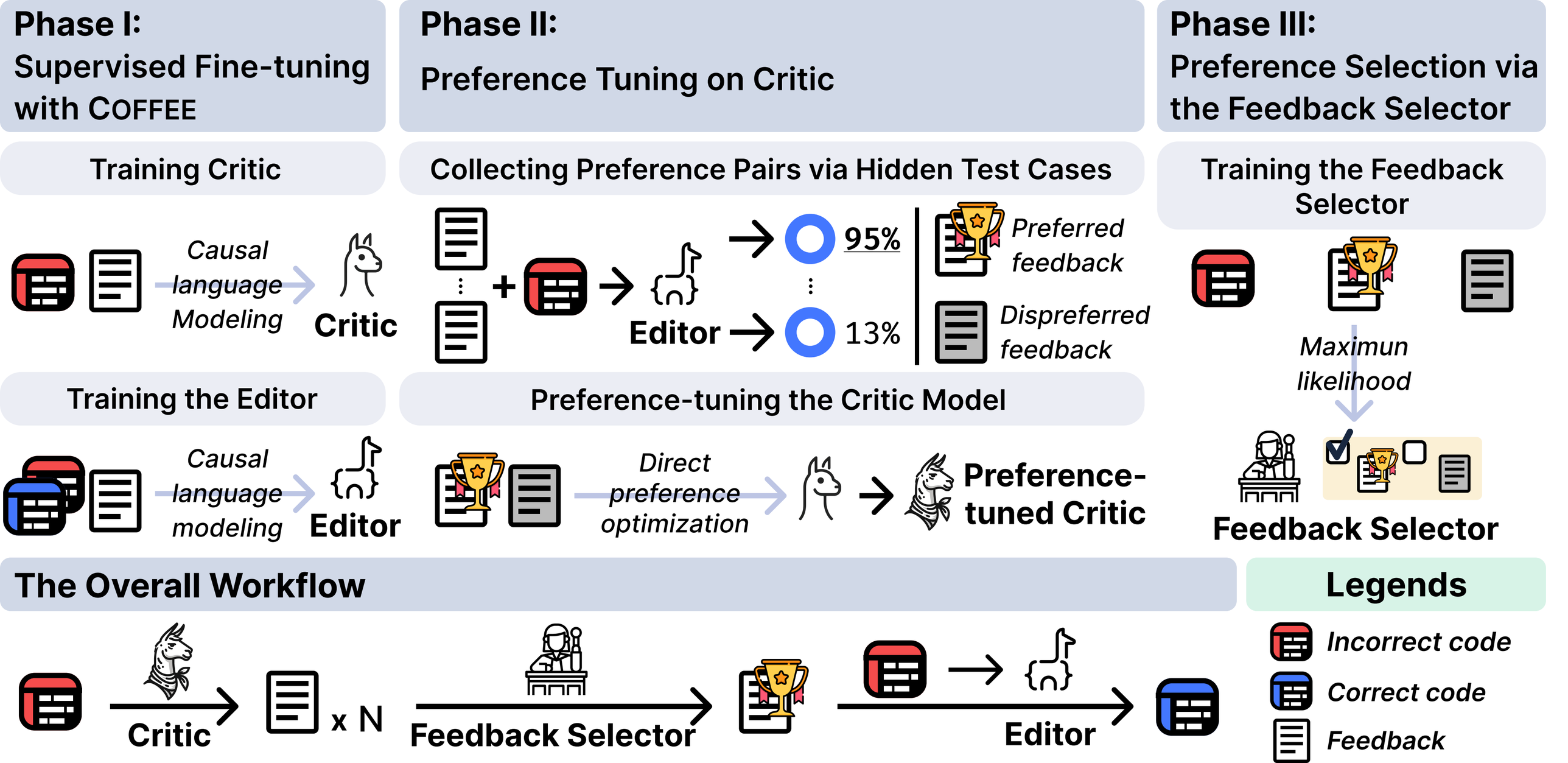
Phase I:
Supervised Fine-tuning with COFFEE
Phase II:
Preference Tuning on Critic
Phase III:
Preference Selection via the Feedback Selector
Training Critic
Causal language Modeling
Critic
Collecting Preference Pairs via Hidden Test Cases
Editor
95%
13%
Preferred feedback
Dispreferred feedback
Training the Feedback Selector
Maximun likelihood
Feedback Selector
Training the Editor
Causal language modeling
Editor
Preference-tuning the Critic Model
Direct preference optimization
Preference-tuned Critic
The Overall Workflow
Critic
x N
Feedback Selector
Editor
Legends
Incorrect code
Correct code
Feedback