APPENDIX A
EXPERIMENTAL RESULTS

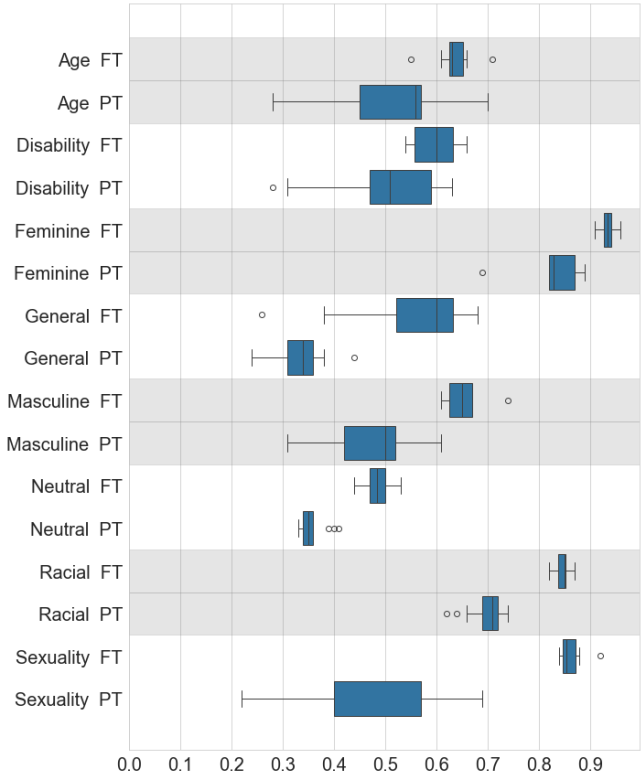

Fig. 6: F_1 Scores Across Various Categories and Model Types

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.66	0.84	0.54
GPT-4o-2024-05-13	pZS	0.70	0.75	0.66
BERT large uncased	FT	0.63	0.87	0.50
RoBERTa base	FT	0.63	0.80	0.51
RoBERTa large	FT	0.61	0.86	0.47
Flan T5 XL	FT	0.71	0.89	0.59
Gemma2-9B	FT	0.65	0.72	0.60
	pZS	0.58	0.72	0.49
	pFS	0.54	0.61	0.49
	pCOT	0.56	0.66	0.49
	pSC	0.57	0.67	0.49
Llama3-8B	FT	0.63	0.91	0.49
	pZS	0.33	0.74	0.21
	pFS	0.56	0.65	0.50
	pCOT	0.45	0.78	0.31
	pSC	0.44	0.76	0.31
Phi3 3.8B 4k	FT	0.55	0.89	0.40
Phi3-7B 8k	pZS	0.64	0.62	0.66
	pFS	0.28	0.93	0.16
	pCOT	0.56	0.47	0.70
	pSC	0.57	0.47	0.71

TABLE VI: Results for Bias Category: Age

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.54	0.79	0.41
GPT-4o-2024-05-13	pZS	0.59	0.62	0.56
BERT large uncased	FT	0.56	0.91	0.40
RoBERTa base	FT	0.63	0.87	0.50
RoBERTa large	FT	0.60	0.88	0.45
Flan T5 XL	FT	0.55	0.89	0.40
Gemma2-9B	FT	0.66	0.95	0.50
	pZS	0.49	0.74	0.36
	pFS	0.38	0.68	0.26
	pCOT	0.53	0.82	0.39
	pSC	0.51	0.81	0.38
Llama3-8B	FT	0.64	0.97	0.47
	pZS	0.42	0.64	0.31
	pFS	0.63	0.60	0.66
	pCOT	0.50	0.68	0.40
	pSC	0.52	0.70	0.41
Phi3 3.8B 4k	FT	0.60	0.97	0.44
Phi3-7B 8k	pZS	0.47	0.72	0.35
	pFS	0.31	0.88	0.19
	pCOT	0.62	0.66	0.59
	pSC	0.63	0.67	0.59

TABLE VII: Results for Bias Category: Disability

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.94	0.95	0.94
GPT-4o-2024-05-13	pZS	0.85	0.87	0.82
BERT large uncased	FT	0.94	0.95	0.93
RoBERTa base	FT	0.93	0.93	0.94
RoBERTa large	FT	0.96	0.97	0.94
Flan T5 XL	FT	0.91	0.92	0.90
Gemma2-9B	FT	0.95	0.99	0.91
	pZS	0.88	0.80	0.97
	pFS	0.81	0.70	0.95
	pCOT	0.83	0.73	0.96
	pSC	0.83	0.73	0.96
Llama3-8B	FT	0.92	0.99	0.86
	pZS	0.84	0.85	0.84
	pFS	0.69	0.55	0.95
	pCOT	0.82	0.79	0.85
	pSC	0.82	0.79	0.85
Phi3 3.8B 4k	FT	0.93	0.99	0.89
Phi3-7B 8k	pZS	0.87	0.81	0.95
	pFS	0.83	0.73	0.96
	pCOT	0.89	0.84	0.95
	pSC	0.89	0.84	0.95

TABLE VIII: Results for Bias Category: Feminine

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.38	0.90	0.24
GPT-4o-2024-05-13	pZS	0.31	0.57	0.21
BERT large uncased	FT	0.60	0.88	0.45
RoBERTa base	FT	0.62	0.75	0.53
RoBERTa large	FT	0.26	0.65	0.16
Flan T5 XL	FT	0.68	0.79	0.60
Gemma2-9B	FT	0.60	0.84	0.46
	pZS	0.44	0.35	0.59
	pFS	0.36	0.25	0.68
	pCOT	0.37	0.26	0.69
	pSC	0.38	0.26	0.69
Llama3-8B	FT	0.67	0.82	0.56
	pZS	0.36	0.30	0.46
	pFS	0.31	0.22	0.54
	pCOT	0.33	0.37	0.30
	pSC	0.33	0.37	0.30
Phi3 3.8B 4k	FT	0.57	0.65	0.51
Phi3-7B 8k	pZS	0.27	0.23	0.35
	pFS	0.24	0.20	0.31
	pCOT	0.35	0.28	0.49
	pSC	0.36	0.28	0.50

TABLE IX: Results for Bias Category: General

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.47	0.32	0.86
GPT-4o-2024-05-13	pZS	0.33	0.21	0.71
BERT large uncased	FT	0.47	0.32	0.88
RoBERTa base	FT	0.50	0.38	0.72
RoBERTa large	FT	0.48	0.33	0.89
Flan T5 XL	FT	0.53	0.37	0.93
Gemma2-9B	FT	0.49	0.34	0.86
	pZS	0.35	0.48	0.28
	pFS	0.27	0.50	0.19
	pCOT	0.39	0.56	0.30
	pSC	0.41	0.60	0.31
Llama3-8B	FT	0.50	0.35	0.90
	pZS	0.34	0.22	0.70
	pFS	0.40	0.31	0.56
	pCOT	0.35	0.23	0.81
	pSC	0.35	0.23	0.81
Phi3 3.8B 4k	FT	0.44	0.30	0.90
Phi3-7B 8k	pZS	0.36	0.29	0.47
	pFS	0.36	0.23	0.81
	pCOT	0.33	0.30	0.35
	pSC	0.34	0.32	0.36

TABLE XI: Results for Bias Category: Neutral

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.61	0.72	0.54
GPT-4o-2024-05-13	pZS	0.52	0.74	0.40
BERT large uncased	FT	0.67	0.82	0.56
RoBERTa base	FT	0.67	0.78	0.59
RoBERTa large	FT	0.64	0.76	0.55
Flan T5 XL	FT	0.74	0.83	0.68
Gemma2-9B	FT	0.66	0.65	0.66
	pZS	0.43	0.29	0.86
	pFS	0.51	0.36	0.86
	pCOT	0.50	0.35	0.90
	pSC	0.52	0.36	0.91
Llama3-8B	FT	0.63	0.62	0.65
	pZS	0.31	0.94	0.19
	pFS	0.42	0.73	0.30
	pCOT	0.35	0.75	0.23
	pSC	0.33	0.74	0.21
Phi3 3.8B 4k	FT	0.61	0.95	0.45
Phi3-7B 8k	pZS	0.50	0.54	0.47
	pFS	0.43	0.71	0.31
	pCOT	0.53	0.58	0.49
	pSC	0.53	0.58	0.49

TABLE X: Results for Bias Category: Masculine

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.85	0.89	0.81
GPT-4o-2024-05-13	pZS	0.64	0.95	0.49
BERT large uncased	FT	0.87	0.87	0.86
RoBERTa base	FT	0.82	0.83	0.81
RoBERTa large	FT	0.85	0.87	0.84
Flan T5 XL	FT	0.83	0.89	0.79
Gemma2-9B	FT	0.86	0.90	0.82
	pZS	0.70	0.81	0.62
	pFS	0.67	0.61	0.74
	pCOT	0.69	0.78	0.62
	pSC	0.69	0.78	0.62
Llama3-8B	FT	0.84	0.89	0.80
	pZS	0.66	0.89	0.53
	pFS	0.73	0.76	0.70
	pCOT	0.71	0.89	0.59
	pSC	0.71	0.90	0.59
Phi3 3.8B 4k	FT	0.85	0.93	0.79
Phi3-7B 8k	pZS	0.73	0.91	0.61
	pFS	0.62	0.95	0.46
	pCOT	0.72	0.86	0.62
	pSC	0.74	0.89	0.62

TABLE XII: Results for Bias Category: Racial

Model	Type	F_1	Precision	Recall
BASELINES				
BERT base uncased	FT	0.87	0.94	0.81
GPT-4o-2024-05-13	pZS	0.40	1.00	0.25
BERT large uncased	FT	0.85	0.97	0.76
RoBERTa base	FT	0.92	0.96	0.89
RoBERTa large	FT	0.86	0.97	0.78
Flan T5 XL	FT	0.88	0.96	0.81
Gemma2-9B	FT	0.85	0.97	0.75
	pZS	0.57	0.97	0.40
	pFS	0.62	0.79	0.51
	pCOT	0.57	0.97	0.40
	pSC	0.57	0.97	0.40
Llama3-8B	FT	0.84	0.95	0.75
	pZS	0.28	0.93	0.16
	pFS	0.55	0.97	0.39
	pCOT	0.22	1.00	0.12
	pSC	0.22	1.00	0.12
Phi3 3.8B 4k	FT	0.84	0.95	0.75
Phi3-7B 8k	pZS	0.64	0.70	0.60
	pFS	0.56	0.89	0.41
	pCOT	0.69	0.79	0.61
	pSC	0.69	0.79	0.61

TABLE XIII: Results for Bias Category: Sexuality

APPENDIX B
PROMPT EXAMPLES

Job descriptions are truncated for brevity. The **Green text** is the large language model output.

PROMPT 3: Zero-Shot (pZS) - Phi3-8k-Instruct

USER **You are an analyst of job postings.**
Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, masculine, racial, sexuality, general, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, general, masculine, racial, sexuality. If no biased language is present, please return zero labels. No further explanation is required.

ASSISTANT **Labels: general**

PROMPT 4: Zero Shot (pZS) - gemma-2-9b-it

USER **You are an analyst of job postings. Your task is to detect implicit bias, if present.**
Analyse the job posting for implicit bias. The bias categories are age, disability, . . . , general. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary . . .
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT **Labels: general**

PROMPT 5: Few-Shot (pFS) - Meta-Llama-3-8B-Instruct

SYSTEM You are an analyst of job postings. Your task is to detect implicit bias, if present.

USER **Analyse the job posting for implicit bias. The bias categories are age, disability, feminine, masculine, racial, sexuality, general. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.**
Job Posting:
About the business \nLumber Punks Axe Throwing is more than an activity, it's a movement. We bring the Punk ethos to everything we do and make sure every guest ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT **Labels: age, disability, masculine, general**

USER Job Posting:
Job Title: Unix System Administrator \nCompany: Dean Ltd \nLocation: Malta \nAbout Us:\nDean Ltd is a leading technology company based in Malta, providing ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT **Labels: age, racial, sexuality**

USER Job Posting:
5-7 years of relevant work experience. \nExcellent verbal and written communication skills. \nOutgoing and proactive personality. \nIdeally ...
===END===

Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: neutral

USER Job Posting:
Experienced Heavy Rigid Drivers \nABOUT THE CLIENT \nAction Workforce has partnered up with a very reputable company based in South Brisbane. This company ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: neutral

USER Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: general

PROMPT 6: Few-Shot (pFS) - Phi3-8k-Instruct

USER **You are an analyst of job postings.**
Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, masculine, racial, sexuality, general, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
5-7 years of relevant work experience. \nExcellent verbal and written communication skills. \nOutgoing and proactive personality. \nIdeally ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: neutral

USER You are an analyst of job postings.
Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, masculine, racial, sexuality, general, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
About the business \nLumber Punks Axe Throwing is more than an activity, it's a movement. We bring the Punk ethos to everything we do and make sure every guest ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: age, disability, masculine, general

USER You are an analyst of job postings.
Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, masculine, racial, sexuality, general, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
Experienced Heavy Rigid Drivers \nABOUT THE CLIENT \nAction Workforce has partnered up with a very reputable company based in South Brisbane. This company ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: neutral

USER You are an analyst of job postings.

Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, masculine, racial, sexuality, general, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
About Ryman Retirement Villages\nRyman Healthcare's Jane Winstone Retirement Village has a strong family and community focus. With a welcoming atmosphere, our village is filled with people ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: feminine, general

USER You are an analyst of job postings.
Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, masculine, racial, sexuality, general, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: general

PROMPT 7: Few-Shot (pFS) - gemma-2-9b-it

USER **You are an analyst of job postings. Your task is to detect implicit bias, if present.**
Analyse the job posting for implicit bias. The bias categories are age, disability, feminine, masculine, racial, sexuality, general. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Job Posting:
5-7 years of relevant work experience. \nExcellent verbal and written communication skills. \nOutgoing and proactive personality. \nIdeally ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: neutral

USER Job Posting:
Job Title: Unix System Administrator \nCompany: Dean Ltd \nLocation: Malta \nAbout Us:\nDean Ltd is a leading technology company based in Malta, providing ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: age, racial, sexuality

USER Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure ...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: general

PROMPT 8: Chain-of-Thought (pCoT) - Meta-Llama-3-8B-Instruct

SYSTEM You are an analyst of job postings. Your task is to detect implicit bias, if present.

USER **Analyse the job posting for implicit bias. The bias categories are age, disability, feminine, general, masculine, racial, sexuality, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.**

Think step by step to determine if there is implicit bias present in the job posting.
Step 1: Check for age bias.
Step 2: Check for disability bias.
Step 3: Check for feminine bias.
Step 4: Check for masculine bias.
Step 5: Check for racial bias.
Step 6: Check for sexuality bias.
Step 7: Check for general bias - bias other than the previous bias categories.
Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: general

PROMPT 9: Chain-of-Thought (pCoT) - Phi3-8k-Instruct

USER **You are an analyst of job postings.**
Your job is to analyse a job posting and determine if it contains implicit bias towards or against certain groups. The bias categories are age, disability, feminine, general, masculine, racial, sexuality, neutral. Return the appropriate label only when absolutely sure that the particular bias category is present in the posting.
Think step by step to determine if there is implicit bias present in the job posting.
Step 1: Check for age bias.
Step 2: Check for disability bias.
Step 3: Check for feminine bias.
Step 4: Check for masculine bias.
Step 5: Check for racial bias.
Step 6: Check for sexuality bias.
Step 7: Check for general bias - bias other than the previous bias categories.
Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no biased language is present, please return zero labels. No further explanation is required.

ASSISTANT Labels: general

PROMPT 10: Chain-of-Thought (pCoT) - gemma-2-9b-it

USER **You are an analyst of job postings. Your task is to detect implicit bias, if present.**
Analyse the job posting for implicit bias. The bias categories are age, disability, feminine, masculine, racial, sexuality, general. Return the appropriate label only when absolutely sure that the particular bias category is present in the text.
Think step by step to determine if there is implicit bias present in the job posting.
Step 1: Check for age bias.
Step 2: Check for disability bias.
Step 3: Check for feminine bias.
Step 4: Check for masculine bias.
Step 5: Check for racial bias.
Step 6: Check for sexuality bias.
Step 7: Check for general bias - bias other than the previous bias categories.
Job Posting:
Role: Graduate Trainee Recruitment Consultant. Location: Bristol. Package: £20-24K Basic Salary. Uncapped Commission Structure...
===END===
Does the job posting contain any implicit bias? Please respond in the format of 'Labels: <labels>' where the possible labels are zero or more from the choices of age, disability, feminine, masculine, racial, sexuality, general. If no bias is detected please return 'Labels: neutral'. No further explanation is required.

ASSISTANT Labels: general

APPENDIX C
SOFTWARE, MODEL CARDS AND DATASETS

This appendix provides a comprehensive list of the software, models, and datasets utilised in the study. See additional references at the end of this appendix section.

A. Datasets

1) Data Gathering: See Appendix E, Figure 8 for the pipeline that gathers these datasets.

- https://www.kaggle.com/datasets/techmap/job-postings-ireland-october-2020
- https://www.kaggle.com/datasets/techmap/job-postings-ireland-october-2021
- https://www.kaggle.com/datasets/techmap/job-postings-ireland-october-2022
- https://www.kaggle.com/datasets/techmap/international-job-postings-september-2021
- https://www.kaggle.com/datasets/techmap/us-job-postings-from-2023-05-05

2) Cleaning & Preprocessing: See Appendix E, Figures 7 and 13 for the pipelines create these datasets.

- https://huggingface.co/datasets/2024-mcm-everitt-ryan/job-postings-raw
- https://huggingface.co/datasets/2024-mcm-everitt-ryan/job-postings-english-clean
- https://huggingface.co/datasets/2024-mcm-everitt-ryan/possible-bias

3) Fine-Tuning and Prompting:

- https://huggingface.co/datasets/2024-mcm-everitt-ryan/benchmark

4) Augmented Taxonomy Bias Terms: See Appendix E, Figure 12 for the pipeline that augments the terms.

- https://huggingface.co/datasets/2024-mcm-everitt-ryan/augmented-taxonomy-terms

5) Trichotomy Phrases:

- https://huggingface.co/datasets/2024-mcm-everitt-ryan/trichotomy-phrases

B. Models

- Google's BERT [1]
- Facebook AI's RoBERTa [2].
- Google's Flan-T5 [3].
- Microsoft's Phi3 [4].
- Meta's LLama3 [5].
- Google's Gemma2 [6].

C. Software: Data Gathering, Cleaning, & Preprocessing, Evaluating

1) Apache HOP resources: See Appendix E for pipeline/workflow visuals.

- Extract, transform, load (ETL): https://github.com/apache/hop
- Potential Bias Filtering: Custom HOP plugin using Stanford CoreNLP [7]. https://github.com/2024-mcm-everitt-ryan/hop-plugins
- HOP Pipeline/Workflow: https://github.com/2024-mcm-everitt-ryan/hop-pipeline

2) Dataset Preprocessing: See Appendix E, Figure 9 for the workflow that preprocesses the dataset. See Appendix E, Figure 10 for the pipeline that detects language. See Appendix E, Figure 11 for the pipeline that cleans HTML and removes accents and diacritical marks.

- Language Detection: https://github.com/pemistahl/lingua
- HTML Cleaning: https://github.com/jhy/jsoup
- Removing accents and diacritical marks: https://commons.apache.org/proper/commons-lang/apidocs/org/apache/commons/lang3/StringUtils.htmlstripAccents-java.lang.String-
- Deduplication: *all-MiniLM-L6-v2* [8] and Facebook AI Similarity Search (FAISS) [9].

3) Data Labelling: See Appendix E, Figure 14 for the workflow that processes annotated Label Studio exports to create a verified annotated dataset.

- https://github.com/HumanSignal/label-studio
- https://github.com/2024-mcm-everitt-ryan/label-studio

4) Docker Container:

- https://github.com/2024-mcm-everitt-ryan/docker
- https://hub.docker.com/r/ep9io/2024-mcm-everitt-ryan

5) Fine-Tuning and Prompting:

- DSPY (experimental) [10].
- Python (3.10)
- Hugging Face Transformers (4.42.3)
- Nvidia CUDA (12.1.1)
- PyTorch (2.2.0)
- Ubuntu (22.04)
- Fine-tuning encoders for multi-label problems [11].

6) Evaluation Metrics:

- SciKit-Learn implementations of precision, recall, F-score, and the exact match ratio [12].

7) Miscellaneous:

- Typesetting: LaTeX
- Graphics: Matplotlib [13] and Seaborn [14].

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186.

[2] Y. Liu, M. Ott, N. Goyal, *et al.*, *RoBERTa: A robustly optimized bert pretraining approach*, https://arxiv.org/abs/1907.11692, 2019. arXiv: 1907.11692 [cs.CL].

[3] H. W. Chung, L. Hou, S. Longpre, *et al.*, *Scaling instruction-finetuned language models*, https://arxiv.org/abs/2210.11416, 2022. arXiv: 2210.11416 [cs.LG].

[4] M. Abdin, S. A. Jacobs, A. A. Awan, *et al.*, *Phi-3 technical report: A highly capable language model locally on your phone*, https://arxiv.org/abs/2404.14219, 2024. arXiv: 2404.14219 [cs.CL].

[5] AI@Meta, *Llama 3 model card*, https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024.

[6] G. Team, "Gemma," 2024.

[7] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[8] Hugging Face, *Sentence-transformers/all-minilm-l6-v2*, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, Accessed: 2024-07-28, 2024.

[9] M. Douze, A. Guzhva, C. Deng, *et al.*, "The faiss library," 2024. arXiv: 2401.08281 [cs.LG].

[10] O. Khattab, A. Singhvi, P. Maheshwari, *et al.*, "Dspy: Compiling declarative language model calls into self-improving pipelines," *arXiv preprint arXiv:2310.03714*, 2023.

[11] N. Rogge, *Fine-tuning bert (and friends) for multi-label text classification*, https://github.com/NielsRogge/Transformers-Tutorials/blob/master/BERT/Fine_tuning_BERT_(and_friends)_for_multi_label_text_classification.ipynb, 2021.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[14] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

APPENDIX D
REGULAR EXPRESSIONS

See Appendix E, Figure 11 for the order execution of the HTML-to-Text regular expressions.

A. HTML-to-Text: Fix tags

```
import static org.apache.commons.lang3.StringUtils.replaceEachRepeatedly

html = job_posting_html

String[] searchList = [
" ",
"&lt;",
"&gt;",
```

```
"DefSemiHidden=\"",
"UnhideWhenUsed=\"",
"\" Name=\"",
"\"/> Name=\"",
" classid=\"",
"table.MsoNormalTable {",
"\\n",
"\\xc2\\xa0",
"\\xe2\\x80\\x99",
"/n",
"/xc2/xa0",
"/xe2/x80/x99",
"\\xc2\\xb7",
"/xc2/xb7"
]

String[] replacementList = [
" ",
"<",
">",
"<Name=\"",
"<UnhideWhenUsed=\"",
"\"/><Name=\"",
"\"/><Name=\"",
"<classid=\"",
"<style>table.MsoNormalTable {",
"<br/>",
" ",
"'",
"<br/>",
" ",
"'",
"  ",
"  "
]

//"&",
//"\"",
//"'"

html = replaceEachRepeatedly(html, searchList, replacementList)

html = html.replaceAll("<{2,}", "<")
```

B. HTML-to-Text: remove excess whitespace

```
import static org.apache.commons.lang3.StringUtils.replaceEachRepeatedly

html = html.replaceAll("\\s{2,}", " ")
html = html.replaceAll(">\\s+<", "><");

String[] searchList = [
" ",
" </p>",
" <p>",
" <li>",
" </li>",
" <div>",
" </div>",
" <i>",
" </i>",
" <b>",
" </b>"
]

String[] replacementList = [
```

```
" ",
"</p>",
"<p>",
"<li>",
"</li>",
"<div>",
"</div>",
"<i>",
"</i>",
"<b>",
"</b>"
]

html = replaceEachRepeatedly(html, searchList, replacementList)
```

C. HTML-to-Text: add periods

```
opening_tags = ["br","ol","ul","li","div","p"]
closing_tags = ["li", "div", "p"]


//for (String tag : opening_tags) {
   // Regular Expression Explanation:
   // (?i) - Case-insensitive matching
   // (</" + tag + ">) - Matches the closing tag (e.g., </li>, </div>, </p>)
   // (.*?) - Matches any characters (non-greedily) before the closing tag
   // (?<!\\.) - Negative lookbehind to check that the last character is not already a period
  // html = html.replaceAll("(?i)([^.])<" + tag + ">", "\$1.<" + tag + ">");
//}

for (String tag : closing_tags) {
   // Regular Expression Explanation:
   // (?i) - Case-insensitive matching
   // (</" + tag + ">) - Matches the closing tag (e.g., </li>, </div>, </p>)
   // (.*?) - Matches any characters (non-greedily) before the closing tag
   // (?<!\\.) - Negative lookbehind to check that the last character is not already a period
   html = html.replaceAll("(?i)([^.])</" + tag + ">", "\$1.</" + tag + ">");

   // "(?i)(?<!\\.)((.*?)</" + tag + ">)", "\$2.</" + tag + ">"

}

 html = html.replaceAll("(?i)([^.])<br>", "\$1.<br/>");
 html = html.replaceAll("(?i)([^.])<br/>", "\$1.<br/>");
```

D. HTML-to-Text: remove excess dots

```
import static org.apache.commons.lang3.StringUtils.replaceEachRepeatedly

text = text.replaceAll("\\.\\s+\\.", "\\.");

String[] searchList = [
".."
]

String[] replacementList = [
"."
]

text = replaceEachRepeatedly(text, searchList, replacementList)
```

E. HTML-to-Text: Split by bad full-stop

```
/*
Split these examples into two sentences:
```

```
with dollies and cartsThis is a safety-sensitive position and
with dollies and cartsThis, is a safety-sensitive position and
This is a sentence with Abc.It should not be joined.
This is a sentence with a full-stop.A should be split into two.
judgment and quality de-cisionsAch-ieve goals, handle assigned workload
management philosophy and decisionsCollaborates with City Manager's Office

(
,?\s
(?:I|A|[A-Z]?[a-z]+-[a-z]{2,}|[A-Z]?[a-z]+)
[.!?]?
)
(
(?:I|A|[A-Z][a-z]+-[a-z]{2,}|[A-Z][a-z]+)
,?\s
)
*/

regex="(,?\s(?:I|A|[A-Z]?[a-z]+-[a-z]{2,}|[A-Z]?[a-z]+)[.!?]?)((?:I|A|[A-Z][a-z]+-[a-z]{2,}|[A-
    Z][a-z]+),?\s)"

text = text.replaceAll(regex, '$1. $2')
```

F. HTML-to-Text: Split by semi-colon

```
def separator = '018cd912-15d4-7d59-b6c9-92932f443496'
def wordWindow = 3

//regex = "((?:I|A|[A-Z]?[a-z]+-[a-z]{2,}|[A-Z]?[a-z]+|,?\\s?){$wordWindow});\\s?((?:I|A|[A-Z
    ]?[a-z]+-[a-z]{2,}|[A-Z]?[a-z]+|,?\\s?){$wordWindow})"
regex = "((?:[A-Z][a-z]*|[a-z]+-[a-z]{2,}|,?\\s?){$wordWindow});\\s?((?:[A-Z][a-z]*|[a-z]+-[a-z
    ]{2,}|,?\\s?){$wordWindow})"

text = text.replaceAll(regex, '$1' + separator + '$2')
text = text.split(separator).collect {it.trim().capitalize()}.join('. ')
```

G. HTML-to-Text: Split by bullet points

```
regex="([a-z]?)((?:\\*|\\+|\\-||)\\s+)((?:[A-Z]?[a-z]+-[a-z]{2,}|[A-Z]?[a-z]+|[0-9]+),?\\s)"
text = text.replaceAll(regex, '$1. $3')
```

H. HTML-to-Text: Split by CAPITAL LETTERS

```
// The dilemma:
// CLIENT RELATIONS ENERGY COORDINATORThe job role requires energy coordination.
// CLIENT RELATIONS ENERGY COORDINATORthe job role requires energy coordination.
//
// GPT-4 is able to handle that split.
// Conventional way not so easy. Perhaps attempt both regex and test which one is most probable
     by
// looking up the words in the dictionary. Two dictionary valid words means it is highly
    probable.

//separator = '018cd912-15d4-7d59-b6c9-92932f443496'

//regex = "([A-Z]{5})([A-Z][a-z]{2,}:?,?\\s)" // Include last capital for group 2
//regex = "([A-Z]{5})([a-z]{2,}:?,?\\s)" // Don't include last capital for group 2
//text = text.replaceAll(regex2, '$1' + separator + '$2')

//text = text.split(separator).collect {it.trim().capitalize()}.join('. ')
```

I. HTML-to-Text: Clean text

```
import static org.apache.commons.lang3.StringUtils.*
import java.util.regex.Pattern

static String clean(v) {

      v = trimToEmpty(v)

      def email = v.contains("@")
      if(email) {
      //    println(v)
      }

      def emailPattern = ~/\b[A-Za-z0-9._%+-]+(@[A-Za-z0-9.-]+\.\s*[A-Za-z]{2,7})\b/

      // Replace the left part of the email address with "***"
      v = v.replaceAll(emailPattern) { match, domain ->
            domain = domain.replaceAll("\\s+", "")
            "***${domain}"
      }

      if(email) {
      //    println(v)
      }

      v = v.replaceAll("\\s+", " ") // Reduce multiple spaces to single space
      v = v.replaceAll("\\s*([,.!?])\\s*", "\$1 ") // Fix punctuation spacing
      v = v.replaceAll("\\s([,.!?])", "\$1") // Remove space before punctuation
      v = v.replaceAll("([,.!?])\\s*\$", "\$1") // Remove trailing space after punctuation

      // Remove excess special characters, mostly used as dividers or attention symbols (e.g.
          #)
      v = v.replaceAll("[`|\"%^&*\\-_=+\\[\\{\\]\\};:'@#~/\\\\|<>]{2,}", "")

      // Reduce excess punctiations to 1
      v = v.replaceAll("([\\?\$!]){2,}", "\$1")

      // Replace punctuation-dot-space ":. " with punctuation-space ": "
      [":", ".","!","?"].eachWithIndex { e, idx ->
            v = v.replaceAll("\\"+e+"\\.\\s", e +" ")
      }

      v = v.replaceAll("(\\d+\\.)\\s(\\d+)", "\$1\$2") // Join digits that are split by a dot
          separator.

      if (v && (v[0] =~ /[`|"%^&*\\\-_=+\[\{\]\};:'@#~\/!]/)) {
         v = v.substring(1).trim()
      }

      // Removing accents and diacritical marks
      // Apache commons might not normalise all ligatures
      // https://www.baeldung.com/java-remove-accents-from-text
      v = stripAccents(v)

      // Keep only the necessary characters and symbols to simplify the text further
      def pattern = "[^0-9a-zA-Z.,;:!?\\\$\\/()\\[\\]{}\\- \"%&']"
      v = v.replaceAll(pattern, "")

      v = trimToEmpty(v)

      // Check if the text ends with a period, if not, add one
      if (!v.isEmpty() && !v.endsWith('.')) {
         v += '.'
      }

      return v
```

```
}

text = clean(text)
```

J. HTML-to-Text: clean_html plugin

https://github.com/2024-mcm-everitt-ryan/hop-plugins/tree/main/plugin-code/html2text-transform

K. HTML-to-Text: html to text plugin

https://github.com/2024-mcm-everitt-ryan/hop-plugins/tree/main/plugin-code/html2text-transform

L. HTML-to-Text: Sentences

https://github.com/2024-mcm-everitt-ryan/hop-plugins/tree/main/plugin-code/stanford-simple-nlp-transform

APPENDIX E
APACHE HOP PIPELINE/WORKFLOW

Apache HOP workflows/pipelines used in the study available at:

https://github.com/2024-mcm-everitt-ryan/hop-pipeline

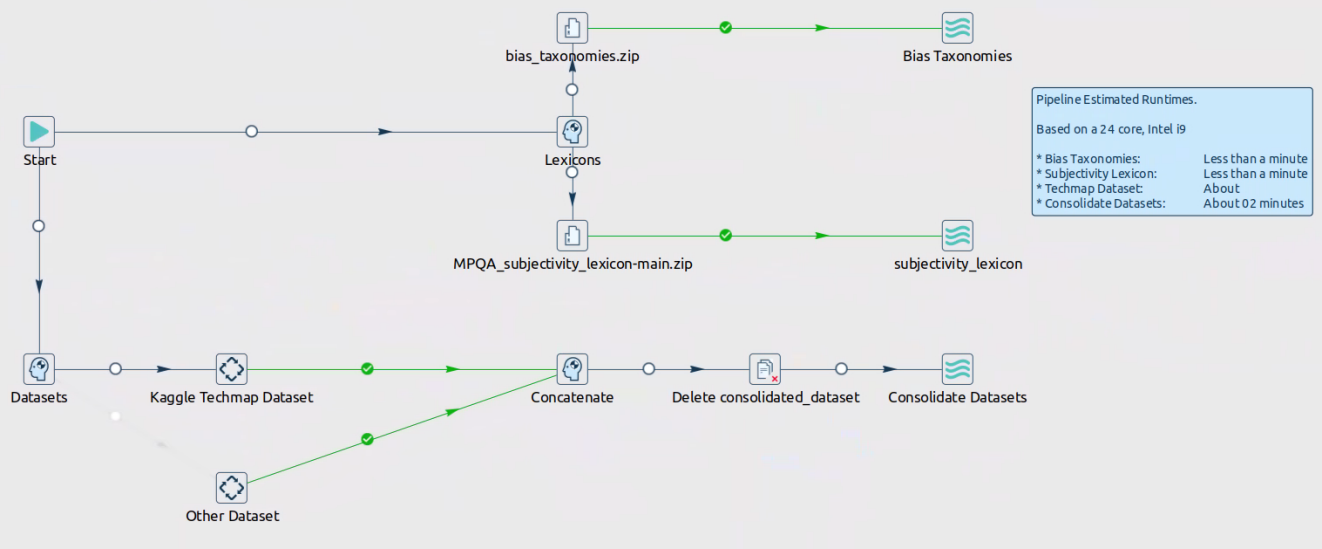

Fig. 7: Workflow: Data Gathering

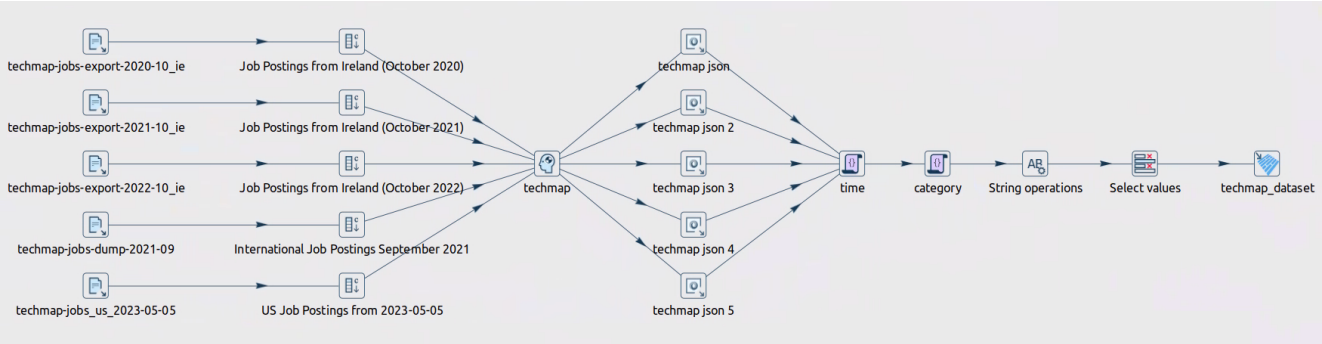

Fig. 8: Workflow: Kaggle Techmap Dataset

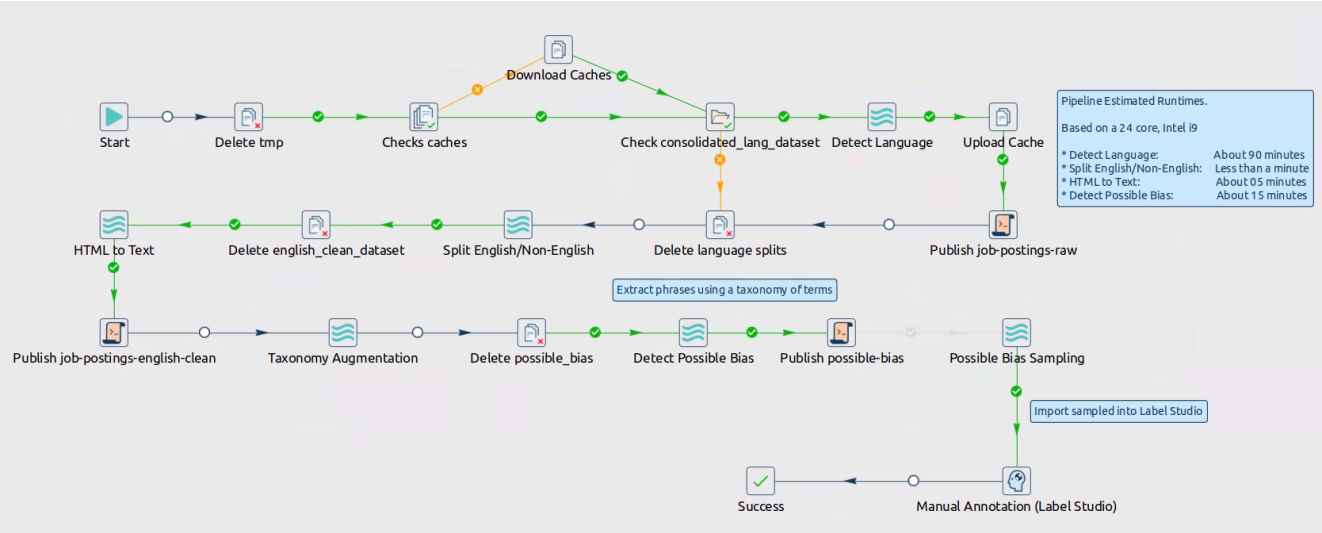

Fig. 9: Workflow: Data Processing

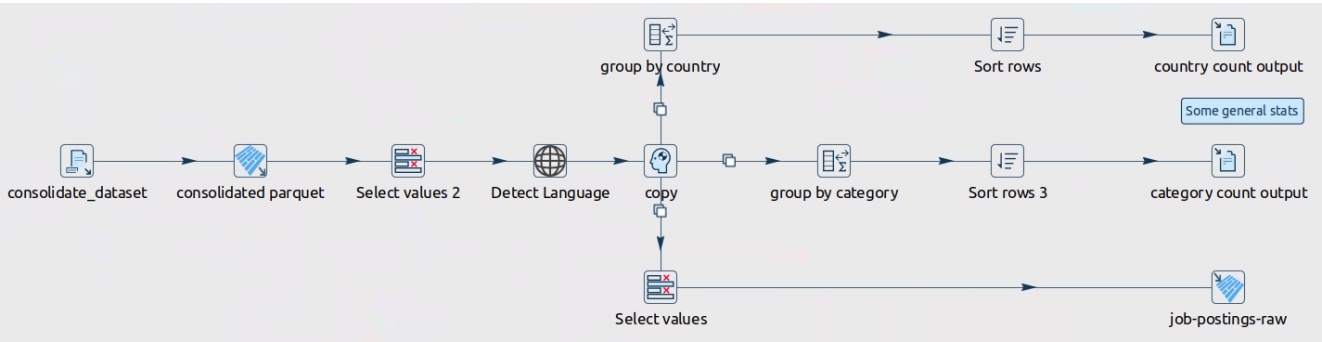

Fig. 10: Pipeline: Detect Language

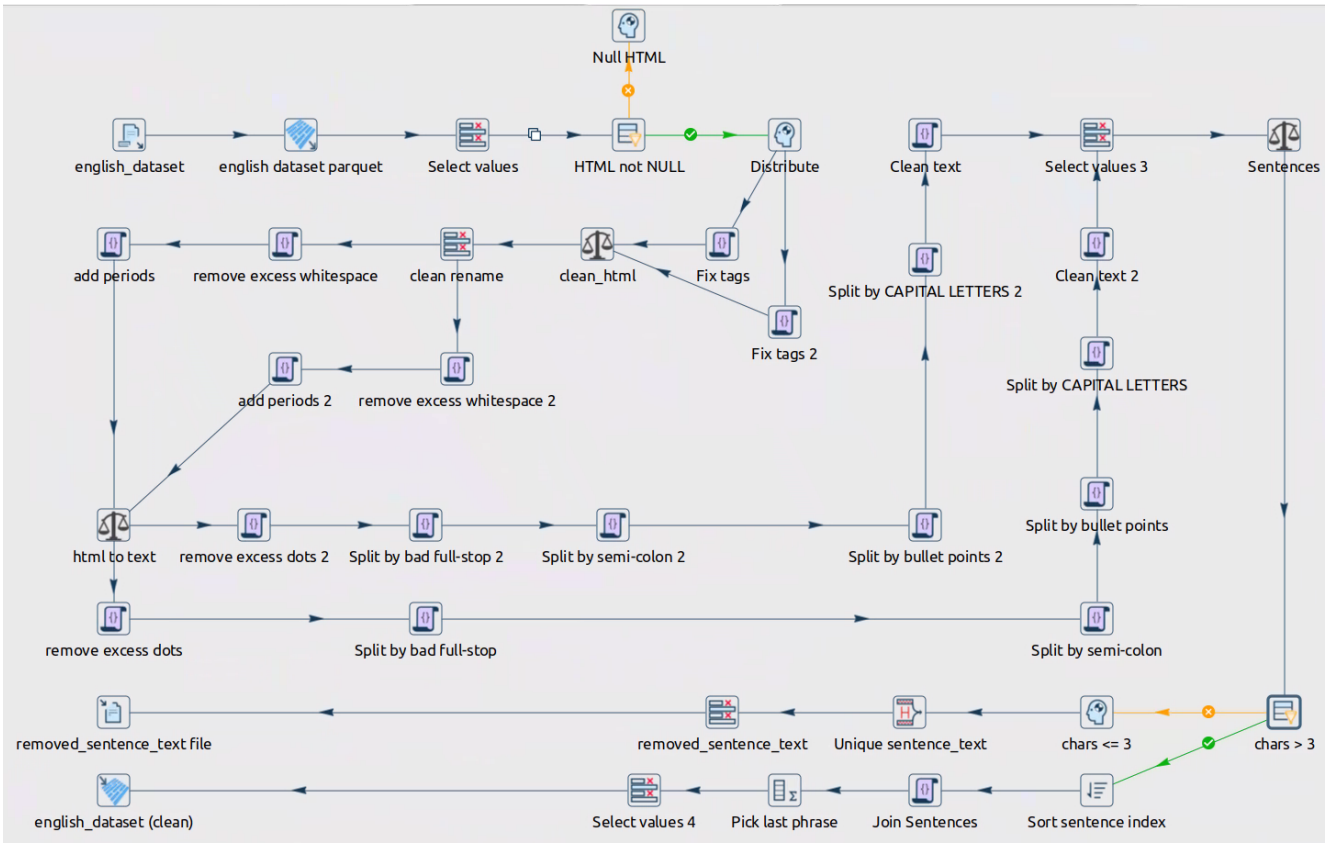

Fig. 11: Pipeline: HTML to Text

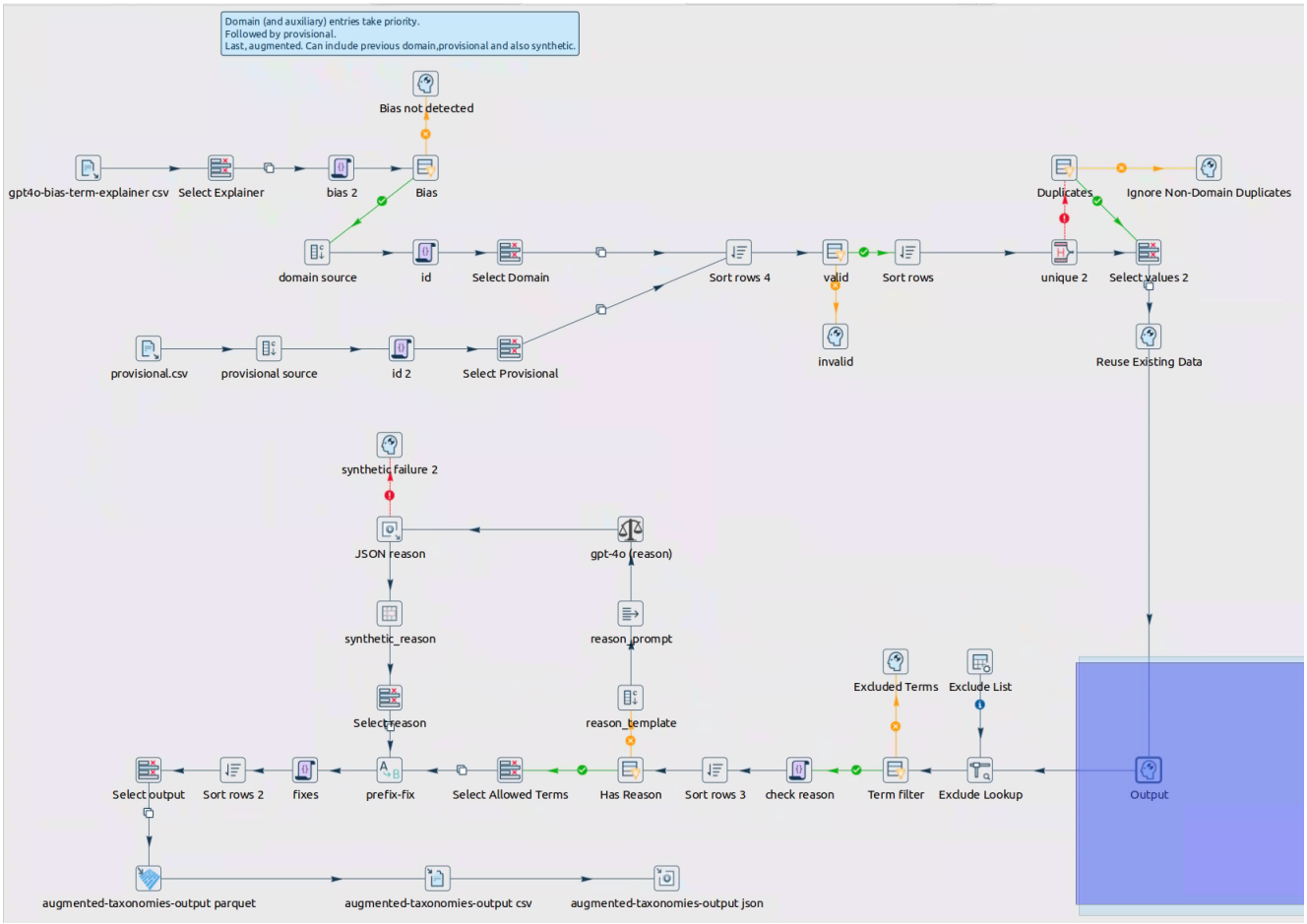

Fig. 12: Pipeline: Taxonomy Augmentation

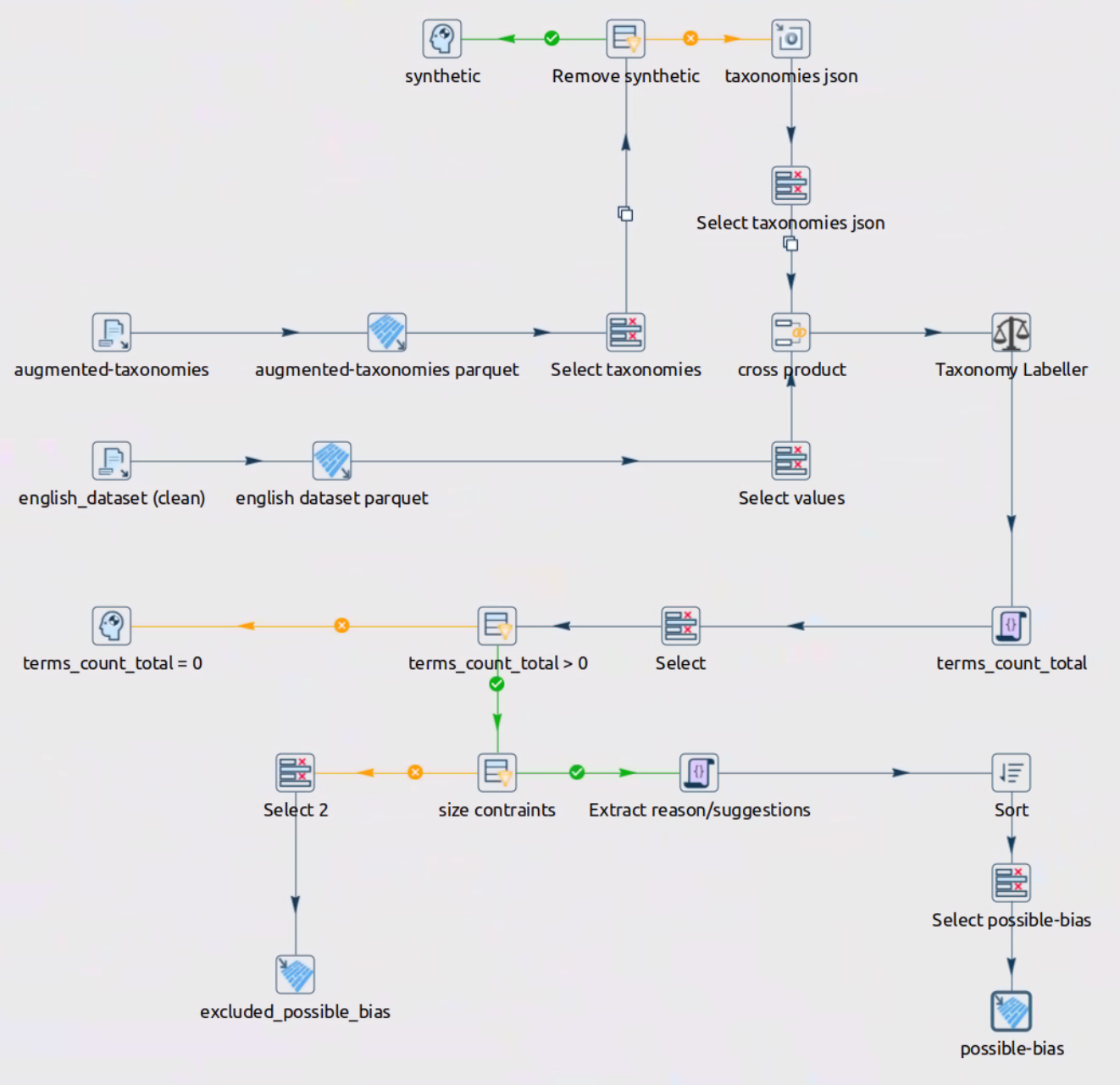

Fig. 13: Pipeline: Detect Possible Bias

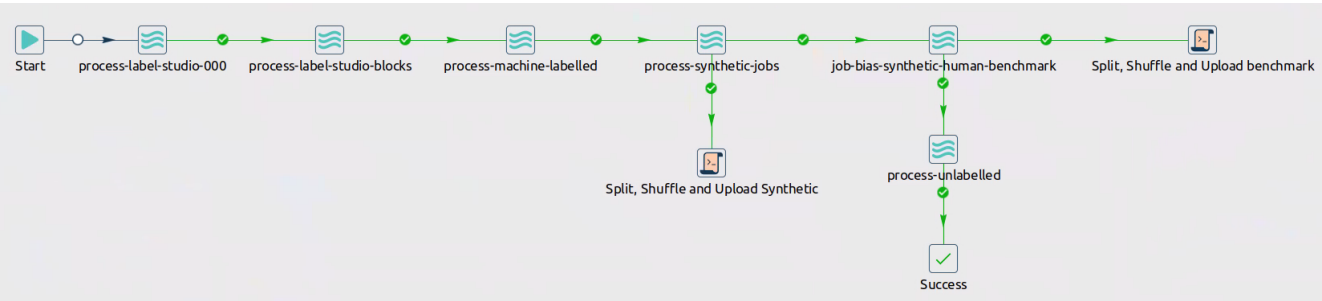

Fig. 14: Workflow: Create Dataset