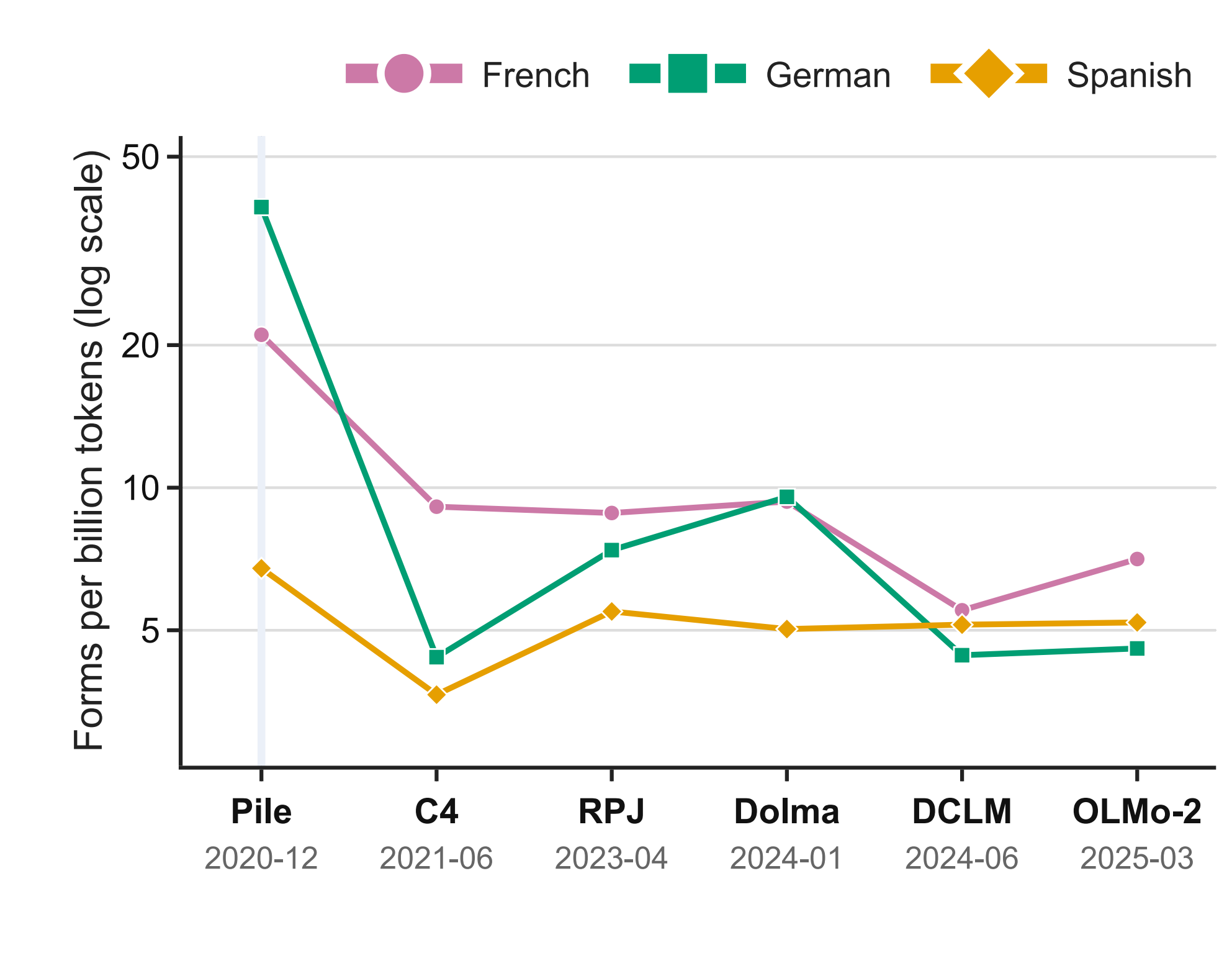

French
German
Spanish
Forms per billion tokens (log scale)
50
20
10
5
Pile
2020-12
C4
2021-06
RPJ
2023-04
Dolma
2024-01
DCLM
2024-06
OLMo-2
2025-03