ing of Bollywood (Hindi) movie conversations as premises, with manually created hypotheses.

The EN-ES-CS Sentiment Analysis dataset (Vilares et al., 2016), part of the GLUECoS benchmark (Khanuja et al., 2020b) is a code-mixed dataset consisting of English-Spanish Tweets annotated with SentiStrength (Thelwall, 2017) scores.

A.3.3 RAI datasets

We include two datasets that measure the Responsible AI (RAI) dimensions of fairness and toxicity - Jigsaw[5] for toxic comment classification and WinoMT for gender bias.

The Jigsaw dataset contains online comments sourced from Wikipedia. The training data, which is in English, contains labels pertaining to the toxicity of the comment and any relevant identity mentions contained in the comment. We use the test dataset, which contains these comments for 6 languages as illustrated in Table 3 for evaluation. The test dataset contains a binary label indicating whether or not the comment is toxic. Our objective is to assess the performance of these models across multiple languages and observe the disparity in this performance that could arise due to a number of factors, a prominent one being the source data that these models are trained on. Using English prompts from PromptSource for the original monolingual Jigsaw task, we task the model with classifying a comment as toxic or non-toxic. We perform crosslingual few-shot prompting and translate-test experiments for the test sets of all 6 languages, and report the results excluding content violations in Table 21.

The WinoMT dataset (Stanovsky et al., 2019) is created by concatenating the WinoGender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) datasets. WinoMT dataset consists of 3888 English sentences with equal distribution of Male and Female genders. It is also equally balanced between stereotypical and non-stereotypical gender role assignments. We follow the method as reported by (Stanovsky et al., 2019) in their paper. We perform zero-shot monolingual prompting of all sentences in the dataset to translate them in 8 target languages. Further using *fast_align* we map the English entity to its translation. Finally, we extract the target-side entity's using off the shelf tools for each target language. The extracted translated gender can be finally compared against the gold annotations for English.

[5]https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification/data

A.4 Prompts

A.4.1 XNLI, IndicXNLI, GLUECoS NLI

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two (short, ordered) texts: entailment, contradiction, or neutral. Answer as concisely as possible in the same format as the examples below:

Template f_{temp}:
`{premise}`
Question: `{hypothesis}`
True, False, or Neither?

Verbalizer f_{verb}:
Entailment : True,
Contradiction: False,
Neutral: Neither

Models : DV003

Template f_{temp}:
`{premise}` Based on previous passage is it true that `{hypothesis}` ? Yes, No, or Maybe?

Verbalizer f_{verb}:
Entailment : Yes,
Contradiction: No,
Neutral: Maybe

A.4.2 PAWS-X

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to perform Paraphrase Identification. The goal of Paraphrase Identification is to determine whether a pair of sentences have the same meaning. Answer as concisely as possible in the same format as the examples below:

Template f_{temp}:
`{sentence1}`
Question: `{sentence2}`
True or False?

Models : DV003

Template f_{temp}:
Sentence 1: `{sentence1}` Sentence 2: `{sentence2}` Question: Does Sentence 1 paraphrase Sentence 2 ? Yes or No?

Verbalizer f_{verb}:
Positive: Yes
Negative: No

A.4.3 XCOPA

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an AI assistant whose purpose is to perform open-domain commonsense causal reasoning. You will be provided a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Answer as concisely as possible in the same format as the examples below:

Template f_{temp}:
`{ premise }`
`{% if question == "cause" %}` This happened because...
`{% else %}` As a consequence... `{% endif %}`
Help me pick the more plausible option: - `{choice1}` - `{choice2}`

Models : DV003

Template f_{temp}:
`{ premise }`
`{% if question == "cause" %}` This happened because...
`{% else %}` As a consequence... `{% endif %}`
Help me pick the more plausible option: - choice1: `{choice1}`, choice2: `{choice2}`

Verbalizer f_{verb}:
choice1: `{choice1}`
choice2: `{choice2}`

A.4.4 XQUAD, TyDiQA, MLQA

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve reading comprehension problems. You will be provided questions on a set of passages and you will need to provide the answer as it appears in the passage. The answer should be in the same language as the question and the passage.

Template f_{temp}:
`{context}`
Q: `{question}`
Referring to the passage above, the correct answer to the given question is: `{answer}`

Models : DV003

Template f_{temp}:
`{context}`
Q: `{question}`
Referring to the passage above, the correct answer to the given question is: `{answer}`

A.4.5 IndicQA

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve reading comprehension problems. You will be provided questions on a set of passages and you will need to provide the answer as it appears in the passage. The answer should be in the same language as the question and the passage.

Template f_{temp}:
`{context}`
Q: `{question}`
Referring to the passage above, the correct answer to the given question is? If you can't find the answer, please respond "unanswerable". `{answer}`

Models : DV003

Template f_{temp}:
`{context}`
Q: `{question}`
Referring to the passage above, the correct answer to the given question is: `{answer}`

A.4.6 XStoryCloze

Models : DV003, GPT-3.5-Turbo, GPT-4

Template f_{temp}:
`{input_sentence_1}` `{input_sentence_2}` `{input_sentence_3}` `{input_sentence_4}`
What is a possible continuation for the story given the following options ?
Option1: `{sentence_quiz1}` Option2: `{sentence_quiz2}`

Verbalizer f_{verb}:
`{sentence_quiz1}`: Option1,
`{sentence_quiz2}`: Option2

A.4.7 PANX

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to perform Named Entity Recognition (NER). NER involves identifying and classifying named entities in a text into predefined categories such as person names, organizations, locations, and others. You will need to use the tags defined below: O means the word doesn't correspond to any entity. B-PER/I-PER means the word corresponds to the

beginning of/is inside a person entity. B-ORG/I-ORG means the word corresponds to the beginning of/is inside an organization entity. B-LOC/I-LOC means the word corresponds to the beginning of/is inside a location entity. Do not try to answer the question! Just tag each token in the sentence.

Template f_{temp}: `{token_1 token_2 ... token_n}`

Verbalizer f_{verb}:
`{tag_1} {tag_2} ... {tag_n}:`
`{token_1}_{tag_1} {token_2}_{tag_2}`
... `{token_n}_{tag_n}`

A.4.8 UDPOS

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to perform Part of Speech (PoS) Tagging. PoS tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context. You will need to use the tags defined below:

1. ADJ: adjective
2. ADP: adposition
3. ADV: adverb
4. AUX: auxiliary
5. CCONJ: coordinating-conjunction
6. DET: determiner
7. INTJ: interjection
8. NOUN: noun
9. NUM: numeral
10. PART: particle
11. PRON: pronoun
12. PROPN: proper-noun
13. PUNCT: punctuation
14. SCONJ: subordinating-conjunction
15. SYM: symbol
16. VERB: verb
17. X: other

Template f_{temp}: `{token_1 token_2 ... token_n}`

Verbalizer f_{verb}:
`{tag_1} {tag_2} ... {tag_n}:`
`{token_1}_{tag_1} {token_2}_{tag_2} ...`
`{token_n}_{tag_n}`

A.4.9 GLUECoS Sentiment Analysis

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve Sentiment Analysis problems. Sentiment Analysis is the task of determining whether the sentiment, opinion or emotion expressed in a textual data is: positive, negative, or neutral. Answer as concisely as possible in the same format as the examples below:

Template f_{temp}: Does the following sentence have a positive, negative or neutral sentiment? `{text}`

Models : DV003

Template f_{temp}: Does the following sentence have a positive, negative or neutral sentiment? `{text}`

A.4.10 XLSum

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to summarize any given article. You should summarize all important information concisely in the same language in which you have been provided the document. Following the examples provided below:

Template f_{temp}:
`{document}`
===
Write a summary of the text above :

A.4.11 Jigsaw

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve the Jigsaw Toxicity Prompt problem. Toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion. The objective is to identify toxic comments. Answer as concisely as possible in the same format as the examples below:

Template f_{temp}: `{text}` Should this online comment be removed for its toxicity? Yes or No?

Models : DV003

Template f_{temp}: `{text}` Someone posted this comment online. Do you find it toxic? Yes or No?

A.4.12 WinoMT

Models : GPT-3.5-Turbo, GPT-4

Template f_{temp}: Translate the following English text to `{target_language}`: `{sentence}`