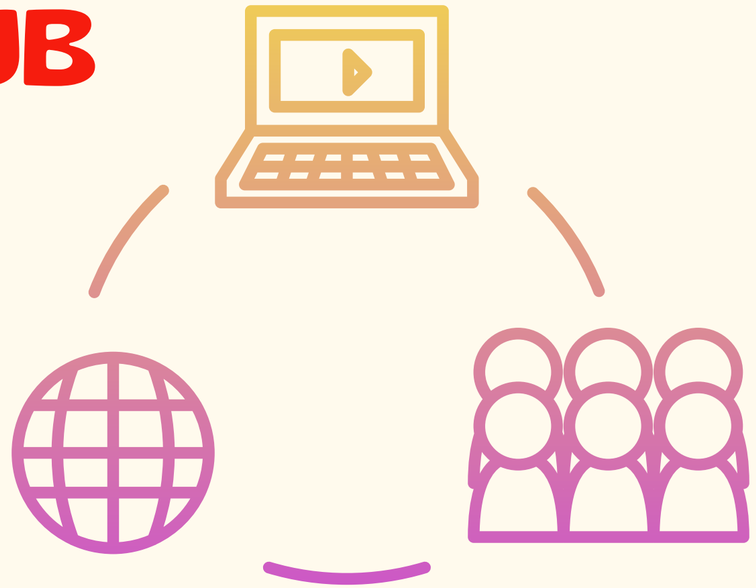
AUTOTRAIN ON CUSTOM DATASETS FROM HUB



DIVE DEEP INTO WORLD OF DATASETS & TRAINING MODELS

CHALLENGE SOLVED: RE-WIRING MODEL FOR YOUR NEED

WHAT IS FINE-TUNING?

RE-TRAINING THE EXISTING LARGE LANGUAGE MODEL WITH NEW SET OF PROMPT + EXPECTED OUTPUT SO THE LLM CAN LEARN THE INSTRUCTION

PROCESS OF TUNING THE LLM

- DECIDING THE TASK AND THE MODEL THAT NEEDS TO BE TRAINED
- CREATE DATASET THAT CAN BE USED FOR THE SPECIFIC TASK
- CREATE HUGGING FACE DATASET N PUSH TO HUB (OPTIONAL)

FOLLOWING STEPS ARE AUTOMATED BY THE "AUTO TRAIN" SERVICE:

- CREATE TRANSFORMERS TRAINER INSTANCE WITH TRAINING ARGUMENTS
- TOKENIZE THE PROMPTS + EXPECTED OUTPUTS AND FEED TO THE TRAINER
- AWAIT THE TRAINER TO COMPLETE AND THEN
- TEST THE NEW MODEL FOR INFERENCE

HTTPS://GITHUB.COM/INSIGHTBUILDER

CHALLENGE SOLVED: HOW TO GET THE DATA

THERE IS LOT OF DATASETS IN HUGGING FACE & KAGGLE

1. **QUESTION ANSWERING:** FINETUNE THE MODEL TO ANSWER QUESTIONS IN A SPECIFIC CONTEXT. YOU CAN USE DATASETS LIKE SQUAD OR TRIVIAQA FOR THIS.
2. **PARAPHRASING:** FINE-TUNE THE MODEL TO PRODUCE PARAPHRASES OF A GIVEN SENTENCE OR PARAGRAPH. YOU CAN USE DATASETS LIKE PARANMT OR QUORA QUESTION PAIRS FOR THIS.
3. **SENTIMENT ANALYSIS:** FINETUNE THE MODEL TO CLASSIFY THE SENTIMENT OF A GIVEN TEXT. YOU CAN USE DATABASES LIKE IMDB OR YELP REVIEWS, OR SIMILAR ONES ON KAGGLE.
4. **TEXT GENERATION:** FINETUNE THE MODEL TO PRODUCE TEXT IN A SPECIFIC DOMAIN, SUCH AS NEWS STORIES, SCHOLARLY PAPERS, OR SOCIAL MEDIA POSTS. YOU CAN USE DATASETS LIKE CNN/DAILY MAIL FOR THIS.
5. **DIALOGUE GENERATION:** FINETUNE THE MODEL TO GENERATE REPLIES IN A CONVERSATIONAL SETTING. YOU CAN USE DATASETS LIKE PERSONA-CHAT OR THE CORNELL MOVIE DIALOGUES CORPUS FOR THIS.
6. **TEXT SUMMARIZATION:** FINE-TUNE THE MODEL TO PROVIDE SUMMARIES OF LONGER TEXTS. YOU CAN USE DATASETS LIKE CNN/DAILY MAIL OR XSUM FOR THIS.

THANKS TO SURAJ520 (GITHUB HANDLE), FOR SHARING THE LIST OF DATASETS HTTPS://WWW.KAGGLE.COM/SURAJ520/CODE

LEVERAGING DATASET LIBRARY

IF YOU ARE THINKING OF EASIER ROUTE:

- VISIT HF TASKS
- LOCATE THE DATASET AT THE BOTTOM OF THE TASK PAGE
- LOAD_DATASET() IN THE COLAB ENVIRONMENT
- USE SELECT() METHOD AND TAKE 500/3000 ROWS
- PUSH THE DATA TO HF HUB & USE IT

HTTPS://GITHUB.COM/INSIGHTBUILDER

AUTOMATE WITH CLARITY : PRACTICE

SOME IMPORTANT TASKS:

- IMAGE CLASSIFICATION
- TEXT QUESTION ANSWERING
- TEXT REGRESSION
- TEXT SUMMARIZATION

HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

REMEMBER TO PRACTICE WITH EXAMPLES

HTTPS://GITHUB.COM/INSIGHTBUILDER