LLM terms (cont...)

- Open weights vs closed weights models: open weight models are those models whos learned weights are shared with everyone, meaning you can download and use it on your device and modify it. Closed weight is the opposite - you can send the LLM a request and it will use the weights locally but you cannot download or access the weights

Generation controls (Like temprature).

- Top-p: also called nucleus sampling, is a setting that guides how a LLM picks the next word by sorting all possible next words by probability, it then finds the smallest group of words whos combined chance adds to p (Ex 0.9) only those words stay in the running Lower p keeps likly words (safer) higher p = creative & more risk of err

- Frequency penalty: tells LLM "stop repeating yourself". as the model writes it keeps track of each word count. a pos freq penalty values lowers the chance of picking a word again if it has seen it many times while a value of 0 turns the rule off but too high=bad, good=0.2

- presence penalty: forces model to choose words it has not chosen so far, each time a word already used it picked again the model gets a score cut, higher penalty = bigger cut pushing the model to pick new words while lower penalty lets it repeat words

- Stopping Criteria: tells LLM when to stop so it does not go forever Common Rules: Max tokens, end of seq token Ex "\n\n" of if it hallucinates.

- Max Lenght: Max tokens a LLM model can use controles the lenght of output balancing cost, speed