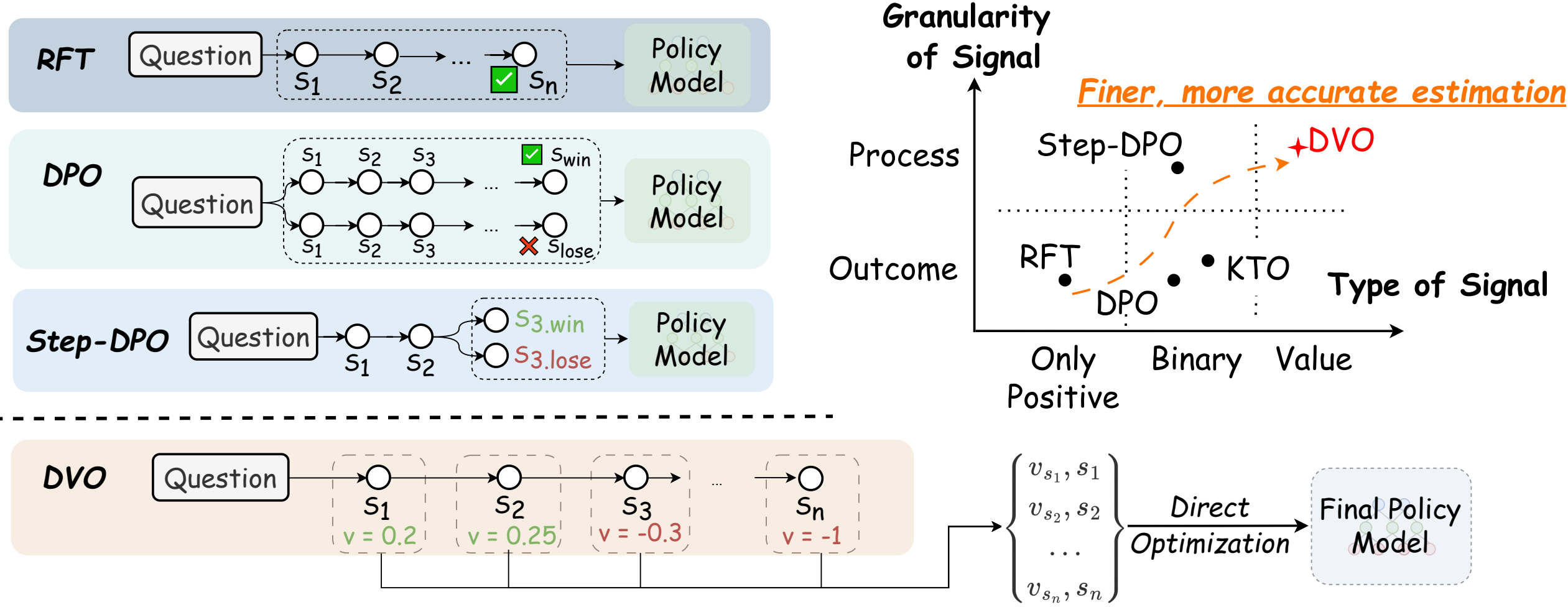
RFT
Question
s_1
s_2
...
s_n
Policy Model
DPO
Question
s_1 s_2 s_3 ... s_{win}
s_1 s_2 s_3 ... s_{lose}
Policy Model
Step-DPO
Question
s_1
s_2
$s_{3.win}$
$s_{3.lose}$
Policy Model
DVO
Question
s_1 v = 0.2
s_2 v = 0.25
s_3 v = -0.3
...
s_n v = -1
v_{s_1}, s_1
v_{s_2}, s_2
...
v_{s_n}, s_n
Direct Optimization
Final Policy Model
Granularity of Signal
Finer, more accurate estimation
Process
Outcome
Step-DPO
DVO
RFT
DPO
KTO
Type of Signal
Only Positive
Binary
Value