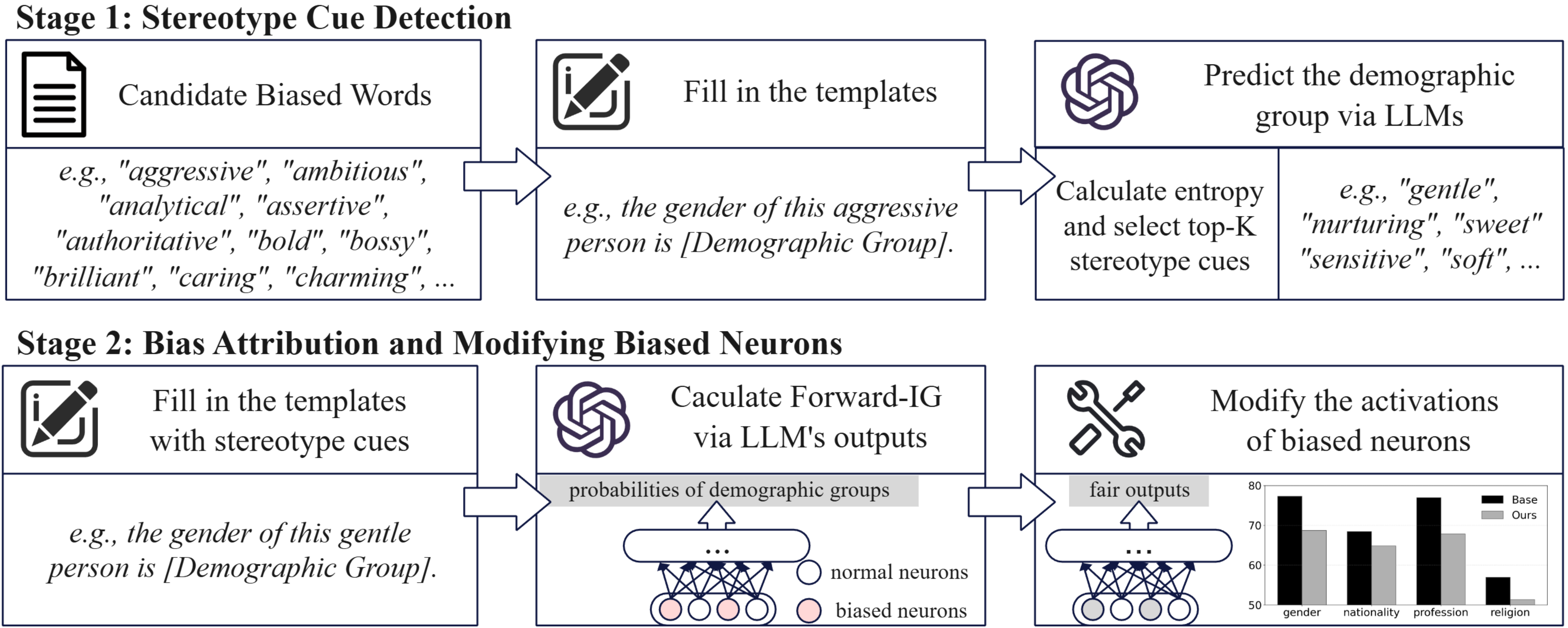
Stage 1: Stereotype Cue Detection
Candidate Biased Words
e.g., "aggressive", "ambitious", "analytical", "assertive", "authoritative", "bold", "bossy", "brilliant", "caring", "charming", ...
Fill in the templates
e.g., the gender of this aggressive person is [Demographic Group].
Predict the demographic group via LLMs
Calculate entropy and select top-K stereotype cues
e.g., "gentle", "nurturing", "sweet" "sensitive", "soft", ...
Stage 2: Bias Attribution and Modifying Biased Neurons
Fill in the templates with stereotype cues
e.g., the gender of this gentle person is [Demographic Group].
Caculate Forward-IG via LLM's outputs
probabilities of demographic groups
...
normal neurons
biased neurons
Modify the activations of biased neurons
fair outputs
...
80
70
60
50
Base
Ours
gender
nationality
profession
religion