The SQuAD (Stanford Question Answering Dataset) is already structured for question-answering tasks, so in most cases, it doesn’t require heavy normalization or data cleaning. However, depending on the exam generator model’s requirements, you might still consider the following steps:

1. **Basic Cleaning**:
   * Ensure there are no null values, though SQuAD generally has well-formatted data.
   * Verify that all entries have the expected fields (context, question, answer).
2. **Data Transformation or Filtering**:
   * If you only need specific types of questions or answers (e.g., only factual questions or certain answer lengths), filter these out.
   * For certain models, you may need to tokenize the context and question fields or preprocess text to remove unwanted characters or symbols.
3. **Data Normalization (Optional)**:
   * If your model benefits from standardized text, consider text normalization (like lowercasing or removing punctuation).
   * For complex NLP tasks, you might also standardize answer formats, e.g., expanding contractions or lemmatizing words.

These steps ensure the dataset is clean, consistent, and prepared to feed into the exam generator model without over-complicating the structure.