Evaluation of LLM (LLMotti) (Also ser los)

Other (confusion matrix, ROC - AUC, log Loss)

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
↑	Persision	when model says its correct how often is it correct ie reliability of correct ans	classification

2) Human (Qualitative) Evaluation → human friendly

Better if	Metric	Example
↑	Helpfullness	Dose it actually answer the question well
↑	Fluency	Dose 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 inovation in LLM leading to efficient scaling

? - MoE is a NN arcetecture 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 not all
3) allows model to be very Big but efficient (instead of lots of 1B model use bit of 1T Model)

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

Notes: • Gating network is a small NN that scores each expert based on input token
• only to k experts chosen K: hyperparam
• output is sum or avg of experts output
• done per token