**Outline for Literature Analysis**

**1. Challenges in Training LLMs and the Need for Synthetic Data**

* **Papers to analyze**:
  + **"On the Diversity of Synthetic Data and its Impact on Training Large Language Models"** (Chen et al.)
  + **"A Comparative Study of Synthetic Data Generation Methods for Grammatical Error Correction"** (White & Rozovskaya)
  + **"Machine Learning for Synthetic Data Generation: A Review"**
* **How to analyze**:
  + Identify key **challenges** in acquiring real data (privacy, cost, availability).
  + Examine how synthetic data **compensates for these challenges** (scalability, control over quality).
  + Discuss whether synthetic data **preserves diversity** and **avoids biases** in LLM training.

**2. General Overview of Synthetic Data Generation Methods**

* **Papers to analyze**:
  + **"Comprehensive Exploration of Synthetic Data Generation: A Survey"**
  + **"Utility Assessment of Synthetic Data Generation Methods"**
  + **"Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study"**
* **How to analyze**:
  + Provide a **broad classification** of synthetic data techniques (GANs, VAEs, statistical models).
  + Highlight trade-offs in terms of **data utility, fairness, and computational cost**.
  + Assess the applicability of different methods to **LLM training vs. structured datasets**.

**3. Specific Comparisons of Synthetic Data Generation for ML Applications**

* **Papers to analyze**:
  + **"Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study"**
  + **"Can Synthetic Data be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms"**
* **How to analyze**:
  + Extract insights on **how well synthetic data preserves statistical similarity** to real-world data.
  + Compare different synthetic data techniques in **real-world case studies**.
  + Discuss **ethical concerns**, such as privacy and fairness.

**4. Connecting to Your Project: SDV vs. Synthcity**

* **Papers to analyze**:
  + **Your project proposal**
  + **Any documentation or benchmarks from SDV and Synthcity**
* **How to analyze**:
  + Identify which synthetic data generation methods **SDV and Synthcity** implement.
  + Compare their reported **accuracy, performance, and limitations** from literature.
  + Discuss which approach is more **suitable for ML applications** and why.