Accountability for Gender Equality in Education

Working Paper 4

# Versatile and expansive or biased and overcomplicated? Reflections on using Large Language Models (LLMs) to review ideas about gender in curriculum and policy documents

Dr Ellen Weavers

Bridging AGEE Project Team, University College London

2025

## Acknowledgements

This Working Paper has been developed from an internal report commissioned by the team working on the AGEE project. Thanks to all of those who contributed to discussions of drafts of this paper – Gonzalo Guerrero Hernandez, Helen Longlands, Eliza Ngutuku, Tim Neuman, Rosie Peppin Vaughan, Caine Rolleston, Elaine Unterhalter, Aaron Benavot and Lee Crawfurd. Thanks for detailed peer review comments from Matthias Eck, Helen Longlands, Tamara Marti Casado, Rosie Peppin Vaughan, Justine Sass and Elaine Unterhalter.

## Follow Up work

This AGEE Working paper has been published as part of preparations for research to be conducted in 2025-2026 documenting the work of Bridging AGEE using AI with partners in Indonesia, Kenya and Malawi. This is part of the wider project supported by the Global Partnership for Education Knowledge and Innovation Exchange (GPE/KIX), a joint endeavour with the International Development Research Centre, Canada. Please see [www.gendereddata.org](http://www.gendereddata.org) for more information.

AGEE

# Contents

AGEE

# List of Figures

**Grey boxes which contain quotes generated by an LLM in response to zero-shot prompts:**

# Acronyms

| | |
|---|---|
| AGEE | Accountability for Gender Equality in Education |
| AI | Artificial Intelligence |
| GAT | Gender Approach Terminology (e.g., gender sensitive, gender transformative) |
| GPE | Global Partnership for Education |
| IDRC | International Development Research Centre |
| IOE | Institute of Education |
| KIX | Knowledge and Innovation Exchange |
| NER | Named entity recognition |
| NLP | Natural Language Processing |
| LLM | Large Language Model |
| SEAT | Sentence Encoder Association Test |
| UCL | University College London |
| UNESCO | United Nations Educational, Scientific and Cultural Organization |
| UNICEF | United Nations Children's Fund |
| WEAT | Word Embedding Association Test |

# Glossary

| | |
|---|---|
| **Bias Detection** | *"Classifying content as biased or non-biased"* within a 'bias dimension' using pre-established lexicons and/or rules (Raza et al., 2024, p.7) |
| **Black box models** | Models where only the inputs and outputs are visible and there is a lack of transparency around the internal 'decision-making processes' and how the outputs are generated (Chintam et al., 2023) |
| **Causal language modelling** | A process of predicting the next token in a sequence based only on preceding tokens, using learned vector representations to generate coherent text **(**Kumar, 2024) |
| **Chain-of-thought prompts** | A prompting technique that requires LLMs to break down complex problems into intermediate reasoning steps, making the model's thought process explicit and often improving performance on complex reasoning tasks (Wei et al., 2022) |
| **Contextual Analysis** | Analysing "*linguistic features and structures that go beyond individual sentences"* to identify context (Zhu et al., 2024, p.1) |
| **Dependency parsing** | The analysis of the grammatical structure of a sentence to identify how words are connected to each other. This enables the identification of which words within sentences depend on others to convey meaning (Crawfurd et al., 2024) |
| **Few-shot prompts** | A form of LLM prompt where examples of task inputs and outputs (called 'shots') are provided, which enables the LLM to learn what is expected 'within the context of the task. 'Few-shot' prompts are named in relation to the number of examples provided (e.g., 'zero-shot' prompts contain no examples; 'two-shot' prompts include two examples) (Lui et al., 2023). |
| **Hallucination** | A challenge with the use of LLMs where they *"produce outputs that, while seemingly coherent, might be misleading, fictitious, or not genuinely reflective of their training data or actual facts"* Quevedo et al., 2024, p.1 |
| **Large Language Models (LLMs)** | Models trained on vast volumes of open text (e.g., internet sources including online books, Wikipedia) which are capable of performing broader and more flexible linguistic analyses than traditional LLMs (Lui et al., 2023) |
| **Lemmatization** | A text pre-processing technique which involves assigning words to their 'lemma' (root word or dictionary form) so that words with similar meanings can be represented by a single word (e.g., better would be represented by good whereas if stemming were applied, it would be reduced to 'bet'. Different pre-processing techniques are required for different types of source documents and different types of analysis (Khyani et al., 2021). |

| | |
|---|---|
| **Named Entity Recognition (NER)** | "*The identification and classification of named entities, such as names of individuals, organizations, locations, dates, and more, within textual data*" (Pakhale, 2023, p.1) |
| **Natural Language Processing (NLP) models** | *"Natural language processing employs computational techniques for the purpose of learning, understanding, and producing human language content. Early computational approaches to language research focused on automating the analysis of the linguistic structure of language and developing basic technologies such as machine translation, speech recognition, and speech synthesis"* (Hirschberg and Manning, 2015, p.261) |
| **Relationship Extraction** | *"Extracting relationships between target nouns, called entities, which bring forth the semantic meaning of the text"* (Swarup et al., 2024, p.167655) |
| **Semantic Search** | *"Processing.... document chunks for retrieval based on semantic similarity with user queries"* (Ghali et al., 2024, p.3) |
| **Sentiment Analysis** | *"The process of recognizing emotions expressed in text. AI comprehends the tone of a statement, as opposed to merely recognizing whether particular words within a group of text have a negative or positive connotation"* (Taherdoost & Madanchian, 2023, p.1) |
| **Stemming** | A text pre-processing technique which involves reducing words to their root form (e.g., walking, walked, walks, and walker would be reduced to the stem 'walk'). Different pre-processing techniques are required for different types of source documents and different types of analysis (Khyani et al., 2021). |
| **Text Classification** | *"Categorizing and organizing text data to facilitate easier management.... extract meaningful patterns and categorize text data, thereby transforming unstructured text into structured, actionable insights"* (Taha et al., 2024, p.1) |
| **Text Coherence Analysis** | *"Assessing the coherence of a text in terms of how its constituent parts – whether at the sentence, paragraph, or document level – come together to form a unified whole that is logically consistent and aesthetically pleasing"* (Barbosa and Campelo, 2024, p.2) |
| **Text Summarization** | *"The process of condensing text into a concise summary, preserving essential information" (*Karanikolas et al., 2023, p.279) |
| **Tokenisation** | A text pre-processing technique which involves breaking text down into smaller units (e.g., words or sentences, depending on the nature of the analysis). Different pre-processing techniques are required for different types of source documents and different types of analysis (Chai, 2023). |

| | |
|---|---|
| **Topic Modelling** | "*Identifying themes by identifying repeated patterns of words and grouping these patterns of words into topics that reflect the content of the documents*" (Churchill and Singh, 2022, p.1) |
| **Transformer based models** | These models differ from earlier NLP approaches as they use self-attention mechanisms that enable parallel processing of input data. This allows them to more effectively capture complex relationships between words across long texts than earlier NLP models (Lui et al., 2023) |
| **Vector analysis** | The analysis of word (or sentence) vectors to detect patterns including relationships or similarities between words/word use or patterns in language (Kumar, 2024) |
| **Word embeddings** | In relation to NLP or LLM analysis, a way of representing words as vectors in a multi-dimensional space, where the distance and direction between vectors reflect the similarity and relationship among the corresponding words (Kumar, 2024) |

# 1. Introduction

## 1.1 Background and purpose of the working paper

This working paper was commissioned as a technical contribution to the collaborative research project **Accountability for Gender Equality in Education: Bridging the Local, National and Global** (**Bridging AGEE**), hosted by University College London (UCL) in partnership with UNESCO (United Nations Educational, Scientific and Cultural Organization) and the University of Malawi. The **Bridging AGEE project**, funded by the Global Partnership for Education Knowledge and Innovation Exchange (GPE-KIX) through Canada's International Development Research Centre (IDRC), is a three-year collaborative research project (2024-2027) that aims to improve how gender equality in and through education is documented and addressed in three countries – Indonesia, Kenya and Malawi – and to explore the broader application of the findings in other countries. The project aims to:

- Build knowledge concerning gender equality in and through education at local, national, and global levels looking at what data exists, what is missing and what is needed to support insights on contextually specific and relevant processes.
- Use new measurement tools and platforms (local, national, and global) to build knowledge on how to expand data systems and improve processes for data collection, uptake and critical reflection regarding gender equality in and through education.
- Inform research about scaling by highlighting how intersectionality such as the interplay between gender, poverty or ethnicity is approached in data at national and local levels and developing indicators that not only describe gender inequality but consider transformative processes that support gender equality.

Building on previous work during the development of the AGEE framework (Unterhalter et al., 2022), the **Bridging AGEE project** includes a strand of activity focusing on developing new data sources and tools relating to gender representation and gender equality within education systems. This working paper was initially developed to support the design of a pilot phase within this strand of work which will explore the potential for using Artificial Intelligence (AI) tools, such as Natural Language Processing (NLP) tools and Large Language Models (LLMs), to analyse gender representation in national curricula and education policy documents. The approaches adopted in the pilot phase will be decided through consultation with National Advisory Groups (NAGs) in each of the countries in which the project is working.

This working paper provides:

- A review of existing research to highlight the potential and limitations associated with using NLP and LLM techniques to analyse gender representation in national curricula and education policy documents.
- Potential frameworks for gender analysis of national curricula and education policy documents to provide evidence-based, structured approaches for applying NLP and LLM techniques.
- An overview of key technical considerations in designing research protocols using NLP and LLM techniques to examine gender representation in national curricula and education policy documents.

The primary intended audience for this working paper is the **Bridging AGEE** project team and their collaborating partners. The working paper is also intended to inform the development of practical guidance papers for discussions within the National Advisory Groups (NAGs), enabling informed decision-making regarding gender analysis of documents from their respective countries under the **Bridging AGEE** project. The document has been published as a working paper with the aim of supporting other research teams who are exploring or undertaking similar applications of NLP and LLM techniques to analyse complex and multifaceted themes like gender and gender equality.

## 1.2 Structure of the working paper

This working paper comprises four main sections:

- **Section 1: Introduction** provides background information on the intended purpose of the working paper, outlines its structure and contents, and clarifies how the terms 'curriculum' and 'policy' are understood and applied within the working paper.

- **Section 2: Previous work using NLP models and LLMs** starts with a summary of the evolution of NLP models and LLMs, as described in the literature, to contextualise changes in their functionality and application over time. Previous work carried out using NLP models or LLMs to analyse gender or other themes in curricula and policy documents is also reviewed alongside work to identify and address gender bias within AI. Key reflections from the literature are discussed.

- **Section 3: Towards analysis — Constructing frameworks and frames of reference** introduces draft frameworks for gender analysis of curriculum and policy documents that have been developed as part of the technical contribution to the **Bridging AGEE** project documented within this working paper. The draft frameworks have been developed to provide evidence-based structures to inform NLP or LLM analysis of these documents. This section also includes analysis of particular characteristics of national curricula and education policy documents that may require specific consideration in for text analysis.

- **Section 4: Technical considerations for curriculum or education policy analysis using NLP or LLMs tools** identifies key decisions involved in designing analytical protocols for gender analysis of these types of documents. This includes decisions related to measuring or mitigating bias within NLP/LLM tools, language-specific considerations, optimising AI and human roles within analytical protocols, and validation strategies. The technical requirements related to preparing documents for analysis, developing training datasets for NLP or LLM tools, or using direct prompting approaches are also considered.

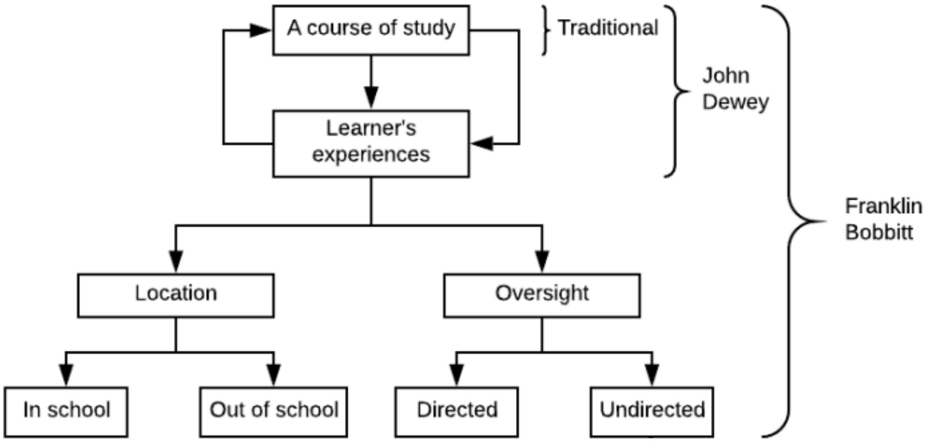
- **Section 5: Conclusion** summarises the potential opportunities and risks of using NLP and LLM tools to analyse gender representation in national curricula and education policy documents.

Throughout the report, grey boxes feature examples of quotes generated by an LLM in response to simple prompts. These responses demonstrate LLMs' ability to interpret and generate nuanced text while simultaneously offering alternative explanations of key LLM characteristics.

## 1.3 Framing 'curriculum' within the working paper

The term 'curriculum' has evolved to encompass multiple, interconnected meanings both in educational theory and practice. It is important, therefore, to establish clearly the meaning of 'curriculum' adopted in this report. The term is understood by some to encompass the "*entire range of experiences a child has within the school*" (Eisner, 1979, p. 28) extending even to experiences beyond the school gate (see *Figure 1).*

*Figure 1: Diagrammatic representation of different conceptualisations of experiential curriculum influenced by Dewey (1916) and Bobbitt (1918)*




*Adapted from Breault & Marshall (2010, p.180)*

These 'experiential' definitions encompass a range of other conceptualisations that are explored extensively in academic literature. These include the hidden curriculum – "*those unstated norms, values, and beliefs embedded in and transmitted to students through the underlying rules that structure the routines and social relationships in school and classroom*" (Kentli, 2009, p.86, summarising Giroux, 1983); the 'enacted curriculum' as delivered by teachers in the classroom; and the 'experienced curriculum' as interpreted by learners (e.g., Clemmons et al., 2022).

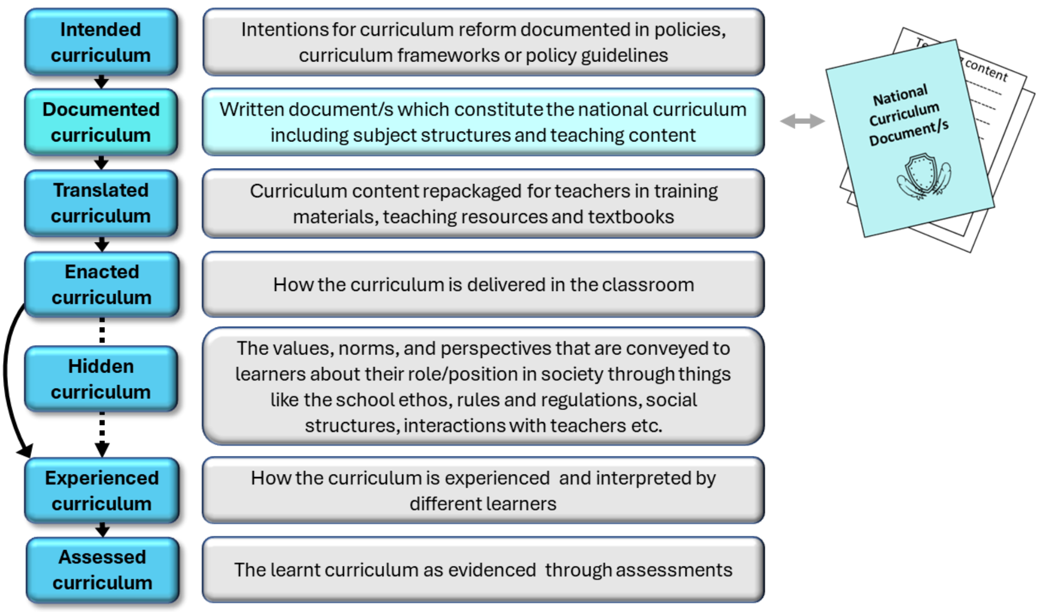
Alternative definitions conceptualise the process of curriculum construction and interpretation at different levels within the system (see *Figure 2*) as forms of social practice influenced by the interplay between curriculum constructs and social constructs. In these types of definitions, curriculum can be described as being *"socially constructed differentially across the layers of the education system as the curriculum is created in various institutional settings"* (UNESCO-IBE, 2025, p.14). UNESCO, for example, adopts a *"holistic definition of curriculum as the multi-layered social practices, including content selection, infrastructure, pedagogy, and assessment, through which education is structured, enacted, and evaluated"* (UNESCO-IBE, 2025, p.14, adapted from Priestley et al., 2021).

In the simplest form, however, the term describes a document setting out required teaching content or a *"plan for providing sets of learning opportunities for persons to be educated"* (Saylor et al., 1981, p. 8). These documents exist at various levels and scales, ranging from curricula documented at international or national levels to those developed by individual schools or teachers.

For the purposes of the intended gender analysis of curricula described within this report, the term 'curriculum' is used to refer specifically to **national curricula in the form of documents**

**setting out the specific content that all learners should be taught.** These national level 'documented curricula' (Clemmons et al., 2022) are developed in response to the 'intended curriculum' described in policy documents or national curriculum frameworks (see *Figure 2*) and so represent the government's expectations for teaching and learning the documents are superseded or archived.

*Figure 2: Representation of the relationship between 'documented' national curriculum documents and a range of levels of interpretation of the curriculum which are influenced to varying degrees by education policy*




The process of content selection and, equally, decisions regarding content exclusion during the development of a 'documented curriculum' means that these documents express what society values (Marope, 2025), or at least what the 'representatives of society' involved in curriculum development value. As such, the development of national level 'documented curricula' can be described as "*a dynamic and transformative articulation of collective expectations of the purpose, quality, and relevance of education and learning to holistic, inclusive, just, peaceful, and sustainable development, and to the well-being and fulfilment of current and future generations*" (Marope, 2025, p.15).

The 'documented curriculum' at the national level also represent what all learners are entitled to be taught and can be presented in various forms. Differences in the nature and structure of national curriculum documents from different jurisdictions introduce complications for comparative analysis using NLP or LLM tools (see **Section 3.2.1**).

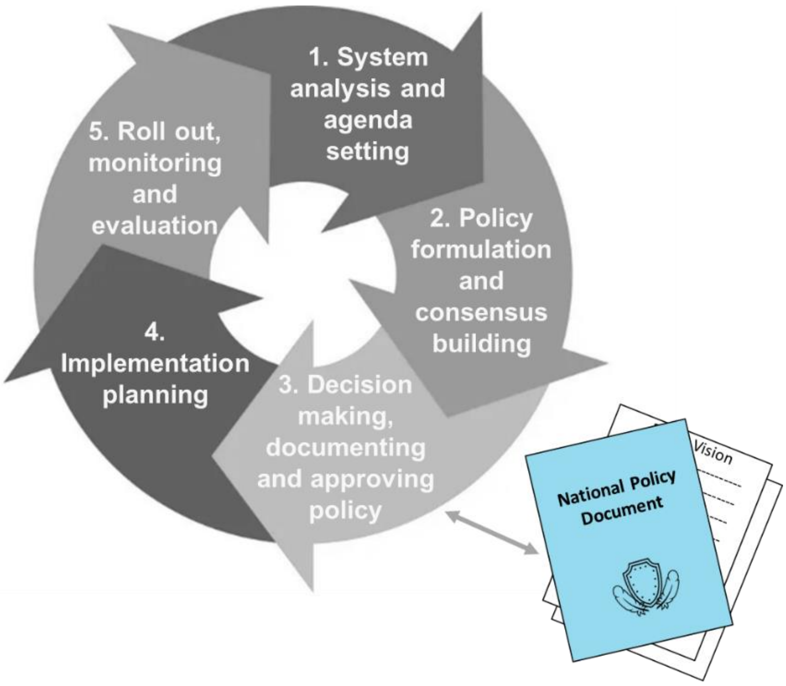
## 1.4 Framing 'policy' within the working paper

*"Policy is rather like the elephant — you recognise it when you see it but cannot easily define it"*
(Cunningham, 1963, p. 229)

The term 'policy' has been broadly applied to describe a range of positions, documents, and actions, including the decision-making process involved in governing, adopted or ratified positions, documented decisions, the documents themselves, or actions taken (or purposeful inaction) (Hill, 2021). According to some interpretations, a policy represents something fixed or documented at a specific point in time. For example, "*a policy is a broad statement that sets out the government's main goals and priorities*" (UNESCO, 2013, p.7). However, alternative conceptualisations consider policy as an ongoing, evolving process — one definition suggests that *"policy is essentially a stance which, once articulated, contributes to the context within which a succession of future decisions will be made"* (Friend, Power, and Yewlett, 1974, p.40), while another proposes that *"a policy may usefully be considered as a course of action or inaction rather than specific decisions or actions"* (Heclo, 1972, p.85).

For the purposes of the intended gender analysis of education policy described within this report, the term 'policy' refers specifically to **governmental policy documents that set out nationally approved policy positions.** These documents range from high level, whole-sector policies (e.g., an Education Policy) to more detailed sub-sector specific documents (e.g., a Teacher Training Policy) (UNESCO, 2013). These types of documents are developed and approved at particular points within an ongoing policy development and refinement process (see *Figure 3*). The policy statements and positions within a policy therefore represent the government's intentions at a specific point in time and remain 'current policy' until the policy document is superseded or archived.

*Figure 3: A representation of the ongoing policy development and refinement process*




*Adapted from Howlett and Giest, 2015*

The term 'policy analysis' also describes a range of functions within the policy development cycle and related academic literature. Schmitt (2012) identifies three broad categories of research — *"a first group of scholars measures the impact of policies by analysing the policies' effects on the target of regulation. A second group of researchers focuses on policy outcomes, namely the immediate consequences of a policy decision, while the third approach refers to policy outputs, namely the content of the decision"* (pp.30-31). The first two categories relate to the work of policy analysts in agenda setting, consultation, and consensus building within policy development (see *Figure 3*). The conceptualisation of policy analysis drawn upon in this report falls within the third category.

# 2. Previous work using NLP models and LLMs

## 2.1 The emergence of Large Language Models

Natural Language Processing (NLP) models are computational systems that analyse and generate language by learning patterns in linguistic structure and semantics through training on large volumes of sample texts. The architecture and functionality of NLP models have evolved significantly over time. The relatively recent development of transformer-based models, including Large Language Models (LLMs), has significantly expanded the analytical capabilities of these models. Transformer-based models differ from earlier NLP approaches in that they use self-attention mechanisms that enable parallel processing of input data. This allows them to capture complex relationships between words across long texts more effectively than earlier NLP models (Lui et al.,2023). LLMs, like ChatGPT or Google Gemini, are models trained on vast volumes of open text (e.g., internet sources including online books, Wikipedia) and are therefore capable of performing broader and more flexible linguistic analyses (Lui et al., 2023). Understanding the nature and timeline of this development pathway from traditional NLP models to LLMs provides useful context for understanding the range and evolution of approaches adopted using NLP models and LLMs to analyse gender concepts in text and documentation.

**Box 1: Example of LLM response to zero shot prompt — comparing NLP/LLM function**

**PROMPT:** Develop a witty quote to explain the difference between the functional abilities of traditional non-transformer-based NLP models and LLMs

**LLM response:** *"Where traditional NLP models are precision tools for specific tasks, LLMs are the Swiss Army knives of language—versatile, expansive, and sometimes overcomplicated, but capable of solving problems you didn't think a knife could handle"*

There is an extensive and rapidly growing body of literature focusing on the development, application, and evaluation of the performance of NLP models. The pace of publication increased as LLMs emerged and became more readily available — for example, the average number of papers published on the open-source arXiv platform that mentioned "large language model" within the title or abstract increased from 0.40 per day to 8.58 per day since the release of Chat GTP (Zhao et al., 2023). The pace of development of NLP models and LLMs has also accelerated over time. Lui et al. (2023) describe a series of paradigm shifts observed in the evolution of the learning models and functional architecture of these tools (see *Figure 4*).

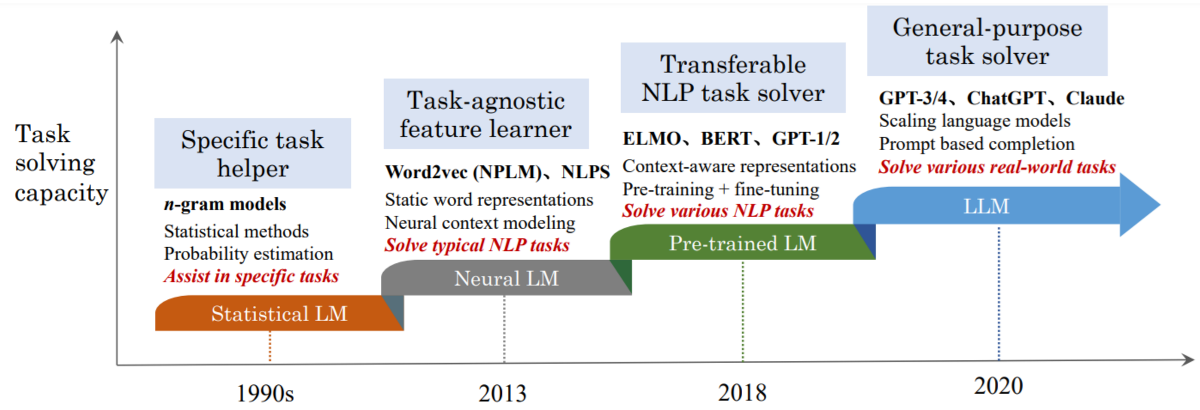
*Figure 4: Summarisation of four learning paradigms in experienced to date during the development of NLP models and LLMs*

| Paradigm | Approach to training/inputting the learning model |
|---|---|
| **~1990s – 2013** **1. Fully supervised learning using labelled data (non-neural network)** | **Feature engineering —** Individual NLP models were specifically designed for a particular task and were trained on a dataset of labelled/annotated input–output examples related to the target task. Researchers then used their expertise to provide structured instructions to enable the NLP model to 'learn' from the data. |
| **2013-2017** **2. Fully supervised learning using labelled data (neural network)** | **Architecture engineering —** As neural network models started to be incorporated within NLP models, the network architecture within the model allowed NLP models to learn during training on a task-specific dataset. Once trained, the NLP models would be able to complete a range of similar tasks, but different network architecture would be required for a significantly different task. |
| **2017-2019** **3. Pre-train, fine-tune** | **Objective engineering —** As the availability and ability to process text increased, NLP models were trained to be language models through pre-training tasks involving predicting the probability of the occurrence of words or phrases within extensive bodies of text. These 'general' language models could then be prepared for particular analytical tasks by the introduction of new objectives and parameters in a 'fine-tuning' training process. |
| **2019-ongoing** **4. Pre-train, prompt, predict** | **Prompt engineering —** As the scale of language models increased, LLMs emerged which can be assigned tasks through the use of prompts resembling those used within the initial training. LLMs can potentially predict the outcome without undergoing any task-specific training by drawing on the patterns learned during initial training on extensive bodies of text. |

*Source: Summarised from Lui et al. (2023)*

These paradigm shifts in the underlying function and scale of NLP models fundamentally changed the nature of the training protocols required and the flexibility with which the models can apply their training. This has resulted in rapid advancements in the 'task-solving capacity' of NLP models and LLMs (see *Figure 5*).

*Figure 5: An overview of the evolution process of the four generations of language models (LM) from the perspective of task-solving capacity set against a timeline of the emergence of related articles and publications*




*From Zhao et al., 2023, p.2*

This timeline is important for understanding the nature of the published literature available concerning gender analysis of and by NLP models and LLMs (see **Sections 2.2** and **2.3**). The rapid pace of development of the technology means that the time lag involved in publishing work has particular implications. Many papers documenting recent research are only available on open-access pre-publication platforms like arXiv. Traditionally, papers like these which had not gone through formal peer-review would not necessarily be included in reviews. However, the average lag between availability as pre-prints and being formally published is estimated to be 12 months for computer sciences (Wang et al., 2020). Given the pace of development with NLP models and LLMs, inclusion of pre-publication papers in related reviews is more common in this sector and has been adopted for this review. Including papers that have not gone through the full peer-review process introduces risks in terms of reporting unreliable or unverified findings. However, steps have been taken to mitigate these risks including labelling pre-publication papers clearly in the references section and incorporating the papers within the framing of a critical review of the methodology and data presented.

Like earlier NLP models, LLMs can conduct a range of analytical functions by deploying core language analysis processes that focus on translating text and patterns within text into numerical representations, which are then applied to analyse or generate text. Examples of these core language analysis processes include (summarised from Kumar, 2024):

- **Tokenization** — the process of breaking text into smaller, processable units (tokens) such as words or word components
- **Word embeddings** — a way of representing words as vectors in a multi-dimensional space, where the distance and direction between vectors reflect the similarity and relationship among the corresponding words
- **Vector analysis** — the analysis of word (or sentence) vectors to detect patterns including relationships or similarities between words/word use or patterns in language
- **Causal language modelling** — a process of predicting the next token in a sequence based only on preceding tokens, using learned vector representations to generate coherent text

*Figure 6* lists examples of core approaches (applicable with both traditional NLP models and LLMs) and examples of potential applications for gender analysis.

*Figure 6: Examples of analytical techniques that can be carried out by NLP models and LLMs (including summative references) and examples of how these could be applied in gender analysis*

| Analytical approach | Description | Example of potential application in gender analysis |
|---|---|---|
| **Sentiment Analysis** | *"The process of recognizing emotions expressed in text. AI comprehends the tone of a statement, as opposed to merely recognizing whether particular words within a group of text have a negative or positive connotation"* (Taherdoost & Madanchian, 2023, p.1) | Recognising and categorising the emotional tone of text (e.g., identifying statements about men or women as positive, negative, or neutral) |
| **Named Entity Recognition (NER)** | *"The identification and classification of named entities, such as names of individuals, organizations, locations, dates, and more, within textual data"* (Pakhale, 2023, p.1) | Measuring the frequency or prominence of male and female entities in a document, analysing roles and relationships |

| | | |
|---|---|---|
| **Text Classification** | "*Categorizing and organizing text data to facilitate easier management.... extract meaningful patterns and categorize text data, thereby transforming unstructured text into structured, actionable insights*" (Taha et al., 2024, p.1) | Classifying text to see the extent of focus on particular gender-related issues or particular gender perspectives |
| **Text Summarization** | *"The process of condensing text into a concise summary, preserving essential information"* (Karanikolas et al., 2023, p.279). | Generating summaries (e.g., policy briefs) of texts highlighting key points relating to gender issues or strategies |
| **Topic Modelling** | "*Identifying themes by identifying repeated patterns of words and grouping these patterns of words into topics that reflect the content of the documents*" (Churchill and Singh, 2022, p.1) | Identifying dominant themes related to gender and clustering relevant language into topics relating to gender |
| **Bias Detection** | "*Classifying content as biased or non-biased*" within a 'bias dimension' using pre-established lexicons and/or rules (Raza et al., 2024, p.7) | Detecting implicit or explicit gender biases in the language or identifying content that reinforces or counters stereotypes |
| **Semantic Search** | "*Processing.... document chunks for retrieval based on semantic similarity with user queries*" (Ghali et al., 2024, p.3) | Identifying sections of the document based on meaning (rather than keyword) in relation to particular gender-related concepts |
| **Relationship Extraction** | "*Extracting relationships between target nouns, called entities, which bring forth the semantic meaning of the text*" Swarup et al., 2024, p.167655) | Identifying relationships between gendered entities in text e.g., exploring how men and women are portrayed in relation to each other or to power, resources or status |
| **Contextual Analysis** | Analysing "*linguistic features and structures that go beyond individual sentences*" to identify context (Zhu et al 2024., p.1) | Exploring how gendered terms are used in different contexts or how gender roles are framed within policies |
| **Text Coherence Analysis** | *"Assessing the coherence of a text in terms of how its constituent parts – whether at the sentence, paragraph, or document level – come together to form a unified whole that is logically consistent and aesthetically pleasing"* Barbosa and Campelo 2024, p.2) | Evaluating if themes are coherently presented throughout a document – for example, if gender equality commitments are consistently reinforced throughout a document |

With traditional non-transformer-based NLP models, these techniques are applied within the boundaries of the word lists or rules directly provided. In contrast, transformer-based NLP models (including LLMs) are built around attention mechanisms. This means that while many foundational processes remain similar to those in earlier NLP models, transformer-based models have greater flexibility and generalizability to unseen tasks and contexts beyond the training materials (Karanikolas et al., 2023). Additionally, LLMs have significantly enhanced text-generative capacities compared to earlier NLP models (Karanikolas et al., 2023).

## 2.2 Previous work using AI/LLMs to analyse curriculum documents

Currently, there is very little academic literature available concerning the use of LLMs or NLP models to analyse gender concepts or representations within curriculum documents. There may be work of this nature underway that relates to school curricula. However, studies that have reached the publication stage tend to focus on higher education curricula and, in most cases, consider a single analytical approach based on NLP capabilities. Takizawa (2023), for example, describes the use of NLP topic mapping of the curriculum for Yale Medical School (United States). While the main purpose of the work was to enable mapping of the extent to which school-wide skills were integrated within different course curricula, the analysis also included a focus on how curriculum content related to gender identity had changed over time.

While there is limited NLP/LLM-based gender analysis literature available concerning school curricula, there is potential crossover in the applicability of techniques adopted within a growing body of work concerning gender analysis of school textbooks. Most of this literature that has reached publication reports on the use of a combination of a small number of NLP techniques. For example, Lucy et al. (2020) combined word embeddings, topic modelling and lexicon-based word association analysis to explore representations of gender, race, and ethnicity in history textbooks in use in Texas between 2015 and 2017. Crawfurd et al. (2024) reported on their use of word co-occurrences and word embeddings to quantify gender references and extract relational information from 1,255 school textbooks from 34 countries. The team also used dependency parsing which involves analysis of the grammatical structure to identify how words connect to each other. This enables identification of which words within sentences depend on others to convey meaning. There are also examples of NLP analysis concerning broader aspects of how gender bias can manifest in textbooks. For example, Skinner and Bromley (2019) carried out content analysis of textbooks concerning to human rights and social movements which included a focus on the extent to which women's rights were foregrounded in textbooks over time.

Combinations of simple NLP approaches have also been applied to the analysis of gender representation in children's literature (e.g., Lewis et al., 2022; Adukia et al., 2022; Adukia et al., 2023). For example, Adukia et al. (2023) analysed longitudinal patterns in representation in terms of gender, race, and age in children's books that had been given an award by the American Association for Library Services to Children since the awards began in 1922. They combined optical character recognition to categorise individuals represented in images with token counts of gendered terms and named entity recognition (NER). To assign gender to fictitious characters via NER, the probability of gender assignment was predicted using data from the Social Security Administration on the frequency of names by sex in America.

Alongside analysis of children's literature, there are also examples of the use of similar NLP methods to analyse gender representation within broader texts including online fiction (e.g., Fast et al., 2016), books (e.g., Hoyle et al., 2019; Boutyline et al., 2023), and books in combination with broader media (e.g., Bolukbasi et al., 2016; Garg et al., 2018).

Beyond analysis of gender representation or related themes, there are examples of NLP and LLM application to analyse different aspects of curricula. In this work, the focus tends to be on the mapping of skills within diverse higher education course curricula (e.g., Ketamo et al., 2019; Russell, 2024; Zamecnik et al., 2024; Murrugarra-Llerena et al., 2022). Due to the time lag involved in completing research and publishing the findings, most of these papers describe the application of NLP techniques as more advanced LLM tools were not yet readily available. Some more recent publications, however, are starting to report on the use of LLM techniques. For example, Chacon (2024) describes the use of ChatGPT to evaluate higher education course curricula in terms of cohesiveness, relevance, and alignment of curriculum content with workplace requirements. There are also examples of work that applies different NLP techniques including text mining (e.g., West, 2017) and thematic analysis (e.g., Macias et al., 2024) to explore the alignment of content with the sustainable development goals. In terms of school level curricula, there are examples of work looking at how teachers use AI in curriculum development (e.g., Karataş et al., 2024) and exploring the extent of alignment between the content of curricula and related textbooks (e.g., Yang et al., 2023). This latter work may provide insight into the potential for comparative analysis of the alignment of national curriculum documents with their related curriculum frameworks.

## 2.3 Previous work using AI/LLMs to analyse policy documents

Currently, there is very limited academic literature concerning the use of AI to identify or analyse gender themes within policy documents. There are examples, however, of use of NLP models to review the degree of association of particular gender terms with key terms identified within the related policy sector (e.g., Carroll et al., 2024). The NLP techniques applied by Carroll et al. (2024) provide a broad overview of potential bias within national policy documents from different countries within a particular sector (e.g., energy, environment). Other policy applications include exploring gender bias in national policies concerning AI (e.g., Venugopal and Rituraj, 2022).

These examples sit within a growing literature base involving the application of AI in policy analysis more broadly beyond considerations of gender. Much of this work relates to the use of AI to extract key information and summarise policy content to support the work of policy analysts (e.g., Safaei and Longo, 2024) in managing the early stages of policy development. As with the recommendations emerging from much of the work on textbooks and curricula, Safaei and Longo (2024) suggest that with the current capabilities of AI, the best results are currently achieved in the development of accurate and compelling policy briefs through a combination of AI and human input.

AI approaches have also been applied to support other aspects of the work of policy analysts who support the development of new policies. Examples include modelling the potential impact of different policy options concerning pandemic controls (Song et al., 2020) or modelling control groups to monitor the actual impact of carbon pricing strategies (e.g., Abrell et al., 2022).

In terms of document analysis of existing policies, there are examples of use of topic modelling and clustering to compare the nature of different national policies concerning particular sectors. For example, Papadopoulos and Charalabidis (2020) used this approach to identify three distinct approaches within national AI strategies, and Żółkowski et al. (2022) applied a similar approach to identify distinct policy frames applied by different European countries in their 10-year climate action plans.

Beyond analysis of policy documents, there are examples of literature concerning the use of NLP models to analyse gender representation in other complex and nuanced forms of text. For example, there are examples looking at gender bias in legal documents (e.g., Ash et al., 2024; Ratovicius et al., 2024; Sevim, 2023) and in the media (e.g., Pair et al., 2021; Singla and Mukhopadhyay, 2022).

## 2.4 Assessing gender bias within AI

One of the challenges in identifying academic literature focusing on AI use in gender analysis of documents is that search terms tend to return papers from the substantial (and growing) body of work concerning the identification of gender biases within AI models themselves (e.g., Bolukbasi et al. 2016; Caliskan et al., 2017; Garg et al., 2018) and their applications (e.g., Salinas et al., 2023; Pavón Pérez, 2022).

As established in a recent study, "*bias in AI can be introduced at any stage of its development, from design and modelling decisions, to data collection, processing, and the context of deployment*" (UNESCO, 2024, p.4). While gender bias is just one form of bias identified within AI outputs, an extensive body of work has emerged focusing on the identification and 'correction' of gender bias. This body of work has the potential to provide useful insights regarding the use of LLMs to analyse gender representation in curricula and policy, or other texts, in two key areas:

- Firstly, raising awareness of the limitations of AI and the potential complications for gender analysis from existing gender bias within LLM tools and models.
- Secondly, in the identification of potential analytical techniques, lexicons, or LLM training models, and prompting structures that may contribute to gender analysis approaches.

While there are a range of sources for the introduction of bias to pre-trained LLMs (Gray et al., 2024, Hall and Ellis, 2023), a significant limitation of LLMs is that from the outset they "*are marred by biases inherited from the data they are trained on*" (Ajith et al., 2023, p.888) – or more simply, "*machine learning is 'bias in, bias out'* " (UNESCO, 2019a, p.130). LLMs undergo extensive 'supervised' or 'unsupervised' training using large volumes of data related to particular fields or broader contexts. For example, JurisLMs and Lawyer LLaMA were trained using legal texts (Lai et al., 2024) whereas GPT-3 was trained on an extensive body of web pages, book catalogues and Wikipedia (Nazir and Wang, 2023). Extensive evidence shows that, in the process of training, LLMs encode social biases (including gender biases) present within their training materials and apply these in their analysis of text (e.g., Hutchinson and Mitchell, 2019; Salinas et al., 2023; Stanczak and Augenstein, 2021; UNESCO, 2024). Research also suggests that using AI to generate text responses further "*amplifies the risks of perpetuating biases in training data and algorithms, thereby reinforcing gender, racial, and cultural stereotypes*" (Wei et al, 2025., p.1).

An extensive body of literature is developing around different approaches to detect gender bias in particular LLM analytical functions, including:

- Evaluating bias in word embeddings by analysing the strength of conceptual associations between gender terms and other lexicons. For example, the Word Embedding Association Test (WEAT) is based on three lists of 'word categories' (careers vs. family activities; maths vs. arts; science vs. arts) (Caliskan et al., 2017). WEAT has been further developed by a variety of researchers. For example, Garg et al. (2018) added two further 'word category' lists (intelligence vs. appearance; physical or emotional strength vs. weakness) and Chaloner and Maldonado (2019) proposed a method to use cluster analysis within source documentation to develop targeted additional 'gender bias word categories'.
- Evaluating bias in sentence encoding models using the Sentence Encoder Association Test (SEAT) which was developed from WEAT (May et al., 2019).

- Testing bias in LLMs predictive capability (based on training materials) for example by:
    - Testing them on a bank of word masking statements to identify potential bias using probability scores. The LLM predicts the probability of which gendered term is behind the masking — for example, "*She is a medical records technician*" becomes "*[MASK] is a medical records technician*" (Sevim, 2023, p.29)
    - Using paired statements (a stereotype and an anti-stereotype) and scoring the percentage of times the LLM predicts a higher likelihood of each being true (with a score of 50 representing an unbiased response). Different approaches to sourcing statements for this approach have been used including crowdsourcing for CrowS-Pairs (Nangia et al., 2020) and sourcing content from Reddit conversations for RedditBias (Barikeri et al., 2021) which contains paired statements concerning gender.
    - Using prompts of incomplete sentences which include gendered terms (e.g., "*the boy is made to be...*" Nozza et al., 2021, p. 2398) and measuring the 'hurtfulness' of the proposed sentence completions. For example, the HONEST tool (Nozza et al., 2021) contains a range of prompts and draws on the classifications within the HurtLex lexicon of harmful language (Bassignana et al., 2018) to score the responses.

These examples represent a small sample of the different models and proposed adaptations described in the literature. The majority of the approaches to identify gender bias focus on elements of bias that are implicit in the text itself due to the direct use of gendered language. However, "*bias metrics within the...literature often do not correlate*" (e.g., Bartl and Leavy, 2024, p. 8; Aribandi et al, 2021). This is understandable, given that different approaches each focus on different specific elements of bias representation in text. Gender bias detection is also more challenging in languages with grammatical gender systems (Derner et al., 2024). This suggests that bespoke tools may be required for different models and languages, which emphasises the limitations of these models for identifying gender bias across the range of complex and multifaceted ways it can manifest in written texts.

## 2.5 Addressing gender bias in AI

A wide range of approaches have also been developed to 'debias' LLMs. For example, regarding text analytics using trainable LLMs, approaches have evolved that involve development of 'fine-tuning' materials for LLMs to use after training on a large volume of text or data. There is extensive academic literature about the technical elements of development and testing of different approaches that address particular elements of gender bias within texts — for example:

- Some approaches focus on 'fine-tuning' training for LLMs to reduce bias demonstrated within pronoun-noun associations — for example, the Gendered Ambiguous Pronouns (GAP) dataset was developed by Webster et al. (2018) to be used in coreference fine-tuning exercises. The dataset contains nearly 9,000 'ambiguous pronoun-name pairs'.
- Some approaches aim to reduce bias through the 'gender neutralisation' of gender-specific terms (e.g., fireman, housewife) in LLM outputs. For example, Thakur et al. (2023) created materials using established vocabulary lists to identify and replace gender-specific terms with gender-neutral versions (e.g., fire fighter). In their 'fine-tuning' materials, they also adapted phrases containing references to one gender (e.g., men or

women) to reference both. Bartl and Leavy (2024) developed this approach further by extracting examples from LLM training materials before applying 'gender-neutralising' adaptions. This enabled the creation of fine-tuning materials that are targeted using the content of documents concerning a particular sector rather than generalised lists.
- Other approaches adopt a more 'directly challenging' approach. Counterfactual Data Augmentation processes, for example, purposefully switch gendered nouns and pronouns to the opposite gender to create materials that challenge stereotypical constructs (e.g., Ghanbarzadeh et al., 2023; Vashishtha et al., 2023; Fatemi et al., 2023)

These examples represent a small sample of the approaches documented within the extensive literature available. One limitation of this current literature base, however, is that it predominantly focuses on bias mitigation in NLP/LLM analysis where the inputs and outputs are in English. When LLMs are applied to analyse gendered languages, however, gender bias identification and mitigation are more challenging (e.g., Derner et al., 2024) as underlying biases can be amplified or obscured by patterns in grammatical gender (see **Section 4.1.2**).

Much of this work also relates to specific analytical techniques where LLMs are applied to analyse specific characteristics of different texts. Therefore, much of the work on addressing gender bias within these analytical models focuses on addressing specific (and relatively simplistic) aspects of how gender bias can be represented. There are also, however, a growing number of approaches that aim to address more complex representations of gender bias in their 'fine-tuning' of LLMs. Zhang et al. (2024), for example, built on previous models to develop the GenderAlign dataset that includes 8000 conversation prompts. These prompts are each paired with two potential responses — a 'chosen' response demonstrating lower levels of gender bias and a 'rejected' response. Each prompt was developed to provide challenge in relation to four categories of bias — "*stereotypes, discriminatory language, sexism in occupational and educational institutions, and bias against marginalized genders*" (Zhang et al., 2024, p.2).

In the literature related to 'fine-tuning' models for NLP/LLM-based textual analysis, the majority of the approaches described tend to demonstrate successful bias reduction in the outputs of particular LLMs. For example, Bartl and Leavy (2024) applied the same protocol with three different LLMs and recorded a reduction in 'gender stereotyping tendencies' in all three models. However, the extent of the reduction varied significantly between the different models, which is consistent with broader findings in the literature. While different fine-tuning strategies have been demonstrated to have impact (e.g., in terms of a reduction in the generation of gendered words, stereotypes, and/or harmful language) different approaches to fine-tuning work better for different LLMs based on their underlying architecture (Bartl and Leavy, 2024).

While the majority of examples refer to approaches for adjusting the training or fine-tuning of NLP models, there is also a growing body of work on addressing gender bias through prompt engineering (e.g., Dwivedi et al., 2023; Sant et al., 2024). Other 'post-process' approaches in the literature explore the potential to 'correct' the outputs of NLP models without addressing the underlying bias in the models (e.g. Jia et al., 2020; Persico et al., 2024).

In summary, there is an extensive and growing body of work concerning the identification and correction of gender bias in different NLP models. However, the remaining challenges are well summarised by Chintam et al. (2023) — "*despite this work, we still lack a proper understanding of how to best measure biases...how biases are implemented in the language model internals...and...what techniques are effective at reducing undesirable downstream behaviour* —

*is data curation more or less effective than filtering output? Is intervening in the model internals feasible?"* (p.1).

Many of these challenges relate to the difficulty of interpreting how NLP models and LLMs process information and arrive at specific outputs. The lack of transparency around the internal 'decision-making processes' means that these tools function essentially as 'black box' models where only the inputs and outputs are accessible. In particular, *"the 'black box' nature of LMs makes it difficult to identify and interpret how bias manifests and propagates in them, especially relying solely on correlational methods"* (Chintam et al., 2023, p. 1). This raises potential challenges for the application of LLMs to conduct analysis of gender representation within text. Currently, different data-based, model-based, or output-based approaches to address gender bias within NLP models and LLMs are validated in relation to generic text analysis processes. However, with the existing ambiguity around how bias manifests within these tools, different approaches to mitigating gender bias may have varying impacts on the accuracy of later analysis of gender themes in texts.

## 2.6 Reflections on previous work

### 2.6.1 Considering the limitations of NLP/LLM analysis

One of the challenges associated with the use of LLMs to analyse gender representation is that fundamental language processing approaches are designed to convert patterns in text into numerical outputs. Consequently, while it is common to find papers that have combined a small number of LLM techniques, these tend to provide evidence concerning a slightly reductive framing of particular elements of how gender bias can be identified in text. The majority of the literature to date focusing on textbooks, for example, has focused on identifying the number of gendered terms in texts and on analysing associations between these terms and language associated with common areas of stereotype (e.g., personal characteristics, home and job roles, achievement, power).

With different projects applying different techniques, there is a risk that this may lead to a fragmented body of evidence. In addition, the reductive nature of traditional NLP text analysis may enable the identification of broad trends but may not contribute effectively towards providing practical, actionable outcomes for policymakers. In contrast, while there are many examples of manual gender analysis of textbooks that focus on one particular element of gender representation, there are also well-established broader frameworks that cover a range of forms of gender representation (e.g., as described in Leach, 2003; UNESCO, 2009).

There are also significant potential limitations to the use of traditional NLP approaches within individual projects. While they provide a repeatable result that can be used to identify broad trends, there are limitations to the effectiveness of context or tone analysis in the most commonly used traditional NLP approaches in the literature. Some simpler methods involving word embeddings that are commonly used in textbook analysis have no contextual component. This means that the relationships implied by particular words being closely associated within documents can misrepresent the actual framing of a document. For example, a word embeddings analysis could be carried out to quantify the association of words relating to gender and words relating to 'home, family and family roles'. If a simple co-occurrence analysis (without any contextual component) was applied to two documents setting out opposite arguments —

for example, one text explaining the value of traditional gender roles, and another text discussing the damaging nature of the same gender role stereotypes — female-associated gender words would be likely to be strongly associated with the words in ‘home, family and family role’ lexicons in both documents.

When these analysis approaches are used in isolation, care needs to be taken in interpreting the results. A specific example is provided in *Figure 7,* where a single word change (of a word not found in either list) in the document changes the entire context of the associations between words in the lists. This example demonstrates how the changing of the word ‘still’ to ‘not’ could fundamentally change the meaning of the text without changing the ‘distance’ between the target words. These two paragraphs would generate an identical or ‘nearly identical’ score in a traditional NLP word embeddings co-occurrence approach based on vector analysis. The example also demonstrates how words from the list (e.g., man and threshold) can be used to mean something different from the ‘expected meaning’, giving a ‘false result’ in terms of the purpose of the analysis.

*Figure 7: Example of a pair of paragraphs that would score identically/nearly identically in simple word embedding analysis (relying on analysis of the vector distance between words on two lists – cosine similarity analysis) but which have entirely different contextual framings due to a single word difference that is not covered in either list (highlighted in bold underlined)*

**List 1:** **["He", "She", "Him", "Her", "His", "Hers", "Sister", "Brother", "Wife", "Husband", "Male", "Female", "Man", "Woman", "Men", "Women", "Boy", "Girl", "Boys", "Girls"]**

**List 2:** **["Home", "family", "household", "threshold", "boundary", "kitchen", "living room", "bedroom", "dining room", "chores", "parenting", "caregiving", "cooking", "cleaning", "laundry", "childcare", "homemaker", "breadwinner", "spouse", "partner", "sibling", "parent", "mother", "father", "child", "son", "daughter", "housework", "nurturing", "budgeting", "shopping", "organizing", "decorating", "maintenance", "gardening", "pets", "babysitting", "hosting", "tenants", "roommates", "privacy", "boundaries", "harmony", "support", "teamwork", "routines", "traditions", "safety", "shelter", "responsibility", "comfort", "warmth", "belonging"]**

**Paragraph 1:** In the modern world it is important that **girls** are **still** encouraged to focus on **family** responsibilities and managing the **home**. Societies must embrace a future of innovation and adaptability without losing their cultural roots. **Man** stands on the **threshold** of a new dawn yet again.

**Paragraph 2:** In the modern world it is important that **girls** are **not** encouraged to focus on **family** responsibilities and managing the **home**. Societies must embrace a future of innovation and adaptability without losing their cultural roots. **Man** stands on the **threshold** of a new dawn yet again.

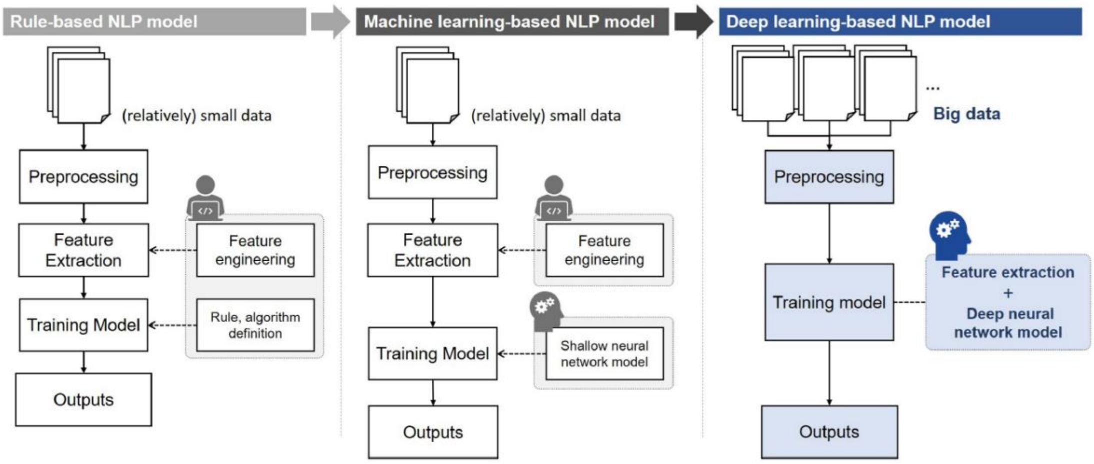
This is an extremely simplified example, and there are NLP techniques of increasing complexity that consider semantic similarity and contextual framings. However, with these traditional text analysis approaches, it is important to recognise exactly what is being measured and what can be inferred from the data — and acknowledge that the quality of traditional NLP analysis relies on the quality of the word lists, framings, and rules provided. Therefore, these materials need to be very carefully designed and tested.

The emergence of transformer-based NLP models like LLMs, however, has fundamentally changed the way in which contextual analysis is applied. These LLMs are trained on vast open-source data sets, which means that they have the ability to apply contextual understanding that has been developed through extensive analysis of ‘everyday’ information rather than a purposefully designed training data set (see *Figure 8*).

Consequently, while LLMs can potentially carry out the same traditional forms of document analysis using the same techniques as traditional NLP models, their capabilities extend beyond this. LLMs can approximate outcomes or carry out contextual analysis by constructing models of context based on patterns identified during their initial training on highly varied content.

**Box 2: <u>Example of LLM response to zero shot prompt – text and contextual analysis</u>**
**PROMPT:** Develop a witty quote about the difference between using traditional text analysis with NLP models and using the emerging LLM contextual analysis capabilities
**ChatGPT response:** *"Working with non-transformer NLP models is like training a dog to fetch only specific sticks—efficient but limited. Working with LLMs is like chatting with a philosopher who not only fetches the stick but also tells you why it's symbolic of human ambition"*

*Figure 8: Comparison between classical NLP models and deep learning-based NLP models*




*From Song et al., 2020, p.569*

To consider the opposite extreme to the data that are generated from technically specific text analysis protocols, LLMs have the predictive text capacity to respond to simple question-answer prompts (see **Section 4.3**). A dialogue-based LLM for example, can provide a response to questions like 'is there any evidence of gender stereotypes in the following text'. A rapid response will be generated which may at first glance seem remarkable. However, one of the key challenges with LLMs is that they operate via a black box model, which makes it difficult to understand how conclusions were reached. While the response may seem detailed and relevant, there is a chance that it may not be fully reflective of the nature of the text itself. LLMs function by applying patterns learned during training, so responses may include generalisations and factually incorrect 'hallucinations' created through predictive text generation rather than actually referring to sources in real time (Kaddour et al., 2023).

**<u>Box 3: Example of LLM response to zero shot prompt – NLP models/LLMs as black box models</u>**
**PROMPT:** Develop a witty quote about the challenges of working with LLMs due to the nature of their black box analytical approach
**LLM response:** *"Consulting an LLM is like asking a parrot for legal advice: it'll repeat something that sounds intelligent, but you're not sure if it will stand up in court"*

The development of detailed prompts (see **Section 4.3**) can assist in reducing the occurrence of generalisation and hallucinations — however, it not possible to fully overcome this challenge due to the functional structure of the LLMs available at this point in time.

There are therefore a number of potential trade-offs to be considered when selecting between the full range of analytical approaches that are possible with the use of LLMs. For example:

- There is a significant trade-off between the technical precision and document-focused nature of the traditional text-analysis approaches and the extent of work required to develop the necessary training materials and procedures.
- Conversely, there is a potential trade-off between the depth of the analysis and its accuracy (particularly in relation to a complex concept like gender). While Ratovicius et al. (2024) suggest that analysis approaches that focus on lexical aspects (e.g. the placing of individual words and their frequency), *"lose crucial information related to the order and sequence of words in the text"* they argue that NLP models suffer from *"the 'Curse of dimensionality', meaning they become less efficient as the dimensionality of the data increases"* (p.3). Beyond the limitations of traditional NLP approaches, LLMs' advanced capabilities in conceptual analysis enable them to identify both direct and implied references within text. However this brings with it a risk of generalisations or hallucinations being introduced which increase the validation workload.
- Another potential trade-off exists in terms of the accessibility and applicability of the outputs from different techniques. While statistical outputs enable direct comparison of documents, they may not necessarily be easy to interpret at the policy level or translate into practical solutions.

In reality, the range of analytical approaches possible form a continuum between the most technical, reductive statistical analysis of text in a document and the more interpretive question-answer responses to unqualified prompts. Given the current capabilities of LLMs at this point in time, it is likely that the best solution for analysis of complex themes like gender representation will involve a combination of different techniques (see **Section 3.2.2**) and a combination of AI and manual analysis (see **Section 4.1.3**). In addition, points made by Steiner (2025) and Dohn (2025) highlight how using AI in conjunction with, and with due acknowledgment of, well-articulated theory such as that associated with the AGEE project can contribute to appreciating what is and is not evident when using these techniques.

### 2.6.2 Validation

In most studies, some form of validation was carried out by comparing manual analysis or coding directly with the outputs of the NLP/LLM. In most of the work concerning textbooks or other books, this involved manual analysis of whole texts rather than a structured sampling approach at assigned points within the full analysis run. For example, Adukia et al. (2023) hand-coded a sample of 30 short stories and poems out of a total sample of over 1,100 books and Crawfurd et al. (2024) analysed a sample of 5 books from different countries and different subjects from a total sample of 1,255 books to validate one process.

Validation and manual checking processes also tend to be carried out in the initial preparation stages, particularly in relation to the generation and categorisations of lexicons or data sets. For example, Lucy et al. (2020) describe various stages of manual checking of both the initial tools (e.g., lexicon construction and categorisation) and analytical outputs (e.g., named entity extractions). Through this process, they identified that NLP processes like topic mapping had the potential to outperform manual coding when dealing with identifying potential 'relational meanings' in large amounts of text. However, they argue that during the analysis process "*subject area expertise remains of central importance, as researchers must thoughtfully attribute meaning to the automatic groupings*" (Lucy et al., 2020, p.7).

The recommendation that the best results are achieved through an intentional combination of NLP/LLM analysis and manual inputs is common in wide-ranging LLM literature, linked to suggestions to rename AI as 'Augmented Intelligence' (e.g., Zheng et al., 2017).

> *With minimal guidance, ChatGPT can substantially enhance the efficiency of domain experts in accomplishing their objectives. The findings suggest a nuanced integration of artificial intelligence (AI) with human expertise" and "signals a future...where the robust capabilities of AI technologies integrate with human creativity and innovation"* (Qiu and Jin, 2024, p.1)

However, in much of the education-related literature, steps involving 'manual input' tend to be described in terms of validation. It is advisable therefore to be clear about the difference between manual steps being included in the analysis protocol in any research design and the processes required to validate the analysis approach. While the latter may involve manual checks, the validation protocols would also need to encompass the outcomes of analytical processes that happen to include manual steps.

Another consideration concerning developing the validation protocol is ensuring that reporting templates or protocols involve consistent documentation of specific references to the documents being analysed (e.g., extracting direct quotes or recording page numbers) alongside analytical outputs (e.g., as recommended in Ratovicius et al., 2024). This would potentially reduce the time required for validation, as checks could be carried out on whether the extracts represent accurate identification of relevant content rather than by carrying out a full document analysis manually. It would also support a process of ongoing validation monitoring, which would enable corrections in real time within the analysis if required (see **Section 4.1.4**).

In some papers, the outcomes of validation are reported directly. In these cases, the validation process may identify questions concerning the accuracy of the NLP/LLM outputs. For example, when Crawfurd et al. (2024) manually coded named entities within five texts from their full set, the findings suggested that only 65% of the words identified as female names by the NLP model were actually female names, compared with 78% for male names. In other papers, however, the reporting of validation measures and outcomes is more ambiguous, which raises questions for about validity of the findings. This emphasises the importance of maintaining transparency and setting pragmatic tolerance targets for validation processes from the outset.

### 2.6.3 Simplistic models of gender

One of the common criticisms of the body of work concerning the use of NLP models in gender analysis is the predominance of binary representation of gender. There are examples of lexicons and training data sets that incorporate more diverse representations of gender and sexuality. For example, in some projects gender noun lexicons have been extended to include non-binary and transgender identities (Bartl and Leavy, 2024). In terms of data sets, RedditBias was developed to incorporate paired statements concerning gender and queerness (Barikeri et al., 2021) and the 'HONEST benchmark' prompt-based data set originally only included binary gender terms but was later extended to include prompts related to LGBTQ+ terminology (Nozza et al., 2022). However, the majority of work in this area uses analytical frames based on a restricted representation of gender. This is an issue that is reflected more broadly in AI work (Collett and Dillon, 2019) and is an ethical issue that needs to be considered in project design.

In later sections of this report, initial drafts of analytical frameworks for curriculum and policy are presented, and the trialling of potential prompting approaches is discussed. Due to the need to simplify the datasets for initial modelling, the trialling of prompts draw from lexicons based on binary representation. However, the draft Gender Framework for Curriculum Text Analysis proposed for further review in **Section 3.2.1** includes a focus on references to diverse representations of gender and sexuality within curriculum documents.

### 2.6.4 Ethical considerations

There are a number of ethical considerations noted in the literature in relation to the use of AI. Alongside issues raised by the predominance of binary representations of gender, the well-documented existence of gender bias within LLMs raises further concerns in relation to perpetuating or reinforcing societal biases. In addition, ethical concerns around fairness, accountability, and transparency when dealing with the 'black box' nature of LLMs have particular implications when LLMs are being applied to compare documents from different nations. The well-documented risks of LLM generalisation and hallucination during analysis would need to be robustly addressed in the design of the analysis protocol. In contrast, the use of more reductive representations of gender bias in more traditional NLP techniques also risks mis-representing gender dynamics or contextual framings within curriculum and policy documents. Transparency is of particular importance in work intended to influence policymaking, particularly given concerns raised in the literature about policymakers' understanding of the nature of AI and its limitations.

Broader concerns about data privacy and consent are not directly applicable to documents that are already in the public domain, but this may need navigating if restricted-access documents were to be included in the dataset for analysis. Other broader concerns raised are applicable to all research using LLMs, and these include:

- Challenges with authorship and attribution, both in terms of how LLMs apply pre-learned patterns without attribution to sources, and in the acknowledgement of LLM use in research.
- The environmental impacts of using LLMs, as the extensive computational resources required to run LLMs have significant environmental implications in relation to energy consumption, resource extraction, waste generation, and water usage.
- The uncertainty about the future capabilities of AI and the potential socio-economic impact of its evolution raises ethical concerns around contributing to the use and development of technology without clear understanding of potential impacts of doing so.

# 3. Towards analysis — constructing frameworks and frames of reference

## 3.1 Gender and curriculum frames of reference

### 3.1.1 Gender analysis and curriculum

*"Curriculum resides in relationship with the concept of gender in complex and multifaceted ways" (Elwood, 2016, p.247)*

Theoretical perspectives on the delicate interplay between gender and various forms of curriculum have been explored in depth (e.g., Paechter, 2000; Hernández et al., 2013; Elwood, 2016). Building on this work is a growing body of literature that considers gender from a more practical perspective, specifically within 'documented' national curricula (e.g., Aikman et al., 2005) or the 'translated curriculum' within textbooks and teaching materials (e.g., UNESCO, 2009). This body of work includes analysis of gender representation through text or document analysis — for example by analysing:

- The use of gendered terminology
- The nature of the relationships between males and females referenced in the content
- The inclusion of direct and implied references to gender stereotypes

The outcomes of this type of work tend to align with practical strategies that can be applied during curriculum development, including:

- Avoiding gendered language
- Including set texts from equal numbers of male and female authors
- Carrying out gender audits during curriculum development
- Including teaching content/curriculum topics relating to gender equality

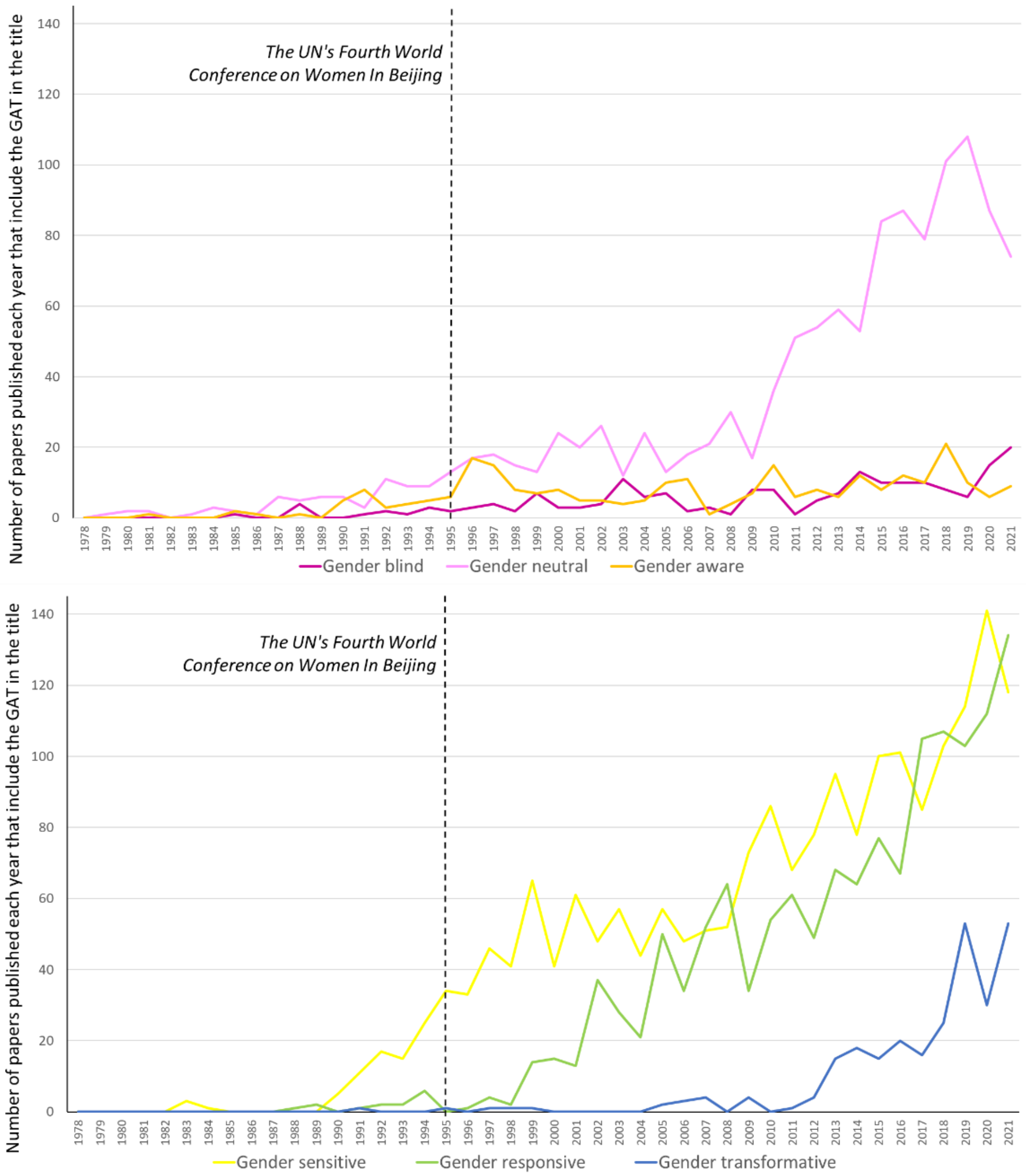
While analysis and auditing of curriculum documents have traditionally been carried out manually, the nature of this work lends itself well to the use of LLMs in analysing curriculum documents.

### 3.1.2 Gender Approach Terminology and curriculum

Over time, a range of 'Gender Approach Terminology' (GAT), including gender-blind, gender-neutral, gender-aware, gender-accommodating, gender-sensitive, gender-responsive, gender-redistributive or gender-transformative, has been used in various sources to describe approaches for addressing gender inequality within planning or policy development processes. For example, early references to the term gender-blind appear in literature concerning constitutional politics (e.g., Levin, 1975), education (e.g., Howard-Merriam, 1979), and law (e.g., Bassis, 1979).

In contrast, other terms emerged in association with particular movements concerning urban and rural planning within the development sector — these include 'gender-aware' (e.g., Mugure, 1983; Lindsay, 1985; Moser, 1985) and 'gender-responsive' (e.g., Awash, 1988; Manyazewal, 1988; Bishaw, 1988). The term 'gender-transformative', which emerged later, initially appeared in relation to development planning or discussions of Freirean (informed by Freire, 1972) and feminist models of education (e.g., Goetz, 1992; Clover, 1995; Mayberry, 1998). However, over the past 12 years, there has been a relatively rapid increase in use of the term across a wide range of development sectors (see *Figure 9*) and the term gender-transformative has increasingly been applied in relation to education programming.

*Figure 9: The number of papers which include particular GAT in their title published each year since 1978 (Data Source: Google Scholar)*




*From Weavers, 2022, p.10*

The adoption of particular GAT in international conferences has also had a significant impact on their prevalence in the literature over time. For example, the term gender-sensitive initially emerged in literature concerning development architecture (e.g., Martin, 1970) and economic modelling (e.g., Varon, 1975) but became an increasingly common term across wide-ranging development sectors as a result of its adoption as a key phrase in the run-up to the UN's 4th Conference on Women, in Beijing in 1995 (Subrahmanian, 2004) (see *Figure 9*). In the resulting

'Beijing Platform for Action', participating governments agreed to "*design, implement and monitor... gender-sensitive policies and programmes*" (UN, 1995, Item 19, p.3) and "*create a gender-sensitive educational system*" (UN, 1995, p.28). As a result, the use of the term 'gender-sensitive' in relation to approaches to education policy and planning grew extensively from this point in time.

While GAT have been in use for several decades now, the definitions of the terms found within the literature from different sectors (e.g., health, education, and urban planning) are varied and, at times, conflicting (Weavers, 2022; Foulds, 2013; Foulds, 2014). Within education, extensive work has been carried out by a range of international development organisations to provide definitions (e.g., UNESCO 2022) and practical guidance on the application of particular terms to education policy and planning (e.g., UNICEF, 2024; UNESCO, 2020; UNESCO 2021; GPE et al., 2017).

Over the last three decades, GAT like 'gender-sensitive', 'gender-responsive' and 'gender-transformative' have increasingly been used to label the intended approach to addressing gender in national curricula. While a number of guidance documents have been developed that set out particular curriculum approaches related to specific GAT (e.g., UNESCO, 2002; Plan International 2016), challenges still remain in establishing a shared understanding of how particular GAT relate to different practical strategies. It is the results of these practical strategies which may be detectable within the text of national curriculum documents (e.g., the avoidance of gendered terms compared with the balancing of gendered terms used) (Weavers, 2022).

While there remains significant variation in the way different GAT are interpreted practically, there is broad acceptance within the literature that there is some form of intuitive progression within GAT. To represent this, a range of GAT 'continuum diagrams' have been developed over time. These were initially used to "*offer categorizations of policy which are used in gender and development theory ...and... alert the user to the level of transformation intended by specific policies*" (World Health Organization, 2003, p.29). These continuum diagrams set out an implied progression between different GAT from 'negative or damaging approaches' through positive or 'intended' approaches, towards more transformative action (see *Figure 10* and **Appendix 1**).

Most of the early forms of continuums originated as frameworks for gender analysis of policies, including the Social Relations Approach (Kabeer, 1994a, 1994b; Kabeer and Subrahmanian, 1996), the 'Gender Responsive Assessment Scale' (GRAS) (World Health Organization, 2011), or the 'Gender Equality Continuum' (IGWG, 2017). More recently, however, GAT continuums have been used to inform sectoral planning and to set organisational strategies within the education sector (e.g., Oxfam, 2017; UNESCO, 2019b; UNICEF, 2024).

While a range of alternative continuums now exist, they tend to share significant similarities (see *Figure 10* or **Appendix 1**) as the majority are the result of adaptation of the original models by a range of authors for different purposes. As a result, each of the continuum models that have emerged have applied or defined GAT in subtly different ways. Today, there remains no 'unifying' continuum. Even where particular versions have been adopted by a range of organisations (e.g., the UNICEF 2019 continuum), it is evident that while "*many organizations use a variant of this continuum, ...they use its terminology in various and different ways*" (Marcusi et al., 2021, p. 12).

*Figure 10: Positioning of GAT on 9 gender approach continuums (see Appendix 1 for the full continuums). Grey terms indicate those terms no longer used frequently in the literature.*

| | Negative | Intermediary | | | Positive/Desired |
|---|---|---|---|---|---|
| 1 | Gender blind | | | | Gender aware |
| | Gender exploitatative | Gender accommodating | | | Gender transformative |
| 2 | Gender blind | Gender neutral | Gender sensitive/aware | | Gender transformative/redistributive |
| 3 | Gender negative | Gender blind | Gender targeted | Gender responsive | Gender transformative |
| 4 | Gender unequal | Gender blind | Gender sensitive | Gender specific | Gender transformative |
| 5 | Gender unaware | Gender aware | Gender sensitive | | Gender transformative |
| 6 | Gender blind | Gender neutral | Gender sensitive | | Gender responsive/transformative |
| 7 | Gender discriminatory | Gender blind | Gender sensitive | Gender responsive | Gender transformative |
| 8 | Gender unequal | Gender blind | Gender aware | Gender responsive | Gender transformative |
| 9 | Gender blind/neutral | Gender sensitive | Gender responsive | | Gender transformative |

*From: 1. IGWG, 2017; 2. Miruka et al., 2013; 3. UNDP, 2015; 4. Dwyer and Woolf, 2018; 5. Brown et al., 2019; 6. Subramanian, 2019; 7. UNICEF, 2019; 8. UNFPA and UNICEF, 2020; 9. UNGEI, 2024*

*From Weavers, 2022, p.36*

Over time, the language used in relation to education programming has evolved in line with movement along the continuums — where previously there was a strong focus on developing gender-sensitive curricula, many of the international organisations within the development sector (including for example, Plan International, Oxfam, Save the Children, UNICEF, UNGEI and UNESCO) now use 'gender-transformative' when setting out their strategy for curriculum reform and education programming more broadly (e.g., UNESCO, 2019b; UNICEF, 2024).

Despite the subtle differences between different GAT continuums, there are fundamental similarities in that each contains similar structures and progression models (see *Figure 11*).

*Figure 11: Implied progression within all GAT continuums (see Figure 10)*

| **Negative** | **Intermediary** | **Positive/Desirable** |
|---|---|---|
| A GAT that is considered to represent a negative approach which would fail to address gender inequality | A number of GAT that describe intentional approaches that are considered to form a pathway towards 'the best' approach | One (or two) GAT that represent a positive (or 'the best') approach to improve gender equality |

***IMPLIED 'DESIRED' PROGRESSION***

Despite the prevalence of GAT continuums, the notion that there is a need to progress through the intermediary gender approaches before being able to achieve the more positive/desirable gender approaches is not fully accepted in the broader literature. However, GAT continuums provide a structure against which the approaches adopted in different contexts can be compared. This in turn enables the identification of potential next steps to enable progression towards the approaches framed as more positive or desirable. Considered in this way, a GAT continuum can provide a useful overarching structure for potential analytical tools like the draft Gender Framework for Curriculum Text Analysis proposed in **Section 3.2.1**.

## 3.2 Gender and curriculum analysis

### 3.2.1 Constructing a Gender Framework for Curriculum Text Analysis

Construction of a comparative framework to potentially inform gender analysis of curriculum documents using NLP or LLM techniques requires two key components:

- An evidence-based way to identify the full range of different approaches or characteristics of a curriculum document that could be designed by intention (or unknowingly implemented) by curriculum developers in relation to gender representation within the text of curriculum documents.
- An overarching (theoretically/conceptually informed) structure within which to comparatively locate different gender-related curriculum document characteristics, which could enable classification or comparison (or potentially 'scoring') of different documents.

Previous work on the use of GAT terms to describe curriculum characteristics provides a useful source for identifying the full range of different characteristics that may be detectable within the text of curriculum documents. In 2022, an extensive literature review was carried out to identify the range of ways in which GAT-related curricula were defined in literature published between 1995 and 2022 (Weavers, 2022). In total, definitions were extracted from 1,049 papers and were then deconstructed to identify any common 'definition components'. The purpose of this work was to explore how consistently the different GAT were understood in relation to curriculum. However, the 63 different 'definition components' identified through the work provide a comprehensive overview of the full range of curriculum characteristics associated with different GAT that have been discussed in the literature over this time period.

The 2022 literature review included definitions of GAT curricula drawing from terms used across the full range of GAT continuums published by different authors (specifically gender-blind, gender-neutral, gender-aware, gender-sensitive, gender-responsive, and gender-transformative curricula). The definition components extracted therefore include examples of what could be considered to be 'negative', 'intermediary' and 'positive/desired' approaches (see *Figure 11*). For example, some of the curriculum characteristics described are considered to be negative or damaging for gender equality (e.g., ***C22:*** *The curriculum reinforces traditional ideas of masculinity and femininity or gender power dynamics*) in contrast to more constructive characteristics (e.g., ***C15:*** *The curriculum content includes representations of women in diverse roles (e.g., doctors, leaders);* or ***C60:*** *The curriculum requires learners to critically analyse beliefs/norms that lead to gender inequality* — see *Figure 12* for an overview of the types of characteristics identified, and the full list is provided in **Appendix 2**).

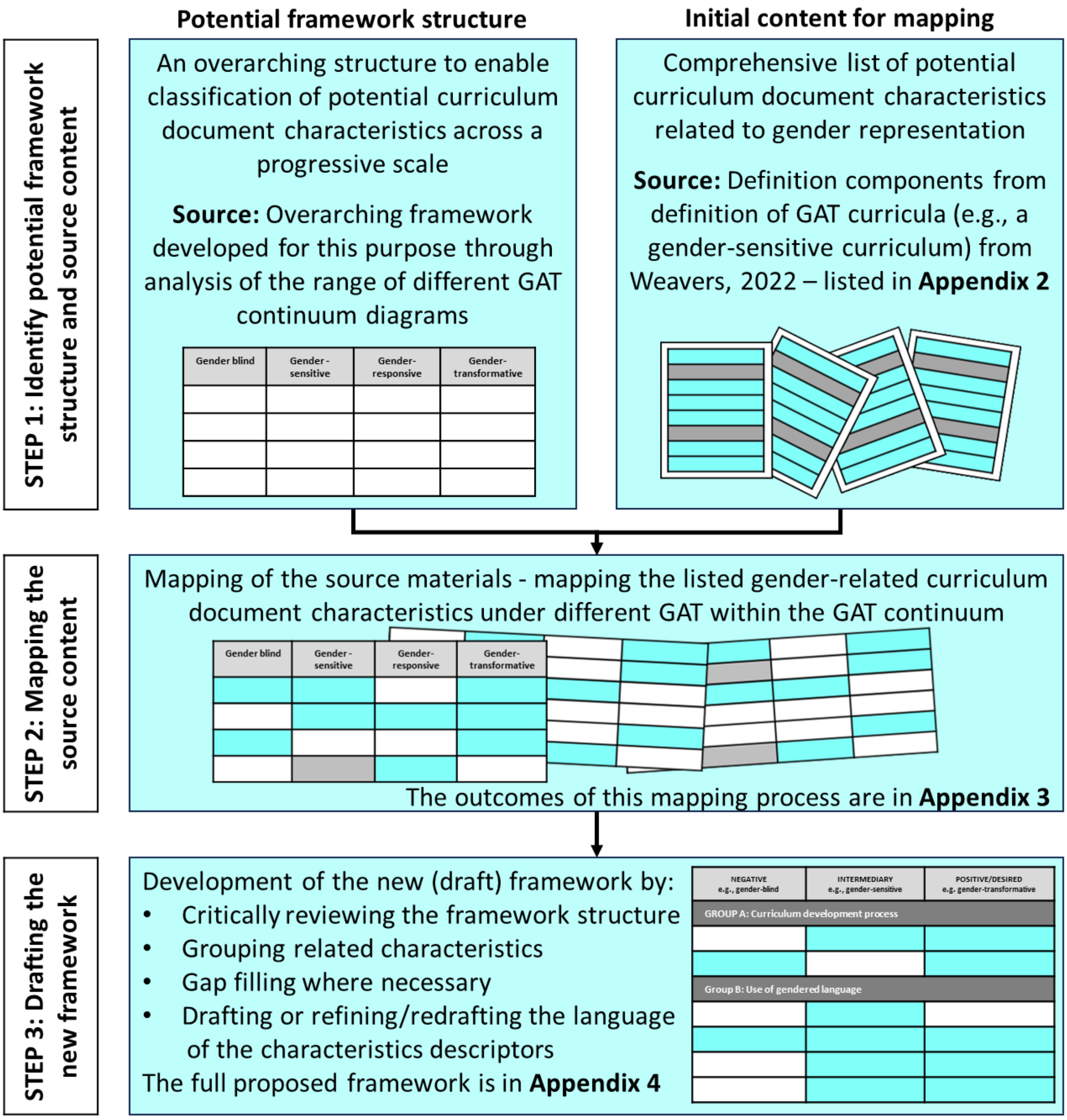
*Figure 12: Overview of the 7 categories into which the 63 identified GAT curriculum 'definition components' were organised*

| **Groupings of GAT definition components for use in curriculum development** |
|---|
| Group A: Definition components related to the curriculum development process |
| Group B: Representations of gender or gender relationships that are purposefully avoided in the curriculum |
| Group C: References to and/or representations of gender or gender relationships that are purposefully included in the curriculum |
| Group D: References to meeting the (different) needs of different learners |
| Group E: General skills, attitudes and/or values |
| Group F: Specific content/topics to be taught |
| Group G: Specific skills, values and/or attitudes related to gender/gender inequality |

*From Weavers, 2022*

Mapping this list of 'definition components' across a GAT continuum structure could therefore provide an evidence-based source to inform the development of a draft Gender Framework for Curriculum Text Analysis (see *Figure 13*).

*Figure 13: Overview of the process used in the development of the draft Gender Framework for Curriculum Text Analysis*




The language used in the definition components (from Weavers, 2022) is representative of the language found in the relevant definitions within the literature. While these components provide a useful evidence-based source to inform the final framework, significant refinement and redrafting were required to develop descriptors for the final framework (Step 3 in *Figure 13*).

The framework structure used in this mapping process (see *Figure 14*) was constructed by identifying the most commonly used GAT terms in the previously published continuums (see *Figure 10*), applying definitions from the UNESCO Glossary of terms related to gender equality and inclusion in education (UNESCO, 2022), and assigning these across the progression structure set out in *Figure 11*.

*Figure 14: Overarching framework developed for mapping of the GAT curriculum definition components (Step 1 and 2 in Figure 13) to inform the development of a Gender Framework for Curriculum Text Analysis.*

| NEGATIVE | INTERMEDIARY | | POSTITIVE/DESIRED |
|---|---|---|---|
| **Gender blind *** | **Gender sensitive** | **Gender responsive** | **Gender transformative** |
| The failure to recognise that the roles and responsibilities of men/boys and women/girls are given to them in specific social, cultural, economic and political contexts and backgrounds. The failure to....take into account these different roles and their diverse needs. This maintains the status quo and will not help transform the unequal structure of gender relations (UNESCO, 2022, p.5) | Acknowledges that differences and inequalities between women and men require attention, but does not attempt to address them (UNESCO, 2022, p.8) | Addressing and responding to the different needs, aspirations, capacities and contributions of women and men (UNESCO, 2022, p.7) | Challenging existing and biased/ discriminatory policies, practices and programmes and affecting change for the betterment of life for all (UNESCO, 2022, p.8) |

*Definitions from UNESCO gender equality and inclusion in education Glossary (UNESCO, 2022)*

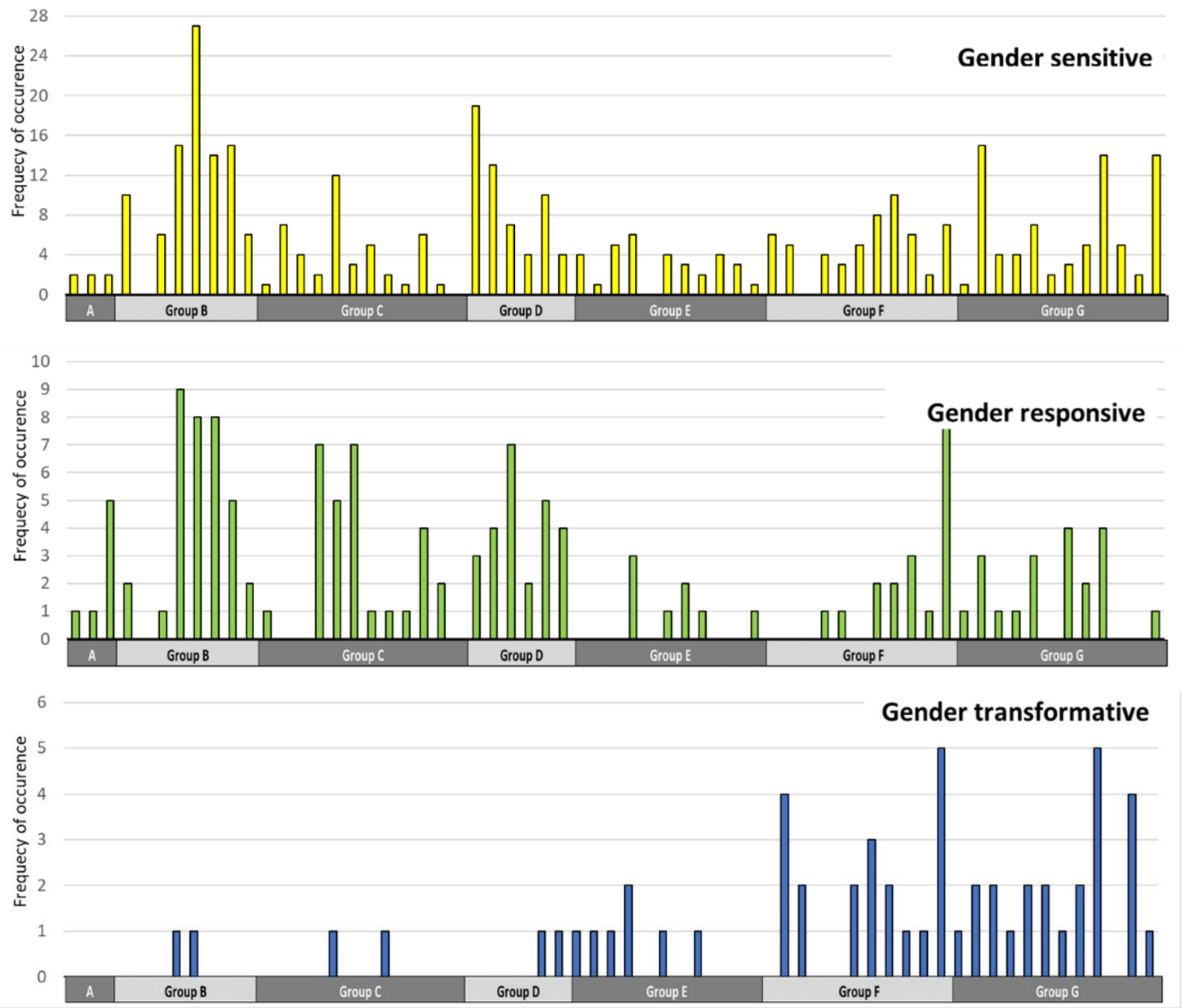
**In the literature, gender-blind can be found in relation to negative approaches (e.g., blind to the different needs or experiences of men/boys and women/girls) or more constructive approaches (e.g., making documents 'blind' to gender by removing gendered language completely from documents to avoid reinforcing bias or existing stereotypes). For this overarching framework, UNESCO definitions have been adopted, and these categorise 'gender-blind' as a negative approach*

The initial mapping of the GAT definition components (in Appendix 2) against the overarching continuum structure (in Figure 14) can be found in **Appendix 3**. This process reinforced some of the challenges regarding the design and use of GAT and continuum diagrams related to them. For example, over time different GAT (e.g., gender-neutral, gender-sensitive) have been adopted by or have been heavily influenced by distinct theoretical conceptualisations of gender within development (e.g., Women in Development, Gender and Development). Consequently, the widely accepted definitions of different GAT are relatively theoretical in nature and can be difficult to apply distinctly at a more practical, detailed level.

In addition, when looking at the more technical or practical definitions of GAT within the curriculum literature, some GAT are so varyingly defined that they lack any 'intuitive meaning' to curriculum developers (Weavers, 2022). For example, the terms 'gender-sensitive' and 'gender-responsive' are described in so many different ways within the literature that the majority of definition components identified (see Appendix 2) have been used in at least one definition of the terms (Weavers, 2022). There is also significant overlap between how the two terms are defined within the literature (see *Figure 15*).

Some definitions of GAT curricula like 'gender-blind' (negative applications) or 'gender-transformative curricula' showed some commonality and grouping around particular definition components (Weavers, 2022). However, for gender-sensitive and gender-responsive there is no clear way at this point in time to distinguish between what the terms mean at a practical level. This was evident within the mapping process as it proved very difficult to confidently allocate different 'definition components' between these two GAT categories.

*Figure 15: Numerical counts of definition components used to describe 'gender-sensitive curricula', 'gender-responsive curricula' and 'gender-transformative curricula' in literature published between 1995 and 2022. The definition component groups A-G are listed in Figure 12, and a full list of definition components is provided in **Appendix 2***




*From Weavers (2022)*

In the development of the draft Gender Framework for Curriculum Text Analysis, this was addressed by using a simplified version of the continuum based around ‘negative’, ‘intermediary’ and ‘positive/desired’ categories (see *Figure 14*). After mapping the original definition components across the continuum, the comparative potential was reviewed and scaled definitions were developed where appropriate to ‘fill the gaps’ across the simplified continuum. Substantial refinement and redrafting of the language were then carried out to ensure that the draft framework represented current thinking around gender and gender equality within and through education. An initial version of this draft Gender Framework for Curriculum Text Analysis can be found in **Appendix 4**, and a summary of the groupings of characteristics within the framework is shown in *Figure 16*.

*Figure 16: Groupings of approaches from the draft Gender Framework for Curriculum Text Analysis to inform thinking for the Bridging AGEE.*

| Groupings of gender approaches from the draft Gender Framework for Curriculum Text Analysis |
|---|
| 1. Curriculum development process |
| 2. Curriculum aims, goals or vision |
| 3. Use of language |
| 4. Gender bias (in mentions) |
| 5. Gender dynamics |
| 6. Gender stereotypes — personal characteristics |
| 7. Gender stereotypes — roles and achievement |
| 8. Recognised figures and gendered perspectives |
| 9. Access, opportunities and relevance |
| 10. Curriculum content — sex education |
| 11. Curriculum content – sexuality |
| 12. Curriculum content — gender equality |
| 13. Curriculum skills — Social skills linked to improving gender relationships and equal participation |
| 14. Curriculum skills — Recognising and analysing inequalities |

This framework is presented at this stage as an initial draft to inform thinking around the potential for analysis of gender representation within curriculum documents. It represents an early draft that would benefit significantly from further review and refinement but is shared at this stage to support thinking about different LLM techniques for use in curriculum analysis (see **Section 3.2.3**). Further consultation and review of this draft framework could also contribute towards providing clarity in the intended meanings of GAT that currently lack consistent definitions at the practical level of curriculum development. This would enable the framework to be structured around individual GAT rather than the current use of the overarching categories from the GAT continuums.

### 3.2.2 Options in constructing an analytical model

**Box 4: Example of LLM response to zero shot prompt – applying LLMs to gender analysis**

**PROMPT:** Develop a witty quote about why it is challenging to use LLMs to analyse complex concepts like gender representation in complex documents like policy and curriculum documents

**ChatGPT response:** *"Using an LLM to analyse gender representation in complex documents is like trying to measure the weight of a shadow with a ruler—it can outline the shape, but capturing the full depth and nuance requires tools designed for subtler dimensions."*

One of the challenges identified with the existing literature in this area (see **Section 2.6.1**) is that in most cases the research focuses on a limited version of the myriad of ways in which gender and gender issues can present within text. This is in part due to the work required to establish particular analytical pathways using traditional NLP methods. With the emergence of LLMs, there is potential to carry out richer contextual analysis but also to explore potential ways to simplify the analytical protocols to enable a broader range of potential representations of gender to be explored.

For example, considering the range of categories within the draft Gender Framework for Curriculum Text Analysis (see *Figure 16* and **Appendix 4**) enables the identification of an array of potential analytical techniques that could be deployed. These include traditional NLP approaches and comparable LLM capabilities (see *Figure 17*). The listed approaches are just examples rather than an exhaustive list, as there are a range of subtly different analytical approaches that could be used or adapted to provide information concerning each of the different categories of gender representation. Comparison of the options related to traditional NLP techniques (which can be applied by LLMs) with the more emergent capacities of LLMs can also assist in identifying where less rigid analytical approaches can be adopted. For example, traditional NLP models would not efficiently identify whether a curriculum document states that a gender audit was carried out — particularly in comparison with an LLM question-response.

*Figure 17: Mapping of examples of potential traditional NLP and LLM techniques that may be applicable to generate data relevant to each of the themes identified in the draft Gender Framework for Curriculum Text Analysis (see* ***Appendix 4*** *for full details of the draft framework)*

*Column A: Whether all curriculum documents would likely include relevant content (Green: very likely; Amber: likely; Red: unlikely)*

*Column B: Potential accuracy of LLM responses (Green: highly accurate; Amber: reasonably accurate; Red: potential accuracy issues)*

| | | A | Examples of potential traditional NLP analysis | Examples of potential LLM analysis | B | Potential nature of data from analysis |
|---|---|---|---|---|---|---|
| **1. Curriculum development process** | **Gender balance in curriculum development** | | Named entity recognition using probability-based gender approximation based on datasets of local names | Context-Aware Entity Recognition using probability -based approximation drawing on initial training | | Counts of male and female developers |
| | **Use of gender audits** | | Semantic search | Question answering | | Frequency of occurrence, Y/N |
| **2. Curriculum aims, goals or vision** | **Promotion of gender equality**<br>**Promotion of equity**<br>**Equal aspirations** | | Lexicon-driven semantic search; information retrieval, topic cluster analysis | Question answering; Contextual understanding (with exemplar material) or context-enhanced semantic search | | Frequency/lists of occurrences, confidence scores, Y/N |
| **3. Use of language** | **Balance of gendered language** | | Simple frequency counting/token matching using provided list | Contextual matching using an initial word list generation and refinement process | | Word counts |
| | **Absence of sexist or demeaning language** | | Lexicon-driven semantic search or application of the Hurtfulness Index | Contextual understanding (with exemplar material) or context-enhanced semantic search | | Co-occurrence of gender/sexist language or scale rating of statements |
| **4. Gender bias (in mentions)** | **Balance in references to individuals** | | Entity extraction | Entity extraction | | Mention counts, scaled responses or rubric responses |
| **5. Gender dynamics** | **Addressing gender/power dynamics** | | Relationship extraction (e.g., based on order of mention, named vs. general, titles and roles, context | Multidimensional Relationship Analysis or hierarchy detection | | Category responses, numerical scaling |

| | | | | | | |
|---|---|---|---|---|---|---|
| **6. Gender stereotypes — personal characteristics** | **Avoiding character-based gender stereotypes** | | Word embedding analysis using word lists associated with gender, personal characteristics and roles/achievement | Stereotype attribute mapping and evaluation — analysis of linguistic patterns, contextual associations, and sentiment to identify and quantify stereotypes in how personal characteristics or roles/achievement are attributed to different groups within a document. | | Statistical data, heat map representations |
| **7. Gender stereotypes — roles and achievement** | **Avoiding role-based gender stereotypes** | | | | | |
| **8. Recognised figures and gendered perspectives** | **Balance in references to recognised individuals** | | Named entity recognition | Named entity recognition, followed by cross referencing of the names with a knowledge graph (to identify known figures and application of a semantic similarity model to match ambiguous references to known figures | | Counts of different categories of recognised figures |
| | **Balance in inclusion of authors or artists** | | | | | |
| | **Balance in gender perspectives represented** | | *Could be approximated by multi-step analysis using a range of NLP techniques* | Possible through a sequence of processes (e.g., linguistic patterns, entity references, and contextual roles) | | Statistical data, proportional count of allocated perspectives |
| **9. Access, opportunities and relevance** | **Equal access to curriculum courses** | | *Possible using keyword searching and proximity analysis but this is complicated less reliable* | Question/answer | | Y/N |
| | **Equal balance of relevant learning/ assessment styles** | | Lexicon-based key word searching | Contextual searches | | Proportional data |
| | **Equal balance interests-related content** | | Lexicon-based key word searching | Contextual searches | | Proportional data |
| **10. CURRICULUM CONTENT — sex education** | | | Keyword/expression-based search; topic modelling/cluster analysis | Question-answer; Contextual semantic search | | Frequency/lists of occurrences, confidence scores |
| **11. CURRICULUM CONTENT — sexuality** | | | | | | |
| **12. CURRICULUM CONTENT — gender equality** | | | | | | |
| **13. CURRICULUM SKILLS Social skills for gender equality** | | | | | | |
| **14. CURRICULUM SKILLS: Recognising and analysing inequalities** | | | | | | |

In terms of constructing an analytical approach to comparatively analyse gender representation in national curriculum documents, there are a number of options that can be explored. Considering the variety of established analytical approaches included as examples in *Figure 17*, there are a range of ways that the framework could be used — for example:

1. To prioritise a particular analytical approach that would be applicable to the vast majority of different styles of curriculum documents (e.g., *Gender role stereotypes*) and explore how to use the emerging capabilities of LLMs to streamline or add contextual richness to this measure (compared with the existing applications in the literature).
2. To identify a couple of different measures of a particular aspect/category of gender representation (e.g., *Use of language*) that can be applied jointly across the majority of curriculum documents to fully explore different elements of the category and generate a measure that overcomes some of the challenges related to the reductive nature of NLP techniques.
3. To select a range of 5-6 standard forms of analysis (potentially combining more traditional and more emergent approaches) that would provide evidence concerning different categories of gender representation (e.g., *Use of language, Gender stereotypes, Recognised figures, Curriculum content, Curriculum skills*) within curriculum documents to create a form of index score or a visual model to enable comparability between nations.
4. To develop a detailed analysis protocol (potentially making use of some of the capabilities of LLMs to approximate more in-depth analysis) designed to develop rich datasets about the documents drawing on the full range of identified categories. This would be likely to include forms of analysis that would not be applicable to all forms of curriculum documents. However, it would generate rich data that could be displayed visually or scored to give direct indications of practical strategies that can be applied to address any issues. Just as the AGEE framework has done, this approach would involve "*using multiple sources of information in well-organised, holistic yet adaptable combination, taking account of the complexity of processes in play, to develop guidance on processes for transformational and sustainable change*" (Unterhalter et al., 2022, p.511).

Another more explorative option would be to test the possibility of building a set of incomplete prompt sentences or sentences with masked words that cover the full range of categories of gender representation identified. This would draw on the approaches used to detect and address gender bias in AI itself (e.g. the GenderAlign database of 8,000 sentence prompts) (see **Section 2.5**). A fundamental question to explore would be whether LLMs can reliably respond to these prompts on the basis of the document content alone (rather than by generalising or drawing on prior knowledge). If so, then clever design of the sentence prompts would enable an overarching tool to be developed that could convey the extent of different forms of bias within the same document through the application of a single analytical protocol. While this would not provide direct numerical responses (e.g., the number of male and female named entities within the document), the protocol could be used to give an overarching 'score'. However, substantial testing would be required to determine the extent to which the results are representative of the curriculum document content in each country context and the extent to which they enable comparability between documents.

If this approach is viable, then it could enable the development of a wide range of statements addressing the different categories in the draft Gender Framework for Curriculum Text Analysis alongside statements aligned to established theoretical frameworks. The statements could be completed and then scored using an LLM in a two-step process. However, significant work would be required to develop the prompt statements and to test whether the results correlate with different forms of gender representation within curriculum documents.

There are a number of factors which will influence which option is most applicable, including:

- The intended use (e.g., direct comparability between nations or targeted recommendations for individual nations)
- The desired format of the outputs (e.g. descriptive text, categorisation, numerical scoring, or visual representation of strengths and weaknesses)
- The extent of time and resource available for the design and development phase
- The extent of time and resource available for the ongoing implementation
- The potential for manual involvement within the analytical protocol in addition to validation processes
- The appetite and capacity to explore novel approaches compared with applying well-documented approaches

### 3.2.3 Implications of curriculum document characteristics for analysis

As established in **Section 1.3**, there is significant variation in the way the term 'curriculum' is understood from different perspectives. Even when the focus is narrowed specifically examine at national-level curriculum documents that set out the prescribed content that is expected to be taught or learnt at each stage of education, there is still huge variety in the form these documents take.

In different countries, the 'national curriculum' can take a range of forms and names. For example, the documents that contain this information may be referred to as the curriculum, the syllabus, the assessment specification, or a set of standards. The way in which the information is packaged also varies significantly. In some countries there is a single curriculum document for each phase of schooling (e.g., the primary and secondary school curriculum documents in England). In other countries there are curriculum documents for each subject area that contain the information for all or for a selection of grades (e.g., by phase in Ghana). In other countries, documents are organised by grade or, in some cases, there are separate documents for each subject in each grade (e.g., Kenya). In some cases there is no separate curriculum document at all, and the textbooks are referred to as 'the curriculum' (e.g., Andra Pradesh, India).

This variability raises questions for the potential analysis of curriculum documents in terms of the extent to which the content analysed from different countries is directly comparable. If particular subjects in particular grades were to be sampled, then choices would need to be made regarding how to deal with the 'surrounding content' in the documents (e.g. introductory sections setting out the aims, goals, and curriculum approach), as in some cases this will be grade-specific and in other cases it will refer to multiple grades.

The decision to sample particular subjects or grades in contrast to analysing the full suite of curriculum documents also has other implications for the range of questions that can be explored. Particular curriculum content that relates strongly to gender (e.g., sexuality education, human rights, gender equality) may be covered in different subjects or different grades in different contexts. Therefore, if understanding the extent to which these topics are covered and how they are framed within the curriculum is important, a sampling approach would not necessarily capture the required information.

In Lesotho, for example, the 2009 Curriculum and Assessment Policy describes the intention for gender equality and equity to be one of the 'organising elements' within an integrated curriculum and for learners to be able to demonstrate gender sensitivity as a result of their education (MoET, 2009). When the curriculum was reformed in line with this policy, content related to gender and gender sensitivity was incorporated into specific grades in the primary curriculum (see *Figure 18*) and specific subjects in the secondary curriculum. Sampling different grades within this curriculum to identify the extent of coverage of topics related to gender equality and gender representation would generate very different results.

The nature (and extent) of the content included in curriculum documents also varies widely among different countries. In most cases, the documents contain at least a few sections that are not directly listing teaching content (e.g., vision, aims and goals, curriculum approaches, intended pedagogy, research evidence justifying the approach, scheduling/teaching time, guidance about assessment or exemplars, explanations of cross cutting themes, skills or core competencies, examples of teaching or assessment activities). Therefore, it is important to recognise that analysis of whole documents will not necessarily be comparing like with like from different contexts.

*Figure 18: Gender-related content in the Lesotho Primary School Curriculum implemented from 2013 (Grades 1-7)*

| Integrated Primary Curriculum Grade | Curriculum content related to gender equity, gender equality and gender sensitivity |
|---|---|
| **Grade 1** | No content relating to gender |
| **Grade 2** | No content relating to gender |
| **Grade 3** | No content relating to gender |
| **Grade 4**<br>(NCDC, 2013) | "***Unit 1: Knowing oneself and relating with others***<br>***Learning Outcome 6 —*** *At the end of this unit learners should be able to: Explain that differences in the ability of girls and boys are more cultural than biological"* (NCDC, 2013d, p.1)<br>Learning activities includes differentiating between characteristics of girls and boys that are culturally or biologically determined and understanding the concept of gender roles, where these ideas come from and how they are passed on through generations<br>***"Learning Outcome 7 —*** *At the end of this unit learners should be able to: Resist negative effects of gender socialisation on their self-image"* (NCDC, 2013d, p.1)<br>Learning activities include creating and acting out scenarios with reversed gender stereotypes; exploring ideas about gender norms, values and beliefs and how these can influence perceived gender roles and self-image. |
| **Grade 5** | No content relating to gender |
| **Grade 6** | No content relating to gender |
| **Grade 7** | No content relating to gender |

*From NCDC (2013)*

The different types of content within curriculum documents can also provide contextual challenges for analysis. To give a simplistic example, a curriculum document may contain a vision section with a strong emphasis on girls' education or gender equality that includes content that both describes the current scenario and sets out a vision for the future. This potentially complicates comparative analysis in two main ways. Firstly, if traditional NLP techniques were applied to this document, the results would be very different from a document that only included the teaching content. Secondly, the results could be ambiguous for this document because it contains content that describes gender inequality as well as content that aspires to gender equality, but none of this may be reflected in the actual teaching content.

There are choices to be made therefore regarding the analysis protocol. It may be important for certain forms of analysis being carried out that the curriculum content is separated out into different types of content within an analysis template beforehand. This could provide interesting insights into the difference in references to gender and gender equality between more 'visionary' content and the practical teaching content included.

An additional consideration regarding the extent of document preparation or templating required before analysis is that different curriculum documents contain different levels of detail and lay this out in various different ways that can complicate text extraction for analysis. The layout of the teaching content pages can provide context for particular content that is then lost in text extraction. For example, the teaching content in the English curriculum documents is presented as a simple list of intended learning outcomes, accompanied by some non-statutory guidance to clarify the level of demand (see *Figure 19*).

*Figure 19: Extract from the Year 4 Science Programme of Study in the English Primary Curriculum showing the full content for 3 individual learning outcomes*

**States of matter**

**Statutory requirements**

Pupils should be taught to:

- compare and group materials together, according to whether they are solids, liquids or gases
- observe that some materials change state when they are heated or cooled, and measure or research the temperature at which this happens in degrees Celsius (°C)
- identify the part played by evaporation and condensation in the water cycle and associate the rate of evaporation with temperature.

**Notes and guidance (non-statutory)**

Pupils should explore a variety of everyday materials and develop simple descriptions of the states of matter (solids hold their shape; liquids form a pool not a pile; gases escape from an unsealed container). Pupils should observe water as a solid, a liquid and a gas and should note the changes to water when it is heated or cooled.

*From DfE, 2013, p.21*

In contrast, the national curriculum documents in Kenya provide a wide range of additional content to accompany and exemplify the learning outcomes, and this is presented in a series of complex tables throughout the document (see *Figure 20*).

*Figure 20: Table structure and surrounding information provided alongside the specific learning outcomes in each sub-strand in the current national curriculum document in Kenya*

| **Strand** | **Sub-strand** | **Specific learning outcomes** | Suggested learning experiences | Key Inquiry Question(s) |
| --- | --- | --- | --- | --- |
| | | | | |
| | | | | |
| | | | | |
| Core competencies to be developed: | | | | |
| Pertinent and Contemporary Issues (PCIs): | | | Values: | |
| Links to other learning areas: | | | Suggested Community Service-Learning Activities: | |

**Assessment rubric**

| **Level / Indicator** | **Exceeding Expectations** | **Meeting Expectations** | **Approaching Expectations** | **Below Expectations** |
| --- | --- | --- | --- | --- |
| | | | | |
| | | | | |

*From the Kenyan Institute of Curriculum Development (2017)*

## 3.3 Initial thoughts on frameworks and frames of reference for analysing policy documents

### 3.3.1 Implications of policy document characteristics for analysis

Policy development is essentially a process of consensus building (see *Figure 3*) and policy documents are complex in nature as they often represent the best-case combination of contrasting interests of particular groups — the "*decisions of individual administrators, of course, must be integrated with decisions of others to form the mosaic of public policy*" (Lindbloom, 1959, p.79).

Like curriculum documents, policy documents are also characterised by the varied range of types of content they contain (see *Figure 21*). It is not uncommon for different sections of a policy document to have been drafted by different individuals or groups.

*Figure 21: An overview of the types of content identified as representing best practice from an international comparison study carried out to inform the redesign of a national education policy development process*

I. FOREWORD
II. PREFACE
III. ACKNOWLEDGEMENTS
IV. CONTENTS
V. ACRONYMS / ABBREVIATIONS
VI. GLOSSARY OF TERMS / DEFINITIONS

| |
|---|
| **SECTION 1: INTRODUCTION**<br>1.1 EXECUTIVE SUMMARY<br>1.2 COUNTRY CONTEXT<br>1.3 STRUCTURE OF THE EDUCATION SYSTEM<br>1.4 LINKS TO THE NATIONAL POLICY CONTEXT<br>1.5 LINKS TO THE NATIONAL DEVELOPMENT PLAN<br>1.6 LINKS TO THE INTERNATIONAL POLICY CONTEXT<br>1.7 ANALYSIS OF CURRENT LANDSCAPE, POLICY RATIONALE AND OBJECTIVES<br>1.8 SCOPE OF THE POLICY<br>1.9 GUIDING PRINCIPLES APPLIED WITHIN THE POLICY<br>1.10 DESCRIPTION OF THE POLICY DEVELOPMENT PROCESS<br>1.11 OVERVIEW OF POLICY FINANCING<br>**SECTION 2: POLICY STATEMENTS**<br>2.1 INTRODUCTION TO POLICY STATEMENTS<br>2.2 POLICY STATEMENTS — AREA 1<br>2.3 POLICY STATEMENTS — AREA 2<br>2.4 POLICY STATEMENTS — AREA 3 ETC<br>**SECTION 3: IMPLEMENTATION, MONITORING AND EVALUATION**<br>3.1 IMPLEMENTATION FRAMEWORK INCLUDING COSTINGS AND ROLES AND RESPONSIBILITIES<br>3.2 MONITORING POLICY IMPLEMENTATION<br>3.3 EVALUATING POLICY IMPLEMENTATION<br>3.4 POLICY REVIEW CYCLE |

This presents some potential challenges for the types of analysis carried out using NLP models or LLMs, as different sections can reference gender and gender issues in completely different contextual framings. Most policy documents include some form of vision or aims (which may set out aspirational intentions regarding gender and gender equality) and a needs analysis or systematic review of the current landscape (which may include critical analysis of current issues with gender equality). Both of these involve different contextual framings from the policy statements themselves and any implementation planning content, which may include (in increasing levels of detail) suggested actions to be taken to address gender inequality.

Another consideration is that policy documents represent the intention of policymakers at a particular moment in time. Therefore, the policy statements or commitments documented within them tend to include a combination of statements where some:

- Describe existing practice already being implemented at the time the policy was approved.
- Describe changes to practice or processes that were intended to be implemented imminently when the policy was approved.
- Describe practice that is aspirational and likely to never be fully implemented (due to budget or system limitations) or, in some cases, never fully intended to be implemented (e.g., Meyer et al., 1997).

This, in combination with the fact that policy statements can become outdated over time, means that there is almost always a gap between the intended practice as described in full in policies and the actual reality on the ground. The size of this gap varies between policy documents, but as an example, a review of the 1994 Education and Training Policy was carried out with a team from the Ministry of Education in Ethiopia in 2019. The content of the still active policy was coded according to whether it described previous practice, current practice or 'still aspirational' practice. On average 32% of the content in the different sections was identified as 'still aspirational' 25 years after the policy was introduced. In any form of policy analysis, it is important to recognise the nature of the content as a statement of intended practice that was documented at a particular point in time.

### 3.3.2 Potential approaches to analysis of policy documents

There are a range of ways in which more traditional NLP approaches and the more recent LLM capabilities could be applied to analyse gender themes and representation within policy documents. If it is desirable to apply a single approach, a more traditional method like topic mapping could be used. However, the challenge with this is that the resulting topics are not always easy to interpret, and manual review and consolidation are required. There is also a risk that gender may not emerge strongly or consistently within the topic mapping of different policies due to the complexity of the concept and the fact that it is sometimes what is not stated directly that is the most telling in relation to gender equality.

Alternatively, an LLM technique involving semantic understanding and contextual analysis could be applied. However, if there are multiple policies to be analysed from each country, this may require extensive processing capacity, and there may be challenges with the consistency and comparability of the more descriptive results generated. There is also a risk of generalisation and hallucination when applying these techniques, which would need to be managed and minimised where possible through fine-tuning training or carefully constructed prompts. In general, traditional NLP methods are most effective for reductive but scalable analysis, while LLMs are able to process more nuanced and interpretive themes. An approach combining traditional and more nuanced analytical techniques has potential benefits for comparative analysis of complex themes like gender or intersectionality.

Rather than applying a general analysis using gender terms alone, more traditional NLP approaches could also be applied within a particular theoretical framing. For example, carrying out analysis of the co-occurrence or the semantic similarity of gender terms using lexicons developed on the basis of the four dimensions within Stromquist's (2002) conceptualisation of empowerment:

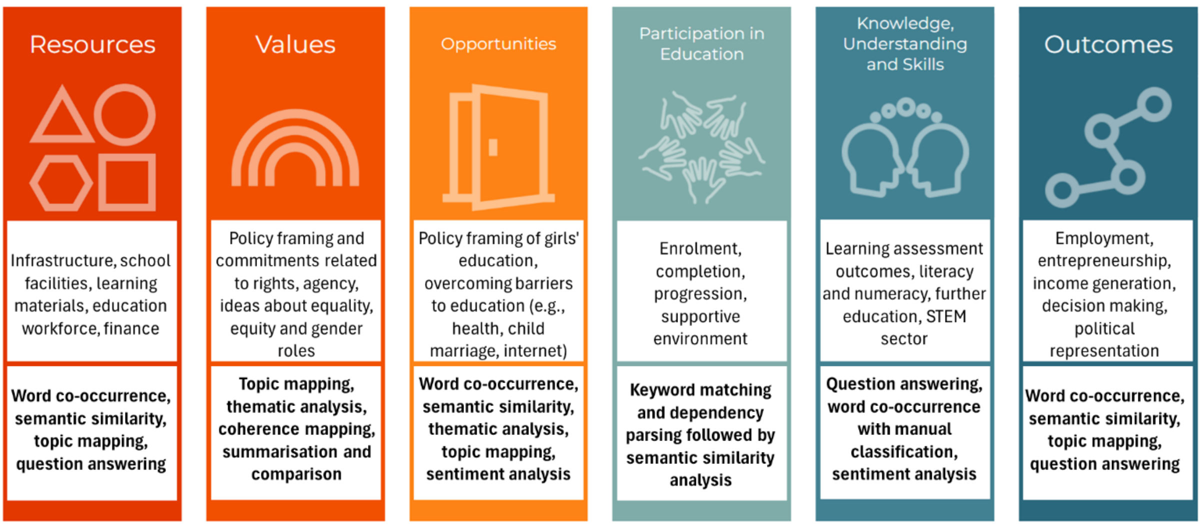
- The cognitive dimension — e.g., a lexicon containing terminology related to understanding of reality including human rights, legal rights, socio-politics, and geo-politics
- The psychological dimension — e.g., a lexicon containing terminology related to self-esteem, confidence, motivation, and autonomy
- The political dimension — e.g., a lexicon containing terminology related to self-expression, advocacy, governance, representation, participation, and awareness of political contexts
- The economic dimension — e.g., a lexicon containing terminology related to employment, employability skills, financial management, banking, and entrepreneurship (Stromquist, 2002).

An alternative approach could involve exploring the extent of '*gender-policy evaporation*' within policy documents (Longwe, 1997). This term refers to how commitments to address gender issues or inequality "*frequently evaporate in planning and implementation processes*" (*Moser and Moser, 2005,* p.15). For example, the governments that signed up to the Beijing Platform for Action following the 4th UN World Conference on Women in 1995 committed to "*design, implement and monitor… gender-sensitive policies and programmes*" (UN, 1995, Item 19, p.3) and "*create a* *gender-sensitive educational system*" (p.28). It is therefore common to find references to 'gender-sensitive' approaches in the initial sections of policy documents setting out the aims, vision, and intended approach (Subrahmanian, 2007). It may be interesting to governments therefore to be able to see how the use of gender-related language changes between these initial sections and other sections containing content like the detailed policy commitments, implementation approaches, and monitoring. This could be applied to specific policies (e.g., non-formal education policies, inclusion policies, or school infrastructure policies) or more generally. One approach would be to use cluster analysis to identify words or meanings associated with gender and intersecting inequalities within different sections of policy documents. Contextual extractions or coherence analysis could then be used to explore the extent to which particular commitments or framings are carried through to the planned actions. This could potentially provide an interesting insight into the consistency of commitments throughout education policy documents — from the vision stage to the detail of planning and implementation.

This approach was applied in the work carried out by Żółkowski et al. (2022) who used NLP techniques to review the framing of key issues within climate policies. In most cases, this approach would require some initial work to organise policy content within a standardised analysis template, which would be relatively challenging given the variety of policy documents to be examined. However, Żółkowski et al. (2022) used an EU document that was drafted by individual countries using the same template structure. The "*similar structure of the documents*" simplified the process, allowing them "*to compare the policy framing in different dimensions separately, which allows for a more fine-grained analysis of policy frames*" (p.2). By separating the content out according to the policy sections, they were able to compare the way key policy issues were framed and described in the content describing the current situation with that in content that described objectives or national targets, and policy commitments for intended implementation.

Where there is capacity for the use of a range of techniques, analysis could be structured around particular frameworks. The AGEE framework (Unterhalter et al., 2022) in particular would provide a strong framing structure.

*Figure 22: Examples of themes and topics related to each domain of the AGEE framework with some examples of NLP and LLM techniques that could be used to analyse policy documents to identify how gender is associated or represented in relation to these themes or topics*



*Adapted from Unterhalter et al., 2022*

*Figure 22* summarises the types of focus areas considered under each of the AGEE domains and includes some initial examples of techniques that could be used to gather evidence related to each theme.

This would provide a more nuanced analysis of gender representation and related actions within policy documents. One of the risks of using a simpler analytical framing is that it presents a reductive view of gender and related challenges. In contrast, while a more nuanced 'framework' approach would require careful design, it would provide richer information which would have the potential to have a greater impact regarding interpretability and translation into actionable change. The challenge, however, would be in designing processes that generate comparable data that are applicable to the range of policy framings adopted by different nations operating within vastly different contexts in relation to gender and gender equality.

# 4. Technical considerations for curriculum or education policy analysis using NLP or LLMs

## 4.1 General considerations

### 4.1.1 Measuring and/or mitigating gender bias

The consideration that needs to be given to acknowledging or addressing gender bias in the design of the analytical process depends to a great degree on the approach and model selected. Due to the broad, open nature of the materials LLMs access during initial training, there is a well-established "*gender bias accuracy gap between LLMs and traditional NLP systems*" (Sant et al., 2024, p.1). In addition to this, while there are techniques that have been developed to address gender bias in traditional NLP techniques, the work on addressing gender bias through prompting is still a developing field. Currently, while a range of prompting techniques have demonstrated some impact in reducing gender bias, there is limited evidence of approaches that work consistently across different models and the mechanisms that explain how particular techniques work in particular contexts are not yet well understood.

This presents a key challenge therefore when using LLMs to interpret and analyse gender themes within curriculum and policy documents. It is difficult to determine whether applying a prompting technique used to reduce gender bias in LLMs could have a distorting effect on the analysis itself. Sant et al. (2024) identified gender bias in a number of LLMs and explored the impact of a wide range of prompting techniques. They identified one particular approach that delivered a significant reduction in bias — a chain-of-thought style prompt including 5 examples (see **Section 4.3.1** for explanations of prompt formats) which all presented anti-stereotypical content and an increased use of female gendered terms. Other approaches, including setting the context by describing the challenge with gender bias or using example structures that had demonstrated impact in other work, were not as successful. It is not clear why this particular structure worked when other approaches didn't or whether the use of only 'anti-stereotypical' examples would influence the nature of data collected about gender representation within the document.

When particular NLP or LLM models and specific analytical approaches are selected for any form of gender analysis, evidence of existing gender bias within the model will need to be considered. It is also important to identify whether any gender bias mitigation strategies have been applied to the training materials used, the internal operation of the model itself, or to the outputs generated. Ethically, it is always important to identify potential gender bias in any NLP/LLM analysis. However, when the focus of the analysis is gender representation within the text, there is potential that existing bias or the related mitigation strategies may distort the data collected. Thorough testing however would be required to understand the impact of any existing model bias relating to girls/women and boys/men, and any potential mitigation techniques on the analysis itself.

### 4.1.2 Looking beyond documents in English

The ability of LLMs to function in multiple languages is advancing rapidly. However, there are issues that need to be addressed in the design of analytical protocols if there is an intention to use LLM techniques to analyse documents like curricula and policies that are only available in other languages.

At the basic level, when manual document analysis is carried out, there are three main choices when dealing with documents in various languages:

- Translating all documents into a base language before analysis
- Using human analysts that are fluent in the range of languages involved to carry out a set analytical protocol (which is only viable for small numbers of languages)
- Translating the analytical tools and training specialists who speak the required languages to carry out the analysis

Each of these approaches has strengths and limitations as there is the need to rely on accurate translation or accurate and comparable human interpretation of the content presented in different languages. When carrying out document analysis using LLMs, similar choices exist, but there are additional considerations around the effectiveness of LLMs in translating or analysing documents in different languages.

In a multi-step analytical protocol, LLMs could initially be used to translate curriculum and policy documents into a base language (e.g., English) before analysis (e.g., Miah et al., 2024). Text translation is a key and growing business-sector application of LLMs. However, while the models are rapidly improving over time, there are particular challenges in accuracy in relation to 'low resource languages' where there has not yet been significant investment in developing large training datasets or tools. Various approaches have been developed to improve the accuracy of translation of lesser known or 'low resource' languages including, feeding human validation data back into the fine-tuning process (e.g., Elshin et al., 2024) or developing new benchmark tools to enrich the range of materials that can be used for training (e.g., Alhanai et al., 2025). While the quality of translation will continue to increase over time, LLM translation of 'low-resource languages' would require significant manual input and validation at this point in time — particularly when the translated text would then be analysed in relation to complex concepts like gender, where loss of any contextual information would potentially skew the analysis.

Alternative approaches involve the use of LLMs that have the capacity to carry out the same analysis on materials in various languages. This highlights however one of the challenges with the lack of transparency around the analytical processes being carried out by LLMs. 'Multi-lingual LLMs' may appear to function as if they are equally fluent in a range of languages. However, recent research has demonstrated that multilingual LLMs tend to process information through representations that are heavily shaped by their base language (which is often English) (e.g., Zhong et al., 2024; Schut et al., 2025). As a result, there tends to be a performance gap in terms of the accuracy of analysis in other languages, which can be hard to quantify as "*multilingual LLMs perform key reasoning steps in a representation that is heavily shaped by English in a way that is not transparent to system users*" (Schut et al., 2025, p.1).

The selection of an appropriate approach to deal with documents in different languages is even more challenging when the intended analysis focuses on a complex concept like gender representation. Various authors have noted challenges in conducting semantic or contextual LLM analysis on texts in other languages or translated versions due to the nuances of meaning in different language structures (e.g., Ketamo et al., 2019). In languages with grammatical gender systems, using LLMs to carry out analysis related to gender representation (in the original language or in translated form) has been found to be far more challenging even in relation to relatively simple analytical approaches like word embedding and cluster analysis (e.g., Ratovicius et al., 2024).

It is also now well documented that gender bias identification and mitigation approaches that work in English do not function as effectively across gendered languages (e.g., Derner et al., 2024). This suggests that it may not be possible to apply consistent tools for gender analysis across documents in different languages. The risks can be mitigated through the input of local expertise in adapting tools and developing culturally informed training materials to provide some level of comparability between the data generated from documents in different languages.

Ultimately, this challenge emphasises the importance of 'drawing the right line' between manual analysis and LLM analysis within research protocols (see **Section 4.1.3**). The use of machine translation of documents or the application of multilingual LLMs that operate within a single base language both risk imposing linguistic and cultural biases that may obscure or misrepresent the original gendered meanings embedded in the original documents. As with most decisions related to the use of NLP or LLM techniques, there will be trade-offs to consider in relation to accuracy, cost-effectiveness, and the degree of adherence to a central principle of only applying NLP or LLM techniques where they are able to facilitate, enhance, or surpass human analytical capabilities. While the translation capabilities of LLMs are developing rapidly, for analysis of documents from low-resource languages in particular, it is likely considerable human input into document translation will continue to be necessary for the foreseeable future.

### 4.1.3 Drawing the right line between NLP models, LLMs, and manual analysis

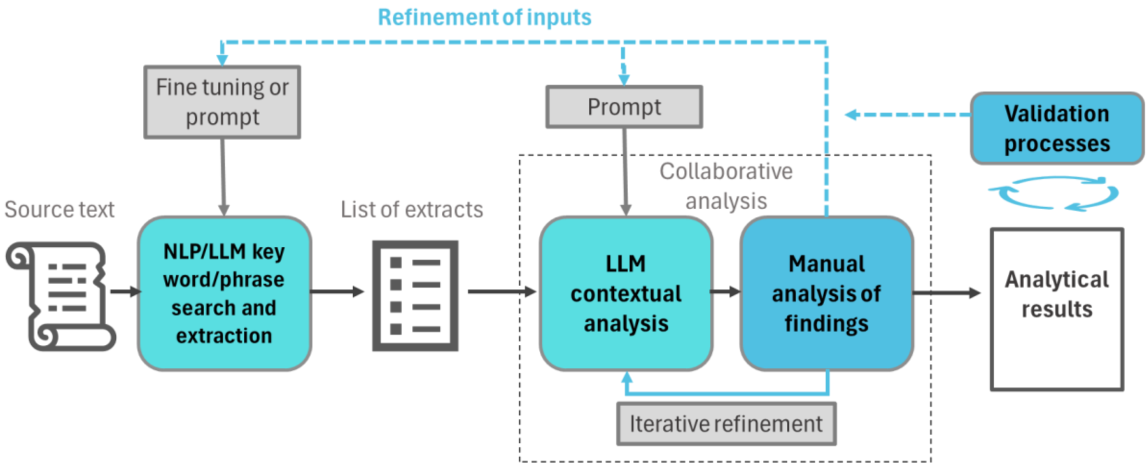
Most of the literature to date that relates to the use of NLP models or LLMs to analyse policy (or curriculum or textbooks) focuses on the power of these tools to conduct analysis that is not currently carried out manually at such a scale because would not be cost-effective to carry out by hand — *"if this text extraction had been done manually it would have required more than 25 researchers working for a year" (Ketamo et al., 2019, p.147)*.

The trade-off with this however is that the more traditional NLP approaches are reductive in nature and elements of nuance, concept, and context can be lost (e.g., Ratovicius et al., 2024). With the emergence and ongoing development of the enhanced contextual capabilities of LLMs, the sector is striving to test the limits of what is possible and find ways to enhance the accuracy of their analytical processes. There has been a tendency therefore for research to focus on the use of LLMs to replace of manual research techniques in their entirety. However to this day, "*arguably one of AI's greatest limitations is in its ability to make sense of human reality. This refers to understanding causality and cultural nuances, that is, unwritten 'rules' that an average person would be able to comprehend and process*" (Wirjo et al., 2022, p.5)

This raises the question — why try to pursue this aim through the use of LLMs at this point in time? Given the functionality available, there is growing recognition that the greatest benefit from the use of LLMs may be in starting from the perspective of a 'human-centred approach' (UNESCO, 2023) with the aim to augment rather than replace the work of specialists. Qiu and Jin (2024) explored the integrated use of ChatGPT by design domain experts and argue that the most effective model is "*a nuanced integration of artificial intelligence (AI) with human expertise, bridging the gap from mere classification models to sophisticated human-analogous text generation systems*" (p.21).

Rather than striving to enhance the accuracy of more complex analysis through LLMs, it is important to consider whether a combined LLM-manual analysis protocol may be more efficient in terms of time and resources. Using LLMs to prepare datasets and run an initial analysis (e.g., semantic search, extraction, and contextual categorisation) which is then reviewed and scored manually (see *Figure 23*) could be far more efficient (and accurate) than putting effort into finding ways to enable LLMs to mimic this to a satisfactory level of accuracy.

*Figure 23: Analytical protocol for a 3-stage approach to contextual analysis involving the extraction of relevant content and initial contextual analysis using an LLM (in aqua) followed by a manual analysis of findings (in blue)*




Errors happen within manual analysis, as in NLP/LLM approaches, but an analytical protocol like this potentially draws on the strengths of NLP, LLM, and manual approaches efficiently, minimising the potential for inaccuracies caused by functional limitations.

### 4.1.4 Building in validation from the outset

In previous work related to NLP/LLM analysis of textbooks, for example, validation processes have tended to focus on ad hoc manual checking of particular stages in the process and a separate analysis of a sample of the full dataset. In most cases, this means a sample of the books being analysed and reviewed in full by experts using a framework developed from the analytical approach. While this can provide a measure of the comparative accuracy of the dataset, it does not provide a way to monitor the analysis in progress to address any issues in real time or identify any drift in the analytical framing over time.

Other options include:

- Designing prompts and reporting protocols that ensure that all decisions/ categorisations are evidenced (e.g., direct quotes and/or page references are recorded alongside 'findings') and 'explained' by the LLM during the analysis process. This would allow for regular checks to be applied to check the accuracy of the data with minimal effort required to cross-check against the documents themselves.
- Developing (or selecting) a 'standardisation text' that is run regularly between other documents throughout the analysis process so that any changes in the results can be used to identify any drift or change in the analytical framing over time.
- Sampling full texts at random for full manual analysis during the analysis process

Validation processes, from initial model testing through to rolling validation of the analysis process, require particular characteristics to be built into analytical protocols, prompts, and reporting templates from the outset.

## 4.2 Preparing datasets

### 4.2.1 Not every dataset speaks the same language

NLP models and LLMs function by applying rules and patterns learned through initial training and later additional guidance given through 'fine-tuning' training materials or prompts (see *Figures 24 and 25,* which demonstrate the range of training or instructional materials required for 'fine-tuning' and prompting-based approaches).

*Figure 24: Overview of the workflow and key training materials required for NLP-based 'fine-tuning' approaches*

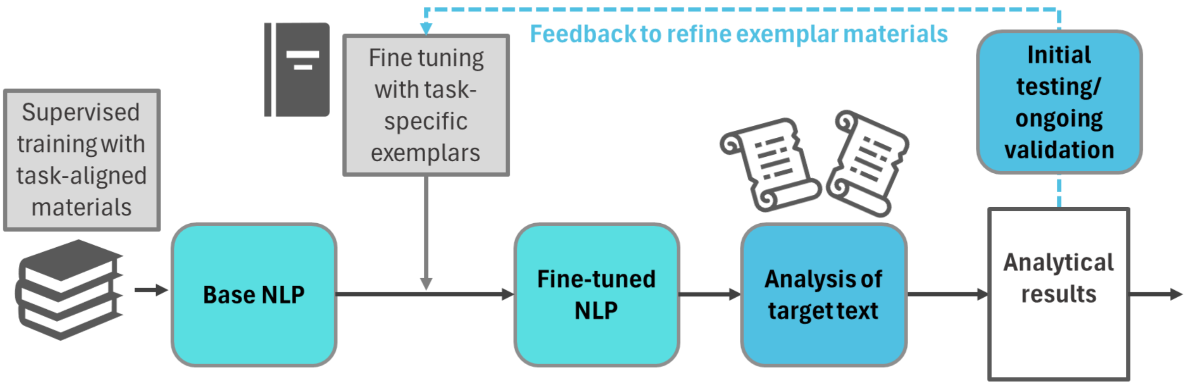
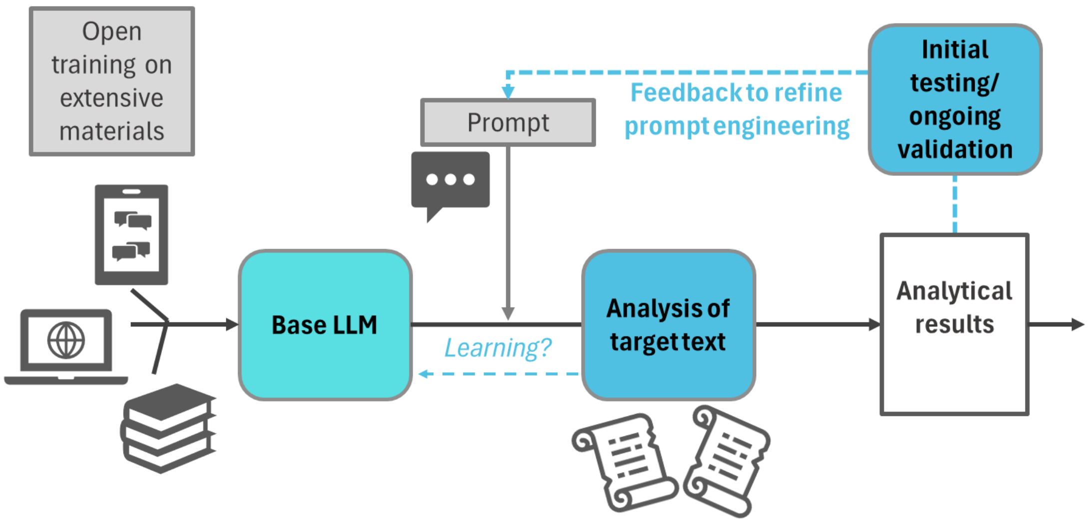



*Figure 25: Overview of the workflow and key training materials required for LLM prompt-based approaches*




The nature and required form of these materials — and the extent to which existing datasets can be used — depends on the nature of the analysis and the type of model selected. Alongside these forms of training and instruction materials, many NLP/LLM processes use thematic lexicons as a tool in the analysis process.

The amount of work required to identify, adapt, or develop these sorts of materials is a key consideration when selecting a model and analytical approach as these require delicate design — particularly when related to complex concepts like gender representation.

**Box 5: Example of LLM response to zero-shot prompt – quality of training materials**
**PROMPT:** Develop a witty quote to explain the importance of developing high quality, targeted training materials for NLP models and LLMs
**ChatGPT response:** *""An NLP is only as eloquent as its dataset, so don't expect Shakespeare from a diet of tweets"*

In general, key considerations include:

- Is there an NLP/LLM that has been trained on a dataset similar or relevant to the types of documents or themes that will be analysed? Or will LLMs that have undergone unsupervised learning on extensive volumes of text be able to interpret the context of the analysis without additional guidance?
- How different is the nature of the analytical task compared with what the most relevant NLP models/LLMs have been prepared for?
- What are the key differences required, and how can these be best introduced (e.g., additional training materials based on particular document types, targeted tasks, or expanded vocabularies; or prompts with relevant exemplars)
- What forms of pre-processing are required for the documents that will be analysed?
- How can the prepared model be effectively evaluated and refined (as required)?

### 4.2.2 Training materials

There are a range of existing datasets relating to different elements of gender and gender representation which can be used (or adapted) in 'fine-tuning' training or for the basis of exemplars in prompts (e.g., the multi-dimensional gender bias classification dataset, Dinan et al., 2020; RedditBias, Barikeri et al., 2021; GECOBench, Wilming et al., 2024; GenderAlign, Zhang et al., 2024). In general, datasets for 'fine-tuning' training take the form of:

- Labelled or annotated texts (e.g., GenderAlign) where responses are predicted and scored against the annotated answers
- Curated collections of statements for predictive question-based approaches (e.g., RedditBias) including alternate gender-paired statements, masked statements, or true-false paired statements
- Domain-specific content (e.g., MIMIC-III, Medical Information Mart for Intensive Care, Johnson et al., 2016) designed to increase the domain-specific vocabulary and range of the model

Alternatively, for specialised tasks, fine-tuning training datasets (or prompt exemplars) can be developed from the body of text or documents that will be analysed. However, the process can be time intensive (e.g., annotating policy documents with a set of gender-related themes, extracting key phrases, and creating masked or paired versions) and requires extensive testing and evaluation.

### 4.2.3 Building, adapting, or adopting lexicons

There are a wide range of lexicons available that have been developed for various purposes, including applying NLP models/LLMs to carry out gender analysis of texts and the detection of gender bias in NLP models/LLMs themselves (e.g., Garg et al., 2018; Pair et al, 2021, Doughman and Khreich, 2022). Lexicons are commonly used in gender analysis approaches that compare the co-occurrence, association, or semantic similarity of terms from a lexicon of gender-related and gendered words with lexicons of domain-specific words (e.g., law, energy sector, finance).

While the majority of the available research uses or adapts pre-existing gender lexicons, it is more common for domain-related lexicons to be developed specifically for the task. Approaches to developing domain-specific lexicons include expert workshops, crowdsourcing, or the generation of word clouds (e.g., from samples of target policy documents or media content related to particular topics or related reports/papers).

Alternatively, a multi-step NLP/LLM process can be designed in which NLP techniques (e.g., topic mapping) are used to analyse the target text to develop a context specific vocabulary list. This can then be applied back to the documents using different techniques (e.g., co-occurrence analysis) to identify patterns in use of the terms or word associations. Takizawa (2023) carried out a full topic mapping of the curriculum documents and then used the data generated to create a model where all mentions related to gender identity were resolved under one 'topic'. This list was then manually refined, before being mapped back across the curriculum documents related to four different years. This approach generated a rich and context-specific range of vocabulary which then allowed detailed analysis of how coverage in the curriculum had changed over time.

### 4.2.4 Preparing curriculum and policy documents for analysis

The nature of the analytical process selected controls the extent to which the curriculum and policy documents would need to undergo text pre-processed before analysis. While there are platforms and tools available to facilitate this, texts may need to go through multiple stages of processing. Forms of pre-processing for traditional NLP analysis can include:

- Text extraction from existing documents — for example, the Kenya curriculum documents are available online only in locked pdfs
- Text sorting or templating — Given the nature of curriculum (and policy) documents, and the range of content types they contain, it may also be necessary to manually sort the contents into standardised templates (e.g., separating out content related to vision and aims, content related to the development process, content related to the overarching curriculum model, subject curriculum content) before text pre-processing occurs.
- Text cleaning, including removing unnecessary characters (e.g., numbers, punctuation, special characters), application of lowercase throughout, removal of stop words (common words like 'the' which don't convey meaning for the particular task), removal of extra spaces
- Tokenisation to break the text down into smaller units (e.g., words or sentences, depending on the nature of the analysis). Reducing words to their root form (stemming) or to their 'dictionary form' (lemmatization) or expansion of contractions may be required for some analytical techniques
- Text vectorisation which converts text into numerical representations (e.g., word embeddings, weighting by frequency)

In contrast, some LLMs can carry out approximations of more technical analysis approaches using uploaded PDF documents or Word documents.

## 4.3 The prompting paradigm for curriculum and policy analysis

### 4.3.1 Forms of prompts

A variety of prompting approaches have been deployed in the use of LLMs (see *Figure 26* for examples), including:

- In-context approaches where examples of task inputs and outputs (called 'shots') are provided, which enables the LLM to learn what is expected 'within the context of the task' (e.g., Radford et al., 2019; Zhao et al., 2021; Lui et al., 2023; Ratovicius et al., 2024). 'Few shot' prompts are named in relation to the number of examples provided (e.g., 'zero-shot' prompts contain no examples; 'two-shot' prompts include two examples). In general, the more 'shots' provided, the better the LLM performs, particularly with complex tasks, with a plateauing of this effect beyond five shots (Lui et al. 2023)
- Chain-of-thought processes where prompts include an example with three components — an input, a chain of thought (which describes intermediate reasoning steps), and an output (e.g., Wei et al., 2022).
- Context-supplying approaches that provide guidance within the prompt to establish the role the LLM is intended to play (e.g., functioning as a gender specialist or a legal specialist), the intended audience, and/or information about the style of the response required.
- Retrieval-Augmented Generation (RAG) prompts, which set out a series of stages starting with retrieval of information from external sources before it is passed to the LLM to carry out an analysis (e.g., prompts that automate the process set out in *Figure 23*). This can assist in reducing generalisations or hallucinations, as accurate and up-to-date information can be retrieved for the analysis rather than relying on an LLMs prior training.

*Figure 26: Examples of different styles of prompts using gender analysis framings*

| |
|---|
| **Zero-shot prompt:** Review the extracted statements from a curriculum document for evidence of gender bias. Categorize each statement as 'biased' or 'not biased' based on the presence of gender stereotypes, unequal representation, or exclusionary language. |
| **Two-shot prompt:** Review the extracted statements from a curriculum document for evidence of gender bias. Categorize each statement as 'biased' or 'not biased' based on the presence of gender stereotypes, unequal representation, or exclusionary language.<br>**Example 1:** 'Girls tend to excel in arts, while boys perform better in science.' → Biased.<br>**Example 2:** 'Students are encouraged to explore all subjects equally.' → Not biased. |
| **Chain-of-Thought prompt:** Review the following statements extracted from a curriculum document for evidence of gender bias. In each case, first check if there is an imbalance in the number of male and female entities referenced. Then check if the statement states or implies any apparent differences based on gender. Then check if the statement references any known gender stereotypes. If a statement demonstrates an imbalance in the referencing or status of each gender, or if it includes stereotypes, categorise it as 'biased'. If there is no evidence of either, categorise it as 'not biased.'<br>**Example1:** 'Girls tend to excel in arts, while boys perform better in science.'<br>**Chain of thought:** This statement mentions an equal number of male and female entities but describes apparent differences between girls and boys. It makes a broad generalization about abilities based on gender. These generalisations reinforce the known stereotype that boys and girls are suited to different subjects. The statement therefore demonstrates gender bias. **Category:** Biased.<br>**Example 2:** 'Students are encouraged to explore all subjects equally.'<br>**Chain of thought:** This statement refers to students in general without differentiating on the basis of gender. The statement doesn't reinforce any known gender stereotypes. This statement does not include any evidence of gender bias **Category:** Not biased.<br>Please categorise the following statements in line with the described process and categorise them as 'biased' or 'not biased' |

**Context-supplying prompt:** You are acting as a gender analyst reviewing statements from a curriculum document for evidence of gender bias. Gender bias includes stereotypes, unequal representation, or language that reinforces traditional gender roles. Please review the following statements for evidence of gender bias. For each statement, please categorize it as 'biased' or 'not biased' in relation to gender.

The appropriate choice of prompt structure depends largely on the complexity of the task in terms of the number of steps required and the level of challenge involved. Developing effective prompts can require multiple rounds of testing and refinement. Increasingly, calibration approaches are being incorporated to enhance the evaluation of prompts and the extent to which they are complied with. This involves measuring a model's ability to predict which questions it can answer correctly (e.g., Kadavath et al., 2022).

### 4.3.2 Iterative development of potential prompts

Using the structure provided by the draft Framework for Curriculum Text Analysis, a number of exploratory prompts were developed, tested, and iteratively improved through the use of ChatGPT and samples of curriculum text. These were prepared and tested in dialogue form using the ChatGPT interface. ChatGPT was selected for this task as it is one of the most advanced LLMs currently available through a public interface. However, the publicly available version has a number of limitations. For example, ChatGPT has strict limits on the allocation of its analytical and data-sharing capacity in relation to individual accounts. This means that in some cases, while it states that it is able to run a technical analysis of a text extract, it is unable to share the resulting data. In these situations, the model tends to resort to generalisation and can, with no indication, create sets of results based on what it predicts the document may contain (based on patterns observed in initial training).

This severely limits the use of the model to fully test prompts on larger samples of curriculum text. However, if a challenge is applied to the incorrect results presented, the model is effective at refining the prompts presented to limit the potential for these issues to occur again. Iterative refinement of prompts can therefore be achieved by running a particular prompt, challenging the errors identified in the response, and then tasking ChatGPT to refine the initial prompt to ensure that the same errors, generalisation, or hallucinations are not introduced when the prompt is run again

**Box 6: Example of LLM response to zero-shot prompt – the prompt refinement process**

**PROMPT:** Provide a witty response to this quote: "*It is becoming clear that ChatGPT is less a magical wish-granting machine than an interpretive sparring partner*" (Bogost, 2023)

*Bogost, I., (2023). Generative Art Is Stupid. And that's how it should be. The Atlantic, January 13th, 2023.*

**LLM response:** "*An interpretive sparring partner, indeed—perhaps the true magic lies not in granting wishes, but in refining them through thoughtful dialogue, shaping raw ideas into something sharper and more purposeful*"

For each form of prompt developed, several rounds of improvement were required before ChatGPT was able to refine the prompts to the point where, for some approaches, responses demonstrated general alignment with manual analysis findings. For other techniques, the results do not reliably align with manual analysis, and results can vary even when the prompt and text remain identical. For the more traditional NLP-style analysis protocols, the curriculum samples used had to be very short as the individual account restrictions on ChatGPT do not allow for testing particular approaches on longer and more challenging texts. However, for more contextual LLM-style analysis, longer documents in a variety of forms could be analysed.

The prompting structure initially drafted combined a range of elements related to the different techniques described in **Section 4.3.1**. Despite this variety, a number of common-sense refinements were identified during this cyclic prompt-refining process that improved the data quality. These refinement approaches align well with suggestions in the broader literature for analysis of more complex texts. These include:

- Providing a clearly defined objective to set the context for the analysis, including a description of the role the LLM is fulfilling (e.g., '*you are acting as a gender analyst…*')
- Developing step-by-step instructions in a similar way in which manual analysis protocols would be described for manual reviewers. This can include using a flow chart to break down complex tasks or breaking the request down into a series of smaller tasks. Individual prompts can then be developed for each sub-stage and linked in a logical sequence.
- Always including the instruction '*Go step by step'* within the prompt
- Repeatedly, at various points in the prompt, emphasising the need to respond based on the content of the target document alone and to avoid using generalisations (e.g., '*Based only on the contents of the target text…*'; '*Ensure the answers relate directly to the content of the text and do not include generalisations*'; '*Do not invent information*')
- Requesting within the prompt that ChatGPT explains how each result was achieved and extracts/quotes the direct evidence from the target text that this was based on (e.g., *Quote the specific phrases from the text that informed each judgement")*. Keeping this information in the form of direct document quotes alongside the analytical outcomes also makes ongoing checking and validation easier.
- Adding a step in each sequence of tasks at the end of each analysis, prompting ChatGPT to check that the extracted examples are real and are drawn directly from the document
- Being specific about the format required for the output (e.g., '*Provide the results in a three-column table where the first column contains the relevant gendered word, the second column contains the number of times the term is used in the document, and the third column provides direct quotes of each of these occurrences with the surrounding text*')

### 4.3.3 Overview of initial testing pf potential prompts

As described in Section 4.3.2, sample prompts were created in order to explore broader elements of gender representation within curricula as set out in the draft Gender Framework for Curriculum Text Analysis. Different forms of indicative prompts were developed for each grouping within the framework to test ChatGPT's ability to run or approximate more traditional approaches NLP tasks, as well as to provide more contextually in-depth analysis. A variety of potential scoring approaches (e.g., judgement-based rubrics, numerical categories) were also included within the different prompts as an exploration of how more comparative data could be generated.

These explorative prompts were tested on a range of curriculum materials (e.g., from a short text extract to a full PDF) depending on the limitations of ChatGPT in particular analytical approaches. Most of the curriculum samples were taken from Ghanaian curriculum documents as the PDF documents available online were not locked. The approaches taken are set out in the table below to give indicative examples of the types of data that can be generated and the limitations presented through the use of ChatGPT. These dialogue-form prompts were tested at this stage to inform thinking rather than as a refined tool. Application of prompting approaches based on these examples would require rigorous testing on full scale text samples — as is the case for many of the stages involved in developing an analytical approach using LLMs.

*Figure 27: Overview of initial prompt testing in relation to the different components of the draft Gender Framework for Curriculum Text Analysis (see Appendix 4)*

| Area of draft GFCA | Form of prompt tested | *Notes on use of ChatGPT or other models* |
|---|---|---|
| ***1. Curriculum development process*** | | |
| **Gender balance of curriculum developers** | Prompt with context and step by step instructions – ChatGPT mimics entity extraction and categorisation | *Traditional NLP approaches would require a model pre-trained on entity recognition and then a database of names and associated genders* |
| **Use of gender audits during development** | A two-stage analysis: i) the creation of a phrase list associated with 'gender audits' ii) the use of a two-shot prompt with contextual information and clarifications | *Fine-tuning training would be required with the phrase list and an extensive list of examples would be required to enable a similar level of differentiation between references to audits and actual use of an audit during development as demonstrated by ChatGPT* |
| ***2. Curriculum aims, goals or vision*** | | |
| **Traditional gender roles** | Zero shot question-answer style prompt | *Broad identification of references to traditional gender roles possible but would require a rubric to score it relative to other contexts.* |
| ***3. Use of language*** | | |
| **Balance of gendered terms** | A two-stage analysis: i) zero-shot word counting using gender word lexicons ii) a step-by step categorisation task using a provided rubric to categorise the document as a whole | *This is challenging for ChatGPT to complete even for small samples of extracted text. Traditional NLP techniques would be applicable or less restricted LLMs.* |
| **Presence of sexist or demeaning language** | A three-shot prompt with detailed, step by step instructions | *This could also be explored using word embeddings and the co-occurrence of gendered words and words identified as sexist or derogatory. This would require two well developed lexicons (but examples exist which could be adapted e.g., the HurtLex lexicon) and would give an idea of broad trends.* |
| ***4. Gender bias (in mentions)*** | | |
| **Balance of male and female gendered terms** | Step by step instructions - entity recognition, key word counts and then classification against a rubric | *This is best done on just the curriculum content to avoid double counting curriculum developers which are covered in a different analysis* |
| ***5. Gender dynamics*** | | |
| **Relationship dynamics between named entities** | *Multi-step process including entity recognition, extraction, conceptual categorisation based on diverse elements* | *This is a complex, multi-step prompt but demonstrates the ability of ChatGPT to apply broad contextual understanding to interpret the nature of relationships. The proportional scoring also demonstrates a different approach.* |

| | | |
|---|---|---|
| ***6. and 7. Gender stereotypes — personal characteristics; roles and achievement*** | | |
| **Personal characteristics** | Zero shot prompt with contextual additions and a confidence level scoring of categorisation | *Chat GTP is limited in the ability to carry this out reliable due to restrictions on data processing* |
| ***8. Recognised figures and gendered perspectives*** | | |
| **Historical figures** | Train of thought prompt with multiple stages based on entity recognition and categorisation | *Extraction from contextual surroundings is reliable but names listed without context would be unreliably extracted* |
| ***9. Access, opportunities and relevance*** | | |
| **Content relevant to interests** | Train of thought prompt in two stages – defining stereotypical interests and then semantic search based on these categories | *This is challenging to frame as it relies on stereotypical notions of the interests of girls and boys.*<br>*ChatGPT is very broad in its application of potential interests* |
| ***10-12. CURRICULUM CONTENT — sex education, sexuality and gender equality***<br>***13-14. CURRICULUM SKILLS — social skills for gender equality, recognising and analysing inequalities*** | | |
| **Content related to gender studies and gender equality** | Two stage process: i) Creation of a lexicon by extracting the 30 most common topic words from an HE gender studies course description ii) Content search using keywords and a topic definition and scoring using a created scale | ***ChatGPT is unable to apply this prompt consistently***<br>*One example is provided here for one subject area – scoring could use a scale as suggested in the prompt or be based on the number of topics/skill areas substantively present based on the framework list.* |

# 5. Conclusion

This paper has summarised some of the potentials and pitfalls associated with using NLP and LLMs to analyse curriculum and education policy documents, by reviewing how meanings are represented and constructed in relation to ideas concerning gender in and through education. A variety of trade-offs are presented with regard to the ways in which researchers make use of the potential of NLP models, LLMs, and reviews by human researchers, and how these may be combined. A number of ways forward emerge for drawing on a theoretical framework based around a gender continuum, or the AGEE framework, to help develop approaches to scanning documents, and using AI scans linked to wider investigations about the presentation of ideas associated with gender. The discussion has stressed the need for consultation, review, transparency, and ethically informed research at all stages.

Carrying out document analysis using NLP or LLM tools can bring significant efficiencies in the processing of extensive documentation, and there is a range of pre-trained models and training materials available related to gender representation that could be applied to the analysis of curriculum and policy documents.

Currently, there are very few examples of literature describing the use of AI to analyse gender representation within curriculum and policy documents. The examples that do exist tend to focus on one or two of the more traditional NLP analytical techniques. There is potential for work based on a richer representation of the different ways in which gender representation can be expressed within text (e.g., the proposed draft Gender Framework for Curriculum Text Analysis and the potential adaptation of the AGEE framework to provide a similar framework for policy analysis), and a wider range of analytical approaches has the potential to significantly contribute to knowledge generation within this research area.

However, whilst there are many potential benefits to the application of NLP/LLM techniques in the analysis of curriculum and policy documents, careful consideration needs to be applied to:

- Addressing the ethical challenges associated with the application of NLP/LLM tools
- Establishing the optimal balance in the use of manual analysis and NLP/LLM analysis within the research protocol
- Addressing inherent biases within NLP/LLM models, including gender bias and the potential overwriting of cultural meaning and contexts through the analysis of content in different languages through operational structures framed within a base language
- Ensuring that the analytical framing is informed by rich contextual understanding of how gender representation manifests within society and the nuances of language in the specific contexts in which different curriculum and policy documents were developed

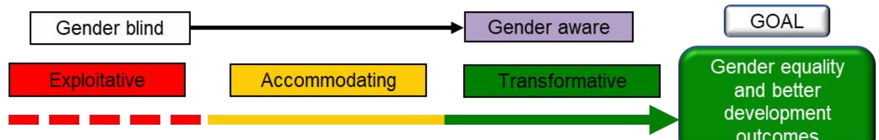
Working with LLMs brings potential challenges but also has the potential to prompt new ways of thinking about curriculum and policy document analysis. The nature of LLM analysis encourages a deep consideration of exactly what is being tested and how decisions around language use in prompts, or training materials, or lexicons influences the analysis outcomes.

This process promotes collaborative working in its truest form in a multitude of ways. For example, the selection of analytical framings and the development of tools are only enriched and improved by the incorporation of wide-ranging perspectives and language related to gender and gender representation. In addition, the iterative refinement of analytical processes and prompts, and the development of analytical protocols that build on the strengths of NLP, LLM, and human abilities, could be argued to represent human-AI collaboration in its most productive form.

As the technology rapidly develops, work of this kind is by its very nature explorative and offers an opportunity to harness the emerging capabilities of LLMs to generate outcomes that can meaningfully inform future thinking about gender representation within curricula and policy on national and international scales.

## Appendix 1: Original visual representations of the different gender approach/equality continuums proposed by a range of authors over time as summarised in *Figure 9* (references within)


**Policy Review**– Kabeer, 1994, p.82

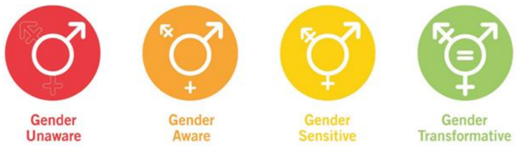



**Gender Equality Continuum Tool** – IGWG, 2013, p.2

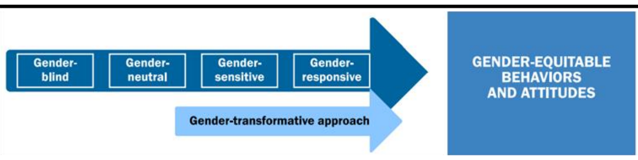

**Gender strategy scanning template** – Miruka et al., 2013, p.29


**Gender Results Effectiveness Scale** – UNDP, 2015, p.4

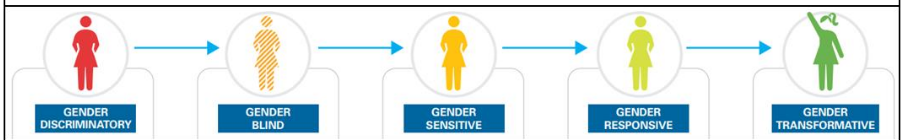

**Continuum of Approaches to Action on Gender and Health** – Dwyer and Woolf, 2018, p.32




**Continuum from gender unaware to gender transformative practice** – Brown et al., 2019, p.9




**Continuum of pedagogical approaches to gender** – Subramanian, 2019, p.13




**Gender Equity Continuum** – UNICEF, 2019, p.13


**Gender Equity Continuum** – UNFPA and UNICEF, 2020, page 1


**Gender Transformative Education continuum** – UNGEI, 2024

From: Adapted from Weavers, 2022

## Appendix 2: Definition components (individual approaches/strategies) identified in definitions of GAT-related curriculum approaches (e.g., of a gender-sensitive curriculum) found in papers published between 1995 and 2022 (Weavers, 2022)

These definition components represent the way in which five different GAT curriculum approaches (gender-neutral, gender-aware, gender-sensitive, gender-responsive and gender-transformative) were described in the relevant papers. While the definition components have been shortened to under 100 characters for analytical purposes, the language used is representative of the language found in papers published between 1995 and 2022. The definition components do not always, therefore, reflect current preferences in language use in relation to gender and gender equality. They do, however, provide a useful overview of the full range of practical strategies identified in the literature in relation to the development of curricula associated with particular GAT – this provides a useful tool in the process of developing a draft framework to be used to inform potential NLP and LLM analysis of these types of practical approaches within curriculum documents.

To create the framework, these definition components from the previous literature review (Weavers 2022) were mapped across a GAT continuum structure (see **Appendix 3**). Thorough gap-filling and rewriting was then carried out to develop the draft framework presented to inform potential NLP and LLM analysis (see **Appendix 4**).

**GROUP A: Development process**

| | |
|---|---|
| **C1** | The curriculum was developed by teams which included equal numbers of men and women |
| **C2** | The curriculum includes set texts/textbooks/curriculum materials developed by female authors |
| **C3** | The curriculum was gender audited during development and/or implementation |

**GROUP B: Representations of gender or gender relationships that are purposefully avoided in the curriculum**

| | |
|---|---|
| **C4** | The curriculum purposefully avoids reinforcing ideas of male superiority or patriarchal ideology |
| **C5** | The curriculum purposefully avoids drawing attention to gender or gender differences |
| **C6** | The curriculum avoids using any sexist language or language that demeans women |
| **C7** | The curriculum content purposefully avoids reinforcing gender bias and/or gender inequalities |
| **C8** | The curriculum content purposefully avoids reinforcing gender stereotypes |
| **C9** | The curriculum avoids reinforcing gender stereotypes about women's roles (e.g., homemaker, caregiver) |
| **C10** | The curriculum avoids reinforcing gender stereotypes about men's roles (e.g., leader, pilot, doctor) |
| **C11** | The curriculum avoids reinforcing stereotypes of the character of women (e.g., shy) or men (e.g., brave) |

### GROUP C: References to and/or representations of gender or gender relationships that are purposefully included in the curriculum

| | |
|---|---|
| **C12** | The curriculum emphasises the value to society of the domestic/ traditional roles performed by women |
| **C13** | The curriculum recognises the importance of women in society and/or development |
| **C14** | The curriculum content incorporates the voices/perspectives of women and girls |
| **C15** | The curriculum content includes representations of women in diverse roles (e.g., doctors, leaders) |
| **C16** | The curriculum content includes examples of women as heroes/leaders/experts (e.g., in history, maths) |
| **C17** | The curriculum content includes representations of men in diverse roles (e.g., childcare, cleaning) |
| **C18** | The curriculum includes varied examples of girls' personalities, some which challenge gender norms |
| **C19** | The curriculum includes varied examples of boys' personalities, some which challenge gender norms |
| **C20** | The curriculum examples include people with distinctive personalities independent of gender norms |
| **C21** | The curriculum includes equal numbers of references to male and females (e.g., historical figures) |
| **C22** | The curriculum only uses unisex/non-gendered language and/or examples (e.g., they, police officer) |
| **C23** | The curriculum reinforces traditional ideas of masculinity and femininity or gender power dynamics |

### GROUP D: Meeting the (different?) needs of different learners

| | |
|---|---|
| **C24** | The curriculum enables equal opportunities and equal attention/treatment of girls and boys |
| **C25** | The curriculum doesn't allow gender discrimination in access to subjects (e.g., D&T, home economics) |
| **C26** | The curriculum content and structure meet the (different) needs of both girls and boys |
| **C27** | The curriculum considers the different learning styles/ assessment approaches of girls and boys |
| **C28** | The curriculum content considers girls' experiences/interests and/or is relevant/engaging for them |
| **C29** | The curriculum content considers boys' experiences/interests and/or is relevant/engaging for them |

### GROUP E: General skills, attitudes and/or values

| | |
|---|---|
| **C30** | The curriculum requires learners to be taught Life Skills |
| **C31** | The curriculum requires learners to be taught conflict resolution skills |
| **C32** | The curriculum enables girls' empowerment/provides opportunities for them to develop agency/autonomy |
| **C33** | The curriculum supports girls to have high aspirations for their educational outcomes/life/career |
| **C34** | The curriculum encourages girls to aspire to traditional female roles within society (e.g., mother) |
| **C35** | The curriculum promotes democratic principles/provides space for girls to express their opinions |
| **C36** | The curriculum includes content that encourages girls to participate in politics |
| **C37** | The curriculum provides opportunities for all learners to develop their ability to be expressive |
| **C38** | The curriculum enables all learners to develop assertiveness, confidence and/or leadership skills |
| **C39** | The curriculum enables development of socio-emotional skills e.g., empathy/understanding other views |
| **C40** | The curriculum enables all learners to develop confidence/aptitude in all subjects (e.g., maths, D&T) |

### GROUP F: Specific content/topics to be taught

| | |
|---|---|
| **C41** | The curriculum requires learners to be taught about HIV and other sexually transmitted infections |
| **C42** | The curriculum requires learners to be taught sex education and about emotional development |
| **C43** | The curriculum requires learners to be taught about sexual consent and when consent can't be granted |
| **C44** | The curriculum requires learners to be taught about sexuality |
| **C45** | The curriculum requires learners to be taught about gay/lesbian/bisexual/other sexual orientations |

| | |
|---|---|
| **C46** | The curriculum requires learners to be taught about sexual abuse and gender-based violence |
| **C47** | The curriculum teaches learners that gender and masculinity/femininity are social constructs |
| **C48** | The curriculum requires learners to be taught examples of gender issues/imbalanced power relations |
| **C49** | The curriculum requires learners to be taught about human rights |
| **C50** | The curriculum requires learners to be taught about women's rights |
| **C51** | The curriculum requires learners to be taught about the importance of gender equality/gender equity |
| **C52** | The curriculum teaches learners that homophobia/ victimisation of people based on sexuality is wrong |

### GROUP G: Specific skills, values and/or attitudes related to gender/gender inequality

| | |
|---|---|
| **C53** | The curriculum enables learners to identify examples of gender inequality/become gender conscious |
| **C54** | The curriculum requires learners to be able to identify gender imbalances in power/ participation |
| **C55** | The curriculum requires learners to analyse how perceptions of gender shape attitudes and behaviours |
| **C56** | The curriculum challenges learners to change any pre-held sexist/stereotypical gender perceptions |
| **C57** | The curriculum provides opportunities for learners to develop equal respect for both genders |
| **C58** | The curriculum enables learners to identify that their abilities are not linked to their gender |
| **C59** | The curriculum helps learners to recognise men and women as equal/ develop equitable relationships |
| **C60** | The curriculum requires learners to critically analyse beliefs/norms that lead to gender inequality |
| **C61** | The curriculum empowers girls to challenge gender norms and/or claim their rights |
| **C62** | The curriculum empowers boys to take action to address gender inequality and/or empower girls |
| **C63** | The curriculum empowers all learners to challenge gender inequality and/or seek gender justice |

## Appendix 3: Mapping of the definition components identified during a previous curriculum-focused GAT review (see Appendix 2) across a GAT continuum constructed from the most commonly used GAT in continuums published by a range of other authors

This mapping was carried out to inform the development of the draft Gender Framework for Curriculum Text Analysis which is set out in **Appendix 4.**

| | Gender blind | Gender sensitive | Gender responsive | Gender transformative |
|---|---|---|---|---|
| | The failure to recognize that the roles and responsibilities of men/boys and women/girls are given to them in specific social, cultural, economic and political contexts and backgrounds. The failure to....take into account these different roles and their diverse needs. This maintains the status quo and will not help transform the unequal structure of gender relations. (UNESCO, 2022, p.5) | Acknowledges that differences and inequalities between women and men require attention, but does not attempt to address them. (UNESCO 2022p.8) | Addressing and responding to the different needs, aspirations, capacities and contributions of women and men (UNESCO 2022, p.7) | Challenging existing and biased/ discriminatory policies, practices and programmes and affecting change for the betterment of life for all (UNESCO 2022, p.8) |
| **1. Curriculum development process** | | | **C1** The curriculum was developed by teams which included equal numbers of men and women<br>**C3** The curriculum was gender audited during development and/or implementation | |
| **2. Curriculum aims, goals or vision** | **C34** The curriculum encourages girls to aspire to traditional female roles within society (e.g., mother) | | | **C32** The curriculum enables girls' empowerment/provides opportunities for them to develop agency/autonomy<br>**C33** The curriculum supports girls to have high aspirations for their educational outcomes/life/career<br>**C61** The curriculum empowers girls to challenge gender norms and/or claim their rights<br>**C62** The curriculum empowers boys to take action to address gender inequality and/or empower girls |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | C63 | The curriculum empowers all learners to challenge gender inequality and/or seek gender justice |
| | | | | C57 | The curriculum provides opportunities for learners to develop equal respect for both genders | C58 | The curriculum enables learners to identify that their abilities are not linked to their gender |
| | | | | C59 | The curriculum helps learners to recognise men and women as equal/ develop equitable relationships | | |
| **3. Use of language** | C22 | The curriculum only uses unisex/non-gendered language and/or examples (e.g., they, police officer) | | | | | |
| | | | | C6 | The curriculum avoids using any sexist language or language that demeans women | | |
| **4. Gender bias (in mentions)** | C5 | The curriculum purposefully avoids drawing attention to gender or gender differences | | C7 | The curriculum content purposefully avoids reinforcing gender bias and/or gender inequalities | | |
| **5. Gender dynamics** | C23 | The curriculum reinforces traditional ideas of masculinity and femininity or gender power dynamics | | C4 | The curriculum purposefully avoids reinforcing ideas of male superiority or patriarchal ideology | | |
| **6. Gender stereotypes — personal characteristics** | | | | C8 | The curriculum content purposefully avoids reinforcing gender stereotypes | | |
| | | | | C11 | The curriculum avoids reinforcing stereotypes of the character of women (e.g., shy) or men (e.g., brave) | C18 | The curriculum includes varied examples of girls' personalities, some which challenge gender norms |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | C19 | The curriculum includes varied examples of boys' personalities, some which challenge gender norms |
| | | | | | | | C20 | The curriculum examples include people with distinctive personalities independent of gender norms |
| **7. Gender stereotypes — roles and achievement** | C12 | The curriculum emphasises the value to society of the domestic/ traditional roles performed by women | C13 | The curriculum recognises the importance of women in society and/or development | C9 | The curriculum avoids reinforcing gender stereotypes about women's roles (e.g., homemaker, caregiver) | C15 | The curriculum content includes representations of women in diverse roles (e.g., doctors, leaders) |
| | | | | | C10 | The curriculum avoids reinforcing gender stereotypes about men's roles (e.g., leader, pilot, doctor) | C17 | The curriculum content includes representations of men in diverse roles (e.g., childcare, cleaning) |
| **8. Recognised figures and gendered perspectives** | | | | | C21 | The curriculum includes equal numbers of references to male and females (e.g., historical figures) | C16 | The curriculum content includes examples of women as heroes/leaders/experts (e.g., in history, maths) |
| | | | | | C2 | The curriculum includes set texts/textbooks/curriculum materials developed by female authors | | |
| | | | | | C35 | The curriculum promotes democratic principles/provides space for girls to express their opinions | C14 | The curriculum content incorporates the voices/ perspectives of women and girls |
| **9. Access, opportunities and relevance** | | | C25 | The curriculum doesn't allow gender discrimination in access to subjects (e.g., D&T, home economics) | C24 | The curriculum enables equal opportunities for girls and equal attention/treatment of girls and boys | | |
| | | | C40 | The curriculum enables all learners to develop confidence/aptitude in all subjects (e.g., maths, D&T) | C26 | The curriculum content and structure meet the (different) needs of both girls and boys | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | C27 | The curriculum considers the different learning styles/ assessment approaches of girls and boys | | |
| | | | | | C28 | The curriculum content considers girls' experiences/interests and/or is relevant/engaging for them | | |
| | | | | | C29 | The curriculum content considers boys' experiences/interests and/or is relevant/engaging for them | | |
| **10. CURRICULUM CONTENT — sex education** | | | C41 | The curriculum requires learners to be taught about HIV/AIDs and other sexually transmitted diseases | | | | |
| | C31 | The curriculum requires learners to be taught conflict resolution skills | C42 | The curriculum requires learners to be taught sex education and about emotional development | C43 | The curriculum requires learners to be taught about sexual consent and when consent can't be granted | | |
| | | | | | C46 | The curriculum requires learners to be taught about sexual abuse and gender-based violence | | |
| **11. CURRICULUM CONTENT — sexuality** | | | | | | | C44 | The curriculum requires learners to be taught about sexuality |
| | | | | | | | C45 | The curriculum requires learners to be taught about gay/lesbian/bisexual/other sexual orientations |
| | | | | | | | C52 | The curriculum teaches learners that homophobia/ victimisation of people based on sexuality is wrong |
| **12. CURRICULUM** | | | | | | | C47 | The curriculum teaches learners that gender and masculinity/ femininity are social constructs |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **CONTENT — gender equality** | | | | | | C48 | The curriculum requires learners to be taught examples of gender issues/imbalanced power relations |
| | | C49 | The curriculum requires learners to be taught about human rights | C50 | The curriculum requires learners to be taught about women's rights | C51 | The curriculum requires learners to be taught about the importance of gender equality/gender equity |
| **13. CURRICULUM SKILLS — Social/society skills** | | | | C37 | The curriculum provides opportunities for all learners to develop their ability to be expressive | C36 | The curriculum includes content that encourages girls to participate in politics |
| | | C30 | The curriculum requires learners to be taught Life Skills | C38 | The curriculum enables all learners to develop assertiveness, confidence and/or leadership skills | | |
| | | | | C39 | The curriculum enables development of socio-emotional skills e.g., empathy/understanding other views | | |
| **14. CURRICULUM SKILLS — Recognising and analysing inequalities** | | | | | | C53 | The curriculum enables learners to identify examples of gender inequality/become gender conscious |
| | | | | | | C54 | The curriculum requires learners to be able to identify gender imbalances in power/ participation |
| | | | | | | C55 | The curriculum requires learners to analyse how perceptions of gender shape attitudes and behaviours |
| | | | | | | C60 | The curriculum requires learners to critically analyse beliefs/norms that lead to gender inequality |
| | | | | | | C56 | The curriculum challenges learners to change any pre-held |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | sexist/ stereotypical gender perceptions |

# Appendix 4: Initial draft of a potential Gender Framework for Curriculum Text Analysis

The draft framework presented here was developed by:

1. Mapping previously identified GAT definition components (in **Appendix 2**) across an overarching GAT continuum (presented in **Appendix 3**)
2. Gap-filling and refining the language to create a more detailed framework identifying **characteristics** of curriculum documents that could be identified through text analysis techniques (potentially using NLP/LLM techniques)

| | **Mapping of characteristics across an overarching GAT continuum** | | | **Likely data availability** |
|---|---|---|---|---|
| | **Negative**<br>Gender-blind<br>Gender-neutral | **Accommodating/Intentional**<br>Gender-sensitive<br>Gender responsive | **Transformative**<br>Gender-transformative | |
| **1. Curriculum development process** | Imbalance in gender those involved in curriculum development | Gender balance/participation in curriculum development | | May not be mentioned in all documents |
| | No gender audits applied during development | | Use of gender audits during development | |
| **2. Curriculum aims, goals or vision** | The curriculum is intended to promote traditional values and gender roles | The curriculum is intended to promote gender equality and equity | The curriculum encourages all learners to challenge gender inequality and/or seek gender justice | *May not be mentioned in all documents* |
| | | Intends to encourage learners to respect men and women as equal/ develop equitable relationships | Intends to support learners to identify that their abilities are not linked to their gender | |
| | The curriculum frames learners' potential in relation to gender and/or traditional roles | The curriculum promotes *agency and* high aspirations for all learners | | |
| **3. Use of language** | Imbalanced/ inappropriate use of gendered language | No gendered terminology or balanced gender terminology | | |
| | Presence of sexist or demeaning language | No sexist or demeaning language | | |
| **4. Gender bias (in mentions)** | Unequal references to male and female individuals | Equal references to male and female individuals | | |
| **5. Gender dynamics** | Promotion of traditional ideas about gender/power dynamics | No presence of traditional ideas about gender/power dynamics | Content that challenges/ counteracts traditional gender/power dynamics | |
| **6. Gender stereotypes — personal characteristics** | Content reinforces or promotes traditional stereotypes about the characteristics of people of different genders | No references to traditional stereotypes about the characteristics of people of different genders | Content includes examples that counter/ challenge traditional stereotypes about the characteristics of people of different genders | |
| **7. Gender stereotypes — roles and achievement** | Content reinforces or promotes traditional stereotypes about gender roles/achievement | No references to traditional stereotypes about gender roles/achievement | Content includes examples that counter/ challenge traditional stereotypes about gender roles/achievement | |
| **8. Recognised figures and gendered perspectives** | Imbalanced gender representation of historical and contemporary named figures (e.g., experts, heroes, leaders) | Balanced gender representation of historical and contemporary named figures | Reference to historical and contemporary named figures that directly challenge gender stereotypes | *More likely in particular subjects* |

| | | | | |
|---|---|---|---|---|
| | Imbalanced gender ratio in authors, poets, artists and performers required to be studied | Balanced gender ratio in authors, poets, artists and performers required to be studied | Inclusion of authors, poets, artists and performers who address issues of gender equality | *More likely in particular subjects* |
| | The perspectives of one gender on key issues is represented more strongly in relation to key issues | The perspectives of both genders are represented in relation to key issues | *All learners are expected and supported to express their views on key issues* | More likely in curriculum materials |
| **9. Access, opportunities and relevance** | Some subjects or courses are restricted to one gender | All learners can select all subjects or courses within the curriculum | | |
| | An unequal proportion of the content is better suited to the learning or assessment style associated with one gender | The content is equally suited to the learning and assessment styles of both genders | | *In all curricula but difficult to identify* |
| | An unequal proportion of the content reflects the interests associated with one gender | The content equally reflects the interests associated with one gender | *Content that traditionally has been more associated with one gender is presented in a more engaging way for the other gender* | *Difficult to identify without drawing on stereotypes* |
| **10. CURRICULUM CONTENT — sex education** | | The curriculum requires learners to be taught about physical and emotional development, and sex | The curriculum requires learners to be taught about sexually transmitted infections including HIV | |
| | | | The curriculum requires learners to be taught about consent | |
| | | | The curriculum requires learners to be taught about gender-based violence | |
| **11. CURRICULUM CONTENT — sexuality** | *The content reinforces stereotypes related to different sexual orientations* | *No reference to stereotypes related to different sexual orientations* | The curriculum requires learners to be taught about sexuality and sexual orientations (e.g., gay, lesbian, bisexual) | |
| | *Uses derogatory language related to different sexual orientations* | *No demeaning language related to sexual orientation* | The curriculum requires learners to be taught that homophobia/sexuality-based abuse is wrong | |
| **12. CURRICULUM CONTENT — gender equality** | | | The curriculum requires learners to be taught that gender, masculinity/ femininity are social constructs that can change | |
| | | | The curriculum requires learners to be taught examples of gender issues/imbalanced power relations | |
| | The curriculum requires learners to be taught about human rights | The curriculum requires learners to be taught about women's rights | The curriculum requires learners to be taught about the importance of gender equality/gender equity | |
| **13. CURRICULUM SKILLS Social skills linked to** | Curriculum emphasises the development of skills for one gender more than the other | Curriculum content related to development of communication/ expression skills for all learners | *The curriculum specifically emphasises encouraging equal participation/ expression* | |

| | | | | |
|---|---|---|---|---|
| **improving gender relationships and equal participation** | | Curriculum content related to development of assertiveness or leadership skills | | |
| | | Content related to development of empathy/ understanding | Content related to development of empathy/ understanding specifically related to gendered experiences | |
| **14. CURRICULUM SKILLS: Recognising and analysing inequalities** | | | Content requires learners to identify examples of gender inequality/become gender conscious | |
| | | | Content requires learners to be able to identify gender imbalances in power/ participation | |
| | | | Content requires learners to analyse how perceptions of gender shape attitudes and behaviours | |
| | | | Content requires learners to critically analyse beliefs/norms that lead to gender inequality | |
| | | | Content requires learners to critically analyse their own pre-held sexist/ stereotypical gender perceptions | |

# 6. References

Abrell, J., Kosch, M., & Rausch, S. (2022). How effective is carbon pricing? A machine learning approach to policy evaluation. *Journal of Environmental Economics and Management*, *112*, Article 102589.

Adukia, A., Chiril, P., Christ, C., Das, A., Eble, A., Harrison, E., & Runesha, H. B. (2022). Tales and tropes: Gender roles from word embeddings in a century of children's books. *University of Chicago Becker Friedman Institute for Economics Working Paper No. 2022-128*.

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2023). What we teach about race and gender: Representation in images and text of children's. *The Quarterly Journal of Economics*, *138*(4), 2225–2285.

Aikman, S., Unterhalter, E., & Challender, C. (2005). The education MDGs: Achieving gender equality through curriculum and pedagogy change. *Gender & Development*, *13*(1), 44-55.

Ajith, S., Rithani, M., & SyamDev, R. S. (2023). Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach. *Seventh International Conference on Image Information Processing* (pp. 888-893). IEEE.

Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025, April). Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(27), 27802-27812.

Aribandi, V., Tay, Y., & Metzler, D. (2021). How reliable are model diagnostics? *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (pp. 1778–1785, Online).

Ash, E., Chen, D. L., & Ornaghi, A. (2024). Gender attitudes in the judiciary: Evidence from US circuit courts. *American Economic Journal: Applied Economics*, *16*(1), 314-350.

Awash, F. (1988). Present Contributions of Women to Various Sectors of Development. *Proceedings of National Workshop on Integrated Gender Responsive Development Planning,* Addis Ababa, Ethiopia.

Barbosa, B. K., & Campelo, C. E. (2024). LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis. *15th Symposium in Information and Human Language Technology* (pp.1-10).

Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021). RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*.

Bartl, M., & Leavy, S. (2024). From'Showgirls' to'Performers': Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs. *arXiv preprint arXiv:2407.04434*.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, Vol. 2253, pp. 1-6. CEUR-WS.

Bassis, L. M. (1979). The psychological test of sex and the transsexual: a legal response. *The Journal of Psychiatry & Law*, 7(2), pp.211-267.

Bishaw, T. (1988). Situation Analysis of Ethiopian Women. Proceedings of National Workshop on Integrated Gender Responsive Development Planning, Addis Ababa, Ethiopia.

Bobbitt, J. F. (1918). *The Curriculum*. Pfew York: Houghton Mifflin Company.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems,* 29.

Dohn, N.B., (2025). 'AI for learning — Why? Searching for pedagogical justifications for using AI for learning'. *Paper presented at ICER, Seoul*, July 2025.

Boutyline, A., Arseniev-Koehler, A., & Cornell, D. J. (2023). School, studying, and smarts: Gender stereotypes and education across 80 years of American print media, 1930–2009. *Social Forces*, 102(1), 263-286.

Breault, D. A., & Marshall, J. D. (2010). *Curriculum, definitions of. Encyclopaedia of curriculum studies*, 1, 179-181.

Brown, S., Budimir, M., Sneddon, A., Lau, D., Shakya, P., & Upadhyay, S. (2019) *Gender Transformative Early Warning Systems: Experiences from Nepal and Peru*, Rugby, UK: Practical Action

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

Carroll, P., Singh, B., & Mangina, E. (2024). Uncovering gender dimensions in energy policy using Natural Language Processing. *Renewable and Sustainable Energy Reviews*, 193, 114281.

Chacon, F. (2024). AI-based curriculum evaluation in Higher Education. In *ICERI2024 Proceedings*, pp. 484-492. IATED.

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural language engineering, 29*(3), pp. 509-553.

Chaloner, K., & Maldonado, A. (2019, August). Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 25-32).

Chintam, A., Beloch, R., Zuidema, W., Hanna, M., & Van Der Wal, O. (2023). Identifying and adapting transformer-components responsible for gender bias in an English language model. *arXiv preprint arXiv:2310.12611*.

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, 54(10s), 1-35.

Clemmons, A. W., Donovan, D. A., Theobald, E. J., & Crowe, A. J. (2022). Using the Intended–Enacted–Experienced Curriculum Model to Map the Vision and Change Core Competencies in Undergraduate Biology Programs and Courses. *CBE—Life Sciences Education,* 21 (1), pp.1-18.

Clover, D. E. (1995). Gender transformative learning and environmental action. *Gender and Education*, 7 (3), pp.243-258

Collett, C., & Dillon, S. (2019). *AI and gender: Four proposals for future research*. Leverhulme Centre for the Future of Intelligence Report, pp. 1-44

Crawfurd, L., Saintis-Miller, C., & Todd, R. (2024). Sexist textbooks: Automated analysis of gender bias in 1,255 books from 34 countries. *PloS one*, 19(10), e0310366.

Cunningham, S. C. (1963). Policy and practice. *Public Administration*, 41(3), 229-238.

Derner, E., de la Fuente, S. S., Gutiérrez, Y., Moreda, P., & Oliver, N. (2024). Leveraging large language models to measure gender representation bias in gendered language corpora. *arXiv preprint arXiv:2406.13677*.

Dewey, J. (1916). *Democracy and Education.* New York: The Macmillan Company

DfE (Department for Education, England) (2013). *The national curriculum in England Key stages 1 and 2 framework document*. London, pp.1-201

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Doughman, J., & Khreich, W. (2022). Gender bias in text: Labeled datasets and lexicons. *arXiv preprint arXiv:2201.08675*.

Dwivedi, S., Ghosh, S., & Dwivedi, S. (2023). Breaking the bias: Gender fairness in LLMs using prompt engineering and in-context learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(4).

Dwyer, E. & Woolf, L. (2018). *Down by the river*, pp.1-52. Victoria: Oxfam

Eisner, E.W. (1979). *The Educational Imagination: On the design and evaluation of school programs*. New York: Macmillan

Elshin, D., Karpachev, N., Gruzdev, B., Golovanov, I., Ivanov, G., Antonov, A., Skachkov, N., Latypova, E., Layner, V., Enikeeva, E., Popov, D. Chekashev, A., Negodin, V., Frantsuzova, V., Chernyshev, A., & Denisov, K. (2024). From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning. In *Proceedings of the Ninth Conference on Machine Translation* (pp. 247-252).

Elwood, J. (2016*). Gender and the Curriculum*. The sage handbook of curriculum, pedagogy and assessment, 247-262.

Fast, E., Vachovsky, T., & Bernstein, M. (2016). Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 10, No. 1, pp. 112-120).

Freire, P. (1972). Pedagogy of the Oppressed. UK: Bloomsbury.

Foulds, K. (2014). Buzzwords at play: gender, education, and political participation in Kenya. *Gender and Education*, 26 (6), pp.653-671.

Foulds, K. (2013). The continua of identities in postcolonial curricula: Kenyan students' perceptions of gender in school textbooks. *International Journal of Educational Development*, 33 (2), pp.165-174.

Friend, J., Power, J. M., & Yewlett, C. J. (2013). *Public planning: The inter-corporate dimension*. Routledge.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635-E3644.

Ghali, M. K., Farrag, A., Won, D., & Jin, Y. (2024). Enhancing Knowledge Retrieval with In-Context Learning and Semantic Search through Generative AI. *arXiv preprint arXiv:2406.09621*.

Giroux, H. (1983). Theories of reproduction and resistance in the new sociology of education: A critical analysis. *Harvard educational review*, 53 (3), pp.257-293.

Goetz, A. M. (1992). Gender and administration. *IDS Bulletin*, 23 (4), pp.6-17.

Gray, M., Samala, R., Liu, Q., Skiles, D., Xu, J., Tong, W., & Wu, L. (2024). Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science. *Clinical Pharmacology & Therapeutics*, 115(4), 687-697.

Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*, 47(7), 1264-1279.

Heclo, H. (1972) 'Review article: Policy analysis'. *British Journal of Political Science*, 2, pp. 83–108.

Hernández, M. D. R. N., González, P. I. R., & Sánchez, S. V. (2013). Gender and constructs from the hidden curriculum. *Creative Education*, 4(12), 89.

Hill, M. (2021). *The public policy process*, Fourth Edition. Routledge.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261-266.

Howard-Merriam, K. (1979). Women, education, and the professions in Egypt. *Comparative Education Review,* 23 (2), pp.256-270.

Howlett, M., & Giest, S., (2015). Policy Cycle. *In: International Encyclopedia of the Social & Behavioral Sciences*, edited by James D. Wright, 2nd ed., 188–292. Oxford: Elsevier, 2015.

Hoyle, A., Wallach, H., Augenstein, I., & Cotterell, R. (2019). Unsupervised discovery of gendered language through latent-variable modeling. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.

Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 49-58).

International Gender Working Group (2017). *The Gender Integration Continuum Training Session User's Guide*, USAid PACE paper, p.1-32.

Jia, S., Meng, T., Zhao, J., & Chang, K. W. (2020). Mitigating gender bias amplification in distribution by posterior regularization. *arXiv preprint arXiv:2005.06251*.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L. & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1), 1-9.

Kabeer, N. (1994a). Gender aware policy and planning: a social relations perspective. In M. Macdonald (ed.) *Gender Planning in Development Agencies: Meeting the Challenges*, U.K.: Oxfam

Kabeer, N. (1994b). Reversed Realities: Gender Hierarchies in Development Thought, London and New York: Verso Press and India: Kali for Women Press

Kabeer, N., & Subrahmanian, R. (1996). Institutions, relations and outcomes: Framework and tools for gender-aware planning. *IDS Discussion Paper*, 357. Brighton: Institute of Development Studies.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., ... & Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Karanikolas, N., Manga, E., Samaridi, N., Tousidou, E., & Vassilakopoulos, M. (2023, November). Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics* (pp. 278-290).

Karataş, F., Eriçok, B., & Tanrikulu, L. (2025). Reshaping curriculum adaptation in the age of artificial intelligence: Mapping teachers' AI-driven curriculum adaptation patterns. *British Educational Research Journal*, 51(1), 154-180..

Kentli, F. D. (2009). Comparison of hidden curriculum theories. *European Journal of Educational Studies*, 1 (2), 83-88.

Kenyan Institute of Curriculum Development (KICD) (2017). *Basic Education Curriculum Framework,* pp.1-145. Nairobi: KICD.

Ketamo, H., Moisio, A., Passi-Rauste, A., & Alamäki, A. (2019). Mapping the future curriculum: Adopting artificial intelligence and analytics in forecasting competence needs. In *Proceedings of the 10th European Conference on Intangibles and Intellectual Capital ECIIC 2019*. Academic Conference Publishing International.

Khyani, D., Siddhartha, B. S., Niveditha, N. M., & Divya, B. M. (2021). An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology, 22*(10), 350-357.

Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10), 260.

Lai, J., Gan, W., Wu, J., Qi, Z., & Philip, S. Y. (2024). Large language models in law: A survey. *AI Open*, 5, p. 181-196.

Leach, F. E. (2003). *Practising gender analysis in education*. Oxfam.

Lewis, M., Cooper Borkenhagen, M., Converse, E., Lupyan, G., & Seidenberg, M. S. (2022). What might books be teaching young children about gender?. *Psychological science*, *33*(1), 33-47.

Levin, B. (1975). " Woman's Place" in the Constitution: The Supreme Court and Gender Discrimination. *The High School Journal*, 59 (1), pp.31-45.

Lindblom, C. (1959). The science of muddling through. *Public administration review*, 19(2), 79-88.

Lindsay, J. (1985). Training in Information Systems Design for Urban and Regional Planners in Developing Countries. *A Concept Paper Prepared for the UNCRD*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35.

Longwe, S.H. (1997). The evaporation of gender policies in the patriarchal cooking pot. *Development in practice*, *7(2)*, 148-156.

Lucy, L., Demszky, D., Bromley, P., & Jurafsky, D. (2020). Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open*, 6(3), 2332858420940312.

Macias, M. V., Kharlashkin, L., Huovinen, L. E., & Hämäläinen, M. (2024, November). Empowering Teachers with Usability-Oriented LLM-Based Tools for Digital Pedagogy. In: *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (pp. 549-557).

Manyazewal, M. (1988). Specific Course of Action and Implementation Strategies: Area of Intervention to Initiate Women's Dimension into Mainstream Development Endeavors. *Proceedings of National Workshop on Integrated Gender Responsive Development Planning*, Addis Ababa, Ethiopia.

Marcusi, R., Samuelsii, F., Jalaliii, S., & Belachewiv, H. (2021). Gender-Transformative Programming. *UNICEF Individual papers*, pp.1-11.

Marope, M. (2025). Reconceptualizing and Repositioning Curriculum in the 21st Century: *A Global Paradigm Shift,* pp.1-42. UNESCO International Bureau of Education.

Martin, C. J. (1970). The Case for a Central Projects Bureau. *Finance & Development*, 7(004).

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. In Long and Short Papers (NAACL HLT 2019 — 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies — Proceedings of the Conference; Vol. 1). *Association for Computational Linguistics (ACL)*, pp. 622-628.

Mayberry, M. (1998). Reproductive and resistant pedagogies: The comparative roles of collaborative learning and feminist pedagogy in science education. Journal of Research in Science Teaching: *The Official Journal of the National Association for Research in Science Teaching*, 35 (4), pp.443-459.

Meyer, J.W., Boli, J., Thomas, G.M. & Ramirez, F.O. (1997). World society and the nation-state. *American Journal of sociology, 103 (1),* pp.144-181.

Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, 14(1), 9603.

Miruka O., Van de Grift E., Kristjanson P., Ferdous N., Jost C. (2013). CCAFS Gender Training and Strategizing Workshop Report. *CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS).* Copenhagen, Denmark.

MoET (Ministry of Education and Training, Lesthoto) 2009. *Curriculum and Assessment Policy: Education for individual and social development.* Maseru: MoET.

Moser, C. & Moser, A., (2005). Gender mainstreaming since Beijing: a review of success and limitations in international institutions. *Gender & Development*, 13(2), 11-22.

Moser, C. O. (1985). *Housing policy and women: towards a gender aware approach*. Development Planning Unit, University College London.

Mugure, E. (1983). *The role of women Special reference Self-help groups in urban development with To housing development* (Doctoral dissertation).

Murrugarra-Llerena, J., Alva-Manchego, F., & Murrugarra-Llerena, N. (2022, December). Improving embeddings representations for comparing higher education curricula: A use case in computing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11299-11307).

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Nazir, A., & Wang, Z. (2023). A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. *Meta-radiology*, 100022.

NCDC (National Curriculum Development Centre) (2013). *Integrated Primary Curriculum — Grade 4 syllabus*, pp.1-127. Maseru: MoGYSR.

Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022). Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Oxfam IBIS (2017). Transformative Education. Oxfam IBIS Concept Paper: *Education and Gender*, Version 1 (March 1st, 2017), pp.1-8. Copenhagen: Oxfam IBIS.

Paechter, C. (2000). Power, gender and curriculum. *Knowledge, Power and Learning*.

Pair, E., Vicas, N., Weber, A. M., Meausoone, V., Zou, J., Njuguna, A., & Darmstadt, G. L. (2021). Quantification of gender bias and sentiment toward political leaders over 20 years of Kenyan news using natural language processing. *Frontiers in Psychology*, 12, 712646.

Pakhale, K. (2023). Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv preprint arXiv:2309.14084*.

Papadopoulos, T., & Charalabidis, Y. (2020, September). What do governments plan in the field of artificial intelligence? Analysing national AI strategies using NLP. In *Proceedings of* *the 13th International Conference on Theory and Practice of Electronic Governance* (pp. 100-111).

Pavón Pérez, Á. (2022, July). Bias in Artificial Intelligence Models in Financial Services. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 908-908).

Persico, G., Rosola, M., & Frenda, S. (2024). Artificial Intelligence and gender-fair language in school books: pedagogical insights from potentialities of using an autocorrect in education. *Scholé: rivista di educazione e studi culturali: LXII, 1, 2024*, 214-229.

Plan International (2016) Gender Transformative Advocacy. In *Girls Champions of Change: Curriculum for Gender Equality and Girls Rights*, pp.1-126. Woking, UK: Plan International.

Priestley, M., Alvunger, D., Philippou, S. & Soini, T. (2021). *Curriculum-making in Europe: Policy and practice within and across diverse contexts*. Bingley: Emerald.

Qiu, Y., & Jin, Y. (2024). ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, *21*, 200308.

Quevedo, E., Salazar, J. Y., Koerner, R., Rivas, P., & Cerny, T. (2024, July). Detecting hallucinations in large language model generation: A token probability approach. *In: World Congress in Computer Science, Computer Engineering & Applied Computing* (pp. 154-173). Cham: Springer Nature Switzerland.

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

Ratovicius, C. J., Diaz-Pace, J. A., & Tommasel, A. (2024, August). Detection of Gender Bias in Legal Texts Using Classification and LLMs. In *2024 L Latin American Computer Conference (CLEI)*, pp. 1-10. IEEE.

Raza, S., Bamgbose, O., Chatrath, V., Ghuge, S., Sidyakin, Y., & Muaad, A. Y. M. (2024). Unlocking Bias Detection: Leveraging Transformer-Based Models for Content Analysis. *IEEE Transactions on Computational Social Systems.*

Russell E, W. (2024, March). Mapping Curricula to Skills and Occupations Using Course Descriptions. In *2024 IEEE World Engineering Education Conference (EDUNINE)*, pp. 1-6. IEEE.

Safaei, M., & Longo, J. (2024). The End of the Policy Analyst? Testing the Capability of Artificial Intelligence to Generate Plausible, Persuasive, and Useful Policy Analysis. *Digital Government: Research and Practice*, 5(1), 1-35.

Salinas, A., Shah, P., Huang, Y., McCormack, R., & Morstatter, F. (2023, October). The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-15).

Sant, A., Escolano, C., Mash, A., Fornaciari, F. D. L., & Melero, M. (2024). The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. *arXiv preprint arXiv:2407.18786*.

Saylor, J. Alexander, W., & Lewis, A. (1981). *Curriculum Planning for Better Teaching and Learning*. London: Holt, Rinehart & Winston

Schmitt, S. (2012). Comparative approaches to the study of public policy-making. In *Routledge handbook of public policy* (pp. 29-43). Routledge.

Schut, L., Gal, Y., & Farquhar, S. (2025). Do Multilingual LLMs Think In English?. *arXiv preprint arXiv:2502.15603.*

Sevim, N. (2023). *Analysis of gender bias in legal texts using natural language processing methods* (Doctoral dissertation, Bilkent University).

Singla, S., & Mukhopadhyay, M. (2022). *Gender Norms Do Not Persist But Converge Across Time*. SSRN 4183488.

Skinner, N., & Bromley, P. (2019). Individual and collective social justice education: Comparing emphases on human rights and social movements in textbooks worldwide. *Comparative Education Review*, *63*(4), 502-528.

Song, J., Lee, J. K., Choi, J., & Kim, I. (2020). Deep learning-based extraction of predicate-argument structure (PAS) in building design rule sentences. *Journal of Computational Design and Engineering, 7*(5), 563-576.

Stanczak, K., & Augenstein, I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Steiner, P. (2025). 'The myth of theory free inference: Challenges of AI and machine learning in educational data analysis'. *Keynote address, ICER, Seoul*, July 2025.

Stromquist, N. P. (2002). Education as a Means for Empowering Women. In J. L. Parpart, S. M. Rai, & K. A. Staudt (Eds.), *Rethinking Empowerment: Gender and Development in a Global/Local World* , pp. 22-38. Routledge

Subramanian, S. (2019). India's Policy on Early Childhood Education: Lessons for a Gender-Transformative Early Childhood in India. *Brookings Institute Echidna Global Scholars Program, Policy Paper*, pp.1-23.

Subrahmanian, R., (2007). Making sense of gender in shifting institutional contexts: some reflections on gender mainstreaming. In *Feminisms in development: Contradictions, contestations and challenges*, pp. 113-121.

Subrahmanian, R. (2004). Making sense of gender in shifting institutional contexts: some reflections on gender mainstreaming. *IDS Bulletin 35.4 Repositioning Feminisms in Development*, pp.89-94.

Swarup, A., Bhandarkar, A., Dizon-Paradis, O. P., Wilson, R., & Woodard, D. L. (2024). *Maximizing Relation Extraction Potential: A Data-Centric Study to Unveil Challenges and Opportunities*. IEEE Access.

Taha, K., Yoo, P. D., Yeun, C., & Taha, A. (2024). Text Classification: A Review, Empirical, and Experimental Evaluation. *arXiv preprint arXiv:2401.12982*.

Taherdoost, H., & Madanchian, M. (2023). Artificial intelligence and sentiment analysis: A review in competitive research. *Computers*, *12*(2), 37

Takizawa, P.A. "Using a topic model to map and analyze a large curriculum." *Plos one* 18.4 (2023): e0284513.

UN (United Nations) (1995*). Beijing Declaration and Platform for Action, Fourth World Conference on Women*, 15 September 1995, A_CONF.177_20 and A_CONF.177_20, pp.1-159.

UNDP (United Nations Development Programme – Independent Evaluation Office) (2015). *Evaluation of UNDP contribution to Gender Equality and Women's Empowerment*, pp.1-36. New York: UNDP

UNESCO (2024). *Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models*. Paris, pp.1-22.

UNESCO (2023). *Guidance for generative AI in education and research*, pp.1-44. Paris: UNESCO.

UNESCO (2022). *Glossary: understanding concepts around gender equality and inclusion in education, tool 1*, pp.1-12. Paris: UNESCO.

UNESCO (2021). *From access to empowerment: operational tools to advance gender equality in and through education*, pp.1-68. Paris: UNESCO.

UNESCO (2020). *Gender-sensitive policy making, teacher development and pedagogy in Myanmar*, p,1-11. Yangon: UNESCO

UNESCO (2019a). *I'd blush if I could: closing gender divides in digital skills through education, pp.1-145. Paris: UNESCO.*

UNESCO (2019b). *From access to empowerment: UNESCO strategy for gender equality in and through education 2019-2025*, pp.1-28. Paris: UNESCO.

UNESCO (2013). *UNESCO Handbook on education policy analysis and programming, Vol. I: Education policy analysis*, pp.1-83. Bangkok: UNESCO.

UNESCO (2009). *Promoting Gender Equality through Textbooks. A methodological guide*. Paris: UNESCO

UNESCO (2002). *Gender sensitivity: a training manual for sensitizing educational managers, curriculum and material developers and media professionals to gender concerns*. Compiled by Namtip Askornkool. Paris, UNESCO.

UNESCO-IBE (2025). *Towards a typology of curriculum policy approaches*, pp.1-88.

UNGEI (2024). *Gender Transformative Education Glossary*. https://www.ungei.org/gender-transformative-education-glossary (*accessed July 2025).*

GPE, UNGEI, & UNICEF (2017). *Guidance for Developing Gender-Responsive Education Sector Plans*, PP. 1-159. Washington: GPE.

UNFPA & UNICEF (2020). *Technical Note on Gender-Transformative Approaches in the Global Programme to End Child Marriage Phase II: A Summary for Practitioners*, pp.1-5. New York, NY: UNICEF.

UNICEF (2024). *Gender Transformative Education: Reimagining education for a more just and inclusive world*, pp.1-23.

UNICEF (2019). *Gender Equality: Global Annual Results Report 2019*, pp.1-62. New York: UNICEF

Unterhalter, E., Longlands, H., & Peppin Vaughan, R. (2022). *Gender and intersecting inequalities in education: Reflections on a framework for measurement*. Journal of Human Development and Capabilities, 23(4), 509-538.

Varon, B. (1975). *Enough of everything for everyone, forever?* Finance & Development, 12 (003).

Venugopal, V., & Rituraj, V. (2022). Beyond Siri and Alexa: gender and AI policy. In *Gender, Diversity and Innovation* (pp. 125-147). Edward Elgar Publishing.

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2020). Double-hard debias: Tailoring word embeddings for gender bias mitigation. *arXiv preprint arXiv:2005.00965*.

Weavers, E. (2022). *What is a gender sensitive or a gender responsive curriculum to you? An exploration of the extent of shared understanding of different Gender Approach Terms (GAT) in relation to school curricula and the implications of their use within the development sector.* Masters Dissertations, UCL., pp. 1-116

Wei, X., Kumar, N., & Zhang, H. (2025). Addressing bias in generative AI: Challenges and research opportunities in information management. arXiv preprint arXiv:2502.10407.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837

West, J. (2017). Validating curriculum development using text mining. *The Curriculum Journal*, 28(3), 389-402.

Wilming, R., Dox, A., Schulz, H., Oliveira, M., Clark, B., & Haufe, S. (2024). GECOBench: A Gender-Controlled Text Dataset and Benchmark for Quantifying Biases in Explanations. *arXiv preprint arXiv:2406.11547*.

Wirjo, A., Calizo Jr, S., Vasquez, G. N., & San Andres, E. A. (2022). *Artificial intelligence in economic policymaking*. APEC Policy Brief, 52.

World Health Organisation (WHO) (2011). *WHO Gender Responsive Assessment Scale: criteria for assessing programmes and policies*, p.1.

World Health Organization (WHO), (2003). *Gender analysis in health: a review of selected tools*, pp.1-83. Geneva: WHO.

Yang, H., Kim, J., & Lee, W. (2023). Analyzing the alignment between AI curriculum and AI textbooks through text mining. *Applied Sciences*, 13(18), 10011.

Zamecnik, A., Barthakur, A., Wang, H., & Dawson, S. (2024, September). Mapping Employable Skills in Higher Education Curriculum Using LLMs. In *European Conference on Technology Enhanced Learning* (pp. 18-32). Cham: Springer Nature Switzerland.

Zhang, T., Zeng, Z., Xiao, Y., Zhuang, H., Chen, C., Foulds, J., & Pan, S. (2024). GenderAlign: An Alignment Dataset for Mitigating Gender Bias in Large Language Models. *arXiv preprint arXiv:2406.13925*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Yingqian Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. *In International conference on machine learning* (pp. 12697-12706). PMLR.

Zheng, N. N., Liu, Z. Y., Ren, P. J., Ma, Y. Q., Chen, S. T., Yu, S. Y., Xue, J.R., Chen, B.D. & Wang, F. Y. (2017). Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering, 18*(2), 153-179.

Zhong, C., Cheng, F., Liu, Q., Jiang, J., Wan, Z., Chu, C., Murawaki, Y., & Kurohashi, S. (2024). Beyond English-centric LLMs: What language do multilingual language models think in?. *arXiv preprint arXiv:2408.10811*.

Zhu, Y., Moniz, J. R. A., Bhargava, S., Lu, J., Piraviperumal, D., Li, S., Zhang, Y., Yu, H. & Tseng, B. H. (2024). Can Large Language Models Understand Context?. *arXiv preprint arXiv:2402.00858*.

Żółkowski, A., Krzyziński, M., Wilczyński, P., Giziński, S., Wiśnios, E., Pieliński, B., Sienkiewicz, J., & Biecek, P. (2022). Climate Policy Tracker: Pipeline for automated analysis of public climate policies. *arXiv preprint arXiv:2211.05852*.