* How do people use LLMs to evaluate responses?
* Other datasets I can test, specifically multi discipline?
* 3 attempts? How many?