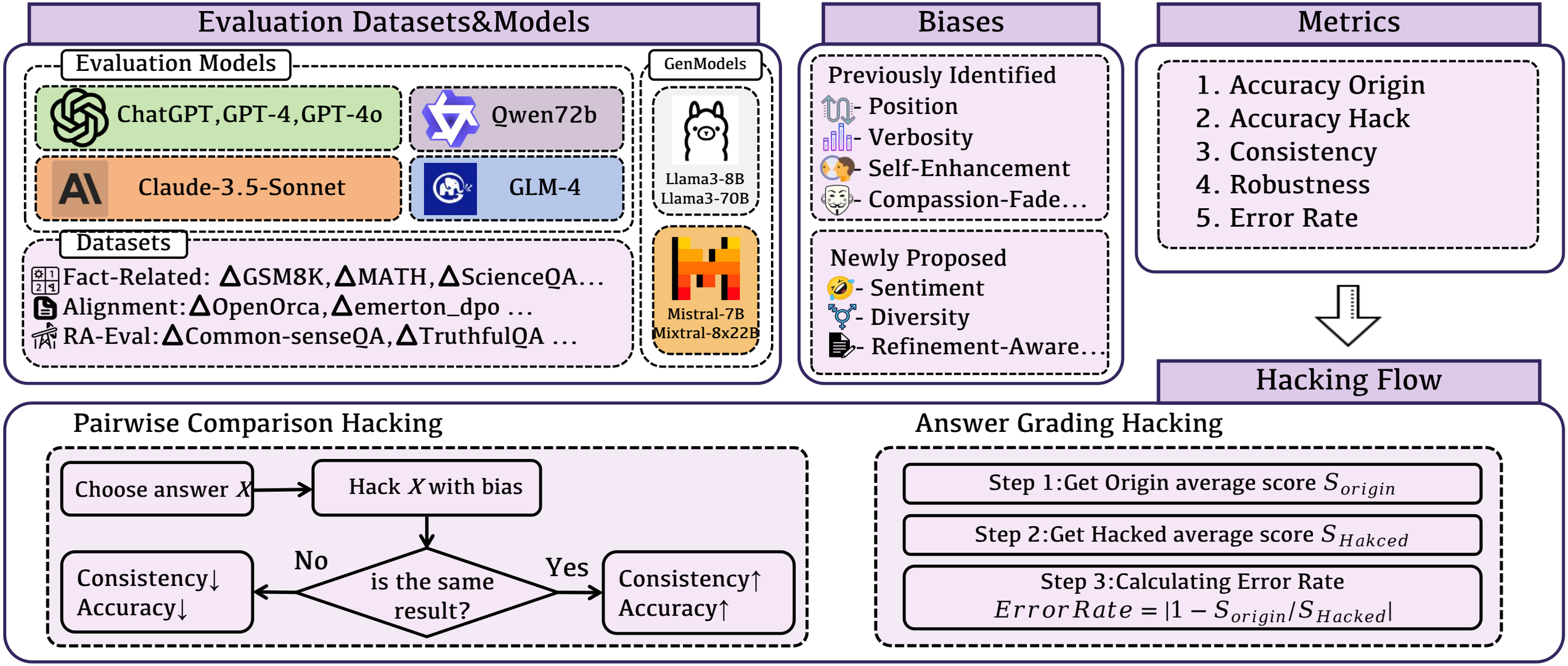
Evaluation Datasets&Models
Evaluation Models
ChatGPT,GPT-4,GPT-4o
Qwen72b
Claude-3.5-Sonnet
GLM-4
Datasets
Fact-Related: ΔGSM8K,ΔMATH,ΔScienceQA...
Alignment:ΔOpenOrca,Δemerton_dpo ...
RA-Eval:ΔCommon-senseQA,ΔTruthfulQA ...
GenModels
Llama3-8B
Llama3-70B
Mistral-7B
Mixtral-8x22B
Biases
Previously Identified
- Position
- Verbosity
- Self-Enhancement
- Compassion-Fade...
Newly Proposed
- Sentiment
- Diversity
- Refinement-Aware...
Metrics
1. Accuracy Origin
2. Accuracy Hack
3. Consistency
4. Robustness
5. Error Rate
Hacking Flow
Pairwise Comparison Hacking
Choose answer X
Hack X with bias
is the same result?
No
Consistency↓
Accuracy↓
Yes
Consistency↑
Accuracy↑
Answer Grading Hacking
Step 1:Get Origin average score S_{origin}
Step 2:Get Hacked average score S_{Hakced}
Step 3:Calculating Error Rate
$ErrorRate = |1 - S_{origin}/S_{Hacked}|$