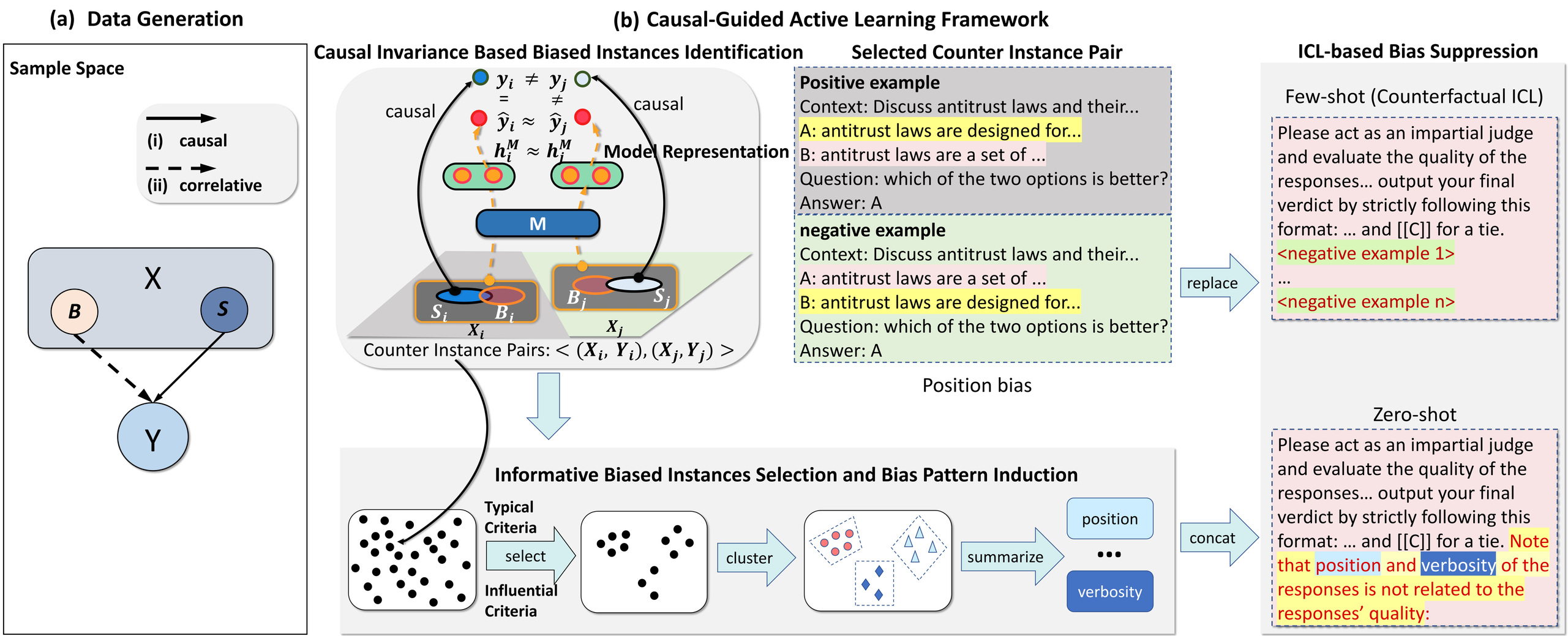
(a) Data Generation
Sample Space
(i) causal
(ii) correlative
X
B
S
Y
(b) Causal-Guided Active Learning Framework
Causal Invariance Based Biased Instances Identification
causal
causal
Model Representation
M
Counter Instance Pairs: < (X_i, Y_i), (X_j, Y_j) >
Selected Counter Instance Pair
Positive example
Context: Discuss antitrust laws and their...
A: antitrust laws are designed for...
B: antitrust laws are a set of ...
Question: which of the two options is better?
Answer: A
negative example
Context: Discuss antitrust laws and their...
A: antitrust laws are a set of ...
B: antitrust laws are designed for...
Question: which of the two options is better?
Answer: A
Position bias
replace
ICL-based Bias Suppression
Few-shot (Counterfactual ICL)
Please act as an impartial judge and evaluate the quality of the responses... output your final verdict by strictly following this format: ... and [[C]] for a tie.
<negative example 1>
...
<negative example n>
Zero-shot
Please act as an impartial judge and evaluate the quality of the responses... output your final verdict by strictly following this format: ... and [[C]] for a tie. Note that position and verbosity of the responses is not related to the responses' quality:
Informative Biased Instances Selection and Bias Pattern Induction
Typical Criteria
Influential Criteria
select
cluster
summarize
position
verbosity
concat