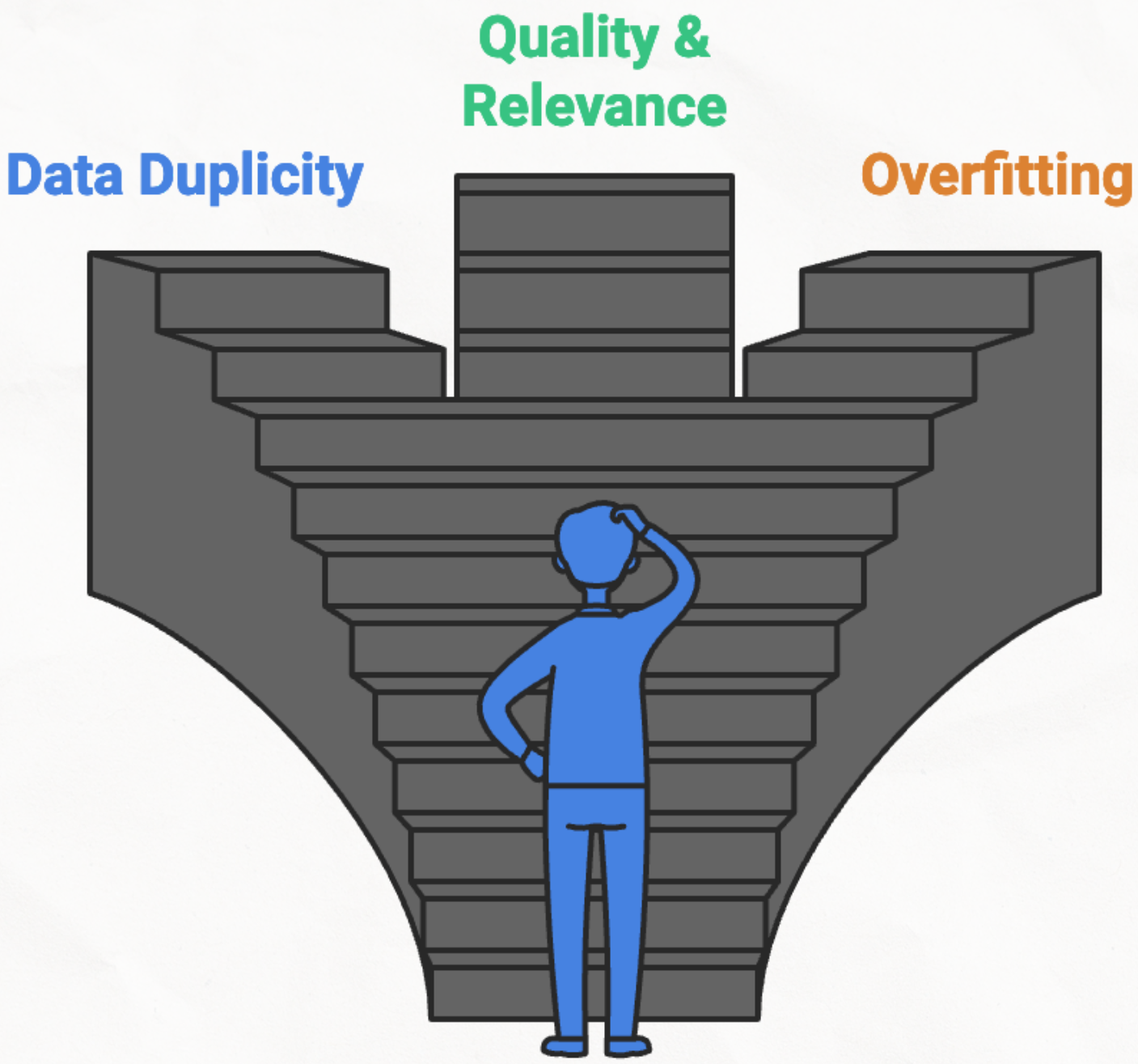
Llama 3.3 for Synthetic Data Generation

Leveraging LLMs to boost the performance of classical ML models (with code)

Bhavishya Pandit

Limitations of Current Techniques



- **Data duplicity**: Oversampling techniques like SMOTE can create repetitive data, limiting the model's ability to learn new patterns and generalize to unseen data.

- **Quality and relevance**: Synthetic data generated by interpolation often lacks real-world complexity, leading to models that struggle to generalize to diverse real-world scenarios.

- **Overfitting**: Limited diversity in synthetic data can increase the risk of overfitting, as models become overly tuned to training data and fail to handle novel examples.

SOLUTION : LLAMA-3.3

- The **Meta Llama 3.3** is an auto-regressive language model that uses SFT and RLHF to align with human preferences for helpfulness and safety, developers can now use its expanded context length of **128k tokens** to produce vast and high-quality data.

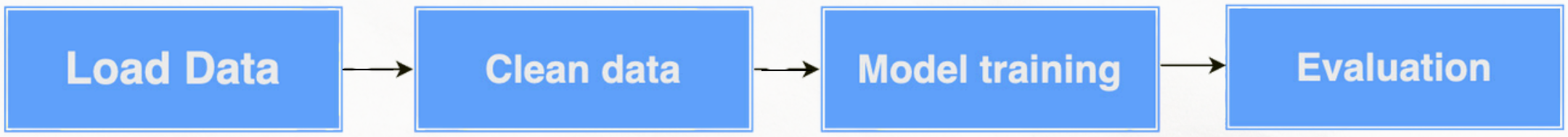
	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowledge cutoff
Llama 3.3 (text only)	A new mix of publicly available online data.	70B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	December 2023

Category	Benchmark	# Shots	Metric	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama-3.3 70B Instruct
	MMLU (CoT)	0	macro_avg/acc	73.0	86.0	86.0
	MMLU Pro (CoT)	5	macro_avg/acc	48.3	66.4	68.9
Steerability	IFEval			80.4	87.5	92.1
Reasoning	GPQA Diamond (CoT)	0	acc	31.8	48.0	50.5
Code	HumanEval	0	pass@1	72.6	80.5	88.4
	MBPP EvalPlus (base)	0	pass@1	72.8	86.0	87.6
Math	MATH (CoT)	0	sympy_intersection_score	51.9	68.0	77.0
Tool Use	BFCL v2	0	overall_ast_summary/macro_avg/valid	65.4	77.5	77.3
Multilingual	MGSM	0	em	68.9	86.9	91.1

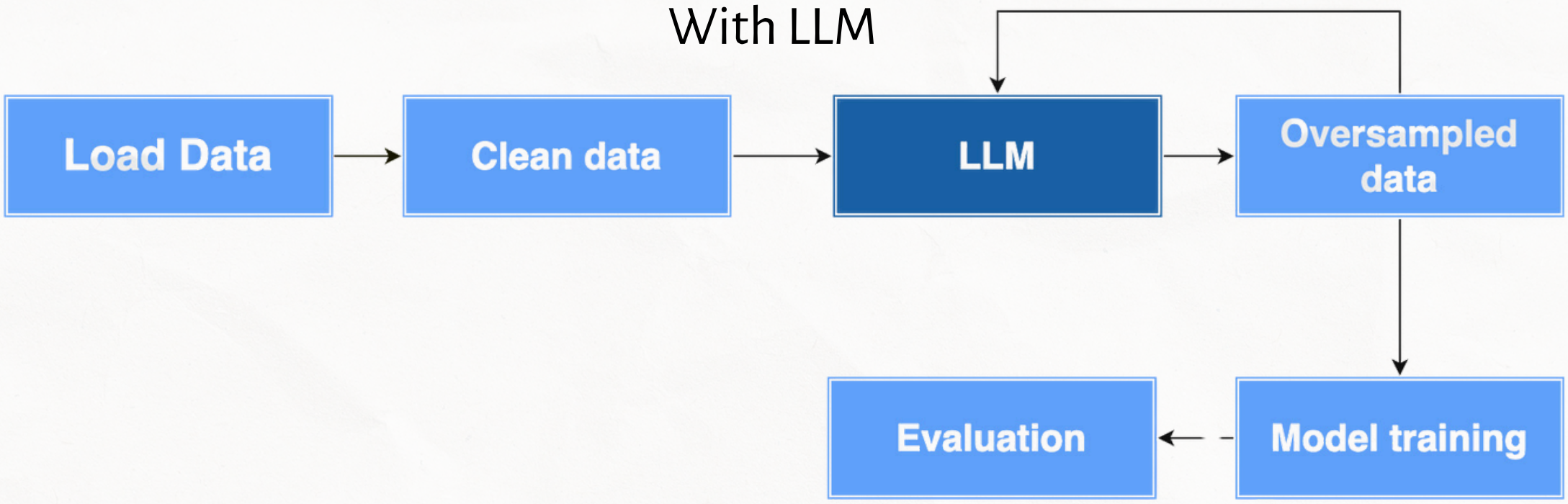
Source:https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

How It Works?

Without LLM



With LLM



Step 1. Load the data

Step 2. Clean the data.

Step 3. Pass the data to LLama 3.3 to generate more data.

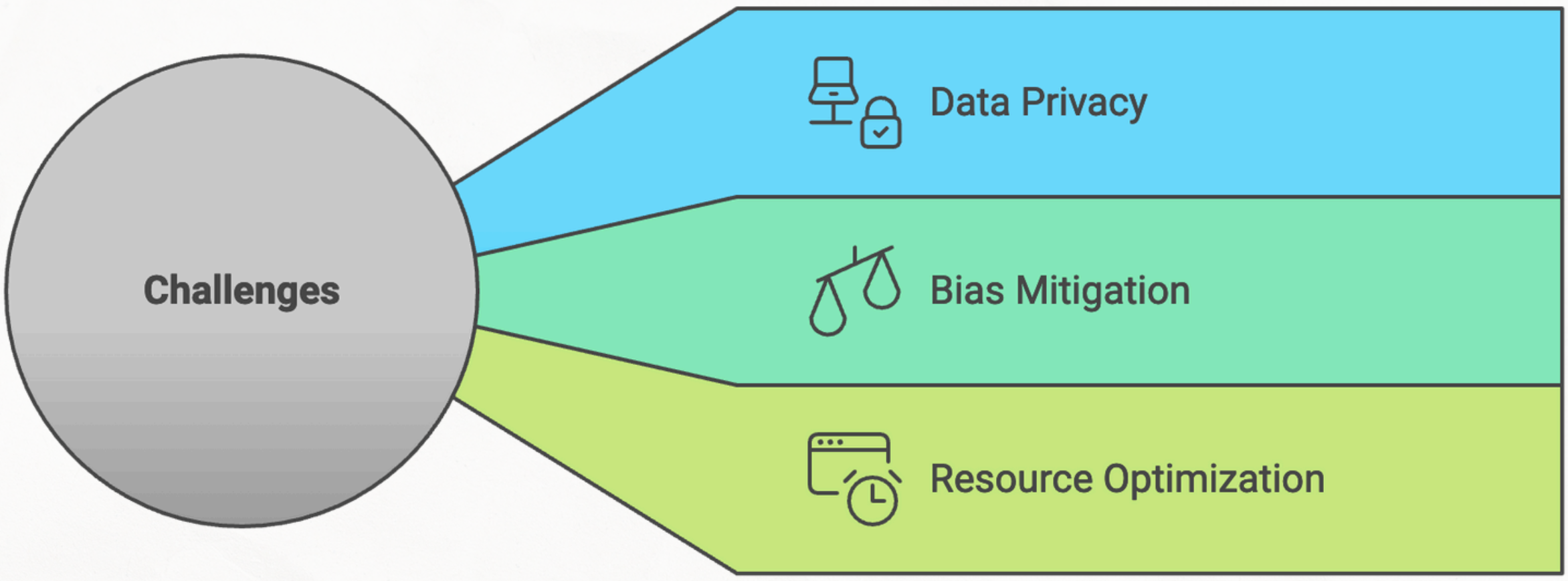
Step 4. Use oversampled data to generate more data

Step 5. Model training.

Step 6. Model evaluation.

For detailed understanding of each step and code refer to IBM developers or check comments

Challenges and Practical Considerations



- **Data privacy**: Protecting sensitive information by using anonymized input data, validating outputs, and carefully controlling prompts.

- **Bias mitigation**: Preventing biased data generation by designing neutral prompts, validating outputs, and employing data preprocessing or prompt engineering techniques.

- **Cost and computational resources**: Optimizing prompt length and generation parameters to reduce computational overhead and ensure efficient use of LLMs.

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST