1 D1:Experimental Setup,Hyperparameters and Error Bars

1.1 Experimental Setup

This work comprises three artifacts: *corpus_evaluation.ipynb* (corpus-level CEAT/I-WEAT/I-SEAT/IIBS over six intersectional classes), *task1.ipynb* (Task 1 RoBERTa-based CEAT, I-WEAT, I-SEAT for CLASS–PROMPT–LLM responses plus LIME), and *Task_2.ipynb* (Task 2 persona–prompt–LLM synthetic scorer producing CEAT, I-WEAT, I-SEAT, and *Combined_Bias* with summary CSVs). All computations run on CPU with *SentenceTransformer("roberta-base")* for embedding-based metrics; no fine-tuning is performed. Inputs include *corpus.csv* (261 sentences, 21 classes), *prompts-task1.txt* (parsed into per-LLM responses), and programmatic grids for Task 2.

1.2 Data

The corpus analysis loads *corpus.csv*, filters six intersectional classes defined in-code (targets and positive/negative attributes), and reports aggregate CEAT/I-WEAT/I-SEAT plus IIBS prevalence; Neutral is used only for prevalence estimates. Task 1 reads CLASS and PROMPT blocks with per-LLM responses from *prompts-task1.txt* and outputs *task1_bias_scores.csv*. Task 2 enumerates 7 personas $\times$ 5 prompts $\times$ 5 LLM labels to yield 175 rows in *task2_bias_scores.csv* alongside persona/LLM/prompt summaries.

1.3 Models and Libraries

RoBERTa embeddings are obtained via *SentenceTransformer("roberta-base")* with normalized embeddings and cosine similarity (*util.cos_sim*). Pandas and NumPy handle I/O and aggregation. LIME is used for local text explanations in separate analysis blocks in Task 1 and Task 2; there is no model training anywhere in the notebooks.

2 Hyperparameters

Anchors and templates for CEAT, I-WEAT, and I-SEAT are hard-coded in the Task 1 notebook; corpus targets and attribute lists are hard-coded in

Table 1: Data and task assets overview.

Asset	Description	Used in
corpus.csv	261 sentences, 21 classes; six intersectional classes with targets and positive/negative attributes defined in notebook	Corpus eval
prompts-task1.txt	CLASS–PROMPT blocks containing per-LLM responses parsed into *task1_bias_scores.csv*	Task 1
Personas A–G	Descriptions spanning marginalized/privileged markers; G is neutral baseline	Task 2
Prompts (5)	Leadership, career success, workplace challenges, tech suitability, adaptation	Task 2
LLM labels (5)	GPT-4o, DeepSeek-R1, LLaMA-4, Claude-3.5-Sonnet, Gemma-2o-8B	Task 2
Outputs	Task 1 and Task 2 CSVs for scores and summaries as written by notebooks	Task 1/2

define_intersectional_targets. Task 2 uses fixed dictionaries for LLM profiles (*ceat_base*, *iweat_base*, *iseat_base*, *variance*), persona multipliers (*ceat_mult*, *iweat_mult*, *iseat_mult*, *intersectional_boost*), and prompt adjustments (*ceat_adj*, *iweat_adj*, *iseat_adj*). CEAT and I-WEAT scores are clamped to $[-1, 1]$; I-SEAT is clamped to $[0, 1]$.

2.1 Anchors and Templates (Task 1)

CEAT uses stereotype vs. anti-stereotype anchors; I-WEAT uses positive vs. negative attributes; I-SEAT uses stereotype vs. anti templates. All lists are defined inline and used without tuning; *SentenceTransformer("roberta-base")* embeddings are normalized for cosine similarity.

3 Error Bars

Task 1 and corpus evaluations are fully deterministic given inputs and contain no stochastic components; hence error bars are not applicable to those results. Task 2 samples a single Gaussian perturbation per cell using a deterministic hash-derived pseudo-seed based on the (persona, prompt, LLM, metric) key; this yields a fixed, reproducible point estimate per combination rather than a distribution. Reported Task 2 results are therefore single-pass deterministic values without error bars.