

Other
(confusion matrix)
ROC - AUC
log Loss

Evaluation of LLM (LLM note)

1) Automatic (Quantitative) Metrics → measurable
→ fast
→ used in model dev

Better if	Metric	What it measures	used for
↓	Perplexity	How surprised is model by real text	language models
↑	Accuracy	correct prediction over total ie how often LLM correct %	classification
↑	BLEU/ROUGE	Overlap with reference texts	translation / summarization
↑	F1 score	precision + recall balance	QA, classification
↑	Exact Match	Exact match with ground truth	QA, code, Math
↑	token lvl accuracy	token by token correctness	fine grained tasks
↑	Precision	when model says its correct how often is it correct ie reliability of correctness	classification

2) Human (Qualitative) Evaluation → human friendly

Better if	Metric	Example
↑	Helpfulness	Does it actually answer the question well
↑	Fluency	Does it sound natural and correct
↑	Factuality	Are the facts accurate
↓	Toxicity/Bias	is the response offensive or biased
↑	Relevance	is the answer on topic

Mixture of Experts (Note experts are not labeled and are probabilistic but can be thought as experts for math, language etc for EX)

- MoE is a innovation in LLM leading to efficient scaling

? - MoE is a NN architecture where only a small subset of the model is active per input rather than using the whole model each time

in LLM - in LLM this means 1) you have many expert sub models (often FF blocks)
2) for each token or input the model activates only a few experts total
3) allows model to be very big but efficient (instead of load of 1B model use 10% of 1B Model)

How? - say a model has 64 experts
- for a token gating mechanism activates top 2 or 4 experts
- output is weighted combination of those expert outputs

Notes: - Gating network is a small NN that scores each expert based on input token
- only top k experts chosen k is hyperparam
- output is sum of any of experts output
- done per token