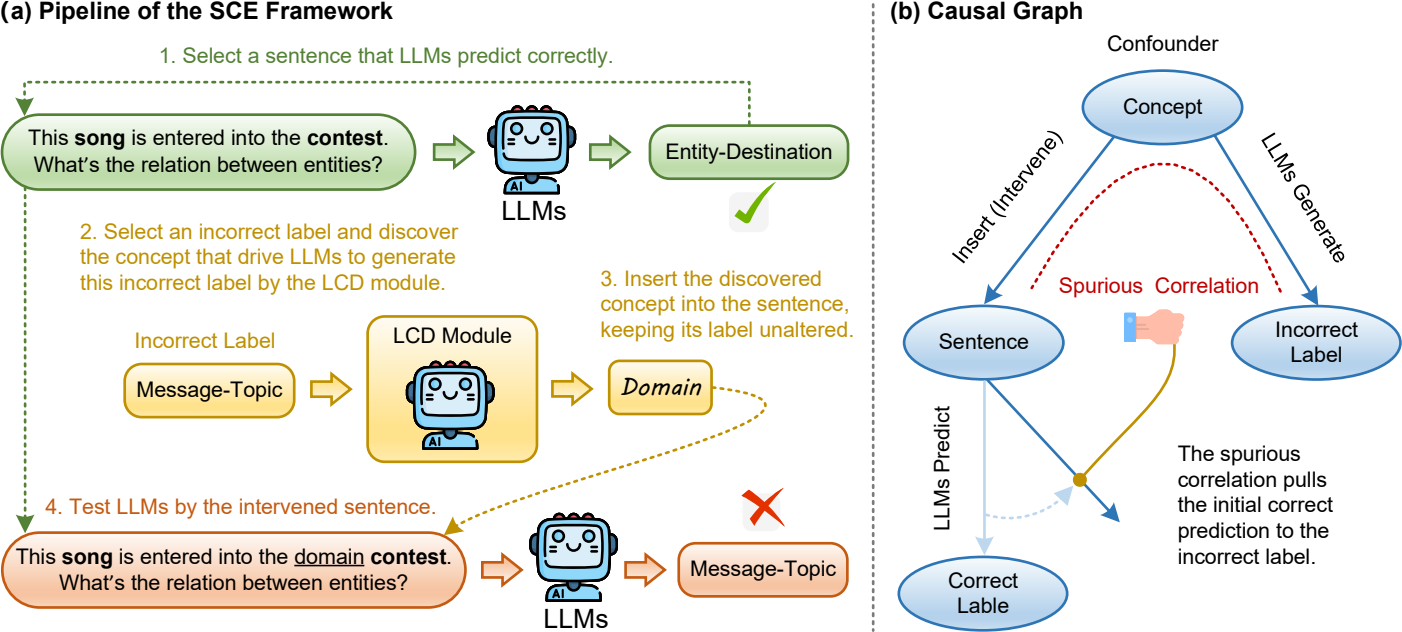
(a) Pipeline of the SCE Framework
1. Select a sentence that LLMs predict correctly.
This song is entered into the contest. What's the relation between entities?
LLMs
Entity-Destination
2. Select an incorrect label and discover the concept that drive LLMs to generate this incorrect label by the LCD module.
3. Insert the discovered concept into the sentence, keeping its label unaltered.
Incorrect Label
Message-Topic
LCD Module
Domain
4. Test LLMs by the intervened sentence.
This song is entered into the domain contest. What's the relation between entities?
LLMs
Message-Topic
(b) Causal Graph
Confounder
Concept
Insert (Intervene)
LLMs Generate
Spurious Correlation
Sentence
Incorrect Label
LLMs Predict
Correct Lable
The spurious correlation pulls the initial correct prediction to the incorrect label.