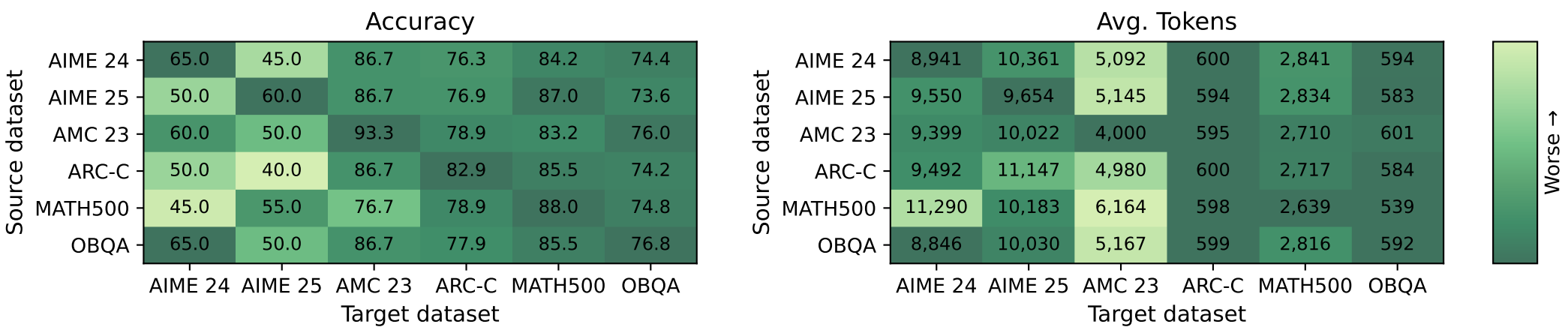
Accuracy
Source dataset
AIME 24
AIME 25
AMC 23
ARC-C
MATH500
OBQA
65.0 45.0 86.7 76.3 84.2 74.4
50.0 60.0 86.7 76.9 87.0 73.6
60.0 50.0 93.3 78.9 83.2 76.0
50.0 40.0 86.7 82.9 85.5 74.2
45.0 55.0 76.7 78.9 88.0 74.8
65.0 50.0 86.7 77.9 85.5 76.8
AIME 24 AIME 25 AMC 23 ARC-C MATH500 OBQA
Target dataset
Avg. Tokens
Source dataset
AIME 24
AIME 25
AMC 23
ARC-C
MATH500
OBQA
8,941 10,361 5,092 600 2,841 594
9,550 9,654 5,145 594 2,834 583
9,399 10,022 4,000 595 2,710 601
9,492 11,147 4,980 600 2,717 584
11,290 10,183 6,164 598 2,639 539
8,846 10,030 5,167 599 2,816 592
AIME 24 AIME 25 AMC 23 ARC-C MATH500 OBQA
Target dataset
Worse →