- **LLM Benchmarking**
  - **Background**
    - Data Contamination
      - Exact Contamination
      - Syntactic Contamination
      - Examples of Each Contamination
      - Significance of Contamination
    - Contamination from LLM Training
    - LLM Benchmarking
  - **Static Benchmarking**
    - Problem Formulation
    - Static Benchmarking Application
      - Math: GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), AIME 2024 (of America, 2024) and CNMO 2024 (Society, 2024).
      - Knowledge: NaturalQuestions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), CMMLU (Li et al., 2023a), MMLU (Hendrycks et al., 2020), BBH (Suzgun et al., 2022), AGI Eval (Zhong et al., 2023), MMLU-Redux (Gema et al.,2024) and MMLU-Pro (Wang et al., 2024b), ControlBench (Darioush et al., 2024), FRAMES (Krishna et al., 2024), and GPQA Diamond (Rein et al., 2023), AlpacaEval (Li et al., 2023c), ArenaHard (Li et al., 2024a).
      - Coding: HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), SWE-Bench (Jimenez et al., 2024; Yang et al., 2025), Codeforces (Code-forces, 2025), Aider (Aider, 2025).
      - Instruction Following: IFEval (Zhou et al., 2023), InfoBench (Qin et al., 2024), C-Eval (Huang et al., 2024).
      - Reasoning: PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2018), C-SimpleQA (He et al., 2024).
      - Safety: RealToxicityPrompts (Gehman et al., 2020), ToxiGen (Hartvigsen et al., 2022).
      - Language: GLUE (Wang, 2018), SuperGLUE (Wang et al., 2019), CLUE (Xu et al., 2020), Typo-fixing (Suzgun et al., 2022).
      - Reading Comprehension: SQuAD (Rajpurkar et al., 2018), QuAC (Choi et al., 2018), BoolQ (Clark et al., 2019).
    - Methods for Mitigation
      - Canary String: BIG-Bench(Jacovi et al., 2023).
      - Encryption: Jacovi et al. (2023), Yang et al. (2023), TRUCE (Chandran et al., 2024) .
      - Label Protection: GLUE (Wang, 2018), SuperGLUE (Wang et al., 2019), HumanEval (Chen et al., 2021).
      - Post-hoc Detection: Direct overlap detection(Touvron et al., 2023), Radford et al., 2019; Brown et al., 2020; Chowdhery et al., 2023, Riddell et al., 2024; Lee et al., 2023; Gunasekar et al., 2023, Li et al., 2024d; Xu et al., 2024), memorization through masked inputs (Ranaldi et al., 2024; Chang et al., 2023), partial completions (Anil et al., 2023; Golchin and Surdeanu, 2024), preference for original over paraphrased test cases (Duarte et al., 2024; Golchin and Surdeanu, 2023; Zong et al., 2024).
  - **Dynamic Benchmarking**
    - Problem Formulation
    - Evaluation Criteria
      - Correctness
      - Scalability
      - Collision
      - Stable of Complexity
      - Diversity
      - Interpretability
    - Existing Work
      - Temporal Cutoff: LiveBench (White et al., 2024), AntiLeak-Bench (Wu et al., 2024), AcademicEval (Zhang et al., 2024a), Live- CodeBench (Jain et al., 2024), LiveAoPSBench (Mahdavi et al., 2025), Forecastbench (Karger et al., 2024).
      - Rule-Based
        - Template-Based: GSM-Symbolic (Mirzadeh et al., 2025), Mathador-LM(Kurtic et al., 2024), Mathador games(Puma et al., 2023), MMLU-CF (Zhao et al., 2024).
        - Table-Based: S3Eval (Lei et al., 2024).
        - Graph-Based: DyVal (Zhu et al., 2024a), NPHardEval (Fan et al., 2024).
      - LLM-Based
        - Benchmark Rewriting: Auto-Dataset (Ying et al., 2024), StructEval (Cao et al., 2024), ITD (Zhu et al., 2024c), VarBench (Qian et al., 2024).
        - Interactive Evaluation: LLM examiner (Li et al., 2023b), LM-as-an-Interviewer (Kim et al., 2024), TreeEval (Li et al., 2024b), KIEval (Yu et al., 2024).
        - Multi-Agent Evaluation: Self-Evolving (Wang et al., 2024a), BENCHAGENTS (Butt et al., 2024).
      - Hybrid: LatestEval (Li et al., 2023d), DARG (Zhang et al., 2024b), C2LEVA (Li et al., 2024c).