ジャンル別の代表的なコーパス

- **Wikipedia**
  - Wikipedia: 英語版の記事を集めた21GBのデータ
- **コード**
  - BigQuery: 様々なプログラミング言語のコードスニペットを含むデータ
- **その他**
  - Pile: 書籍・WEB・コード・論文・SNSなど多様なソースから構成される800GBのデータ

C4 (Colossal Clean Crawled Corpus)

- C4には、以下の5種類のデータが含まれる

1.**en**: 英語のc4データセット

2.**en.noclean**: 重複や不適切な言葉の除去を行っていないデータセット

3.**realnewslike**: RealNewsで使用されるドメインのコンテンツのみを含めたデータセット

4.**webtextlike**: OpenWebTextに含まれるURLからのコンテンツに絞り込んだデータセット

5.**multilingual**: 101の言語を含むデータセット