学習に使用されるデータ

ジャンル別の代表的なコーパス

- **書籍**
  - Book Courpus: 様々なジャンル/トピックの11000冊以上の書籍から構成
  - Project Gutenberg: 小説・エッセイ・詩・演劇・歴史・科学・哲学など70000以上の書籍から構成
- **CommonCrawl** (ペタバイトケールのデータ量を含む最大級のウェブクローリングデータベース)
  - C4: CommonCrawlにクリーニングを行った800GBのデータ
  - RealNews: CommonCrawlからニュースを抽出した120GBのデータ
- **Reddit**
  - OpenWebText: Reditで高く評価された投稿を集めた38GBのデータ