データソース

- 既存のLLMは主に、多様な公開テキストデータセットを混合して使用している
  - ❖ 一般的なデータ
    - Webページ：LLMが汎化能力を獲得するのに有用。品質は玉石混交のためフィルタリングが重要
    - 会話テキスト：質問応答タスクの性能向上に有用。会話をツリー構造に変換する処理が効果的
    - 書籍：物語的で守備一貫したテキストを生成するのに有益な可能性
  - ❖ 特殊なテキストデータ
    - 多言語テキスト：言語理解と言語生成の多言語能力を向上できる
    - 科学的な文章：科学的タスク・推論タスクの性能を向上できる
    - コード：複雑な推論タスクの性能を向上する可能性

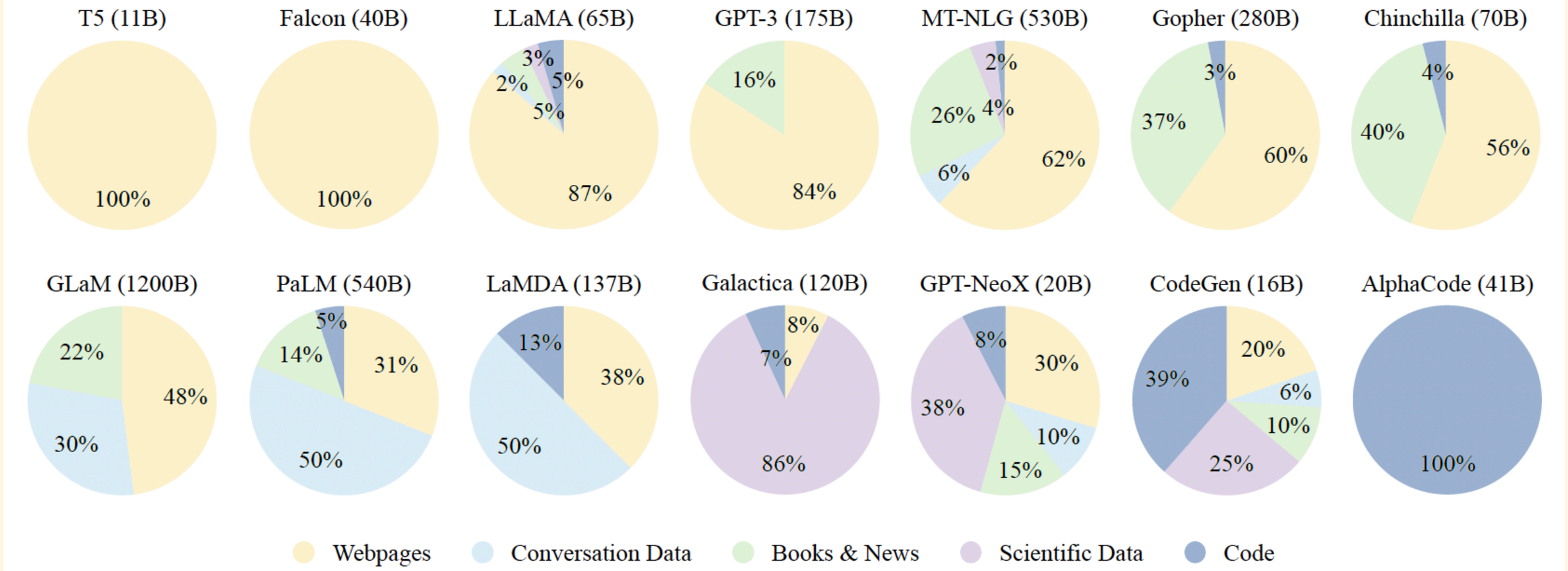

Fig. 5: Ratios of various data sources in the pre-training data for existing LLMs.