**Classifying Genetic Essentialist Biases using Large Language Models**

Ritsaart Reimann1,2,5, Kate E. Lynch1,3,5, Stefan A. Gawronski1,4,5, Jack Chan6, Paul E. Griffiths1,5

**Abstract:** The rapid rise of generative AI, including LLMs, has prompted a great deal of concern, both within and beyond academia. One of these concerns is that generative models embed, reproduce, and therein potentially perpetuate all manner of bias. The present study offers an alternative perspective: exploring the potential of LLMs to detect bias in human generated text. Our target is genetic essentialism in obesity discourse in Australian print media. We develop and deploy an LLM-based classification model to evaluate a large sample of relevant articles (n=26,163). We show that our model detects genetic essentialist biases as reliably as human experts; and find that, while genes figure less prominently in popular discussions of obesity than previous work might suggest, when genetic information is invoked, it is often presented in a biased way. Implications for future work are discussed.

**Keywords**: Genetic Essentialism; Bias; Obesity; LLMs; NLP; Automated bias detection

**Corresponding author**: Ritsaart Reimann. Email: reimann.ritsaart@gmail.com

1 Department of Philosophy, The University of Sydney, New South Wales, Australia

2 Department of Philosophy, Macquarie University, New South Wales, Australia

3 School of Historical and Philosophical Studies, The University of Melbourne, Victoria, Australia

4 School of History and Philosophy of Science, The University of Sydney, New South Wales, Australia

5 Charles Perkins Centre, The University of Sydney, New South Wales, Australia

6 Sydney Informatics Hub, The University of Sydney, New South Wales, Australia

**Funding and Acknowledgements:** This research was supported under Australian Research Council's Discovery Projects funding scheme (project number FL170100160), and rendered possible through the University of Sydney’s involvement in the Language Data Commons of Australia (https://doi.org/10.47486/HIR001), which received investment from the Australian Research Data Commons (ARDC). The ARDC is funded by the National Collaborative Research Infrastructure Strategy (NCRIS). We also thank Monika Bednarek (University of Sydney) for permission to use the Australian Obesity Corpus in this study.

**Classifying Genetic Essentialist Biases using Large Language Models**

**Abstract:** The rapid rise of generative AI, including LLMs, has prompted a great deal of concern, both within and beyond academia. One of these concerns is that generative models embed, reproduce, and therein potentially perpetuate all manner of bias. The present study offers an alternative perspective: exploring the potential of LLMs to detect bias in human generated text. Our target is genetic essentialism in obesity discourse in Australian print media. We develop and deploy an LLM-based classification model to evaluate a large sample of relevant articles (n=26,163). We show that our model detects genetic essentialist biases as reliably as human experts; and find that, while genes figure less prominently in popular discussions of obesity than previous work might suggest, when genetic information is invoked, it is often presented in a biased way. Implications for future work are discussed.

**1. Introduction**

There is concern across disciplines that genetic information is communicated and interpreted erroneously. Though most human traits have a genetic basis (Turkheimer, 2000), genetic effects are typically modest and highly contingent on environmental factors (Tabery, 2014). Gene-trait relations also tend to be pleiotropic and probabilistic, with complex interactions between many genes (marginally) increasing or decreasing the likelihood of various outcomes (Lynch, 2021). There are exceptions to this. So called ‘Mendelian traits’, or traits with a strong genetic explanation (Turkheimer, 1998), are those where mutations in just one or a few genes significantly affect the expression of a single trait (Lynch, 2021). Errors occur when strong genetic explanations are attributed to traits with a modest, or weak, genetic basis. These errors include exaggerated claims about genetic effects (Holtzman, 1997), unwarrantedly deterministic thinking about gene action (Nelkin & Lindee, 1995), an overemphasis of power attributed to DNA (Holtzman, 1999), and the idea that genes are part of a special human “essence” (Conrad, 1997).

Explanations as to why genes are conceived and portrayed in these ways typically appeal to the theory of psychological essentialism: A ubiquitous cognitive bias whereby people believe that certain entities (such as people, social groups, or species) have an underlying, unobservable essence or set of inherent characteristics that determine their properties and define group membership (Griffiths, 2002; Dar-Nimrod & Heine, 2011; Linquist et al., 2011).

The belief that genes provide this essence is termed *genetic essentialism*, and there is ample experimental evidence that genetic explanations indeed elicit essentialist thinking (Gould & Heine, 2012; Castéra & Clément, 2014; Carver et al., 2017; Machery et al., 2019; Yaylaci et al., 2021). These results, which derive from standardised questionnaires and implicit association tests, have been interpreted by Dar-Nimrod and Heine (2011) as suggesting a cognitive framework of genetic essentialist (GE) biases.

This framework posits that four distinct biases are elicited when genetic information is presented: Gene action is interpreted as *deterministic*–where the effects of a gene are thought to be inevitable or immutable[[1]](#footnote-1); to have a *specific aetiology*–where traits are thought to arise from the action of single genes, exemplified in claims about “the gay gene” or “the thin gene”; *naturalism*–where traits with a genetic basis are thought to be natural, and the naturalistic fallacy is invoked to moralise about such traits; and that genetic traits imply *homogenous and discrete* group membership (henceforth *homogeneity*)–where people with the same genes are thought to be of the same homogenous and discrete group, in virtue of their shared genetic essence.

In distinguishing between these four sub-components of genetic essentialism, Dar-Nimrod and Heine’s (2011) framework introduces a more nuanced understanding of the various ways in which beliefs about genetic effects misfire. That this framework also yields the best available understanding of how genetic essentialism manifests as a psychological phenomenon is evidenced both directly by a validated psychometric scale (see Dar-Nimrod et al., unpublished manuscript) and indirectly by its explanatory success in closely related fields (e.g. Aspinwall et al. 2012; Lebowtiz & Ahn 2014; Haslam & Kvaale 2015; Turnwald et al. 2019; Herd et al.,. 2019; Yaylaci et al. 2019; Harden 2021; Dar-Nimrod et al. 2021). Our work extends these applications in two ways. Our primary objective is to develop a semi-automated classification tool capable of detecting essentialist biases across large volumes of text. Dar-Nimrod and Heine’s (2011) framework is particularly well-suited to this task because it gives us four clearly-defined categories, or classes, for such a model to aim at (see §2 for task, class, and data specification). Our second objective is to use this model to investigate the role of the media in facilitating and entrenching essentialist beliefs, specifically in the context of obesity discourse.

While some have argued that the media reflects, rather than shapes public attitudes towards genetics (Condit, 2011), others have argued that biased media reporting is responsible for these attitudes (Nelkin & Lindee, 1995). This claim rests on the assumption that the media inaccurately reports on genetics research, exaggerating scientific interpretations of genetic results in a process of ‘genohype’ (Holtzman, 1999). This assumption, in turn, has been the object of both quantitative and qualitative investigations. As with qualitative analyses of media bias in other domains (e.g., Lukin, 2005), qualitative approaches typically enlist quasi-formal discourse-analytical methods and frameworks to critically evaluate a small number of articles; deconstructing focal passages that encode exaggerated, deterministic, and essentialist claims about genes (e.g., Nelkin & Lindee, 1995; McCombs, 2014). Central to this level of analysis are lexical, stylistic, and more broadly editorial features of the target text(s), including the overarching frame, tone, and stance with which findings from genetics research are communicated (e.g., Conrad, 1997; Conrad & Markens, 2001). A consistent conclusion that emerges from this literature is that reports on genetics research are becoming both more prevalent and increasingly essentialist and deterministic in character.

In contrast, quantitative approaches tend to find that media reporting on genetics is largely accurate. Bubela and Caulfield (2004), for instance, estimate that only 11% of articles on genetics research exaggerate the science. In an earlier study, Condit et al. (1998) found that while discussions of genetics in the media have indeed increased over time, attributions of genetic determinism and genetic causation have become less prevalent (between 1915-1995). These and related studies, which arrive at similar conclusions (Loo, et al., 2001), enlisted human coders to assess small samples of articles (up to 972) for features relating to determinism, causation, exaggeration, and accuracy. To date, a large-scale analysis of the language used in print media, in reference to the best current psychological framework for understanding the interpretation of genetic information (i.e., the framework developed by Dar-Nimrod & Heine, 2011), has been missing from the literature.

The present study offers such an analysis by leveraging recent advances in artificial intelligence (AI), large language models (LLMs), and natural language processing (NLP) to develop a semi-automated classifier that enables us to evaluate a significantly larger sample of news articles (n = ~26,000). We focus specifically on articles dealing with obesity; a pressing health concern in many parts of the world (Roth et al., 2004; Swinburn et al., 2011; Meldrum et al., 2017), and one that is potentially exacerbated by essentialist media discourse. Dar-Nimrod et al. (2014), for instance, have found that essentialist explanations of obesity negatively affect obesity-related health behaviours. Looking specifically at dietary choices, the authors show that exposure to genetic information that portrays gene-action deterministically prompts less healthy nutritional decisions (see also Ahn & Lebowitz, 2018). Important to note is that deterministic explanations of obesity find little support in the scientific literature (cf., Conrad & Markens, 2001; Christensen et al., 2010; Tabery, 2014). Like most other traits, obesity has a weak genetic basis, with many genes exerting modest effects. Lifestyle factors such as diet and exercise are, moreover, in all but exceptional cases,[[2]](#footnote-2) primary determinants of how obesity-related genes are expressed (Loos & Yeo, 2022). Discussions that emphasize genetic effects over environmental influences, then, not only misrepresent the science, but risk inhibiting the very behaviours by which obesity can be kept at bay. Getting clear on the extent to which popular discussions of obesity embed essentialist beliefs is therefore an object of pressing concern.

Our work also contributes to the state of the art in automated bias detection and further expands the use of NLP–including LLMs–as emerging methods for answering philosophical questions. That NLP methods offer novel insights into philosophical issues is evidenced by the wide range of domains in which these tools have been fruitfully applied, including work on vaccine hesitancy (Quintana et al., 2022), online social movements (Klein et al., 2022), and self-regulation (Abedin et al., 2023). Common to each of these contexts is the assumption that latent patterns in natural language are a useful proxy for underlying psychological states; or, at the level of groups, culturally entrenched attitudes.

Our discussion thus far suggests that genetic essentialism straddles this distinction–existing simultaneously as a *psychological bias* to essentialise genes (Dar-Nimrod & Heine, 2011), and as a *biased mode of discourse* that elicits and potentially exacerbates this tendency (Nelkin & Lindee, 1995). Taking the interactions between these two faces of genetic essentialism seriously suggests that a thorough understanding of this phenomenon calls for an ecological theory of bias writ large. That is, a theory on which biases emerge at the interface between individual agents and their socio-material niche, including their informational environment (e.g., de Carvalho & Krueger, 2023). A theory along these lines foregrounds the recursive relationships between cognitive processes and cultural artefacts and resources, focusing specifically on how these artefacts and resources both reflect and reinforce our cognitive and behavioural routines, and how these reciprocal loops lead to the gradual entrenchment of bias at both a cognitive and cultural level. Seen in this way, the cognitive tendency to essentialize genes is both a cause and a consequence of an essentialist discourse, i.e., a cultural resource, that both reflects and further engrains essentialist thinking.

Since neither our methodological nor empirical contributions hang on accepting this view, we do not motivate it further here, save to say that by acknowledging the mutuality between what are often and to our mind mistakenly seen as mutually exclusive explanations, an ecological theory of bias resolves disagreements of the kind typified in debates over whether the media shapes (e.g., Nelkin & Lindee, 1995) or reflects (e.g., Condit, 2011) public attitudes towards genetics.[[3]](#footnote-3)

More germane to our present aims is that advances in computational methods have fostered considerable interest in the use of NLP for developing classification tools capable of detecting biases automatically and at scale (for review, see Hamborg, Donnay, & Gipp, 2019). A general typology of work in this area distinguishes between domain-specific classifiers—typically keyed to topical lexica consisting of, for instance, racial or sexist slurs (e.g., Badjatiya et al., 2017; Anzovino et al., 2018; Ahmed et al., 2022)—and general purpose models that detect bias by extracting broader linguistic and textual features, including emotive phrasing, hyperbole, hedging, a variety of rhetorical devices, and higher-order document elements such as tf-idf[[4]](#footnote-4) (e.g., Recasens et al., 2013; Lim et al., 2018; Spinde et al., 2021).

While the ‘black-box’ nature of LLMs precludes us from commenting with certainty on which (if any) of these elements our model incorporates, the multilayered attentional architecture that is typical of generative language models suggests that they operate across a range of hierarchies: identifying latent patterns in both lower-level relations (e.g., between discrete tokens) andamong more abstract, higher-level representations (Starace et al., 2023). We have, in addition, implemented several strategies expressly aimed at making our classifier sensitive both to *specific lexica* associated with genetic essentialism, and *broader linguistic features* characteristic of biased reasoning more generally.

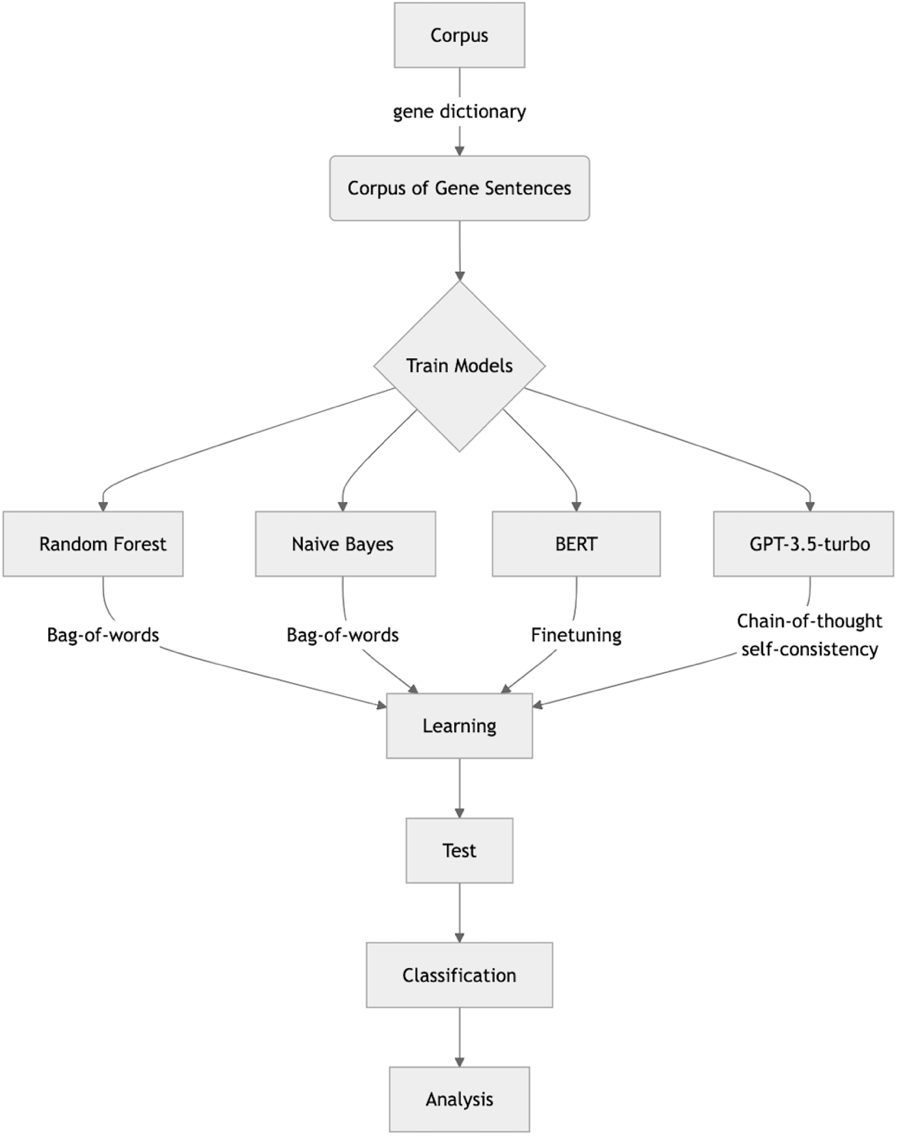
We provide a more detailed discussion of these strategies, as well as our models and workflow, in the following section (§2). In doing so, we hope that our work will be of use for future research aimed at leveraging advances in AI to aid in automated bias detection. Though the present study isn’t the first to move from well-documented concerns aboutbiases *within* LLM generated text to the potential of LLMs for *detecting* bias in human generated language, it does push previous efforts in this field forward by moving from biases with relatively explicit and extensively studied linguistic footprints–such as hate-speech (e.g., Zhao et al., 2021) and harmful stereotypes (e.g., Liu, 2024)–to a set of constructs whose linguistic signature is yet to be discerned, let alone shown to be machine readable. Our analysis of genetic essentialism in the context of Australian obesity discourse is therefore in the first place a proof-of-concept: demonstrating how recent developments in NLP—and LLMs specifically—can be deployed to automatically identify and therein potentially stymie the spread of erroneous beliefs about genes more generally.

As a proof-of-concept, then, our overarching aim is to explore the efficacy of LLMs in detecting, differentiating between, and correctly classifying the four biases introduced above. Assuming satisfactory results, we then deploy the best performing model to answer the following four research questions:

1. How pervasive are genetic essentialist biases in discussions of obesity among Australian news outlets?
2. Are some biases more prevalent than others?
3. Do they occur together? And if so, are there meaningful patterns of co-occurrence?
4. Are the four biases distributed evenly across outlets? Or does frequency differ by venue?

**2. Methods**

In the interest of open science, our data, code, and supplementary materials are available via the Open Science Framework.[[5]](#footnote-5) In the interest of guiding future research, Figure 1 illustrates a general outline of our workflow. Preliminary tests showed that performance for all models improved substantially by transforming our task from a multi- to binary class problem. We therefore ran each model four times as a binary classifier, classifying sentences as either biased or neutral for each of our four focal biases. The following sections provide a detailed discussion of each step.



**Figure 1** General Workflow

**2.1 Corpus and Human annotation**

We retrieved our corpus from CQPweb: a web-based corpus analysis system that hosts a collection of curated corpora as well as a number of open-source corpus analysis tools.[[6]](#footnote-6) We focused exclusively on the ‘Australian Obesity Corpus’, consisting of 26,163 articles from twelve major Australian news outlets, spanning 2008 to 2019 (Bednarek et al., 2023). As noted in the corpus manual (Vanichkina & Bednarek, 2022), the outlets include one national newspaper, plus one or two newspapers from each state and territory. These outlets are also broadly representative of the mainstream media landscape in Australia, reflecting the concentration of media ownership there, with seven of the twelve outlets owned by News Corp. One implication of this concentration of ownership is that the same articles are occasionally published by different venues, and that our corpus therefore contains a number of duplicates. For the purpose of model development, all duplicates were removed. We did, however, reincorporate them in the subsequent analysis, as an article's originality has no bearing on its contribution to the overall composition of the Australian media landscape, which is the ultimate target of our investigation.

With this corpus in hand, we developed a gene lexeme consisting of 23 (stemmed) gene-related terms to extract gene-related sentences. Of these 4303 sentences, two of the authors (redacted for review) independently annotated a sample of 800, coding each for the presence or absence of our four focal biases. Note that a single sentence could be coded for multiple biases, and that only those sentences where coders were in complete agreement (685) were kept for model training, validation, and testing. Both coders have extensive knowledge of genetic essentialism and the genetics of obesity more generally. Both also have a background in philosophy, specifically social epistemology (redacted for review) and the philosophy of science (redacted for review).

As is standard practice in the field (e.g., McAllister et al., 2021), we proceeded in a piecemeal, iterative fashion: breaking our sample into smaller chunks and coming together at regular intervals–every 200 sentences–to discuss disagreements. This allowed us to gradually refine our coding rules and arrive at increasingly clear characterizations of each bias. The determinism and specific aetiology biases were relatively easy to identify from the outset, as both are frequently instantiated via a relatively small set of relatively fixed lexical structures. The phrase ‘in the genes’, for instance, was both commonplace and reliably indicated that the target sentence portrayed genes deterministically, as did specific terms used to describe genetic effects (e.g., genes ‘program’, ‘code’, or ‘control’ for behaviour). Naturalness and homogeneity were more difficult to pin down, in part because there are no simple syntactic structures common to these cases, and in part because the phenomena that these biases speak to are inherently harder to evaluate: moving from epistemological claims about genetic effects, which are relatively easy to assess, to normative claims about genetically ‘caused’ behaviours (naturalness) and the implications of genes for identity (homogeneity). Both our lexeme and coding scheme are publicly available (for further details, see also §1 and §2 of the Appendix).

**2.2 Models and Training**

We trialled five classification models: Two traditional machine learning algorithms–Naive Bayes and Random Forest–and three transformer-based LLMs–Google’s BERT and two implementations of OpenAI’s GPT. Common to all these models is that they work by encoding tokens, which are basic units of text such as words or characters, into a sequence that represents the meaning and context of each token. The models are then trained on these sequences to identify patterns corresponding to semantic, syntactic, and other linguistic features. This is achieved by transforming raw text into numerical representations, which enable the models to detect statistical regularities. In identifying these regularities, each model develops a probabilistic understanding of text, capturing relevant linguistic relationships that can be used to categorize or predict new instances.

While LLMs are widely regarded as a major breakthrough in natural language processing, recent reviews of the relevant literature suggest that their primacy over more traditional machine learning methods, specifically with respect to classification tasks similar to the one presented here, may be less pronounced than is commonly assumed (e.g., Xu et al., 2024). Naive Bayes and Random Forest were included as benchmarks against which to test these positions. Naive Bayes, briefly, assumes conditional independence between all features (i.e., tokens) and, during training, assigns probabilities to each feature describing the likelihood of that feature belonging to a particular class (in our case, a particular bias). When asked to classify a new instance (in our case, a new sentence), it calculates the probability of each class given the observed features and selects the class with the highest probability. Random Forest contrasts this strictly probabilistic approach by implementing an aggregative, or ensemble, architecture. During the training phase, the model builds multiple decision trees by randomly sampling different subsets of features, such that each tree associates different features with each class. During the test phase, each instance is passed through each tree, and each tree makes a prediction. The final classification reflects the majority decision. By aggregating over different trees, Random Forest has a better handle on interactions among features, enabling it to identify higher-dimensional relations between them.

For the present task, both Random Forest and Naive Bayes were used with bag-of-words as the input feature, chosen for its simplicity and short training time. We trained and validated both models on 60 and 20 percent of our annotated sentences respectively, and used the remaining 20% of our annotated data to test performance. Because bag-of-words treats each token as an independent feature, it disregards nuances such as word-order, semantic relationships, and other contextual elements.

The promise of LMMs, particularly with respect to natural language tasks, resides in their ability to represent linguistic regularities at various, progressively higher, and hierarchically organized levels of abstraction: moving from, for instance, local syntactic structures and word-level dependencies to semantic and contextual relationships that span entire documents. These more sophisticated representations rest on at least two key innovations that distinguish transformer-based models from more traditional machine learning algorithms. The first is that transformer-based models generate ‘context-aware’ token embeddings, i.e., abstract semantic numerical vector representations in which the value of each token is influenced by the values of other tokens within the same text (Peters et al., 2018). These token embeddings provide the initial representations that are then operated on by the model’s attention layers, which are designed to dynamically adjust and contextualise individual token vectors by applying a set of weights learnt during the pre-training stage (Vaswani et al., 2017).[[7]](#footnote-7)

Of note for the present purpose is that whereas Google’s BERT is pre-trained using the *masked* language modelling objective, OpenAI’s GPT models are pre-trained with a *causal* objective. The critical difference between these approaches lies in what parts of the surrounding context inform the models’ predictions: while GPT models attend *only* to precedingtokens and are in that sense ‘backward-looking’, BERT effectively ‘looks both ways’, enabling it to capture richer, more complex relationships (Devlin et al., 2019).[[8]](#footnote-8) Although this should, intuitively, make BERT especially well-suited to the present task, a potential advantage of GPT models relative to BERT is their compatibility with a number of ‘prompting strategies’ that have been found to significantly improve performance across a range of tasks (e.g., Wei et al., 2022; Wang et al., 2022). To test the efficacy of these strategies, we ran both a ‘basic’ and ‘augmented’ version of GPT-3.5-TURBO, or what are commonly referred to as ‘zero-shot’ and ‘few-shot’ implementations. We discuss both in greater detail below. A more detailed discussion of how we implemented BERT can be found in the appendix (§3).

*GPT-3.5-TURBO—zero-shot*

In this implementation we provided minimal context, only briefly defining each bias based on the definitions developed in our coding schema (see Appendix §4 for details). We then prompted the model to classify the sentences in our test data into one of two classes for each of the 4 biases: biased or non-biased. Temperature and top\_p, which modify variability in output, were set to 0 and 1 respectively.[[9]](#footnote-9) By setting the temperature to 0, we effectively removed all randomness. This was done in the interest of reproducibility.

*GPT-3.5-TURBO—few-shot*

In this implementation, we added extensive contextual information alongside a basic definition for each class. We also implemented two specific prompting strategies. The first is commonly referred to as ‘Chain of Thought’ (Wei et al., 2022), and involves supplementing what are considered standard prompts (e.g. definitions of terms) with examples illustrating intermediate steps in reasoning: ‘explaining why’ a particular input leads to a corresponding classification.

We leveraged our coding scheme and general expertise to engineer chains of thought that unambiguously emphasised the kinds of textual features that would ground the classification of a sentence as biased. These features included both specific lexicaassociated with each class, and broader syntactic structuresindicative of essentialist biases across the board. To avoid biasing the model, we included an equal number of chains of thought for non-biased sentences. Table 1 gives a number of examples. Full details can be found in the appendix (§4).

|  |  |
| --- | --- |
| **Specific Lexica** | **General Syntactic Structures** |
| Sentence: | Sentence: |
| “Your genes act as a blueprint for all the proteins in your body, which control your physiology and biology.” | “Researchers say that obesity is caused by a hungry gene.” |
| Chain of Thought: | Chain of Thought: |
| “The sentence mentions the phrases ‘blueprint’ and ‘control’ and asserts that genes act as a ‘blueprint’ that ‘controls’ various aspects of our physiology and biology. These phrases are indicative of the ‘determinism’ bias. Therefore, this statement is classified as ‘determinism’.” | “The sentence names a particular gene - the ‘hungry gene’. This implies that it is the gene ‘for’ being hungry. It is also an instance of the general format ‘the [name of trait] gene’. Therefore, this sentence is classified as ‘specific aetiology’.” |

**Table 1** Examples of GPT Prompts

Worth adding for researchers with an interest in deploying these methods for similar tasks is that we refined our prompts iteratively by repeatedly running the model on random samples of the training data. By prompting the model to justify its outputs, we gained insight into when and where our chains of thought were misfiring, enabling us to refine them further. This process also revealed interesting inconsistencies in the model’s responses. In particular, we encountered a substantial number of cases in which the reasoning did not correspond to the classification, i.e., where the model reasoned to one conclusion, yet classified the sentence as if the opposite were true. These cases accounted for a significant number of misclassifications. We eventually confirmed that this issue could be addressed by reversing the default setting, prompting the model to generate its ‘reason’ first, and thenpredict the class.

The second technique, known as ‘Self Consistency’ (Wang et al., 2022), builds on the first by getting the model to generate several chains of thought in response to the same query and, via a voting mechanism, return an output that reflects the majority decision. For the present task, we asked the model to generate three chains of thought for each query. For three of our four biases–determinism, naturalness, and specific aetiology–we implemented the standard version of this protocol: classifying a sentence according to a simple majority. For the homogeneity bias, we adjusted the classification threshold: predicating a positive classification on a unanimous prediction. Note finally that temperature and top\_p were set to 1 and 0.8 respectively, introducing some degree of randomness into the model. Without this randomness there is no benefit to implementing the self consistency strategy, as responses would be (near) identical.

**2.3 Evaluation Metrics**

As is standard in the literature (for review, see Hamborg et al., 2019), precision, recall, and F1 were taken as the preferred evaluation metrics (see table Table 2 for details). While there are no hard and fast rules for which of these values researchers should aim to optimise, the notion of ‘inductive risk’–i.e., the risk of error in accepting or rejecting a hypothesis, data, or model (Douglas, 2000; Winsberg, 2012)–offers some guidance for balancing false negatives and false positives, the minimization of which involves optimising for recall and precision respectively (Karaca, 2021). For reasons that we return to in the discussion, the present context is one in which optimising for precision strikes us as prudent, though not at any cost.

The challenge of striking the right balance between these criteria and establishing an appropriate benchmark for each is compounded by the absence of a direct corollary to our analysis. There are, however, two literatures that offer guidance. First, a review of previous work on automated bias detection shows that precision, recall, and F1 vary widely from one model and task to the next, and suggests that task complexity partially defines what constitutes good performance. At one end, domain-specific classifiers trained on binary tasks–such as categorising an article as containing racist slurs or not–regularly achieve scores in the range of 0.7-0.9 across key performance metrics (for review, see Fortuna & Nunes, 2019). At the other end, general purpose, multi-class models (e.g., Recasens et al., 2013) rarely achieve F1 scores greater than 0.4, with considerable trade-offs between precision and recall (for review, see Spinde et al., 2021). Relevant to our work as well are previous studies that have used LLMs for other kinds of classification tasks, including sentiment analysis (Adoma et al., 2020), toxic language (Zhou, 2021), and more complex psychological constructs (Stavropoulos et al., 2023). Taken together, these two literatures suggest that scores in the range of 0.7-0.8 for each metric constitutes good performance, relative to our task and aims.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Description** | **Formula** |
| Precision | Ratio of correctly predicted positive instances from total predicted instances in a positive class. |  |
| Recall | Ratio of positive instances that are correctly classified. |  |
| F1 Harmonic mean between recall and precision values. | | |

**Table 2** Evaluation Metrics. TP = True Positives, FP = False Positives, FN = False Negatives

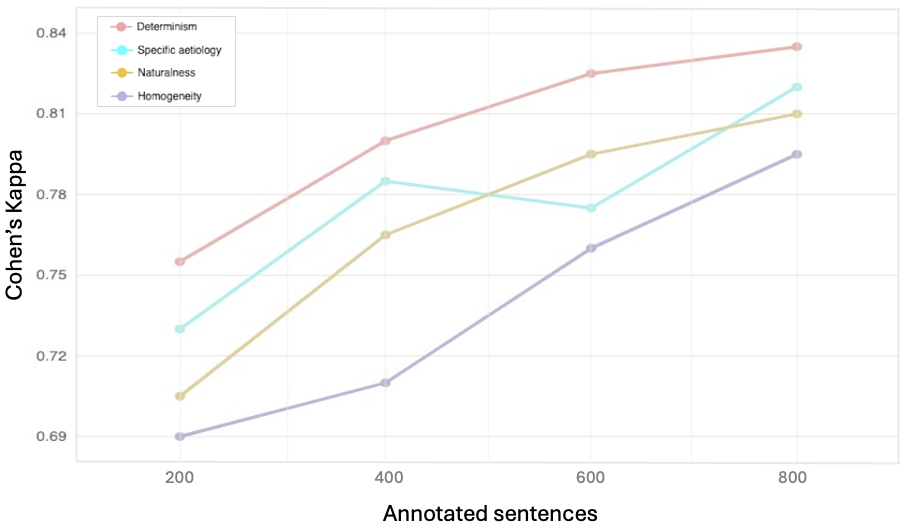
**3. Results**

In this section we present our results. We begin with inter-rater reliability, then evaluate model performance, and conclude by running the best performing model on our corpus. Recall that our analysis of the Australian Obesity Corpus serves primarily as a proof-of-concept. Accordingly, and although our results potentially speak to a broader set of issues, we limit our attention to the four research questions introduced in §1:

1. How pervasive are genetic essentialist biases in discussions of obesity among Australian news outlets?
2. Are some biases more prevalent than others?
3. Do they occur together? And if so, are there meaningful patterns of co-occurrence?
4. Are the four biases distributed evenly across outlets? Or does frequency differ by venue?

**3.1 Human Annotation**

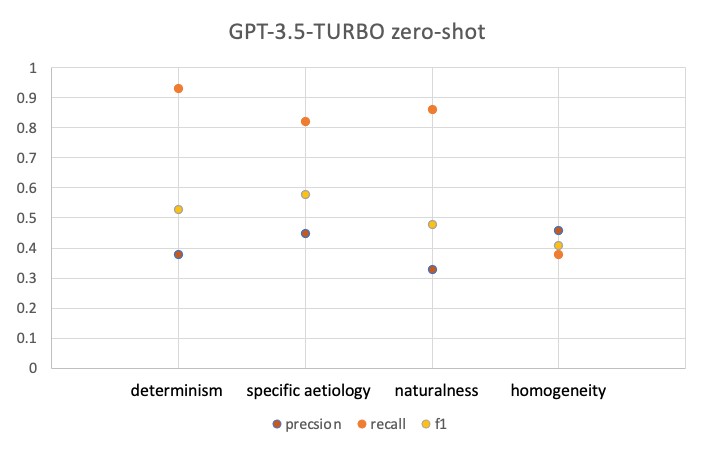
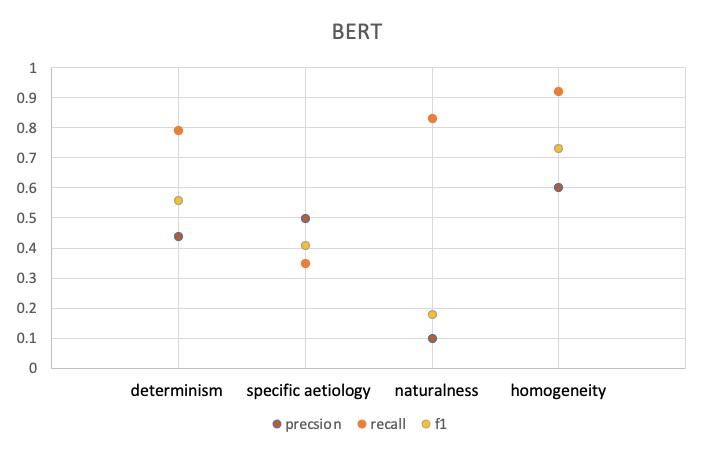
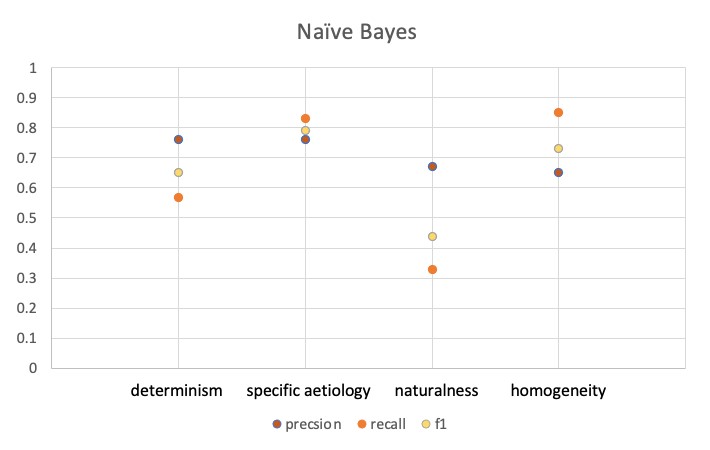
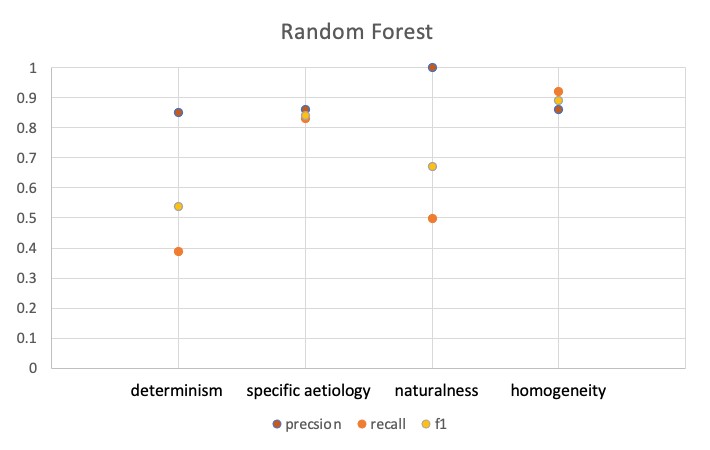
As stated, we split the 800 sentences selected for human annotation into four batches of 200. Cohen’s kappa for inter-rater reliability increased from 0.72 for the first round of coding to 0.83 for the final round. Of this sample, 685 sentences were in complete agreement and were used for training, validation, and testing (see Appendix §5).

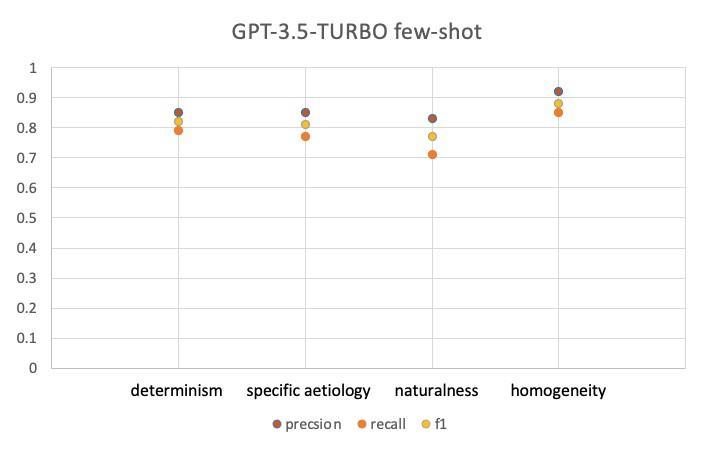


**Figure 2** Changes in interrater reliability over successive rounds of coding.

**3.2 Model Performance**

Figure 3 illustrates that performance varied considerably across models and classes. Naive Bayes and Random Forest struggled to identify and correctly classify sentences containing the naturalness and determinism bias. BERT also struggled with both these classes, and more so with specific aetiology than any of the other models. The zero-shot implementation of GPT-3.5-TURBO demonstrates considerable trade-offs between precision and recall, with scores for all metrics decreasing across classes as we move from determinism to homogeneity. Finally, GPT-3.5-TURBO few-shot showed the most consistent and, on average, best performance across all metrics and biases. With this model in hand, we turn to our corpus.

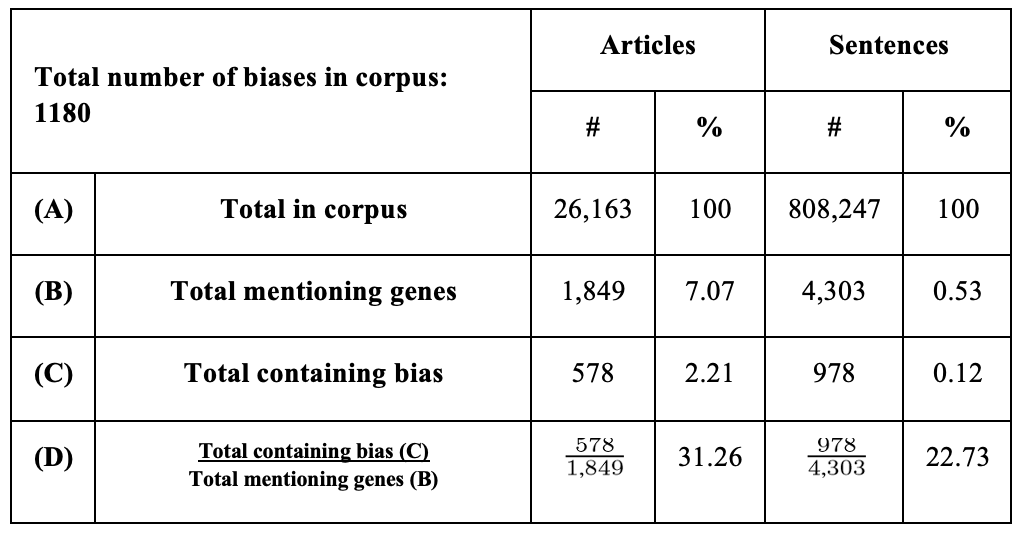




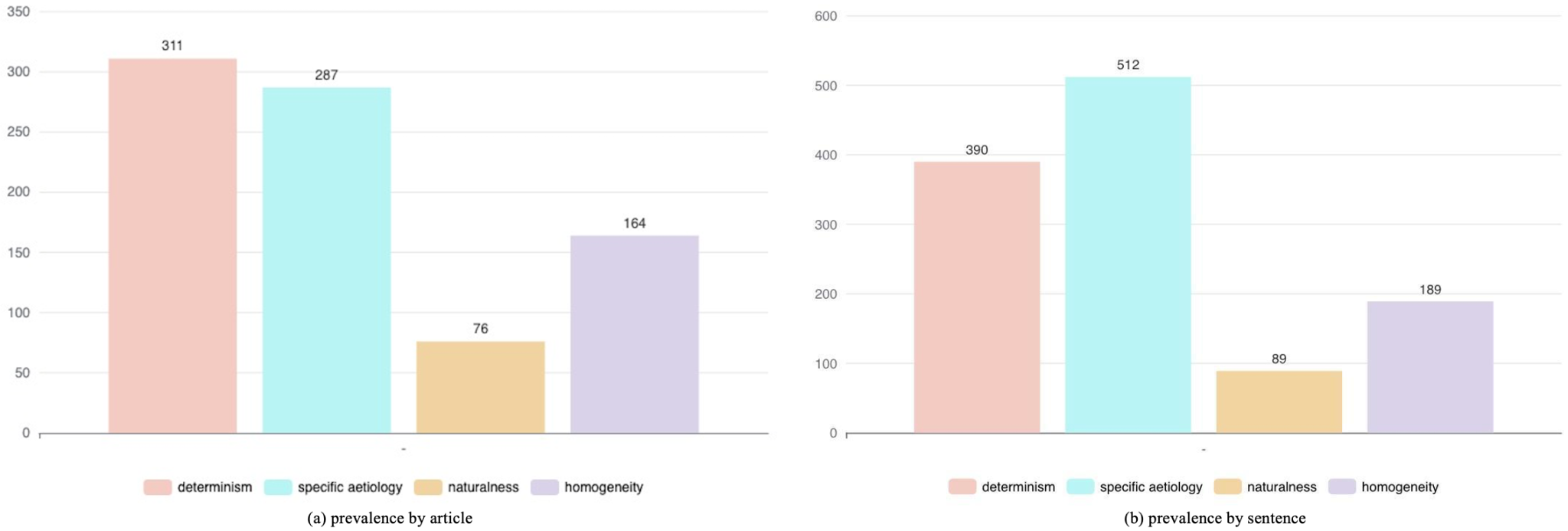
**Figure 3** Precision, Recall and F1 scores for all models

**3.3 Corpus analysis**

Our response to *RQ1* is informed by the summary statistics given in Table 3. We found considerable variation in the prevalence of each bias (see Figure 4), addressing *RQ2*. Specific aetiology is the most prevalent at the sentence level, accounting for 43.4% of biases in our corpus and occurring in ~12% of gene sentences. Homogeneity and determinism contribute 16 and 33 percent of total biases, and occur in ~4.4% and ~9% of all gene sentences, respectively. Naturalness makes up the remaining 7.6% and is found in ~2% of gene sentences. Things look slightly different at the article level, where the determinism bias is the most prevalent. We expand on this result in §4.2.2.



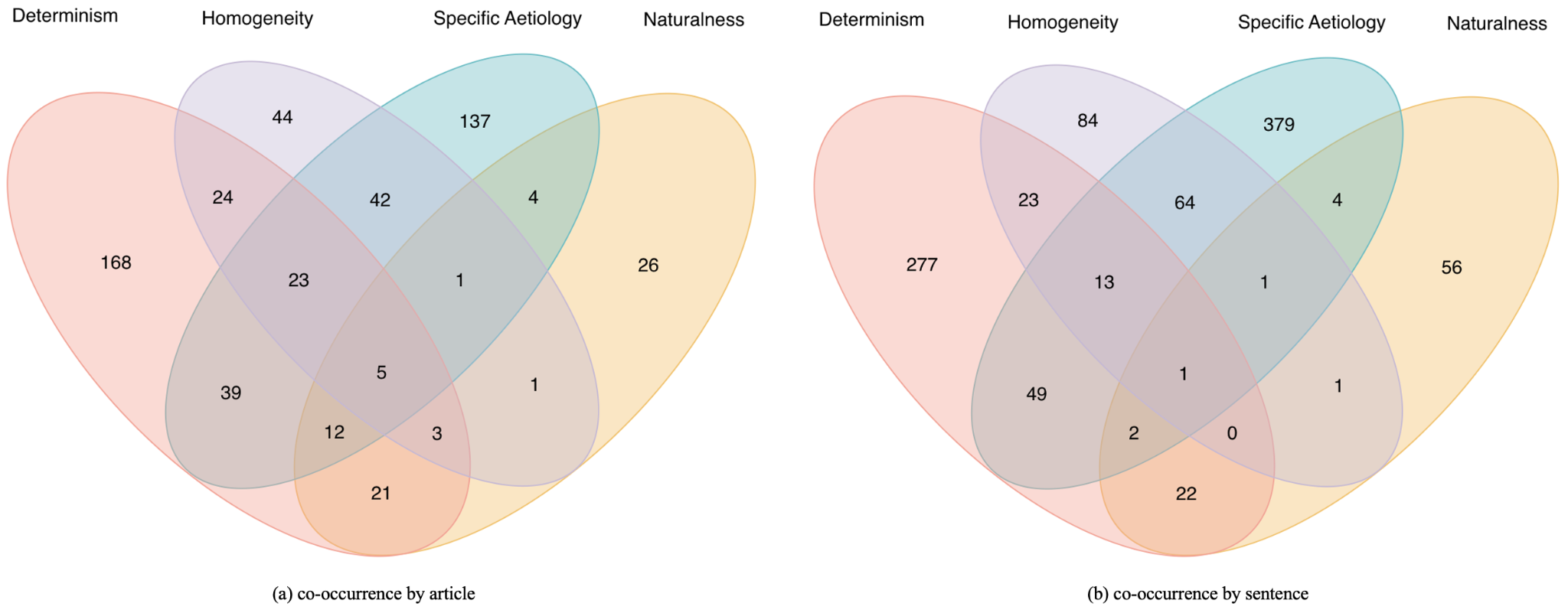
**Table 3** Total number of biases detected by GPT3.5-TURBO few-shot, summarised at the article and sentence level.



**Figure 4** Prevalence of each bias at the (a) article and (b) sentence level.

To answer *RQ3*, Figure 5 shows all instances of co-occurrence at both the sentence and the article level. At both levels, specific aetiology and homogeneity appear together most frequently, followed by specific aetiology and determinism, and determinism and homogeneity. Specific aetiology, determinism, and homogeneity also frequently appear together, both within single sentences and within the same articles. Naturalness co-occurs most often with determinism at both the sentence and the article level.

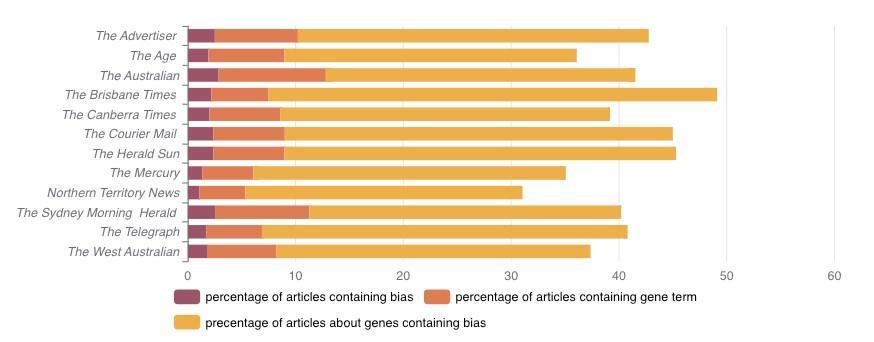
Finally, The Sydney Morning Herald, The Herald Sun, The Courier and The Australian account for more than half of all biases (Figure 6a), addressing *RQ4*. This initially suggests a skewed distribution, with one third of vendors contributing more than the remaining two thirds combined. Figure 6d, on the other hand, indicates that most venues communicate bias at roughly similar rates, suggesting that the asymmetries in Figure 6a are an artefact of the composition of our corpus, with some outlets contributing significantly more articles than others. Figures 6b and 6c tell a similar story at the sentence level: though some outlets mention genes more frequently than others (Figure 6b), the average number of biases per gene sentence is largely similar for each venue (Figure 6c).



**Figure 5** Co-occurence of biases at the article (a) and sentence (b) level. Non-overlapping sections indicate no co-location, overlapping sections indicate two or more biases present in the same article/sentence. Pink ellipse and its intersections = determinism, purple = homogeneity, blue = specific aetiology, yellow = naturalness.

A graph with blue bars

Description automatically generated with medium confidence



(d) Percentage of articles containing gene terms (yellow), containing bias (dark red), an articles containing gene terms that also contain bias (orange) by outlet.

**Figure 6** Results by publication venue.

**4. Discussion**

The subtle and often implicit nature of media bias presents a formidable challenge, both to lay audiences, and to researchers interested in pre-empting the downstream consequences of a misinformed public. The present study explored the potential of LLMs in addressing this challenge, focusing on genetic essentialist biases in discussions of obesity among Australian print media. Here, we briefly comment on the implications of our findings for understanding genetic essentialism in media discourse. Given that this is a preliminary analysis of news articles from just one nation discussing just one trait, we provide only a limited interpretation of what our results say about the pervasiveness of genetic essentialist biases in the media more generally.

This isn’t to say that our results have no bearing on these broader questions. On the contrary: the model developed here provides a proof-of-concept for the methods by which we can arrive at these more general conclusions. We therefore begin by discussing the performance of our models, then offer a brief and preliminary interpretation of the results found thereby, and conclude by suggesting directions for future work.

**4.1 Model Performance**

Overall, performance varied widely across classes, both within and between models. Naive Bayes and Random Forest achieved reasonable to high scores on all metrics for both the specific aetiology and the homogeneity bias. Naive Bayes also demonstrated reasonable performance for the determinism class, but struggled to identify true instances of naturalness. Random Forest, conversely, performed better for naturalness than it did for determinism, though even for this bias it achieved impressive precision. In fact, Random Forest consistently outperformed all other models on this metric, generating the least false positives for three of the four classes. Precision, however, came at the cost of recall, particularly for the determinism bias, where Random Forest retrieved the fewest number of true positives.

Both BERT and the zero-shot implementation of GPT-3.5-TURBO exhibited the opposite trend: achieving high recall across most classes, but at the cost of precision, and hence returning a substantial number of false positives. The only model to avoid this trade-off was the few-shot implementation of GPT-3.5-TURBO, which identified all four of focal biases roughly as reliably as human experts, indicated by the convergence between its scores (see §3.2) and Cohen’s kappa for inter-rater reliability (see §3.1). Notable as well is the improvement of the few-shot protocol relative to its zero-shot implementation: increasing precision and F1 scores across all classes by an average of 45 and 28.5 points respectively.

These results corroborate previous work on ‘prompt engineering’ (e.g., Wei et al., 2022; Wang et al., 2022), confirming that relatively simple strategies such as chain of thought and self consistency can significantly increase the rate at which GPT-models generate accurate outputs. Of interest to researchers intending to implement these methods for similar classification tasks is that we augmented these two strategies with two additional prompting procedures, both of which further improved performance.

Recall that upon analysing the model’s justifications for its classifications, we encountered a substantial number of cases in which the model misclassifiedthe sentence, in oppositionto the reasoning given. This issue was addressed by prompting the model to generate its ‘reason’ first, and thenpredict the class. Ad-hoc analysis showed that this simple tweak increased F1 scores by an average of 11.5 points across all biases (see Appendix 6.1). While the opacity of LLMs precludes us from commenting with certainty on why this adjustment achieved the desired effect, it is reasonable to suppose that in the process of predicting each next token, the model's attentional architecture places *more weight* on *more recent tokens*. If this is the case, then getting the model to generate its reason first stands to increase the consistency of the classification because the preceding tokens, providing the reason, correlate strongly with the corresponding class. This, in turn, improves performance simply by virtue of the fact that the reasoning typically gets it right, such that generating the reason first increases the rate, or likelihood, of a corresponding and therefore correct classification.

Recall as well that we asked the model to return three classifications per query, utilizing the self consistency strategy. As noted in §2.1, determinism, naturalness, and specific aetiology were classified according to the standard implementation of this protocol, i.e., a majority decision. For the homogeneity bias, we adjusted the confidence threshold for a positive classification by predicating a positive response on a unanimousverdict. Thus, for this bias, *only* those queries for which *all three* responsescame back positive were classified as biased.[[10]](#footnote-10) This adjustment was motivated by the comparatively higher number of false positives that our model returned for the homogeneity class, and the intuition that increasing the number of votes required for a positive classification would counteract this issue. Additional analysis confirmed our intuition, with precision and F1 increasing by 29 and 13 points respectively following this adjustment (see Appendix §6.2). This increase in precision did come at the expense of reduced recall, owing to a greater number of false negatives. Though this trade-off is unsurprising, we are, to the best of our knowledge, the first to show that LLMs allow researchers to manually manipulate which of these two types of errors will be more prevalent by adjusting the number of votes required for a positive classification.

While the generalizability of these two strategies—call them *sequencing* and *thresholding*—calls for further empirical investigation, their marked effects in the present context underscore both the intricacies and the potential payoffs of prompt engineering, a feature that is unique to LLMs. Nor is there a principled reason for thinking that the performance of our model could not be improved further by further prompt refinements, suggesting that additional research into the use-case of LLMs as automated bias detection tools may yield even better results. We hope that some of the strategies discussed here will be of general use to researchers with an interest in identifying complex psychological constructs in large volumes of natural language.

**4.2 Genetic Essentialist Biases in Australian Print Media**

**4.2.1 Pervasiveness in discussion of obesity among Australian news outlets**

Discussions of obesity in our news corpus placed relatively little emphasis on genes, with less than 10% of articles mentioning gene-related concepts (Table 3). This result supports those who are sceptical of the “genoyhype” phenomenon, confirming that a focus on genetics is not as pervasive as once thought (Bubela & Caulfield, 2004). However, among articles that *do* mention genes, almost one third (31.26%) contained at least one of our focal biases; suggesting that although Australian print media don’t appear to fixate on genetic causes of obesity, when genetic explanations are invoked, they often encode essentialist beliefs. This supports recent empirical work on genetic essentialism as a cognitive phenomenon (Dar-Nimrod & Heine, 2011; Dar-Nimrod et al., 2014; Ahn & Lebowitz, 2018). However, whereas previous studies in this area illustrate how essentialist biases are elicited when individuals are presented with genetic information (e.g. in the form of vignettes), our results indicate that how genetic information is communicated in the media is also reflective of essentialist thinking about genes.

This finding has serious implications for science communication and its effects on the public, as there is ample evidence that individual tendencies to essentialise genes can result in prejudicial attitudes, negative self-beliefs, a perceived lack of control over one’s own destiny, and unhealthy behavioural choices (Bastian & Haslam, 2006; Dar-Nimrod & Heine, 2011; Dar-Nimrod et al., 2014; Ahn & Lebowitz, 2018). Hence, if media communication of genetics serves to emphasise and further entrench these biases, it could further contribute not only to an inaccurate picture of gene-action, but to negative social and health outcomes. We explore these issues in greater detail below.

**4.2.2 Prevalence of each Bias**

Figure 4 indicates that there is considerable variation in the prevalence of each bias at both the article and the sentence level. Comparing prevalence at these levels of analysis reveals an interesting trend: whereas specific aetiology is the most prevalent in terms of overall instances–or sentences–at the article level, the determinism bias appears more frequently. This tells us that although any *given* article about genes is more likely to contain the determinism bias, if that article is about a *specific* gene, the specific aetiology bias is likely to predominate. In the interest of space, we focus our discussion on the sentence level, where specific aetiology is followed closely by determinism, and then homogeneity and naturalness, which appear notably less frequently than the other two.

Specific aetiology classifications were largely comprised of “gene for” (e.g., “the gene for obesity”) and “X gene” (e.g., “the fat gene”) claims, which indicate that a single gene is portrayed as responsible for the trait of interest. Although there are a handful of specific genes implicated in syndromic obesity (Kaur et al., 2017), media discussions almost exclusively concern non-syndromic obesity. In these cases, obesity is associated with many mutations in many genes that are each thought to make a modest contribution, in combination with the effects of the environment (Loos & Yeo, 2022). Media reporting of “genes for” obesity therefore fundamentally misrepresents gene-action in these contexts.

Like specific aetiology, determinism is an oft emphasised element of genetic essentialism in media reporting (Nelkin & Lindee, 1995; Conrad, 1997). As noted, this bias involves the belief that genetically caused traits will inevitably come about no matter the circumstance; exemplified in discourses about one’s genes determining their fate. Such discourse finds little support in the scientific literature: all genes associated with non-syndromic obesity are highly sensitive to environmental factors such as diet and exercise (Loos & Yeo, 2022), and even patients with syndromic conditions can implement behavioural strategies to keep obesity at bay (Chung, 2012). That deterministic beliefs have been shown to negatively affect the implementation of these strategies (Dar-Nimrod et al., 2014; Ahn & Lebowitz, 2018) makes the prevalence of this bias particularly worrying. It also suggests that discussions of obesity that portray gene-action deterministically can result in a self-fulfilling prophecy, such that media reporting to this effect may itself be contributing towards a more obese society.

Another element of genetic essentialist thinking is that groups of people thought to share the same genes are perceived to be homogeneous and discrete, due to their shared genetic essences. These shared essences are thought to constitute category membership, eliciting stereotype endorsement (Bastian & Haslam, 2006). People who believe in a genetic basis of race, for instance, tend to amplify similarities between individuals classified as the same race, and differences between those of different races (No et al., 2008; Yaylaci et al., 2021). Our results indicate that a nontrivial proportion of gene sentences (~4.4%) imply the homogeneity bias, suggesting that media reporting about the genetics of obesity (occasionally) casts obese individuals as a homogeneous and discrete group. Consider, for instance, the following statement from our corpus, which describes Sumo wrestlers as being “*of a genotype* that allows them to lay down a lot of relatively healthy subcutaneous fat”, and then goes on to say that this is “why they don’t experience the same complications *that other* obese people do” (emphasis ours, published in *The Sydney Morning Herald*). While this particular example may seem relatively benign, it clearly illustrates our broader concern that claims about the genetics of particular groups can serve as a basis for stereotype endorsement. Of particular concern is that such stigmatization can lead obese individuals to internalize and enact negative self-concepts, such as being inherently lazy or weak-willed (Puhl & Heuer, 2010; Pearl et al., 2015). And our corpus indeed contained several statements that are likely to prompt such beliefs, describing obese people, to give just one example, as part of a “*tribe…*.without the restraint gene” (emphasis ours, published in *The Age*). As and when these sorts of beliefs take hold, the homogeneity bias, much like the determinism bias, potentially sets up a self-fulfilling obesity prophecy (Puhl & Brownell, 2003; Madon et al., 2018).

Consider next the naturalness bias, which was the least prevalent in our corpus. Recall that this bias is closely related to the naturalistic fallacy–where people believe that something is morally good or normatively acceptable in virtue of its natural existence. As and when this fallacy is prompted by exposure to genetic information, the naturalness bias is thought to moderate the negative effects of other essentialist beliefs by lending an air of moral permissibility to genetically caused traits. To give an example, people who believe in a genetic basis of sexual orientation are more accepting and hold less prejudicial attitudes towards non-heterosexuals (Jayaratne et al., 2006). Similarly, gay men who accept a genetic explanation for their own sexuality tend to have less self-stigma (Morandini et al., 2015). The comparatively low scores for the naturalness bias (89 out of the 978 biassed gene sentences across the corpus), then, could be indicative of a general perception of obesity as a negatively valenced trait (Flint et al. 2015), not acceptable or normally described as ‘natural’ under an essentialist framework. Indeed, obesity has been described as “...an *unnatural* or excessive amount of fat…” (Ferdowsy et al. 2021, emphasis added).

A broader issue that these results speak to concerns the epistemic component of ‘obesogenic environments’–a popular framework for connecting the various environmental conditions under which obesity is thought to thrive. Although this framework is deliberately broad, and expressly leaves room for social and cultural drivers of obesity (e.g., Swinburn et al., 1999), the idea that obesity-related (mis)information comprises an additional dimension of the obesogenic niche has received only scant attention (Kirk et al., 2010). Moreover, where information has been an object of study, its targets have been cursory: referring, for the most part, to nutritional labels and food advertisement (e.g., Burton et al., 2006; Halford et al., 2007). The evidence reviewed above suggests that interventions at this level are unlikely to influence individuals whose background beliefs about the genetics of obesity are essentialist in character. Developing a richer epistemology of obesity, and situating this epistemology within our broader understanding of obesogenic environments, strikes us as an important, philosophical project.

**4.2.3 Patterns of Co-occurrence**

Recall that sentences could be classified for the presence of more than one bias. Figure 5 shows all instances of co-occurrence at both the sentence and the article level. For brevity, we will discuss only pairwise combinations within single sentences. Specific aetiology and homogeneity occur frequently together, as do homogeneity and determinism, and determinism and specific aetiology. That the homogeneity bias frequently appears with both determinism and specific aetiology tells us that beliefs about the aetiology of specific genes often support inferences about group membership, and that group membership is often portrayed as a genetically determined feature of individual identity. This makes sense conceptually, given that the inferential potential of genes as markers of group membership seems to assume both a specific and deterministic notion of genetic causation. The pairing of specific aetiology and determinism, in turn, suggests that specific genes are often presented as having deterministic effects, and that such effects are often attributed to specific genes. This again makes sense, as claims that posit particular genes as responsible for particular traits presumably don’t get off the ground without a strong notion of genetic causation already in place.

Previous work by Dar-Nimrod and colleagues (unpublished manuscript)–validating the genetic essentialist framework as a psychological scale–has found strong correlations between these biases. Our results are broadly consistent with theirs, but also raise additional insights. First, consonant with an ecological theory of bias along the lines outlined in §1, we find that essentialist biases don’t just cluster together at a psychological level, but also regularly co-occur within media discourse. Second, whereas the mentioned study presented experimental subjects with vignettes that carefully individuated each bias, our results suggest that, in the ‘wild’, these biases are frequently encountered within a single statement. This raises questions regarding possible interaction effects. One issue to consider is whether the prejudicial attitudes associated with the homogeneity bias are exacerbated by its co-occurrence with either or both specific aetiology and determinism. One hypothesis is that the more fine-grained the basis of genetic discrimination (specific aetiology), and the more immutable genetic categories are perceived to be (determinism), the more numerous and pronounced potential fault lines between groups become. If this is right, then statements that embed the homogeneity bias alongside either or both determinism and specific aetiology stand to amplify prejudice, not just in the context of obesity, but in other domains as well. The interactions between specific aetiology and determinism are of equal interest. It is, again, conceivable that these biases augment one another, such that claims about the deterministic effects of specific genes negatively implicate health behaviours above and beyond either deterministic claims, or claims that attribute obesity to specific gene mutations. These are empirical questions ripe for further investigation.

**4.2.4 Distribution of biases across outlets**

As noted, Figure 6a shows marked variation in the raw number of biases per venue, suggesting that some outlets are significantly more biased than others. This result, however, fails to account for both the composition of our corpus, and the prevalence of discussions about genetics across outlets. A fair assessment considers these finer points. Figures 6b and 6d show that the proportion of articles and sentences referring to genetics differs considerably from one venue to the next. When this is accounted for, the percentage of articles and sentences about genetics that contain biases is relatively uniform across outlets, with three notable exceptions (see Figures 6c and 6d). At one end, and despite being the *least* biased outlet in absolute terms, relative to its contribution to our corpus, as well as the number of sentences and articles about genes, the Brisbane Times emerges as the *most* biased outlet by a considerable margin. In fact, our results show that almost one in two articles by this outlet that discuss the genetics of obesity contain one or more of our focal biases, suggesting that patrons of the Brisbane Times with an interest in this issue are routinely exposed to essentialist beliefs about the genetic basis of obesity. Conversely, and despite being the *most* biased venue in absolute terms, when normalized for the relevant criteria, the Sydney Morning Herald is among the least biased outlets, bested only by the Northern Territory News.

Apart from these extremes, we find only modest variation in the rate at which different venues communicate essentialist biases, though genetics discourse is notably more prevalent among some outlets. In both relative and absolute terms, the Australian and the Sydney Morning Herald appear to devote most attention to this topic. The Northern Territory News and the Mercury on the other hand only rarely invoke genetics in discussions of obesity, suggesting regional differences within Australia’s broader media landscape.

**5. Conclusion and future directions**

While there is ample evidence that exposure to genetic information elicits essentialist beliefs in individuals (Gould & Heine, 2012; Castéra & Clément, 2014; Carver et al., 2017; Machery et al., 2019; Yaylaci et al., 2021), there remains considerable disagreement about the prevalence of genetic essentialism in media discourse (Condit et al., 1998; Condit, 2011; Bubela & Caulfield, 2004; Holtzman, 1997). These disagreements are, at least in part, predicated on an empirical ambiguity: absent a large-scale analysis, who is to say how pervasive this phenomenon really is?

The present study set out to fill this gap by leveraging recent developments in artificial intelligence (AI), large language models (LLMs), and natural language processing (NLP) to systematically evaluate a significantly larger sample of relevant articles. Of the five models trained and tested, the few-shot implementation of GPT-3.5-Turbo consistently achieved impressive results. This model in fact detected biases as effectively as human experts, suggesting that the methods developed in this study could be of general interest for future research into classifying biases and other complex constructs across large volumes of text (see also Lupo et al., 2023; Dunivin, 2024). Of particular interest to future researchers in this domain was the finding that prompting the model to generate its reason before rendering its classification improved performance substantially, and that adjusting the number of votes required for a positive classification allows optimising for precision or recall respectively.

While the few-shot implementation of GPT-3.5-TURBO was the optimal model for the present task, it is worth reiterating that this conclusion was drawn based on evaluative factors, including a preference for false negatives over false positives (i.e. optimising precision). This preference was due to the inductive risk involved in overestimating the prevalence of genetic essentialism in Australian obesity discourse, which could support false accusations of bias. Other classification tasks may evaluate inductive risk differently, preferring to overestimate false positives. Whatever the aims of a particular study, it is important that researchers using these kinds of classification models are aware of and explicit about the goals and values embedded in their research.

As for the present study, there are various other contexts in which the investigation of genetic essentialist biases will be relevant and interesting, including additional obesity corpora outside Australia (see e.g., Brookes & Baker, 2021; Coltman-Patel, 2023) and corpora comprised of alternative sources. Doing so would shed light on both regional differences, and differences between more or less formal and regulated discourses. A broader project made possible by our model, then, would involve studying genetic essentialist biases across other news channels (e.g., television and radio transcripts), social media, science communication, and academic and health literature intended for public audiences, including government and private web pages providing health information. One might assume that genetic essentialism is less pervasive in the scientific literature, though there is evidence that the biases discussed here can influence health professionals (Lane et al., 2023), and even researchers working on genetics (Lynch, et al., 2019).

Additional contexts for future investigation also include discussions of other traits, including physiology, disease traits, mental health, cognition, personality, and sexuality. Whether genetic essentialism is prevalent across various types of media, and whether the degree of genetic essentialism present when communicating about genes is trait-specific are empirical questions that warrant future research.

An important question, and potential limitation, raised by these prospects is how easily the present model–prompts included–can be repurposed to detect genetic essentialist biases in other contexts? One possibility is that our model is keyed to linguistic features unique to discussions of obesity. If this is the case, then investigating genetic essentialist biases for other traits would require considerable modification, at least at the level of prompt engineering. In anticipation of this issue, we deliberately kept many of our prompts as general as possible, targeting syntactic and lexical structures that we believe to be indicative of essentialist biases across a broad range of contexts. Recall for instance the example given in §2.2, where the phrase ‘gene for’ is encoded into the chain of thought as explaining why the focal sentence contains the specific aetiology bias. While the gene in this particular example is the ‘gene for’ appetite, it could just as well be the ‘gene for’ homosexuality, or intelligence, *or any other trait.*

In making these and other empirical insights more readily available, our model may also shed light on broader conceptual questions. One interesting and open question is whether biological entities other than genes are essentialized in similar ways? And, if so, whether we can view genetic essentialism as a particular instantiation of biological essentialism more broadly? More broadly still, do people essentialize non-biological entities? And what does this tell us about the nature of human cognition? Though answering these questions would require developing additional, topic-specific lexica, once developed, these lexica could be easily plugged into our pipeline. That the technical steps implemented in the present study potentially bear on these broader questions points to the yet more general insight that even a discipline as remote as philosophy does well by availing itself of advances in computational methods and taking seriously the results generated thereby.

As for the results presented here, our analysis gives cause for both optimism and concern. We have, on the one hand, shown that advances in NLP, and LLMs specifically, stand to further advance the state-of-the-art in automated bias detection. This is of critical importance, given the increasingly dense, polluted, and fast-paced nature of our informational environments. Looking at one such environment, and focusing specifically on genetic essentialist biases, we also found that Australian print media do not appear to fixate on genetic effects, at least in the context of obesity discourse. This too makes for an optimistic conclusion.[[11]](#footnote-11) This optimism, however, calls for qualification. For although genes figure less prominently in popular discussions of obesity than previous work might suggest, when genes are mentioned, there is a high chance that their effects are presented in a biased way. The downstream consequences of these biases–specifically with respect to prejudice and the implementation of positive health behaviours–underline the urgency of our findings.

# **References**

Abedin, E., Ferreira, M., Reimann, R., Cheong, M., Grossmann, I., & Alfano, M. (2023). Exploring intellectual humility through the lens of artificial intelligence: Top terms, features and a predictive model. *Acta Psychologica*, *238*, 103979.

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for TextBased Emotion Recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 117–121). Retrieved 2023-12-19, from [https://ieeexplore.ieee.org/ abstract/document/9317379](https://ieeexplore.ieee.org/abstract/document/9317379) doi: 10.1109/ICCWAMTIP51612.2020.9317379

Ahmed, Z., Vidgen, B., & Hale, S. A. (2022). Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, *11*(1), 8. Retrieved 2023-12-

19, from [https://epjds.epj.org/articles/epjdata/abs/2022/01/13688\_2022\_Article\_319/13688\_2022 \_Article\_319.html](https://epjds.epj.org/articles/epjdata/abs/2022/01/13688_2022_Article_319/13688_2022_Article_319.html)

Ahn, W. K., & Lebowitz, M. S. (2018). An experiment assessing effects of personalized feedback about genetic susceptibility to obesity on attitudes towards diet and exercise. *Appetite*, *120*, 23-31.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *Natural language processing and information systems: 23rd international conference on applications of natural language to information systems, nldb 2018, paris, france, june 13-15, 2018, proceedings 23* (pp. 57–64).

Aspinwall, Lisa G., Teneille R. Brown, and James Tabery. (2012). The double-edged sword: does biomechanism increase or decrease judges' sentencing of psychopaths?. *Science* 337( 6096), 846-849.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion* (pp. 759– 760). ACM Press. Retrieved 2023-12-19, from <http://dl.acm.org/citation.cfm?doid=3041021.3054223> doi:

10.1145/3041021.3054223

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, *42*(2), 228-235. Retrieved from [https://www.sciencedirect.com/science/article/pii/ S0022103105000478](https://www.sciencedirect.com/science/article/pii/S0022103105000478) doi: https://doi.org/10.1016/j.jesp.2005.03.003

Bednarek, M., Bray, C., Vanichkina, D. P., Brookes, G., Bonfiglioli, C., Coltman-Patel, T., … Baker, P. (2023). Weight stigma: Towards a language-informed analytical framework. *Applied Linguistics*, amad033.

Brookes, G., & Baker, P. (2021). *Obesity in the news: Language and representation in the press*. Cambridge University Press.

Bubela, T. M., & Caulfield, T. A. (2004). Do the print media “hype” genetic research? a comparison of newspaper stories and peer-reviewed research papers. *Cmaj*, *170*(9), 1399–1407.

Burton, S., Creyer, E. H., Kees, J., & Huggins, K. (2006). Attacking the obesity epidemic: the potential health benefits of providing nutrition information in restaurants. *American journal of public health*, *96*(9), 1669-1675.

Carver, R. B., Castéra, J., Gericke, N., Evangelista, N. A. M., & El-Hani, C. N. (2017). Young adults’ belief in genetic determinism, and knowledge and attitudes towards modern genetics and genomics: the puggs questionnaire. *PloS one*, *12*(1), e0169808.

Castéra, J., & Clément, P. (2014). Teachers’ conceptions about the genetic determinism of human behaviour: A survey in 23 countries. *Science & Education*, *23*, 417–443.

Christensen, K., Jayaratne, T., Roberts, J., Kardia, S., & Petty, E. (2010). Understandings of Basic Genetics in the United States: Results from a National Survey of Black and White Men and Women. *Public Health Genomics*, *13*(7-8), 467-476.

Chung, W. K. (2012). An overview of monogenic and syndromic obesities in humans. *Pediatric blood & cancer*, *58*(1), 122-128.

Coltman-Patel, T. (2023). *(mis) representing weight and obesity in the british press: Fear, divisiveness, shame and stigma*. Springer Nature.

Condit, C. M. (2011). When do people deploy genetic determinism? a review pointing to the need for multi-factorial theories of public utilization of scientific discourses. *Sociology Compass*, *5*(7), 618–635.

Condit, C. M., Ofulue, N., & Sheedy, K. M. (1998). Determinism and mass-media portrayals of genetics. *The American Journal of Human Genetics*, *62*(4), 979–984.

Conrad, P. (1997). Public eyes and private genes: Historical frames, news constructions, and social problems. *Soc. Probs.*, *44*, 139.

Conrad, P., & Markens, S. (2001). Constructing the ‘gay gene’ in the news: optimism and skepticism in the us and british press. *Health:*, *5*(3), 373–400.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive determinism of DNA. *Psychological bulletin*, *137*(5), 800. Retrieved 2023-12-19, from <https://psycnet.apa.org/journals/bul/137/5/800/>

Dar-Nimrod, I., Zuckerman, M., & Duberstein, P. (2014). Smoking at the workplace: Effects of genetic and environmental causal accounts on attitudes toward smoking employees and restrictive policies. *New Genetics and Society*, *33*(4), 400– 412. Retrieved 2023-12-19, from <https://www.tandfonline.com/doi/full/10.1080/14636778.2014.951993>doi: 10.1080/14636778.2014.951993

Dar-Nimrod, I., Kuntzman, R., MacNevin, G., Lynch, K., Woods, M., & Morandini, J. (2021). Genetic essentialism: The mediating role of essentialist biases on the relationship between genetic knowledge and the interpretations of genetic information. *European Journal of Medical Genetics*, 64(1), 104119.

de Carvalho, F. N., & Krueger, J. (2023). Biases in niche construction. *Philosophical Psychology*, 1–31.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of science*, *67*(4), 559–579.

Dunivin, Z. (2024). Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks.*arXiv preprint at arXiv:2401.15170*.

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*, 100053.

Fleising, U. (2001). In search of genohype: a content analysis of biotechnology company documents. *New Genetics and Society*, *20*(3), 239–254.

Flint, S. W., Hudson, J., & Lavallee, D. (2015). UK adults’ implicit and explicit attitudes towards obesity: a cross-sectional study. *BMC obesity,* 2, 1-8.Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. Retrieved 2023-12-19, from <https://dl.acm.org/doi/10.1145/3232676> doi: 10.1145/3232676

Gould, W. A., & Heine, S. J. (2012). Implicit essentialism: Genetic concepts are implicitly associated with fate concepts. *PloS one*, *7*(6), e38176.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological review*, *109*(1), 3.

Griffiths, P. E. (2002, 12). What is Innateness? *The Monist*, *85*(1), 70-85. Retrieved from [https://doi.org/10.5840/ monist20028518](https://doi.org/10.5840/monist20028518) doi: 10.5840/monist20028518

Halford, J. C., Boyland, E. J., Hughes, G., Oliveira, L. P., & Dovey, T. M. (2007). Beyond-brand effect of television (TV) food advertisements/commercials on caloric intake and food choice of 5–7-year-old children. *Appetite*, *49*(1), 263-267.

Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, *20*(4), 391–415.

Harden, K. P. (2021). “Reports of my death were greatly exaggerated”: Behavior genetics in the postgenomic era. *Annual review of psychology,* 72(1), 37-60.

Haslam, N., & Kvaale, E. P. (2015). Biogenetic explanations of mental disorder: The mixed-blessings model. *Current Directions in Psychological Science,* 24(5), 399-404.

Herd, Pamela, Jeremy Freese, Kamil Sicinski, Benjamin W. Domingue, Kathleen Mullan Harris, Caiping Wei, and Robert M. Hauser. "Genes, gender inequality, and educational attainment." *American Sociological Review* 84, no. 6 (2019): 1069-1098.

Holtzman, N. A. (1997). Testing for genetic susceptibility: What you see is not what you get. *Accountability in Research*, *5*(1-3), 95–101.

Holtzman, N. A. (1999). Are genetic tests adequately regulated? *Science*, *286*(5439), 409-409. Retrieved from [https:// www.science.org/doi/abs/10.1126/science.286.5439.409](https://www.science.org/doi/abs/10.1126/science.286.5439.409) doi: 10.1126/science.286.5439.409

Jayaratne, T. E., Ybarra, O., Sheldon, J. P., Brown, T. N., Feldbaum, M., Pfeffer, C. A., & Petty, E. M. (2006). White Americans’ Genetic Lay Theories of Race Differences and Sexual Orientation: Their Relationship with Prejudice toward Blacks, and Gay Men and Lesbians. *Group Processes & Intergroup Relations*, *9*(1), 77–94. Retrieved 2023-12-19, from [http:// journals.sagepub.com/doi/10.1177/1368430206059863](http://journals.sagepub.com/doi/10.1177/1368430206059863) doi: 10.1177/1368430206059863

Karaca, K. (2021). Values and inductive risk in machine learning modelling: the case of binary classification models. *European journal for philosophy of science*, *11*(4), 102.

Kaur, Y., de Souza, R. J., Gibson, W. T., & Meyre, D. (n.d.). A systematic review of genetic syndromes with obesity. *Obesity Reviews*, *18*(6), 603-634. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/obr.12531>doi: <https://doi.org/10.1111/obr.12531>

Kirk, S. F., Penney, T. L., & McHugh, T. L. (2010). Characterizing the obesogenic environment: the state of the evidence with directions for future research. *Obesity reviews*, *11*(2), 109-117.

Klein, C., Reimann, R., Quintana, I. O., Cheong, M., Ferreira, M., & Alfano, M. (2022). Attention and counter-framing in the black lives matter movement on twitter. *Humanities and Social Sciences Communications*, *9*(1), 1–12.

Kronfeldner, M. E. (2009-06). Genetic Determinism and the Innate-Acquired Distinction in Medicine. *Medicine Studies*, *1*(2), 167–181. Retrieved 2023-12-19, from <http://link.springer.com/10.1007/s12376-009-0014-8> doi:

10.1007/s12376-009-0014-8

Lane, A. S., Lynch, K. E., Arnold, M., Dar-Nimrod, I., Morandini, J., Gawronski, S. A., & Griffiths, P. E. (2023). The undue influence of genetic information on senior medical students’ treatment decisions. *BMC Medical Education*, *23*(1),

938. Retrieved 2023-12-20, from [https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-023 -04895-w](https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-023-04895-w) doi: 10.1186/s12909-023-04895-w

Leboeuf, C. (2020). The embodied biased mind. *An introduction to implicit bias: Knowledge, justice, and the social mind. Routledge*.

Lebowitz, M. S., & Ahn, W. K. (2014). Effects of biological explanations for mental disorders on clinicians’ empathy. *Proceedings of the National Academy of Sciences,* 111(50), 17786-17790.

Lim, S., Jatowt, A., & Yoshikawa, M. (2018). Understanding characteristics of biased sentences in news articles. In *Cikm workshops* (pp. 121–128).

Linquist, S., Machery, E., Griffiths, P. E., & Stotz, K. (2011). Exploring the folkbiological conception of human nature. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1563), 444–453.

Liu, Y. (2024). Quantifying Stereotypes in Language. *arXiv preprint arXiv:2401.15535*.

Loo, L. K., Byrne, J. M., Hardin, S. B., Castro, D., & Fisher, F. P. (1998). Reporting medical information: Does the lay press get it right? *Journal of General Internal Medicine*, *13*(1), 60

Loos, R. J., & Yeo, G. S. (2022). The genetics of obesity: from discovery to biology. *Nature Reviews Genetics*, *23*(2), 120–133.

Lukin, A. (2005). Mapping media bias: a multidimensional affair.[public right to know conference, august 2004.]. *Australian Journalism Review*, *27*(1), 139–155.

Lupo, L., Magnusson, O., Hovy, D., Naurin, E. & L. Wangnerud (2023). Towards human-level text coding with LLMs: The case of fatherhood roles in public policy documents. *Retrieved from:* [*https://arxiv.org/abs/2311.11844*](https://arxiv.org/abs/2311.11844)

Lynch, K. E., Morandini, J. S., Dar-Nimrod, I., & Griffiths, P. E. (2019). Causal reasoning about human behavior genetics: synthesis and future directions. *Behavior Genetics*, *49*, 221–234.

Lynch, K. E. (2021). The meaning of “cause” in genetics. *Cold Spring Harbor Perspectives in Medicine*, *11*(9), a040519.

Machery, E. (2022). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(1), e1569.

Machery, E., Griffiths, P., Linquist, S., & Stotz, K. (2019). Scientists’ concepts of innateness: Evolution or attraction. *Advances in experimental philosophy of science*, 172–201.

Madon, S., Jussim, L., Guyll, M., Nofziger, H., Salib, E. R., Willard, J., & Scherr, K. C. (2018). The accumulation of stereotype-based self-fulfilling prophecies. *Journal of Personality and Social Psychology*, *115*(5), 825.

McAllister, L., Daly, M., Chandler, P., McNatt, M., Benham, A., & Boykoff, M. (2021). Balance as bias, resolute on the retreat? updates & analyses of newspaper coverage in the united states, united kingdom, new zealand, australia and canada over the past 15 years. *Environmental Research Letters*, *16*(9), 094008.

McCombs, M. E. (2014). *Setting the agenda: Mass media and public opinion.* (2nd ed. ed.). Wiley.

Meldrum, D. R., Morris, M. A., & Gambone, J. C. (2017). Obesity pandemic: causes, consequences, and solutions—but do we have the will?. *Fertility and sterility*, *107*(4), 833-839.

Morandini, J. S., Blaszczynski, A., Ross, M. W., Costa, D. S., & Dar-Nimrod, I. (2015). Essentialist beliefs, sexual identity uncertainty, internalized homonegativity and psychological wellbeing in gay men. *Journal of counseling psychology*, *62*(3), 413.

Nelkin, D., & Lindee, M. S. (1995). *The DNA mystique: The gene as a cultural icon*. Freeman.

No, S., Hong, Y.-y., Liao, H.-Y., Lee, K., Wood, D., & Chao, M. M. (2008). Lay theory of race affects and moderates Asian

Americans’ responses toward American culture. *Journal of personality and social psychology*, *95*(4), 991. Retrieved 2023-12-19, from <https://psycnet.apa.org/record/2008-12903-016>

Ojea Quintana, I., Reimann, R., Cheong, M., Alfano, M., & Klein, C. (2022). Polarization and trust in the evolution of vaccine discourse on twitter during covid-19. *PLos One*, *17*(12), e0277292.

Pearl, R. L., Puhl, R. M., & Dovidio, J. F. (2015). Differential effects of weight bias experiences and internalization on exercise among women with overweight and obesity. *Journal of health psychology*, *20*(12), 1626-1632.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Puhl, R., & Brownell, K. D. (2003). Ways of coping with obesity stigma: review and conceptual analysis. *Eating behaviors*, *4*(1), 53-78.

Puhl, R. M., & Heuer, C. A. (2010). Obesity stigma: important considerations for public health. *American journal of public health*, *100*(6), 1019-1028.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1650–1659).

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. arXiv e-prints, page. *arXiv preprint arXiv:2002.12327*.

Roth, J., Qiang, X., Marbán, S. L., Redelt, H., & Lowell, B. C. (2004). The obesity pandemic: where have we been and where are we going?. *Obesity research*, *12*(S11), 88S-101S.

Schuster, M., & Nakajima, K. (2012). Japanese and Korean Voice Search..

Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., & Donnay, K. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, *58*(3), 102505. Retrieved 2023-12-19, from [https://www.sciencedirect.com/science/article/ pii/S0306457321000157](https://www.sciencedirect.com/science/article/pii/S0306457321000157)

Starace, G., Papakostas, K., Choenni, R., Panagiotopoulos, A., Rosati, M., Leidinger, A., & Shutova, E. (2023). Probing llms for joint encoding of linguistic categories. *arXiv preprint arXiv:2310.18696*.

Stavropoulos, A., Crone, D., & Grossmann, I. (2023). Shadows of wisdom: Classifying meta-cognitive and morally-grounded narrative content via large language models.

Swinburn, B. A., Egger, G., & Raza, F. (1999). Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Preventive medicine*, *29*(6), 563-570.

Swinburn, B. A., Sacks, G., Hall, K. D., McPherson, K., Finegood, D. T., Moodie, M. L., & Gortmaker, S. L. (2011). The global obesity pandemic: shaped by global drivers and local environments. The Lancet, 378(9793), 804-814.

Tabery, J. (2014). Beyond versus: The struggle to understand the interaction of nature and nurture. mit Press.

Turkheimer E. Heritability and biological explanation. Psychological Review. 1998;105:782–791. doi: 10.1037/0033-295X.105.4.782-791.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. Current directions in psychological science, 9(5), 160-164.

Turnwald, Bradley P., J. Parker Goyer, Danielle Z. Boles, Amy Silder, Scott L. Delp, and Alia J. Crum. (2019). Learning one’s genetic risk changes physiology independent of actual genetic risk." *Nature Human Behaviour* 3(1): 48-56.

Vanichkina, D. P., & Bednarek, M. (2022). Australian obesity corpus manual. *Open science framework*. Retrieved from <https://osf.io/h6n82>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., … Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., … others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Winsberg, E. (2012). Values and uncertainties in the predictions of global climate models. *Kennedy Institute of Ethics Journal*, *22*(2), 111–137.

Xu, H., Lou, R., Du, J., Mahzoon, V., Talebianaraki, E., Zhou, Z., Garrison, E., Vucetic, S. & Yin, W. (2024). LLMs' Classification Performance is Overclaimed. *arXiv preprint arXiv:2406.16203*.

Yaylaci, S., Roth, W. D., & Jaffe, K. (2021). Measuring racial essentialism in the genomic era: The genetic essentialism scale for race (GESR). *Current Psychology*, *40*(8), 3794–3808. Retrieved 2023-12-19, from [https://link.springer.com/](https://link.springer.com/10.1007/s12144-019-00311-z)

[10.1007/s12144-019-00311-z](https://link.springer.com/10.1007/s12144-019-00311-z) doi: 10.1007/s12144-019-00311-z

Zhao, Z., Zhang, Z., & Hopfgartner, F. (2021). AComparativeStudyofUsingPre-trainedLanguageModelsforToxicComment Classification. In *Companion Proceedings of the Web Conference 2021* (pp. 500–507). ACM. Retrieved 2023-12-19, from <https://dl.acm.org/doi/10.1145/3442442.3452313> doi: 10.1145/3442442.3452313

Zhou, X. (2021). *Challenges in automated debiasing for toxic language detection*. University of Washington. Retrieved 2023-12-19, from [https://search.proquest.com/openview/c3d92b4e2d5fbe76a7a3b44be59176af/1 ?pq-origsite=gscholar&cbl=18750&diss=y](https://search.proquest.com/openview/c3d92b4e2d5fbe76a7a3b44be59176af/1?pq-origsite=gscholar&cbl=18750&diss=y)

**APPENDIX**

All data, results, and analysis included here are also available via the Open Science Framework:

*https*://*osf*.*io*/*r*2*cg*6/?*view*\_*only* = 5749*e*65*fa*4*a*246*fdabba*0754*a*5*e*3*e*52*f*.

Note that we have also included a link to our notebooks and code in the ‘notebooks and code’ folder. These are publicly available, and should enable others to both replicate and extend our work with minimal effort.

# **1 . GENE LEXEME**

To extract sentences mentioning genetics-related concepts from our corpus, we developed a gene lexeme, consisting of 23 (stemmed) gene-related terms. These terms are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| allel\* | chromosom\* | cistron | deoxyribonucleic acid | DNA |
| gene | genes | genetic\* | genic | genom\* |
| mutat\* | nucleotide | oncogen\* | polygen\* | polymorphism |
| transgene | transgeneic | transgenes | wild-type | wild type |
| homozygous | heterozygous | hereditary |  |  |

**T A B L E 1**

**2. CODING RULES**

General Rules:

1. Sentence must be about gene(s) affecting trait(s), not about trait(s) affecting gene(s)
2. Code each statement for all four biases, indicating presence of bias with (1) and absence of bias with (0)
3. *Do not* code for negations (e.g., your genes are *not* to blame = 0)
4. *Do not* code for fictive scenarios, hypotheticals, counterfactuals, or other rhetorical devices.

**Determinism:** When a trait is portrayed as present due to a fixed set of underlying processes that are *inevitable* and *beyond control*, or when a trait is portrayed as *determined, destined* or *fated* to develop due to genes. The core of this bias is that if a trait has a genetic cause, *it will develop no matter what*.

1. When a trait is presented as *inevitable* or *immutable* due to gene(s), code as bias (1).
2. When genes are portrayed as *‘hardware’* or *‘blueprint’*: *controlling* or *programming* for the expression of a trait, code as bias (1).
3. When *fate*, *destiny*, *blame*, and *luck* are invoked, as when someone is lucky to have certain genes, or destined to be a certain way based on their genes, code as bias (1).
4. When a trait or behaviour is portrayed as *‘in the genes’* e.g., obesity is *in* her genes, code as bias (1).
5. When gene trait relation is described with moderating adverbs, e.g., genes *mostly*, *often*, or *sometimes* lead to a trait, code as non-biased (0)
6. When gene trait relation is described in probabilistic or statistical terms, as when 70% of people with a gene have a trait, or a gene increases the risk of a trait by 30%, code as non-biased (0).
7. When genes are said to *predispose* or *relate to*, or are *associated with* or *involved in*, code as non-biased (0).

**Specific Aetiology:** When a *specific* trait is portrayed as having a *specific* gene as its cause, and other causal factors such as environmental influences on gene-expression are downplayed: presenting specific genes, if not as the sole cause, then the most important one. This is often seen in claims about the “gene for” things like obesity, intelligence etc.

1. When specific genes are described as the causal factor - e.g., the *key*, *critical*, or *crucial* cause – for particular traits, code as bias (1)
2. Whenever a gene is described as the ‘x gene’ – e.g., the ‘obesity gene’ – or the ‘gene for x’ – e.g., the ‘gene for obesity – code as bias (1)
3. Code as bias *even if* the ‘x-gene’ or ‘gene for x’ is portrayed as only sometimes leading to or merely being *associated with* some trait. For example, if the obesity gene is said to lead to obesity in 70% of cases, or increase the risk of obesity by 30%, still code as bias (1)

**Naturalness:** When a trait is portrayed as ‘natural’ due to the presence of a gene. ‘Natural’ in this context has two, closely related referents. The first is moralistic and involves the naturalistic fallacy. The second is normative in a broader sense and involves portraying some trait as ‘normal’ due to some gene, usually in light of its ‘adaptive’ or ‘evolutionary’ value.

1. Whenever a statement about genes runs afoul of the naturalistic fallacy, code as bias (1).
2. Whenever a trait is portrayed as ‘normal’ due to some gene, code as bias (1).
3. Whenever a trait is moralized or normalized by appealing to evolutionary concepts, specifically selection pressures acting on genes, classify as bias (1). For instance, when gluttony is portrayed as normal or morally permissible because we evolved under conditions of famine and feast, code as bias (1).

**Homogeneity:** When genetic similarities and differences within and between groups of people are identified and portrayed as being definitive and constitutive of these groups.

1. When genes are invoked to categorize groups, or portrayed as constitutive of some group, code as bias (1).
2. When physiological and behavioural similarities are attributed to genetic similarities, code as bias (1).
3. When similar genes are identified between different species - e.g., humans and apes are genetically similar - do not code as bias (0).
4. When the described similarities are between members of the same family – e.g., she and her sister had the same genes – do not code as bias (0).

**3. BERT**

As illustrated in Figure (1), the [CLS] token initiates a sequence, while the [SEP] token separates one sentence from the next. Since our analysis focuses on single sentences, we ‘padded’ the tokens following [SEP] with [PAD] tokens. For the inputs, we tokenized our text using the wordpiece tokenizer algorithm (Schuster & Nakajima, 2012) and fitted text tokens between the [CLS] and [SEP] tokens. For the outputs, the final hidden layer’s [CLS] token was used to represent the sentence embedding. A feedforward layer with a sigmoid activation function was attached on top to map the 768-dimensional sentence embedding onto a probability space between 0 and 1, where a probability of > 0.5 corresponds to the presence of a bias.

A screenshot of a computer

Description automatically generated

**FIGURE 1** BERT token embedding (Devlin et al., 2018)

Since we trained four binary classifiers—one for each bias—we used binary cross entropy as our objective function.This function was optimised using mini-batch gradient descent and the mean loss of the batch was used. We again implemented a 60:20:20 training-validation-test split: training the model on 60% of our annotated sentences, using the next 20% for validation, and then testing performance by prompting the model to predict the classes of the remaining 20%. We also leveraged early stopping—a common technique that terminates training when the mean loss no longer decreases after a set number of steps—to avoid overfitting.

**4. GPT PROMPTS**

All prompts for all implementations of GPT-3.5-TURBO can be found in the ‘prompts’ folder available via the Open Science Framework (see link at top of page).

**5. DATA, HUMAN ANNOTATION, AND INTER-RATER RELIABILITY**

All data, including annotated sentences used for training and testing, can be found in the ‘data and annotation’ folder available via the Open Science Framework (see link at top of page).

**6. MODEL PERFORMANCE AND OPTIMIZATION**

Performance of all models can be found in the ‘model performance’ folder available via the Open Science Framework (see link at top of page).

**6.1 Sequencing:**

Table 2 shows changes in performance upon implementing the ‘sequencing’ strategy i.e., prompting the few-shot implementation of GPT-3.5-TURBO to generate its reasoning before generating the classification. Note that whereas recall decreases slightly across three of the four classes, precision and F1 scores increase substantially.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **GPT-3.5-TURBO few-shot** | | | **GPT-3.5-TURBO few-shot-sequenced** | | |
| **precision** | **recall** | **F1** | **precision** | **recall** | **F1** |
| **determinism** | 0.84 | 0.75 | 0.79 | 0.85 | 0.79 | 0.82 |
| **specific aetiology** | 0.61 | 0.91 | 0.73 | 0.85 | 0.77 | 0.81 |
| **naturalness** | 0.46 | 0.86 | 0.6 | 0.83 | 0.71 | 0.77 |
| **homogeneity** | 0.57 | 0.92 | 0.7 | 0.92 | 0.85 | 0.88 |

**T A B L E 2** Sequencing

**6.2 Thresholding:**

Table 3 shows changes in performance upon implementing the ‘one vote neutral equals neutral’ strategy for the sequenced-few-shot implementation of GTP-3.5-Turbo. Note that homogeneity is the only class for which this implementation of the model improved performance. In particular, and although recall decreases slightly, both precision and F1 increase significantly. For the other three classes, precision remains roughly constant, while recall and F1 decrease considerably.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Majority prediction** | | | **One vote neutral equals neutral** | | |
| **precision** | **recall** | **F1** | **precision** | **recall** | **F1** |
| **determinism** | 0.85 | 0.79 | 0.82 | 0.83 | 0.54 | 0.65 |
| **specific aetiology** | 0.85 | 0.77 | 0.81 | 0.82 | 0.61 | 0.69 |
| **naturalness** | 0.83 | 0.71 | 0.77 | 0.8 | 0.57 | 0.66 |
| **homogeneity** | 0.63 | 0.92 | 0.75 | 0.92 | 0.85 | 0.88 |

**T A B L E 3** Thresholding

1. A strict or strong sense of genetic determinism amounts to genetic fatalism: the idea that a phenotypic outcome will develop no matter the environment of the individual. i.e., there is, in principle, no environmental or background condition that can be modified to prevent a genetic outcome (say a disease caused by a gene mutation) from occurring. A weaker sense of genetic determinism is the idea that a gene produces its effect across most and/or a wide range of normal environmental backgrounds. This amounts to having a flat or ‘rigid’ reaction norm (Kronfeldner, 2009-06). We refer to genetic determinism in the latter sense, see §2 of Appendix for full details of how we operationalized this bias. [↑](#footnote-ref-1)
2. By exceptional cases we are referring principally to syndromic obesity. Conditions such as Prader-Willi syndrome (PWS), for instance, have a strong genetic basis, and the expression of this and similar conditions is (typically) not contingent on environmental factors. It would, nevertheless, be a mistake to characterize even conditions such as PWS as *deterministically* implicated in the development of obesity, as even here lifestyle factors such as diet and exercise are effective remedial strategies. Worth noting as well is that PWS–the most common form of syndromic obesity–has an incident rate of approximately one in 15,000-25,000 live births (Chung, 2012). [↑](#footnote-ref-2)
3. Worth noting as well is that, in combination with an understanding of biases as embodied perceptual habits (e.g., Leboeuf, 2020), an ecological theory of bias along the lines outlined here sidesteps a number of criticisms recently levelled at more traditional conceptualizations of bias (Machery, 2022), particularly those that treat bias as a subset of implicit attitudes (e.g., Greenwald et al., 2002). [↑](#footnote-ref-3)
4. tf-idf stands for 'term frequency-inverse document frequency', and provides a statistical measure of the relevance of a token in a document relative to a collection of documents. [↑](#footnote-ref-4)
5. https://osf.io/r2cg6/?view\_only = 5749e65fa4a246fdabba0754a5e3e52f [↑](#footnote-ref-5)
6. https://cqpw-prod.vip.sydney.edu.au/CQPweb/ [↑](#footnote-ref-6)
7. This description is, needless to say, highly truncated. Another important ingredient that sets (most) transformer-based models apart from (most) traditional machine learning methods is that transformers process input sequences in parallel rather than sequentially, which enables them to model dependencies across entire texts simultaneously. For a comprehensive introduction, see Vaswani et al. (2017). For a more recent review, we suggest Roger et al. (2020). [↑](#footnote-ref-7)
8. There are, of course, many other differences between these two models. For a detailed discussion of BERT, we refer the reader to Devlin et al. (2019). For a detailed discussion of GPT, see Radford et al. (2019) [↑](#footnote-ref-8)
9. Top\_p takes values between 0 and 1. Reducing top\_p reduces the number of tokens within the probability threshold, i.e., the number of tokens that are taken into consideration as candidates for the next output. Temperature takes values between 0 and 2. Increasing temperature increases the entropy or ‘randomness’ of the probability distribution across candidate tokens toward a uniform distribution, making diverse outputs more likely. [↑](#footnote-ref-9)
10. This adjustment required adding an additional post-processing step. This step is included in our notebook, which is publicly available (see Appendix). [↑](#footnote-ref-10)
11. A less optimistic interpretation is that print media may instead be blaming people and their choices. [↑](#footnote-ref-11)