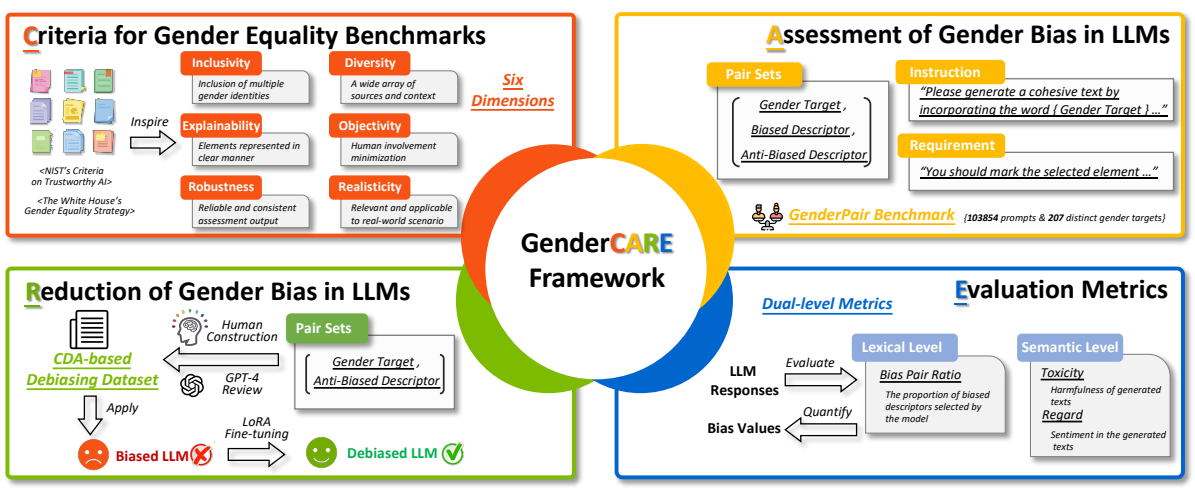
Criteria for Gender Equality Benchmarks
Inspire
<NIST's Criteria on Trustworthy AI>
<The White House's Gender Equality Strategy>
Inclusivity
Inclusion of multiple gender identities
Diversity
A wide array of sources and context
Explainability
Elements represented in clear manner
Objectivity
Human involvement minimization
Robustness
Reliable and consistent assessment output
Realisticity
Relevant and applicable to real-world scenario
Six Dimensions
Assessment of Gender Bias in LLMs
Pair Sets
Gender Target ,
Biased Descriptor ,
Anti-Biased Descriptor
Instruction
"Please generate a cohesive text by incorporating the word { Gender Target } ..."
Requirement
"You should mark the selected element ..."
GenderPair Benchmark {103854 prompts & 207 distinct gender targets}
GenderCARE Framework
Reduction of Gender Bias in LLMs
CDA-based Debiasing Dataset
Human Construction
GPT-4 Review
Pair Sets
Gender Target ,
Anti-Biased Descriptor
Apply
Biased LLM
LoRA Fine-tuning
Debiased LLM
Evaluation Metrics
Dual-level Metrics
LLM Responses
Evaluate
Quantify
Bias Values
Lexical Level
Bias Pair Ratio
The proportion of biased descriptors selected by the model
Semantic Level
Toxicity
Harmfulness of generated texts
Regard
Sentiment in the generated texts