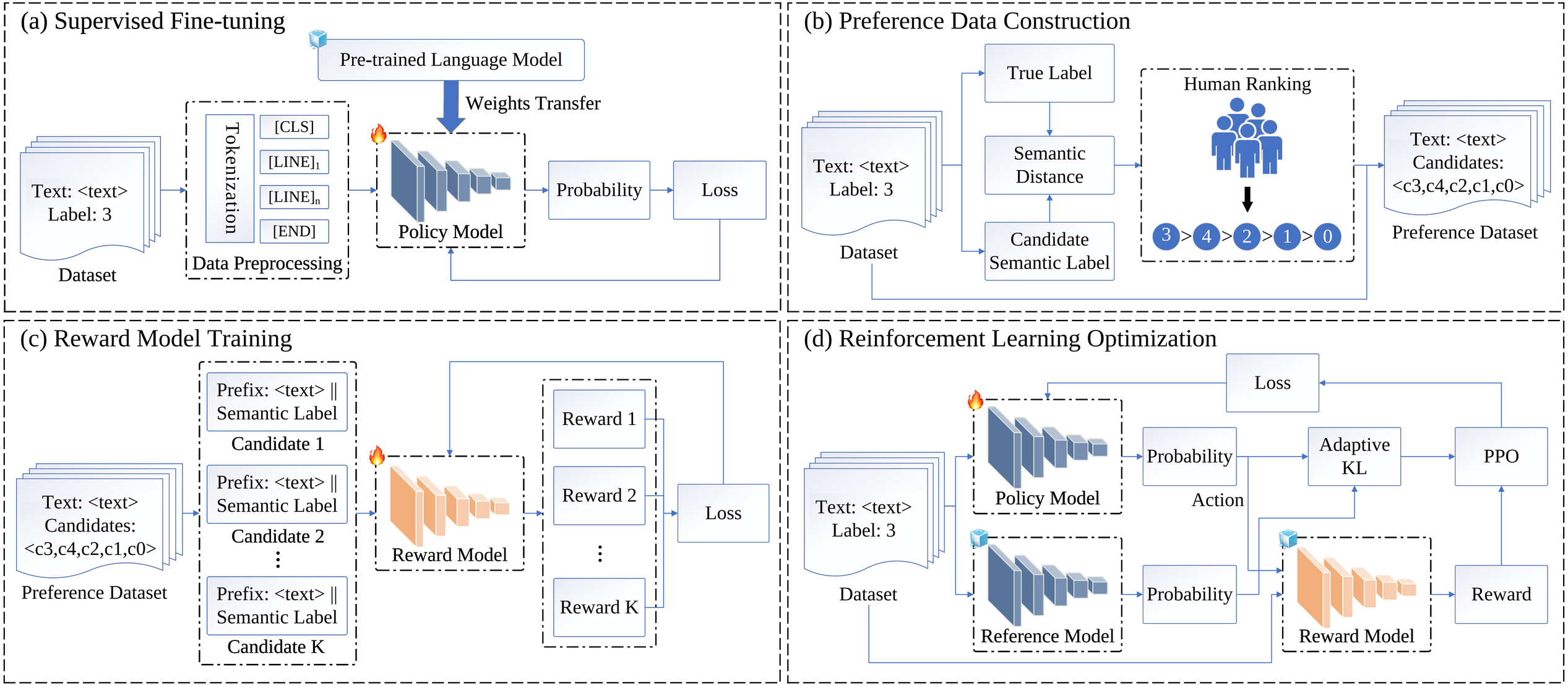
(a) Supervised Fine-tuning
Pre-trained Language Model
Weights Transfer
Text: <text>
Label: 3
Dataset
Tokenization
[CLS]
[LINE]1
[LINE]n
[END]
Data Preprocessing
Policy Model
Probability
Loss
(b) Preference Data Construction
Text: <text>
Label: 3
Dataset
True Label
Semantic Distance
Candidate Semantic Label
Human Ranking
3>4>2>1>0
Text: <text>
Candidates:
<c3,c4,c2,c1,c0>
Preference Dataset
(c) Reward Model Training
Text: <text>
Candidates:
<c3,c4,c2,c1,c0>
Preference Dataset
Prefix: <text> || Semantic Label
Candidate 1
Prefix: <text> || Semantic Label
Candidate 2
Prefix: <text> || Semantic Label
Candidate K
Reward Model
Reward 1
Reward 2
Reward K
Loss
(d) Reinforcement Learning Optimization
Loss
Text: <text>
Label: 3
Dataset
Policy Model
Reference Model
Probability
Action
Probability
Adaptive KL
PPO
Reward Model
Reward