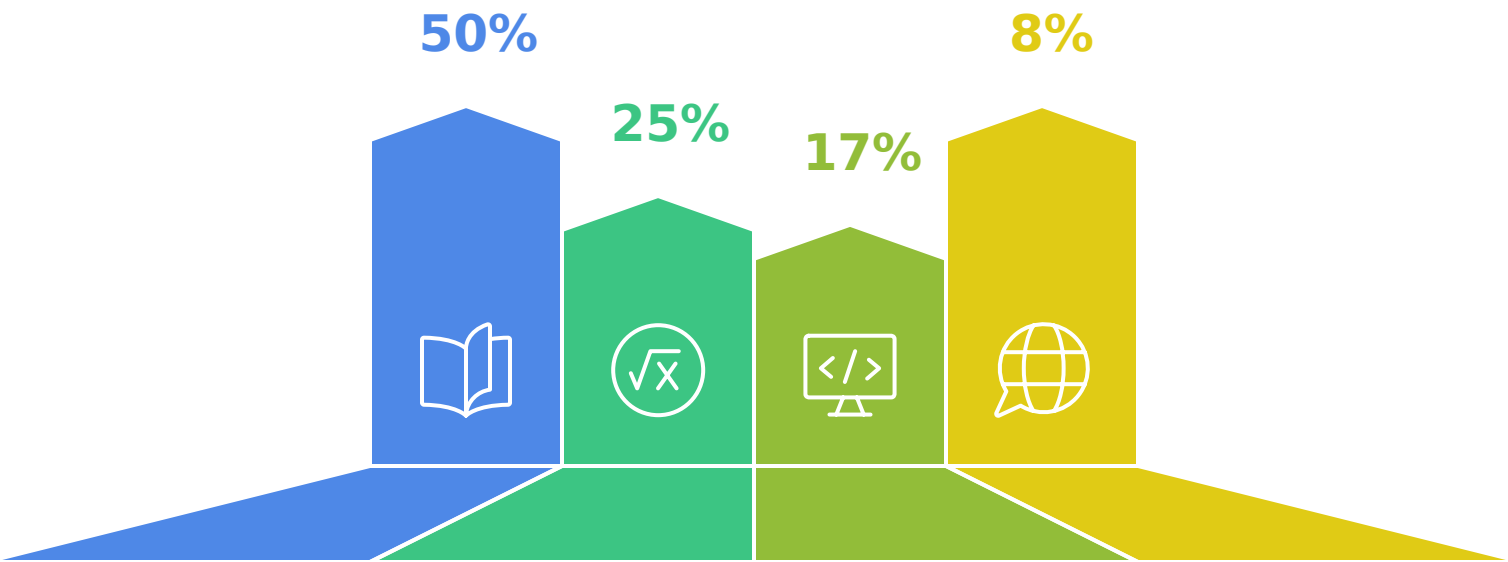
Composition of Final Data Mix for Language Model



General Knowledge

Broad range of common information

Mathematical and Reasoning

Logical and analytical content

Code

Programming and software-related data

Multilingual

Diverse linguistic content