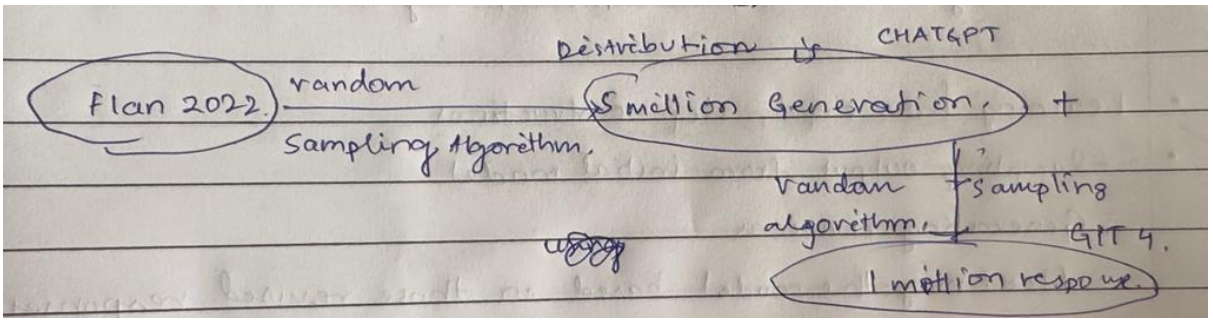
Orca 13B Parameter Model

1.Explanation Tuning- Augmenting{query,response} pairs anf leveraging system instructions.

2.Evaluation-

- Auto Evaluation with GPT4- on existing evaluation sets such as (Vicuna,WizardLM and the Awesome Prompt Collection[ChatGPT])
- Academic Benchmarks like Big-Bench Hard(BBH) and TruthfulQA
- AGI Eval for professional exams such as GRE,GMAT etc
- Safety Evaluation with ToxiGen

3.Scaling Tasks and Instructions:- From Flan 2022



4.Instruction Tuning-For language only tasks instruction tuning has been shown to improve the zero-shot and few shot performance models

- **Drawback-**Limited Task Diversity

5.Dataset Construction-Flan v2 is a sub collection of datasets(**CoT**[Chain of Thought],**NIV2,Flan2021,T0,Dialog**(None was used))

- **Zero shot CoT-** Contains total 18 tasks(math problems,natural language inference,etc)
  - ❖ 18 tasks contains 150k queries
- **NIV2-** 1560 tasks and approximately 5 million queries randomly sampled 300 queries from each task
- **Flan2021-**Contains total 142 tasks (To have a diverse and representative subset we generate 1million queries from each task)
- **T0-**Contains a total of 193 tasks
  - ❖ Out of which Big Bench Hard is excluded as it is used for benchmarking
  - ❖ Sampled 2 million out of 25.7 million queries

- **Training-** Orca is first trained on 5 million responses of ChatGPT(GPT 3.5Turbo).Second round training on 1million responses of GPT-4 augmentations.[Viewed as form of progressive learning or curriculum]
- **Tokenization-**Llama BPE(Byte Pair Encoding)
  - ❖ To deal with variable length sequences we add a **Padding** Token

- **Packing-** A technique to utilize efficiently the computational resources

- ❖ Concatenating multiple inut examples ina single sequence till the max_token_limit=2048(in Orca-13B Case)
- ❖ 2.7 examples per sequence length(**Packing Factor**)

- **Compute-** 20 Nvidia A100 GPUs
  - ❖ 160 hrs — 5 million (GPT 3.5 turbo) — 4 epochs
  - ❖ 40 hrs — 1 million(GPT4) —- 4 epochs