1. **Evaluate <u>current vanilla LLMs</u>** on **related tasks** and **end task** to show they are good for evaluations:
    1. <u>partial task</u>: confusion detection (use specific prompt, we need human scores of confusion degree, P, R,F1)
    2. <u>partial task</u>: completness of options (use new specific prompt asking model to list options and use GT options from plausibleQA or manually made, P,R,F1)
    3. <u>partial task</u>: explain why the options are wrong (use new specific prompt where you give all the options but now ask for using them to decrease confusion). In this case we test only theexplainability of LLM answer when options are given.
    4. test <u>end task using different prompts</u>: baseline, clarify-doubts, clarify-doubts-unrestricted, CFE, cfe-unrestricted
        1. completeness
        2. explainability
        3. confusion
2. **Create dataset of good answers**

- manully create all answers
- semi-manual: GPT automatically --> human manually selects best ones

5. **Finetune LLM** (or via agentic approach) with good answers to teach it how to produce good answers
6. **Evaluate results** (manually and also partially with automatic means) of those finetuned LLMs along with results of vanilla LLMs on how good answers they make for **confusing questions only**