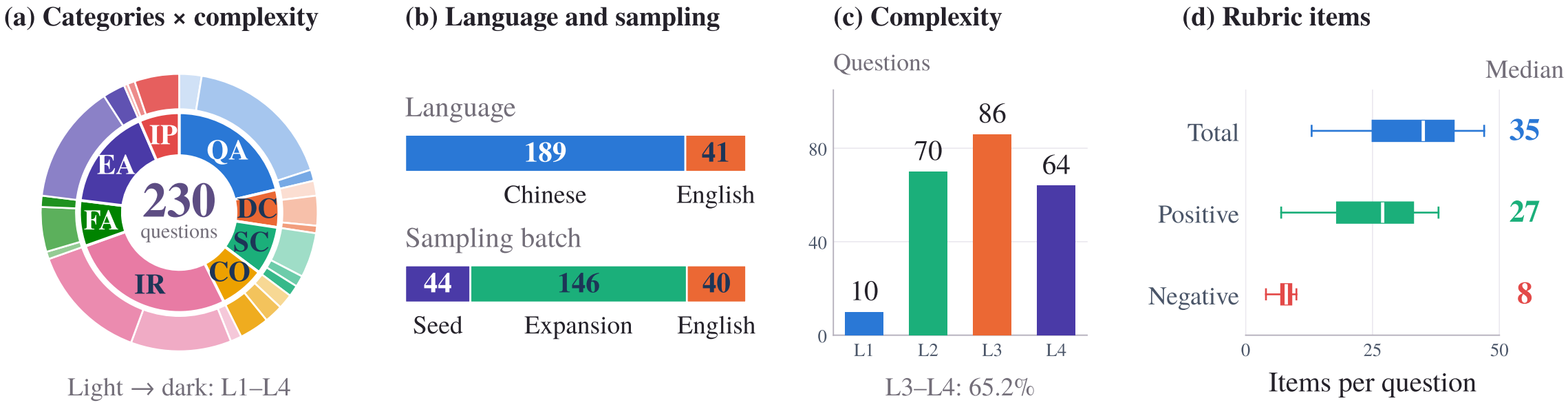

(a) Categories × complexity
IP
QA
EA
DC
230
questions
FA
SC
CO
IR
Light → dark: L1–L4
(b) Language and sampling
Language
189
41
Chinese
English
Sampling batch
44
146
40
Seed
Expansion
English
(c) Complexity
Questions
80
40
0
10
70
86
64
L1
L2
L3
L4
L3–L4: 65.2%
(d) Rubric items
Median
Total
35
Positive
27
Negative
8
0
25
50
Items per question