1. Related work
    a. Blind/oblivious unlearning
2. Dataset
    a. 2 datasets - TOFU and MUSE
        i. variations - 5
            1. showcasing the percentage of overlap b/w forget and retain distributions - 0%, 25%, 50%, 75%, 100%
            2. Tradeoff b/w unlearning performance (multiple metrics) and overlap percentage
3. Model selection
    a. Based on
        i. Size
        ii. Family
        iii. Chat, Instruction, and MoE version of LLMs

        Selected models:
            1. Phi-1.5
            2. Llama2-7B
            3. Mistral-7B-instruct
            4. DeepSeek-R1-0528-Qwen3-8B
            5. Gemma-7B-instruct
            6. Gpt-oss-20b
            7. Need to add a model around 14B size to cover different model sizes
    b. Baseline
        i. GA
        ii. GD
        iii. KLMin
        iv. DPO/NPO
        v. GUARD
4. Evaluation metrics
    a. Automatic
        i. Utility
            1. MMLU
            2. MT-bench
        ii. Unlearning performance
            1. ROUGE
            2. Truth ratio
            3. Conditional probability
            4. MIA AUC
    b. LLM-based
        i. Different aspects to assess
    c. Human evaluation/error analysis
        i. Guidelines and rubrics
5. Ablation

    a. Adversarial attack
        i. From MLP's training data
        ii. Decoding Activation vectors
6. Limitations:
    a. Applicable only to Open-source models
    b. Performance depends on the overlap b/w data distributions
    c. MLP training data – generalizability – need to find appropriate public data for MLP training
    d. Synthetic test set – check TOFU limitations