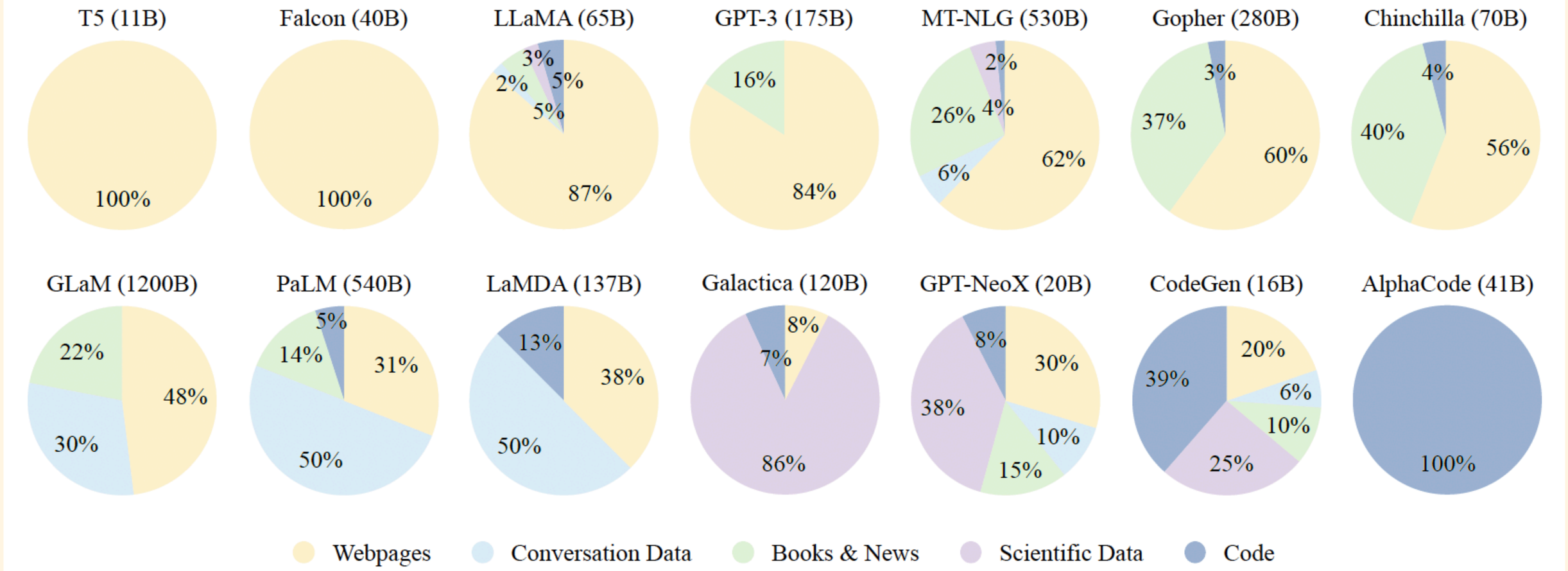

Fig. 5: Ratios of various data sources in the pre-training data for existing LLMs.

データ前処理: ①品質フィルタリング

機械学習ベースで品質を測定する方法と、ルールベースの方法がある
前者は口語なども削除されてコーパスの多様性が低下する可能性がある

- 言語ベース

  学習に不要な言語のテキストをフィルタリングする

- メトリックベース

  生成されたテキストに関する評価指標 (perplexity: テキストの流暢さ など) に基づいて不自然な文章を検出する

- 統計ベース

  句読点や文長などの統計的特徴を利用して低品質なデータをフィルタリングする

- キーワードベース

  HTMLタグや定型文、不快な言葉などのノイズや不要な要素を特定して削除する