

Article

Evaluating Sentiment and Factuality of Offshore Wind Technological Trends Using Large Language Models

Holly Freed, Konstantina Vogiatzaki * and Stephen Roberts 

Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK; hfreed.online@gmail.com (H.F.); stephen.roberts@eng.ox.ac.uk (S.R.)

* Correspondence: konstantina.vogiatzaki@eng.ox.ac.uk

Abstract

The urgent pursuit of net-zero emissions presents a critical challenge for modern societies, necessitating a speedup of transformative shifts across sectors to mitigate climate change. Predicting trends and drivers in the integration of energy technologies is essential to addressing this challenge, as it informs policy decisions, strategic investments, and the deployment of innovative solutions crucial for transitioning to a sustainable energy future. Despite the importance of accurate forecasting, current methods remain limited, especially in leveraging the vast, unlabelled energy literature available. However, with the advent of large language models (LLMs), the ability to interpret and extract insights from extensive textual data has significantly advanced. Sentiment analysis, in particular, has just emerged as a vital tool for detecting scientific opinions from the energy literature, which can be harnessed to forecast energy trends. This study introduces a novel multi-agent framework, EnergyEval, to evaluate the *sentiment* and *factuality* of the energy literature. The core novelty of the multi-agent framework is found to be the use of heterogeneous energy-specialised roles with different LLMs. This investigation, using both multiple persona agents and different LLMs, provides a bespoke collaboration mechanism for multi-agent debate (MAD). In addition, we believe our approach can extend across the energy industry, where deep application of MAD is yet to be exploited. We apply EnergyEval to the case of UK offshore wind literature, assessing its predictive performance. Our findings indicate that the sentiment predicted by the EnergyEval effectively aligns with observed trends in increasing the installed capacity and reductions in Levelised Cost of Energy (LCOE). It also helps us to identify key drivers in offshore wind development. The advantage of employing a multi-agent LLM debate team allows us to achieve competitive accuracy compared to single-LLM-based methods, while significantly reducing computational costs. Overall, the results highlight the potential of EnergyEval as a robust tool for forecasting technology developments in the pursuit of net-zero emissions.

Keywords: offshore wind; artificial intelligence; transformer architecture; large language models



Academic Editor: Lars Johanning

Received: 9 September 2025

Revised: 13 October 2025

Accepted: 28 October 2025

Published: 4 November 2025

Citation: Freed, H.; Vogiatzaki, K.; Roberts, S. Evaluating Sentiment and Factuality of Offshore Wind Technological Trends Using Large Language Models. *Energies* **2025**, *18*, 5816. <https://doi.org/10.3390/en18215816>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The transition to a low-carbon economy is a critical energy challenge worldwide, and it must happen rapidly. For instance, the UK has committed to reducing its greenhouse gas emissions to net-zero by 2050 [1], with other countries setting similar goals. Achieving these targets is demanding, requiring the deployment of innovative technologies, strategic government support, and sustained investment [2], making it essential for

decision-makers—including political leaders, local governments, and investors—to stay informed about the status of emerging and retiring technologies to chart the most effective path toward net-zero.

Building on the rapid advancements in Artificial Intelligence (AI), particularly Large Language Models (LLMs), the potential of AI to help accelerate the transition to a low-carbon economy is increasingly evident. Since the launch of OpenAI's ChatGPT [3], AI applications have expanded significantly, with notable examples including BloombergGPT for financial sentiment analysis [4] and Med-PaLM for medical question-answering [5]. Companies like Google DeepMind and Meta are developing cutting-edge LLMs, such as Gemini [6] and LLaMA [7], which can generate convincing outputs across text, images, and video by learning comprehensive world representations [8]. Despite their complexity, LLMs exhibit capabilities like reasoning [9] and learning from minimal examples [10]. As AI technology evolves, LLMs present a huge potential to be powerful tools in driving our fast transition to a sustainable, net-zero future; however, their potential and performance is yet unexplored.

The key question of this paper is to evaluate the accuracy and reliability of LLMs in interpreting data from diverse sources—such as research articles, policy reports, and industry news—to generate actionable insights that influence the energy landscape. To address this, we have developed a novel multi-agent framework, EnergyEval, which assesses the sentiment of energy literature from various sources and links it to sector progress. Specifically, we use LLMs to track and classify both sentiment and subjectivity over time in the technical energy literature, focusing on aspects like cost, efficiency, and technology design. We then examine how such sentiment analysis can enhance the forecasting of net-zero technology costs and development, so supporting informed investment decisions and identifying key drivers of change. The accurate classification of subjectivity is crucial for differentiating between 'hope' and 'fact', as the perceived positivity of emerging technologies may not always reflect their actual maturity and investment potential.

Recent research has increasingly examined the role of machine learning in enhancing the reliability and security of energy systems. For example, Ref. [11] provides a comprehensive review of the vulnerabilities of machine learning approaches in IoT-based smart grids, highlighting the potential risks arising from data integrity issues and adversarial manipulation. Complementing this, Ref. [12] proposes a formal physics-constrained robustness model to strengthen the data security of low-carbon smart grids, addressing key challenges related to model trustworthiness and cyber-physical consistency. Together, these studies emphasise the need for transparent and interpretable AI frameworks in the energy sector—an objective that the present work advances through the introduction of a multi-agent LLM framework capable of evaluating sentiment and factuality in offshore wind technology trends.

The novelty of the proposed multi-agent framework lies in the integration of heterogeneous, energy-specialised roles powered by different LLMs. This dual investigation, into multiple persona-based agents and diverse LLM collaborations, represents an innovative advancement in collaborative mechanisms for multi-agent debate (MAD) systems. The lack of heterogeneity among agents has been identified as a key limitation in prior MAD research—one that this study directly addresses. Furthermore, the application of the MAD approach within the energy sector constitutes an additional contribution, as such implementations have not yet been explored in this domain.

This paper begins by collating a dataset of offshore wind specific literature for analysis using LLMs. This is then followed by an optimisation study into the best performing prompt design. Once the prompt strategy and dataset have been selected, the LLM is executed to carry out the specific prompts in order to evaluate the polarity and subjectivity

of the data. Further analysis into patterns between identified offshore wind characteristics will be explored to demonstrate the validity of LLMs on the energy-specific literature against offshore wind time-series data. This approach is covered in greater detail in Section 4.

2. The UK Paradigm

In this paper, we use data from the UK energy landscape as a test case to demonstrate our tool. However, the methodology is broadly applicable and can be implemented across various energy ecosystems, ranging from local communities to entire countries.

2.1. Road to Net-Zero

The UK has cut carbon emissions by 48.7% as of 2022 compared to 1990 levels [13], but needs an additional 51.3% reduction to achieve net-zero by 2050. In 2022, 20.7% of the UK's primary energy came from low-carbon sources, up from 12% in 2012, with wind contributing 20% of this mix [14]. Wind energy surpassed gas in electricity generation in 2023, and with increasing electrification in heating and transport, the UK aims for a net-zero electricity mix by 2035, crucial for broader decarbonisation efforts [13]. Despite an 87.2 % decline in fossil fuel use since 2012 due to coal phase-out [13], new oil and gas projects continue to be approved to ensure short-term energy security. This focus on fossil fuels presents challenges to achieving net-zero.

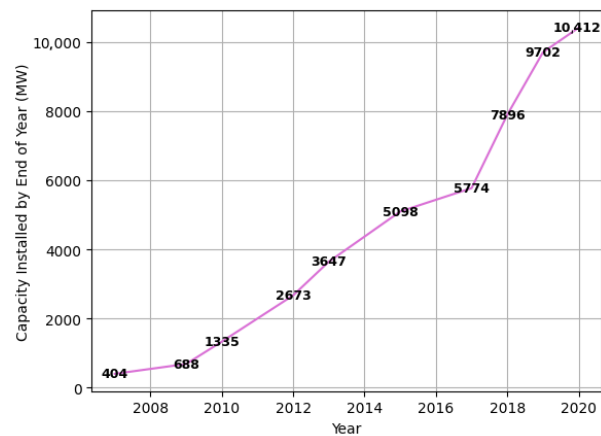
The UK's energy strategy, outlined in the 2020 *Ten Point Plan* [15], emphasises the expansion of offshore wind, development of low-carbon hydrogen, and advancement of nuclear power. By 2030, the UK aims to quadruple its offshore wind capacity to 40 GW—sufficient to power every home today [16]—including 1 GW of floating offshore wind. Achieving this goal will require the installation of more wind turbines, innovation in next-generation turbine technology, and the implementation of energy storage solutions to address the intermittency of wind power [16].

2.2. Offshore Wind Success

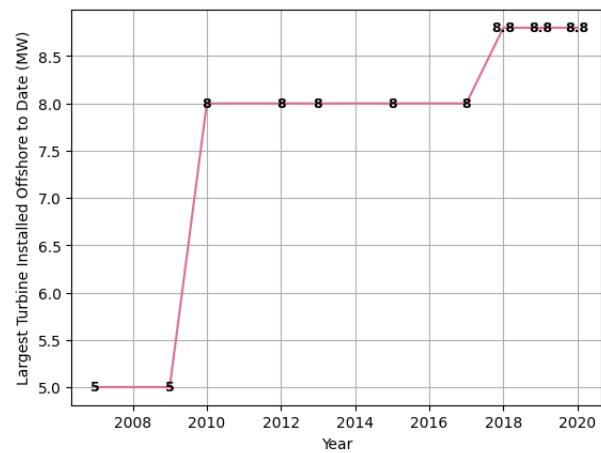
Over recent decades, the UK has become the second largest offshore wind market in the world [15]. Leveraging the combination of long coastline, shallow waters and strong winds, the UK is tailored to generating ample energy with offshore wind turbines. In developing offshore wind, the UK has benefited from targeted government support, such as early-stage development funding and the Contract for Difference (CfD) [2], along with strong partnership between industry and government, and innovations in larger turbines. Enabled by these factors, the UK has successfully deployed 10GW of installed capacity, as of 2020, since the first offshore wind farm in 2000 [17]. Offshore wind will become a critical source of renewable energy for a net-zero future; the UK has strong plans to expand this capacity and deploy pioneering turbine technologies.

Figure 1 illustrates the evolution of installed capacity, turbine size and lifetime cost—Levelised Cost of Electricity (LCOE)—of offshore wind farms in the UK from 2007–2020. Below we list the key factors that have contributed to reducing the cost of wind power based on the available literature, as it is interpreted by “humans” (the authors).

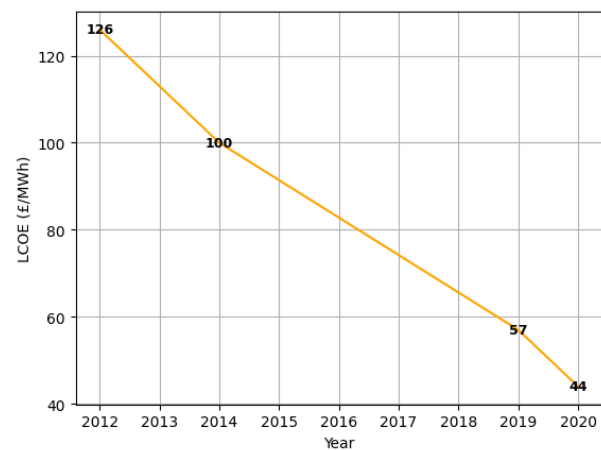
- **Developments in Turbine Design:** Greater power outputs have been enabled through larger turbines: increases in height and rotor diameter have allowed more wind energy to be captured at higher altitudes, where wind speeds are stronger and more consistent. Following a combination of advanced turbine designs and their cost reductions, the capacity of the largest UK turbines increased from 2 MW in 2000 to 8.8 MW in 2020 [17]. The total maximum power the UK can theoretically produce is referred to as *installed capacity*.



(a)



(b)



(c)

Figure 1. (a) Capacity installed by end of year. (b) Largest turbine installed offshore. (c) Levelised cost of electricity.

- **New Construction Techniques:** More efficient construction techniques, such as easier-to-install foundation structures and larger cable sizes, along with process standardisation have further reduced construction, assembly and installation costs [18].
- **Government Support:** Offshore wind in the UK is a policy success story. Consistent government support has led to improved investor confidence, catalysing advancements in research, supply chains and deployment. To incentivise investments with large-scale renewable energy, the UK government established the Contract for Differ-

ence (CfD) scheme, which has minimised investor risks from volatile wholesale prices. The CfD has elicited a healthy pipeline of offshore wind projects bidding into each auction [2]; this has fostered strong competition between developers and has resulted in a reduction of 10–21% on overall project costs [2].

3. Large Language Models

Mathematically, a language model is a probability distribution over a word sequence $x^{(1)}, x^{(2)}, \dots, x^{(n)}$, from which it can be inferred, to determine how likely the next word $x^{(n+1)}$ will be: $P(x^{(n+1)} | x^{(n)}, \dots, x^{(1)})$. With the advent of the transformer architecture [19], huge amounts of Internet data and high compute power, current language models have evolved to reach hundreds of billions of parameters, trained on trillions of words. These are known as large language models (LLMs), where ‘large’ refers to the vast number of model parameters typical in these models. In the following sections we will provide a brief overview of some concepts relating to LLMs where necessary for the presentation of our methodology and results.

One of the first major LLMs was the OpenAI GPT-1 [20] (Generative Pre-Trained Transformer). The transformer, trained initially on a corpus of unsupervised data from over 7000 unique unpublished books [8], while GPT-3 [10] reached a training volume of 175 B parameters on 45 TB of training data [8]. The innovation that arose from increasing the model parameters and dataset size was that the model began to learn natural language processing (NLP) tasks without the need for supervised training (fine-tuning) [8] and computationally expensive fine-tuning processes.

3.1. Emergent Properties

Emergent properties are properties that arise from large-scale models, and are not present in smaller models. In other words, ability ‘emerges’ as scale increases. It is (a rather unexpected) ability to perform a task well during inference time, which the model has not been specifically trained for.

One key emergent ability of LLMs is *zero-shot learning*, where the model performs well across tasks without explicit examples or parameter updates (i.e., no task-specific fine-tuning). Additionally, LLMs exhibit *few-shot learning*, an extension of zero-shot learning, which includes a small number of demonstrations in the prompt, also known as in-context learning (ICL). A visualisation of the process can be seen in Figure 2.

Key questions around ICL include: How does the model learn without gradient updates, and why does this occur in large-scale models? One explanation is that ICL leverages latent concepts the model has already learned during training [21]. Larger models have a higher chance of mapping in-context prompts to these latent tasks, improving performance. Another theory suggests models use gradient descent on prompts to learn new tasks [22], aligning with the idea that large models can override semantic priors to acquire new knowledge. This ability may reduce the need for specialised fine-tuning, which relies on extensive labelled data [10], often difficult to obtain.

A related emergent ability is *zero-shot reasoning* [9], which involves step-by-step reasoning, or chain of thought (CoT) prompting, to reach the answer without task-specific examples. This method significantly enhances performance without few-shot examples, emphasising the power of designing prompts over collecting fine-tuning datasets.

In summary, learning from LLMs can be achieved through pre-training or in-context learning. In-context learning leverages the prompt itself to learn a new task. This significantly contrasts the pre-training approach where the scale of training data is a key factor for effective learning. LLMs unleash the ability to learn from much less data.

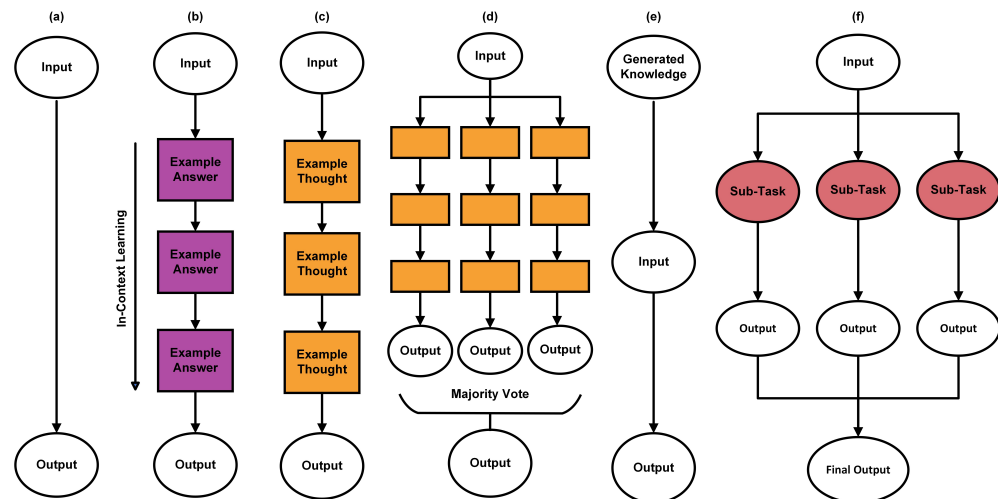


Figure 2. Illustration of various prompt engineering techniques including (a) **standard prompting**, which accepts only a task instruction, (b) **few-shot prompting** or in-context learning, where a number of demonstrations without reasoning are given in the prompt, (c) **chain-of-thought**, where demonstrations are given with reasoning, (d) **self-consistency**, the majority vote over multiple chain-of-thought prompts (e) **generated knowledge**, where the previous output from the model is incorporated into the input prompt, and (f) **least-to-most**, where the input prompt is deconstructed into individual tasks.

3.2. Prompt Engineering

One important aspect in the use of LLMs is the process of prompting. In prompting, an input instructs a pre-trained language model on the desired task, where the model then responds without further training. The design of the prompt heavily influences the relevance and coherence of the generated output.

Figure 2 details the core strategies, which include Comprehensive Description, few-shot prompting, self-consistency, least-to-most, Tree of Thoughts, Retrieval Augmentation, and role prompting. We include a brief overview of these methods in the Appendix A.

3.3. Dangers of Stochastic Parrots

Despite the impressive capabilities of LLMs, some machine learning scientists [23] argue that LLMs do not understand the meaning of language, only that they repeat (manipulations of) words without any underlying semantic understanding. Referred to in these critiques as *stochastic parrots*, LLMs present a number of dangers that limit their real-world applications, including generating incorrect responses and introducing biases, as well as requiring significant energy for training.

Hallucinations occur when a language model produces outputs that are linguistically coherent yet factually incorrect, often presenting false or unsupported information with high confidence. Recent research by Kalai et al. (2025) [24] suggests that such hallucinations arise because current training and evaluation paradigms inherently reward confident guessing rather than calibrated uncertainty. In other words, models are optimised to generate the most probable continuation of text rather than to express doubt or abstain when information is lacking.

This can result in misleading information, which is especially dangerous in cases such as healthcare where correct information is critical. LLMs hallucinate because their training objective is to predict the highest-probability continuation of words [8]. Consequently, probabilistically linking words produces plausible language but may not give a factually correct answer. LLM correctness is also limited by the training dataset [23]: if elements of the data are incorrect, the LLM may stochastically repeat this information. Another factor

responsible for hallucinations is that LLMs generate one token at a time, which makes it difficult to complete the entire response correctly if the model does not know how it will end [8]. However, there are ways to reduce the impact of hallucinations. This includes building more capable models [8], such as GPT-4, and using certain prompting strategies discussed previously, such as asking the model itself about the accuracy of its own input [8].

Another danger that can arise within LLMs is the generation of biases, stereotypes and harmful content. Due to LLMs mirroring their training language, biases may be introduced by the model sampling the highest probability definition, which could be a stereotype or popular biased opinion. Fine-tuning LLMs against this behaviour is one way biases can be mitigated [8].

3.4. Environmental and Financial Costs

LLMs require significant amounts of resources in both training and operating [8]. To run these computations in data centres: energy is needed for power, water is needed for cooling, and money is needed to both finance the large amounts of electricity and purchase the expensive hardware [8]. As models get larger, so do these requirements. Therefore, it is essential for providers to develop energy-efficient models [23], but also for individuals to be conscious of which LLMs to use. Considering the amount of carbon dioxide emitted throughout the process [25], this is especially important for the purpose of this project.

3.5. Sentiment Analysis

Sentiment analysis (SA), also known as opinion mining, is a field of NLP aiming to analyse opinions, attitudes, and emotions from text. SA is a classification problem, where, for most applications, the sentiment is categorised as either positive (1), neutral (0), or negative (−1)—known as the *polarity*. In some cases, sentiment can be classified by the degree of *subjectivity*, which quantifies the amount of factual information contained within the text. The more subjective the text is, the lower the amount of factual information, and the higher the level of personal opinion. These classes include subjective (1), partially subjective (0.5), and objective (0). These classifications are shown in Figures 3 and 4 for polarity and subjectivity, respectively.

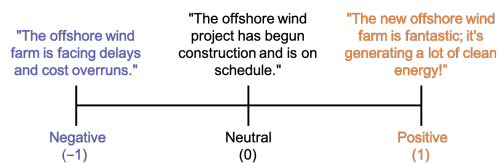


Figure 3. Polarity classifications with examples.

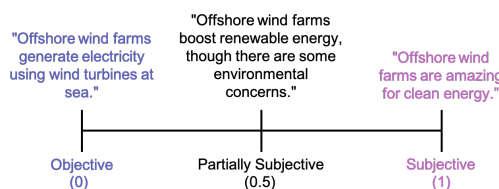


Figure 4. Subjectivity classifications with examples.

Sentiment analysis comprises different degrees of analysis, the object of sentiment. These levels are listed below [26].

- **Document Level** analyses the sentiment of an entire document. This assumes that each document expresses opinions on a single topic, and thus not suitable for evaluating multiple topics.

- **Sentence Level** analyses the sentiment of individual sentences within a document. This enables a way for more granular understanding of sentiment transitions within a document.
- **Aspect Level** analyses the target of opinion (the aspect) instead of language constructs in documents and sentences.

There are also different types of outputs that can be extracted from the dataset. The most common method, known as *Sentiment Classification*, assigns sentiment classes to the input data. For polarity sentiment, these classes can vary from binary classification, positive or negative, to finer graduations such as strongly positive, mildly positive, and so on. This is likewise the case for subjectivity sentiment. Other sentiment analysis techniques include *Aspect-Based Sentiment Analysis* (ABSA) [27], which additionally extracts the object of the sentiment, and *Multifaceted Analysis of Subjective Texts* (MAST) [27], which detects the human emotion present in the input data.

3.6. LLMs vs. SLMs

Over the past year, LLMs have become a highly promising model for sentiment analysis due to demonstrating impressive performance on unseen tasks without the need for supervised training. Due to the specialised nature of sentiment analysis tasks, training the language model to evaluate technical jargon is crucial for accurate results. To achieve this, research has focused on examining in-context learning (ICL) with pre-trained LLMs and fine-tuned small language models (SLMs) as two methods for creating a specialised model. Below, we briefly discuss the two approaches, explaining why LLMs are the most suitable models for this work.

Early studies evaluated the performance of LLMs on various sentiment analysis tasks [27], and how their ability compares to smaller, domain-specialised models [27]. Regarding the benefits of SLMs, the results show that SLMs outperform LLMs with more in-depth tasks like ABSA [27]. LLMs further fall short to SLMs in these tasks due to significant variation with prompt design and the number of few-shot examples [27]. Overall, SLMs are superior to LLMs for fine-grained structured sentiment analysis tasks. However, for simpler SC tasks, LLMs match and exceed SLMs with minimal sensitivity to prompt design [27]. Further to this, LLMs also show greater performance with unseen sentiment analysis tasks under few-shot settings [27]. This could be due to the phenomenon of emergent properties, resulting in the difficulty for SLMs to learn new patterns with a limited number of examples. In order for the performance of LLMs to surpass SLMs, greater research into advanced methods for ICL needs to be explored. With [27] only evaluating few-shot prompting, the potential of LLMs for sentiment analysis is still a vastly undiscovered domain.

3.7. Multi-Agent Debate: LLM Collaboration

Looking to the future of sentiment analysis and LLMs, current questions shift towards leveraging the full potential of existing pre-trained LLMs, instead of fine-tuning SLMs. Unlike in earlier papers [27,28], the need to avoid fine-tuning all together has become a desirable pursuit in research due to eliminating the requirement for massive data acquisition. Improving upon few-shot prompts in [27,28], current studies have achieved more consistent performance by using a 'team' of pre-trained LLMs to decide on the sentiment by debating with each other, combined with strategic prompting techniques [22,29]. Known as *multi-agent debate* (MAD), this approach has been shown to improve the truthfulness of generated content, minimising the effect of hallucinations [30]. This shift to advanced prompt engineering also enables sentiment analysis to become more accessible and affordable for domains with limited, labelled datasets since task-specific training is not needed, presenting advantages over machine learning and SLMs.

3.7.1. Society of Mind

While a single powerful LLM can perform many tasks well, multi-agent debate enhances factual accuracy through collaboration [30]. This is due to the ability of LLMs to learn in context and adapt during inference, reducing bias and instability by avoiding reliance on a single perspective [31]. This approach mirrors human evaluation processes and social collaboration behaviours [31,32]. The *Society of Mind* theory, where intelligence arises from a society of simpler agents, has inspired new MAD designs, including diverse role prompts and thinking patterns.

3.7.2. Principal Components

MAD consists of multiple language model instances (known as agents) communicating with each other through debate in order to reach an agreed consensus. There are three components that define the MAD process [31]:

- **Debater Agents:** These individual LLM instances take input from previous agents and output new perspectives in group discussions. As part of a communication network, information flows between debaters. A key feature of debater agents is their distinct roles, which ensure varied viewpoints.
- **Communication Strategy:** This refers to how chat history is maintained. Various configurations are emerging, some using Summariser agents to overwrite previous chat history [31] and others employing Reflector agents [32] to review responses and extract lessons. Figure 5 illustrates examples of popular communication strategies.
- **Diverse Roles:** Stemming from the parallel between LLMs and social behaviours, having diverse personas for each debater is vital for effective MAD [31]. Designing for these different roles is a core optimisation challenge in understanding what traits and skills need to be distributed to each debater. In [29] approaches for designing these **heterogeneous LLM agents** based on prior domain knowledge of the types of errors that occur in the sentiment analysis tasks are presented.

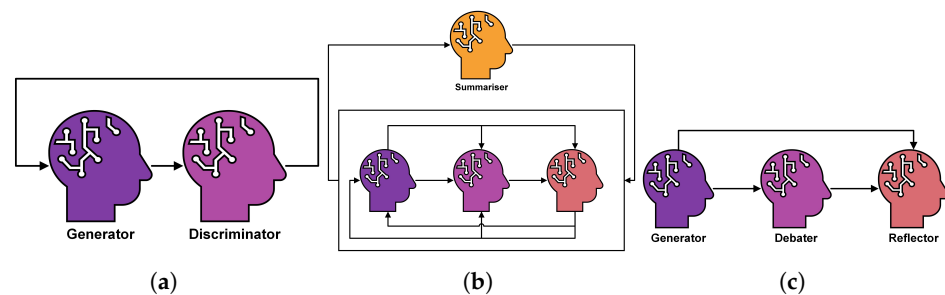


Figure 5. Illustrations of various communication strategies proposed in the MAD literature; including the generator–discriminator framework suggested in [33], simultaneous, and one-by-one debate in [31]. (a) Generator–discriminator. (b) Simultaneous debate. (c) One-by-one debate.

3.7.3. Groupthink Phenomenon

A limitation of MAD is when misguided answers of debater agents sway the correct solutions of other agents [32]. This is the detriment of the ‘groupthink’ phenomenon, where a group of individuals reach a consensus without objective critique, resulting in an irrational outcome. The challenge of MAD is for agents to converge on the right answer, given a varied debate. This phenomenon has arisen from a lack of diversity in the debate process: agents with similar perspectives tend to prioritise harmony over a critical analysis of divergent views [32]. As discussed, refs. [29,31] prevent this with diverse role prompts: [31] recruits famous experts in their prompt design to extract the skills of such professionals for the specific task, and [29] defines each agent to check for a certain error in order to yield better accuracies. Ref. [32] approaches this issue by assigning an individual trait to each

agent. Two contrasting traits are used, over-confident and easy-going, to exhibit typical contrasting behaviours. In collaboration with this, thinking patterns, debate or reflection, are also used in the communication strategy. The results indicate that easy-going agents and debate-dominant thinking patterns are more likely to reach a consensus [32]—which could be because different viewpoints are exchanged and then appreciated to acquire.

4. Research Methodology

To analyse the scientific opinions with offshore wind in the UK, a four-layer approach using LLMs and sentiment analysis is employed to determine the cost, capacity and turbine size trends in offshore wind technology as illustrated in Figure 6. The layers are summarised below, while more details of how each layer is executed are presented in the section relevant to the experimental set-up.

- **First Layer—Dataset** The first layer sources the literature relevant to offshore wind energy in the UK, and processes these data in an interpretable format for labelling. The dataset analysed in this study comprises UK-based energy literature collected from 2010 to 2022, encompassing a total of approximately 50 documents. These include a diverse range of government and industry reports, academic journal articles, and policy briefings focusing on offshore wind technology, energy transition, and sustainability. Data are sourced primarily from open-access repositories, government archives, and Scopus-indexed publications to ensure representativeness across academic and policy domains. While the dataset provides a comprehensive view of the UK energy landscape, we acknowledge a potential bias toward English-language and UK-centric publications, which may limit broader international generalisability. Nevertheless, this focus enables a more detailed and contextually grounded assessment of offshore wind discourse within the UK energy sector.
The data are then gathered across a wide time range and breadth of sources to analyse the time-series trends and limit bias, respectively. Pre-processing converts the PDF documents into tokens of text (as detailed in Section 5.1.2).
- **Second Layer—Labelling:** Following data collection and pre-processing, two consistency-based prompting techniques using pre-trained LLMs are compared using a ground truth dataset constructed from a randomly sampled subset of the overall test data. Popular GPT-3.5 and GPT-4 models are used, alongside open-source models Llama, Mixtral and Falcon. Self-consistency and multi-agent debate prompts are designed and contrasted. These consistency-based techniques are essential for reducing hallucinations to improve the reliability of results. The best-performing approach will be selected for labelling the offshore wind dataset, required for performing sentiment analysis. Polarity, subjectivity, category and aspect opinions are chosen to be extracted from the data in order to capture the fine-grained details of offshore wind features. Hence, this implements an aspect-based sentiment analysis task. Best performing refers to the design that yields the most accurate results against the ground truth dataset and that requires the lowest computational cost since energy-efficient models are needed for the future of LLMs.
- **Third Layer—Offshore Wind Development and Trends:** From the labelled offshore wind dataset, the third layer involves presenting the evolution of polarity and subjectivity, along with the positive and negative keywords, to verify offshore wind becoming more positive and objective with time, following the UK's successful deployment.
- **Fourth Layer—Tracking Sentiment Analysis:** Lastly, the fourth layer entails tracking the sentiment development relevant to cost, capacity, and turbine size features.

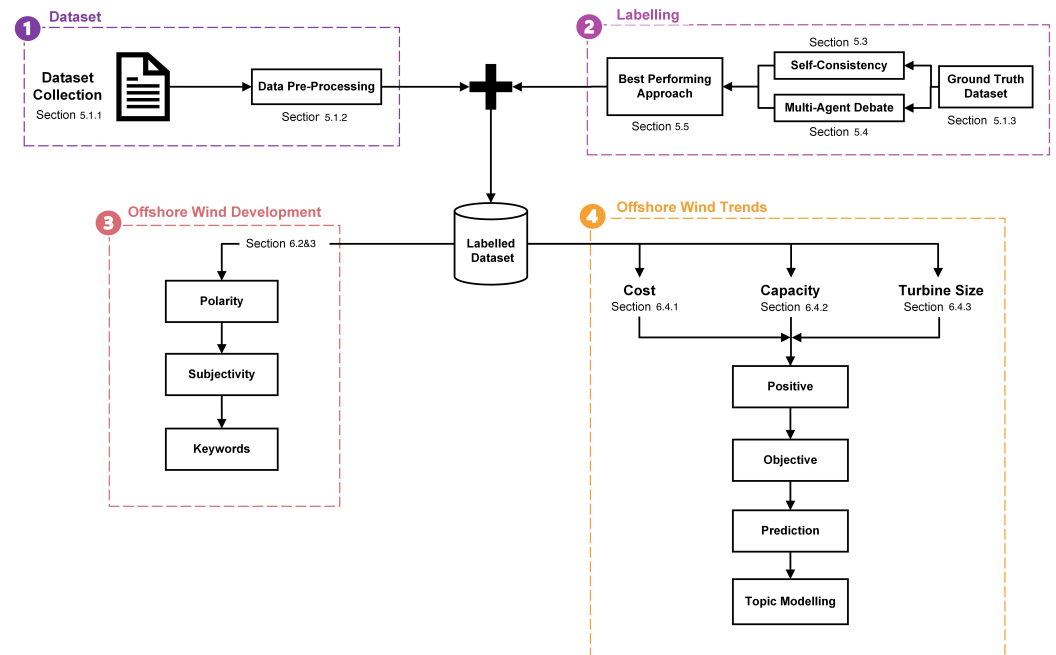


Figure 6. Solution architecture for designing the best-performing LLM prompting strategy on energy literature, forecasting UK offshore wind trends in cost of electricity, installed capacity and turbine size installation using this novel technique, and visualising the impact of these features on offshore wind.

This allows for comparison with real offshore wind data of the levelised cost of electricity, installed capacity, and turbine size installation, to understand whether the model accurately forecasts these trends from the given literature, and what type of sentiment measurement predicts this the most. For each feature, the real data are compared against the normalised positive and objective frequency in that year. This will indicate whether the model captures the opinions from scientific professionals of each feature, and whether these positive opinions correctly reflect the trends themselves. Further to this, the evolution of policy, technology, market and supply-chain sectors will be visualised for each feature—to gain greater insight into the influence of each driver on offshore wind development.

5. Experimental Setup

5.1. Dataset

5.1.1. Dataset Collection

A dataset totalling 5033 sentences (tokens) is collected for the detection of sentiment towards the cost, capacity, and turbine size development of offshore wind in the UK, which is listed in Table 1. The criteria for selection ensure a sufficient temporal range of documents, published by reliable organisations, which cover a holistic breadth of offshore-wind-related topics.

Table 1 consists of technical reports within the years 2008–2020. The dataset was reviewed to ensure that the topics of policy, technology, supply chain, and market are well-distributed and representative of the offshore wind industry in the UK each year. The sources include academic, government, industrial, and consultancy reports, chosen for their reliability, as they are written by scientific and engineering experts. Unlike social media or news articles, this dataset focuses solely on professional and academic literature to minimise the risk of misleading information and ensure accuracy in sentiment analysis.

Table 1. List of the available UK offshore wind literature from 2008 to 2020, which holistically discuss offshore wind policy, technology, supply-chain and market sectors, and is selected from data-driven organisations. This table indicates the document year, title, and organisation. The number of sentences (the chosen token element) is also included to show the approximately equal distribution in each year.

Year	Document	Organisation	Tokens
2020	Policy, Innovation and Cost Reduction in UK Offshore Wind [34]	Carbon Trust, UCL	560
2019	Developments in Wind Power [18]	HM Government	145
	The UK Offshore Wind Industry: Supply Chain Review [35]	Offshore Wind Industry Council	601
2014	Wind Energy: Implications of Large-Scale Deployment on the GB Electricity System [36]	Royal Academy of Engineering	847
	UK Offshore Wind in the 2020s: Creating the Conditions for Cost Effective Decarbonisation [37]	Green Alliance	269
	The UK Offshore Wind Supply Chain: A Review of Opportunities and Barriers [38]	Offshore Wind Industry Council	584
2012	Wind Energy in the UK: State of the Industry Report [39]	Renewable UK	345
2010	Great Expectations: The Cost of Offshore Wind in UK Waters [40]	UK Energy Research Centre	668
2008	Offshore Wind Power: Big Challenge, Big Opportunity [41]	Carbon Trust	1014

5.1.2. Data Pre-Processing

Prior to prompting the LLM, the PDF documents from each report need to be transformed into a structured format for sentiment analysis, known as text pre-processing. Observation of the dataset in Table 1 shows that each sentence expresses a sentiment about a specific aspect, which can vary from sentence to sentence.

Therefore, sentence-level sentiment analysis is most suitable for this application, as it captures the sentiment of individual aspects rather than aggregating multiple opinions as in paragraph or document-level analysis. This approach provides a more precise understanding of sentiment and avoids limitations related to context length. The goal of data pre-processing is to convert the PDF documents into a dataframe of consecutive sentences.

The core steps in data pre-processing include the following:

- **Extract Text:** Using the Python `fitz` library, the text within the PDF document is extracted.
- **Noise Reduction:** This removes special characters that are not `[a-zA-Z0-9, .%-:;]`, HTML tags, header titles, references, and page numbers that appear throughout the data.
- **Tokenisation:** Using the Python `punkt` tokeniser model from the `nltk` library, the string of text is segmented into a list of consecutive sentences.
- **Handling Noisy Text:** The final step in the data pre-processing pipeline is to mitigate the impact of remaining noisy characters, including duplicate newline commands and negligible sentences.

5.1.3. Ground Truth Dataset

In order for the LLM sentiment analysis pipeline to be designed, a test dataset was randomly sampled from Table 1 to evaluate the range of designs. To evaluate the model design, ground truth responses for polarity and subjectivity were measured by 3 annotators for a fair, unbiased assessment. Pre-labelling of polarity and subjective-objective classifications was required in order to evaluate the LLMs against a ground truth dataset for a quantitative assessment of the performance. The final ground truth answer was taken as the mode of 1 human, 1 LLM and 1 human-LLM debate mixture. The LLM response was taken

from an ensemble of multiple prompts across all the LLMs listed in Table 2, with the final answer being the ‘majority vote’ or mode of the ensemble. The human–LLM mixture was constructed by a human evaluating the answers given by the GPT-3.5-Turbo model (The GPT-3.5-Turbo model was chosen for this since it gave the most understandable response for a human to evaluate upon observation of the results) in Table 2. By using the full ensemble of LLMs combined with human verification, it was possible to capture multiple, diverse and human-driven perspectives to polarity and subjective–objective classifications for an accurate ground truth baseline.

Table 2. Properties of the selected large language models for design—including model name, provider, number of model parameters, tokens trained on, fine-tuning, and pricing for accessing the model.

Model Name	Provider	Parameters	Training Tokens	Fine-Tuned?	Pricing
GPT-3.5-Turbo-1106 [3]	OpenAI	175B	-	No	\$2/1M tokens
GPT-4-Turbo-Preview [42]	OpenAI	1.8T	13T	No	\$40/1M tokens
Llama-2-7b-Chat [43]	Meta	6.74B	2T	Yes	Free
Mixtral-8x7b-Instruct [44]	Mistralai	46.7B	-	Yes	Free
Falcon-7b-Instruct [45]	TII	7B	1.5T	Yes	Free

5.2. Selection of Large Language Models

To investigate the performance of different LLMs on our sentiment analysis application, a diverse range of LLMs have been selected and are listed in Table 2. This breadth of LLMs is important for our experimentation; both to address the performance of multiple agents based on different LLMs from [32], and to determine whether low-cost, open-source models are competitive with larger, more expensive models.

Following the strong performance and cost effectiveness within the GPT series, the ChatGPT model (GPT-3.5-Turbo) was used as one of the chosen LLMs. In addition, OpenAI’s latest-trained GPT-4 model (GPT-4-Turbo) was utilised—this is the most expensive LLM in Table 2, trained though on a significantly greater number of training tokens. The Llama, Mixtral, and Falcon models were also tested because they have fewer model parameters than the GPT series. This was intended to enable greater understanding of the behaviour of different-sized LLMs under different prompt techniques. These models are all fine-tuned to dialogue use cases, which makes them suitable and ready-to-use for our sentiment analysis question-answering tasks.

The GPT-3.5-Turbo and GPT-4-Turbo closed-source models will be accessed through the OpenAI API using their Python library. For the open-source models, the `huggingface_hub` Python package will be used alongside the `langchain LLMChain` framework to pass the prompts into the model.

5.3. Self-Consistency

This section investigates the performance of the self-consistency prompt technique, the first consistency-based prompting technique to be evaluated, across the range of selected LLMs. These experiments determine the models that perform with the highest accuracy on the polarity and subjectivity tasks. The method was carried out by calling the API, passing through the prompt and then repeating over three iterations to calculate the final modal answer. In designing the chain-of-thought prompt for self-consistency, Figure 7 illustrates the potential CoT designs: zero-shot, one-shot, and two-shot. The task instruction was created using ChatGPT to ensure it could be understood from an LLM level. Context limitations required the polarity and subjectivity tasks to be separated, implementing a least-to-most structure.

From observation of the experimental results, it was evident that reasoning prompts significantly improve the accuracy of both polarity and subjectivity outputs. As demon-

strated in Figure 7, few-shot CoT provides a structured model output compared to the zero-shot prompt, which can be explained by few-shot CoT leveraging reasoning skills learnt from training. It should be noted that CoT exemplars do not influence the quality of the content, i.e., how the sentiment is analysed. Two-shot CoT was found to overly follow the structure of the examples given, perhaps because the freedom to reason was constrained by the structure dictated by the CoT exemplars. This led to less accurate results, implying in-depth reasoning is highly important to reach an accurate conclusion. Hence, the one-shot CoT prompt in Figure 7 was selected to evaluate the different LLMs.

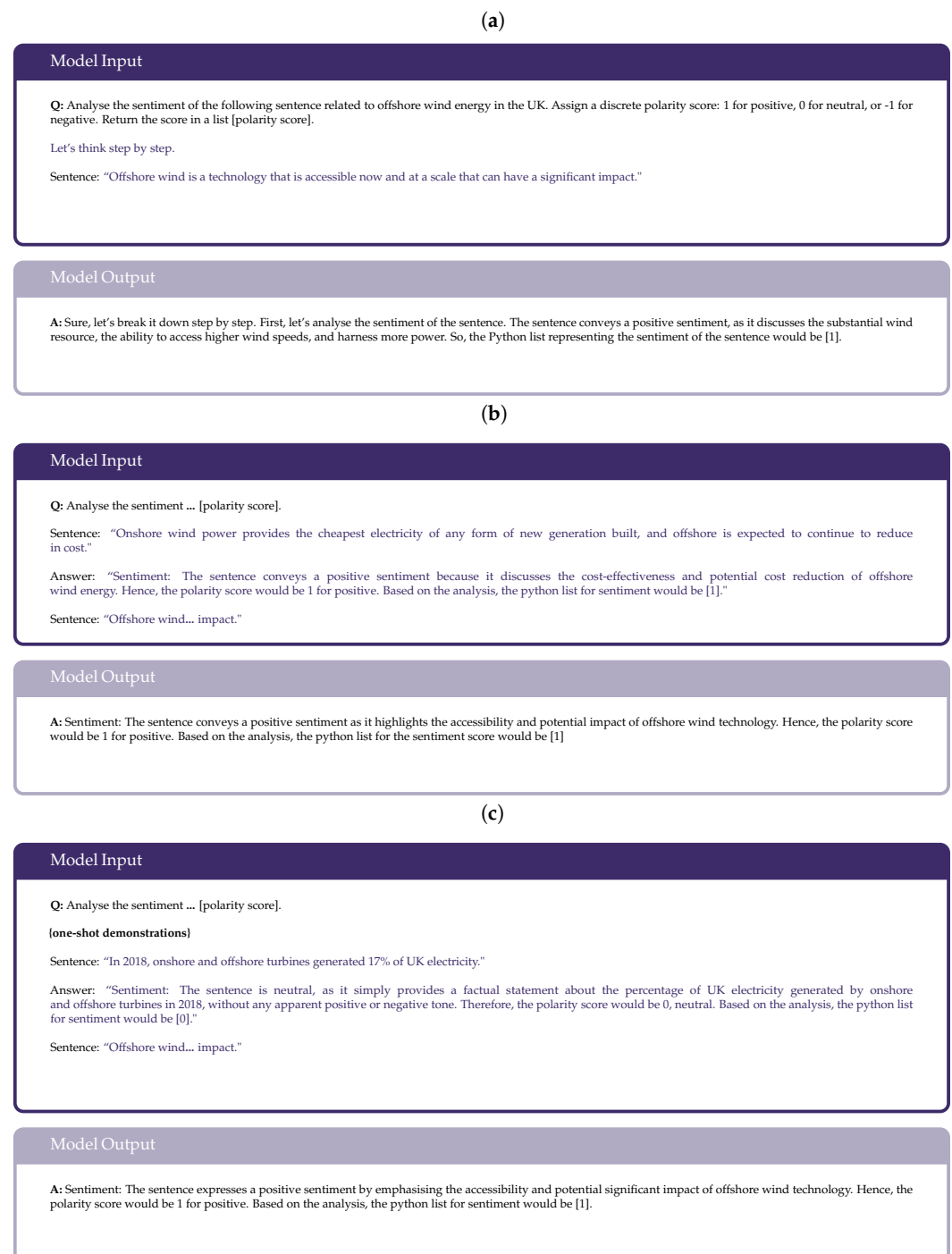


Figure 7. Illustrations of various communication strategies proposed in MAD literature; including the generator–discriminator framework suggested in [33], simultaneous and one-by-one debate in [31]. (a) Zero-shot. (b) One-shot. (c) Two-shot.

Generator LLM Agent

The accuracy of each LLM using self-consistency prompting is listed in Table 3, and was calculated using Equation (1) across all the sentences in the test dataset. This experiment was carried out using the previous method, investigating one-shot CoT only. Table 3 reveals that GPT-3.5-Turbo and GPT-4-Turbo perform the best on polarity and subjectivity tasks, respectively. Below discusses the capabilities of the highest-performing LLMs on these tasks and the limitations of self-consistency:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

GPT-3.5, Mixtral, and Falcon perform the best on polarity tasks. Table 3 shows the percentage of correctly answered predictions for self-consistency across a range of LLMs. For the polarity task, GPT-3.5, Mixtral and Falcon perform with the highest accuracy. This is because these models correctly understood the meaning of positive and negative terminology for offshore wind energy. For instance, rather than labelling ‘reduced’ as negative, these models are able to detect the sentence ‘Increasing turbine size, cheaper finance, and more efficient construction and operations have reduced costs substantially.’ correctly as positive. This advanced technical understanding could be due to the mixture of pre-trained data within these models, enabling greater keyword mappings. In addition, a high-level of reasoning is also demonstrated in their answers, which may be an explanation for their high performance. These emergent reasoning capabilities could be attributed to their large model sizes, notably for GPT-3.5 and Mixtral which have the highest accuracies. Surprisingly, the most expensive and largest model, GPT-4, performs the least well. The cause of this is due to the model over-analysing the sentences—by being overly descriptive, the justifications became too complex and diverted from the simpler, more obvious answers.

GPT-4 and GPT-3.5 perform the best on subjectivity tasks. For the subjectivity task, GPT-4 and GPT-3.5 have the highest accuracy. This is due to correctly identifying factual and hopeful keywords, such as ‘suggested’ and ‘expected’. Falcon and llama models struggled to understand this, using definitions of subjectivity as ‘favourable’, ‘emotion’ or ‘speaker’s opinion’. Unlike in the polarity task, GPT-4 performs comparably better because subjectivity analysis is less focused on offshore wind, meaning the scope for analysis is more direct and obvious since it purely relies on the language presented, and not any wider context. Mixtral performs the best amongst the open-source models, due to understanding the meaning of subjectivity well for this application.

Table 3. Accuracy of self-consistency prompting on polarity and subjectivity tasks, **bold** indicates the highest-accuracy result: GPT-3.5 is the best-performing model for polarity, and GPT-3.5 and GPT-4 are statistically equivalent and thus the best-performing models for subjectivity.

Model	Polarity Accuracy	Subjectivity Accuracy
GPT-3.5	0.79 ± 0.01	0.61 ± 0.03
GPT-4	0.69 ± 0.01	0.65 ± 0.06
Llama 2	0.61 ± 0.02	0.47 ± 0.01
Mixtral	0.73 ± 0.09	0.50 ± 0.02
Falcon	0.71 ± 0.01	0.21 ± 0.05

Performance of polarity tasks is limited by the lack of offshore wind specific reasoning. Despite strong performance shown in some LLMs to understand technical jargon, the models remain limited by a lack of specialist knowledge for offshore wind in the UK. Difficulties include misinterpreting the sentiment of sentences by analysing from the wrong perspective. This arises from the challenge of locating enough latent concepts for the model to learn the optimal, correct output. For instance, ‘Wind has grown rapidly in the past 10 years’

is falsely assigned as neutral due to being seen as factual instead of implying the successful technological progression of offshore wind. Other challenges include being specific to offshore wind in the UK: *'Biomass has significant barriers to delivery'* is consistently and incorrectly interpreted as negative. These issues can be improved by adapting the prompt design to leverage different perspectives—this is explored in the following section.

Performance of subjectivity tasks are limited by misunderstanding the definition of subjectivity for technology development. The challenges associated with the subjectivity task include incorrectly identifying words that indicate uncertainty or objectivity, as well as limited reasoning for the decision. This corresponds to the same issue in the polarity task: the optimisation of latent concepts is crucial to acquire new perspectives and reasoning skills.

5.4. Multi-Agent Debate

This section investigates the performance of the multi-agent prompt technique across the top-performing LLMs from self-consistency: GPT-3.5-Turbo, GPT-4-Turbo and Mixtral. Multi-agent debate involves a number of parameters that require tuning: the agent roles, communication strategy and choice of LLM for each agent. The generator for MAD is chosen to be the Mixtral model for both tasks, due to obtaining the highest accuracy amongst all open-source models. Implementing the one-by-one debate strategy, the following details the selection of role prompts, debater, and reflector LLM agent.

5.4.1. Debate Metrics

Multi-agent debate introduces intricate dynamics through the interaction of multiple LLMs and diverse roles. This involves numerous hyperparameters for tuning the quality of debate. Therefore, this complex system requires unique ways of assessing a model's approach to effective interactive reasoning. Table 4 lists the metrics used to analyse debating behaviour in the subsequent experiments. Below, we explain the significance of these metrics.

- **Accuracy** is important to quantify how well the model performs against the ground truth dataset, a metric already used for the self-consistency results.
- **Answers Changed** is used to measure the confidence of the model in changing the previous model's answer. If the answers changed is low, this could either mean an accurate consensus has been reached between the two agents, or it could imply that the debater model struggles to impart its own opinion. Likewise, if the answers changed is high, this could mean the previous model's answers were mostly incorrect or that the debater agent is overly confident. This behaviour significantly affects the debating performance and understanding such metric is highly important when designing the prompt and selecting the LLM.
- To differentiate these varying interpretations, the **Answers Changed when Correctly Parsed** metric indicates the loss incurred by the debater agent from reevaluating the previous model's correct decisions. A low percentage is desirable since this means the debater model leads to more accurate conclusions. Contrasting this, a high percentage implies that arbitrary changes have been made by the debater without a strategic basis—this would require design improvement to enhance the model's decision-making accuracy.

Table 4. Metrics for analysing debating behaviour in multi-agent debate prompting.

Metric	Description
Accuracy (ACC)	Percentage of correctly answered predictions
Answers Changed (AC)	Percentage of predictions that changed from the previous debate round
Answers Changed When Correctly Parsed (ACCP)	Percentage of predictions that changed from the previous debate round, given correctly parsed answers

5.4.2. Heterogeneous Role Prompts

As explored in [29,31], heterogeneous role prompts are critical for the objective critique of opinions in order to reach an accurate consensus. To improve upon the limitation of specialised knowledge in self-consistency, this section investigates the performance of energy-specific role prompts for the debater agent in MAD by comparing them to homogeneous prompts. The homogeneous prompts contain only the debate task, without the role. These roles are designed using the error-based approach discussed in [29], using the limitations in each task as domain knowledge to acquire new capabilities. Unlike in CoT prompts, role prompting acts to guide the quality and diversity of the model's output—essential for improving reasoning.

Given the lack of energy-specific reasoning in the polarity tasks, agents specialised in aspects of offshore wind energy are created for the polarity task, shown in Figure 8. For the subjectivity task, Figure 9 translates the issue of identifying hope from fact into a prompt instruction. Figure 10 qualitatively shows how these heterogeneous role prompts can accurately reach the desired answer through critical debate: highlighting the ability to exploit their specialised perspective, leverage in-depth reasoning, and mitigate the effects of groupthink.

The Policy Agent: *Pay special attention to any government support, policies, or incentives*
The Technology Agent: *Pay special attention to any technological advancements*
The Market Agent: *Pay special attention to any growing demand for offshore wind energy*
The Supply-Chain Agent: *Pay special attention to any supply chain stages mentioned*
The Environment Agent: *Pay special attention to any environmental impacts mentioned*

Figure 8. Heterogeneous role prompts for the polarity task, where each agent specialises in different energy sectors, designed using the error-based approach in [29].

The Hope Agent: *Pay special attention to elements of hope or factualness in the language.*

Figure 9. Heterogeneous role prompt for the subjectivity task, targeted towards mitigating the limitation in misunderstanding the definition of subjectivity.

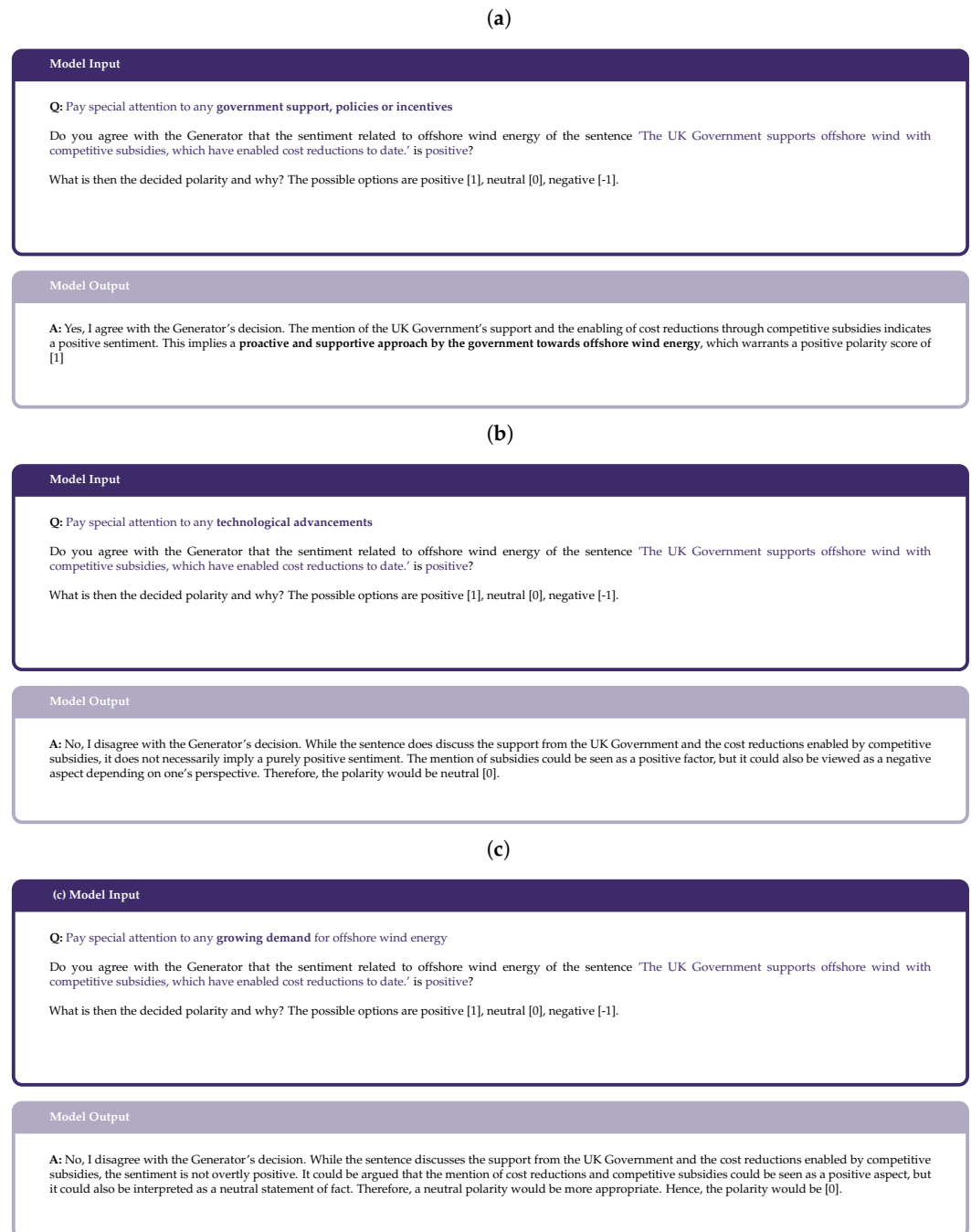


Figure 10. Demonstrations of the improvement in specialised heterogeneous role prompts, where the policy expert reaches the correct answer for the given policy-related input sentence. (a) Policy. (b) Technology agent. (c) Market agent.

5.4.3. Debater LLM Agent

The designs of these heterogeneous prompt designs are tested on the GPT-3.5-Turbo, GPT-4-Turbo, and Mixtral models to deduce which LLM performs with the highest accuracy and lowest ACCP on the debate tasks—this is the model that decides on the most accurate conclusions. This experiment uses one debate round—further rounds are investigated. However, this leads to a reduction in accuracy across all LLMs and is hence not further explored (Table 5 presents these results).

GPT-3.5 and GPT-4 perform well as debaters for both polarity and subjectivity tasks. As shown in Table 5, GPT-3.5 and GPT-4 both exhibit the highest accuracy across both polarity and subjectivity tasks after debate. Beyond accuracy, the optimal debater must also ensure

that it acts to improve upon the generator's answers. Ideally, the debater changes only the answers that are incorrect. This is identified by an ACCP of 0. For the polarity task, GPT-3.5 and GPT-4 change approximately 10% of the generator's values. This is adequate to improve performance but ways of increasing this would be highly desirable. Importantly, GPT-3.5 incorrectly changes less than 50% of the generator's answers, which implies that GPT-3.5 is most capable of reaching accurate conclusions compared to the other LLMs. However, it should be noted that this shows marginal performance improvement on the generator, and ways of enabling lower arbitrary reevaluations is essential for effective polarity debate. Contrasting this, the answers changed in the subjectivity task are significantly greater, and the loss incurred by incorrect changes is remarkably lower. GPT-4 obtains the lowest ACCP and highest accuracy compared to the other LLMs; hence, this is the best debater for subjectivity. Despite this, GPT-3.5 performs with an equivalently high accuracy at a much lower cost. In conclusion, GPT-3.5 is chosen to be the debater model for both polarity and subjectivity tasks due to its strong capability in strategic reevaluations and low cost.

Table 5. Accuracy (ACC), answers changed (AC) and answers changed when correctly parsed (ACCP) metrics of tested LLMs as debaters in MAD for polarity and subjectivity tasks. **Bold** indicates the best-performing result (\uparrow indicates higher values are better, and \downarrow that lower values are better).

Model	Polarity			Subjectivity		
	ACC \uparrow	AC	ACCP \downarrow	ACC \uparrow	AC	ACCP \downarrow
GPT-3.5	0.77 \pm 0.01	0.10 \pm 0.06	0.47 \pm 0.06	0.66 \pm 0.02	0.52 \pm 0.04	0.31 \pm 0.03
GPT-4	0.75 \pm 0.03	0.13 \pm 0.06	0.61 \pm 0.18	0.70 \pm 0.07	0.31 \pm 0.05	0.11 \pm 0.14
Mixtral	0.67 \pm 0.01	0.11 \pm 0.02	0.90 \pm 0.16	0.47 \pm 0.01	0.15 \pm 0.01	0.41 \pm 0.03

Mixtral performs poorly as a debater and struggles to reach consensus. Evident in Table 5, Mixtral can be observed to be a highly defective debater—it makes arbitrary changes to the generator's answers across both tasks, causing accuracy to counter-actively reduce. Possible explanations for this deficient behaviour is that Mixtral struggles to retain prompt information as a debater, identify technical jargon and is limited by context length. These factors prevent Mixtral from reaching justified conclusions.

The results from Mixtral indicate a poor ability as a MAD debater. Additional research is conducted to investigate the debater ability of the remaining SLMs used throughout these experiments. Results show that debater SLMs give lower accuracy than single SLMs, suggesting that the composition of debaters in MAD is not suited to SLMs.

5.5. Sentiment Analysis Pipeline

Following experiments from self-consistency and multi-agent debate, a summary of the best-performing results is given in Table 6. Three prompt strategies from self-consistency and multi-agent debate are compared in order to determine the pipeline that most accurately and affordably performs sentiment analysis over the sampled dataset.

Table 6 indicates that the multi-agent debate design performs with the highest subjectivity accuracy and second-highest polarity accuracy at a significantly lower cost than the GPT-3.5 self-consistency framework. This MAD design improves upon the smaller Mixtral self-consistency model via debate with GPT-3.5 in both polarity and subjectivity, along with minimal deviation between repeated results. In addition, MAD requires fewer API calls, which reduces the total costs and computational complexity to run sentiment analysis. As the environmental and financial costs of language models increase, designing a system of LLMs that require fewer resources whilst performing at a competitive level with larger, more-capable models becomes highly important. In conclusion, the MAD Mixtral GPT-3.5 sentiment analysis pipeline performs most optimally compared to the

self-consistency results, achieving competitive accuracy performance at a significantly lower cost.

Table 6. Comparison of accuracy and cost for polarity and subjectivity tasks between GPT-3.5 self-consistency, Mixtral self-consistency, and MAD. **Bold** indicates the highest-accuracy result.

Prompt Strategy	SC GPT-3.5	SC Mixtral	MAD Mixtral GPT-3.5
Polarity Accuracy	0.79 ± 0.01	0.73 ± 0.09	0.77 ± 0.01
Subjectivity Accuracy	0.61 ± 0.03	0.50 ± 0.02	0.66 ± 0.02
Polarity Cost (/1M tokens)	\$6	\$0	\$2
Subjectivity Cost (/1M tokens)	\$6	\$0	\$2

5.6. Introducing EnergyEval

Figure 11 illustrates the novel multi-agent debate team designed specifically for evaluating the sentiment, subjectivity, and topics within the technical energy literature—called EnergyEval. This particular design has been applied to offshore wind energy in the UK. The framework has the following characteristics:

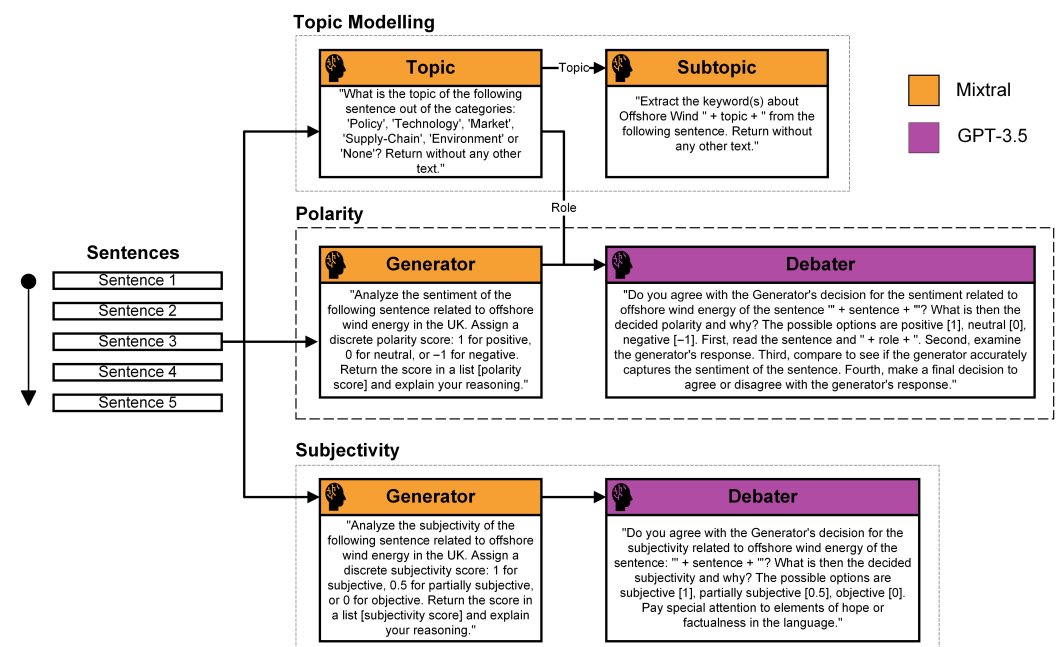


Figure 11. EnergyEval sentiment analysis pipeline—illustrating the topic modelling, polarity and subjectivity task, along with least-to-most and homogeneous role prompting in the debater and use of the Mixtral generator and GPT-3.5 debater.

- Multi-agent debate in the evaluation of sentiment and subjectivity, leveraging heterogeneous energy-specialised role prompts to improve upon accuracy and offshore wind-specific reasoning.
- Effective debate has been encouraged through a least-to-most prompt design in the polarity task, structuring an unbiased and logical debate dialogue to reduce hallucinations and the detriment of groupthink.
- Mixtral and GPT-3.5 are employed in this design as the generator and debater, respectively, in which the use of different LLMs has not been explored yet in MAD sentiment analysis [32].
- The topic modelling task follows a single-prompting strategy using the Mixtral model due to the simplicity of its instruction.

6. Results and Analysis

6.1. Overview

This section executes the EnergyEval model on the offshore wind dataset in Table 1 aiming at predicting trends and drivers of wind power use from 2008 to 2020. The motivation for this comparison is the feedback loop that exists between scientific opinion in a specific time and deployment in later time. As evidence for technology maturity increases, so will positive scientific opinion, which then influences the tangible decisions of stakeholders. Figure 12 explains the process used to predict these trends from EnergyEval. In this, the dataset is first grouped based on four topics, and then labelled based on polarity and subjectivity measurements. These topical groupings include (a) policy, (b) technology, (c) market, and (d) supply chain. The key drivers of Offshore Wind Development through the years in each of the above topics are identified, and then extracted from each sentence. Based on the rate of appearance in the labelled data, the most-frequent drivers are selected to be analysed. These include (a) cost, (b) capacity, and (c) turbine size.

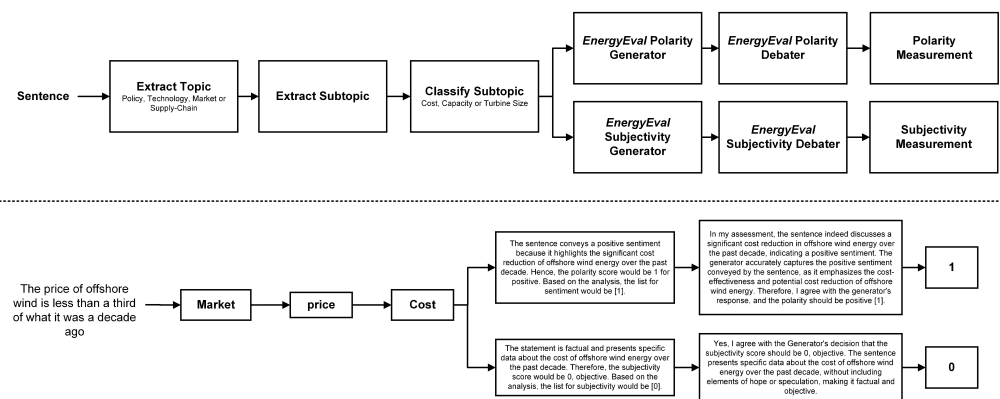


Figure 12. Pipeline used to predict the offshore wind trends from EnergyEval, along with an example.

The next sections investigate whether sentiment analysis using EnergyEval can reflect the actual trends of these drivers. This includes a comparison with trends in levelised cost of electricity, installed capacity, and turbine size installation in the UK from 2008 to 2020.

6.2. Task 1: High Level Trends

To begin with, this section analyses the high-level sentiment trends from the labelled offshore wind dataset, produced by EnergyEval. Since 2000, offshore wind in the UK has grown into a mature, affordable technology and has become a critical source of renewable energy in the UK's route to net-zero. Therefore, this success in deployment indicates an increase in positive and objective sentiment.

To verify that this impressive growth is present in the literature, Figure 13 depicts these trends from the frequency of polarity and subjectivity classifications. The frequency is measured by the number of occurrences in each sentiment label—representing the popularity of opinion that EnergyEval extracted. This seeks to quantify a global opinion about offshore wind from the local, sentence-level sentiment evaluations. For the frequencies to range between 0 and 1, the *minmax* scaling method in Equation (2) is used to transform the data in Figure 13 to this scale:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

Looking at Figure 13a, we note that EnergyEval demonstrates a gradual increase in positive sentiment with time (yellow line), and associated reduction in negative sentiment (blue line). This confirms the expected increase in positivity. There is no observable trend

6.4. Task 3: Analysis of the Drivers

6.4.1. Cost

Figure 15a compares the positive sentiment, measured from the *cost*-associated sentences (blue and purple lines), with the trend in levelised cost of electricity (yellow line labelled as “LCOE”)

$$|(\text{Positive} \cap \text{Objective})|\text{Cost}| = \frac{|\text{Positive} \cap \text{Objective} \cap \text{Cost}|}{|\text{Cost}|} \quad (3)$$

$$|\text{Cost}|(\text{Positive} \cap \text{Policy})| = \frac{|\text{Positive} \cap \text{Policy} \cap \text{Cost}|}{|\text{Positive} \cap \text{Policy}|} \quad (4)$$

For offshore wind in the UK, from the EnergyEval sentiment analysis, the proportion of positive sentiment in *cost* (purple line) is used to compare with the LCOE data in Figure 1. The proportion of positive and objective sentiment is also included in this comparison (blue line) to determine whether factual scientific opinion from EnergyEval more accurately corresponds to real development. These frequencies are calculated using Equation (3), and then normalised using Equation (2).

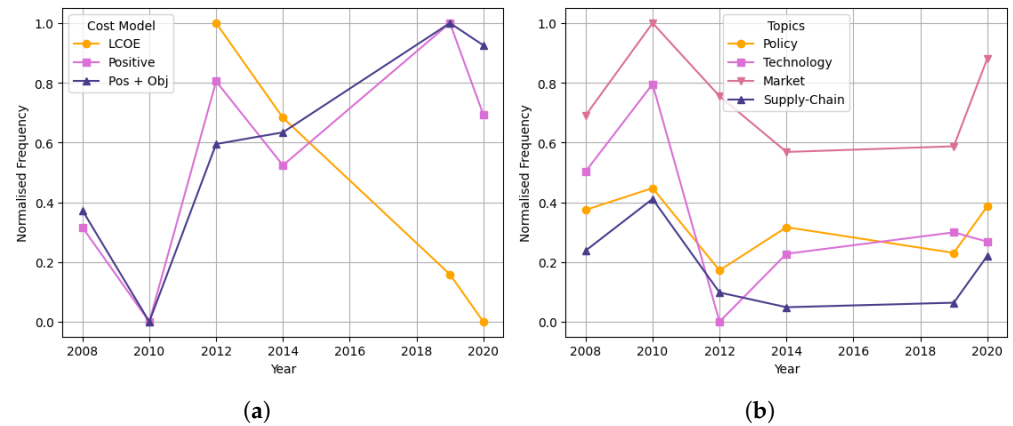


Figure 15. EnergyEval: Cost-associated positive sentiment (a) and frequency of topics (b).

Figure 15b presents the evolution in the frequency of *cost*-associated sentences that contributed to the positive sentiment of each offshore wind sector. Calculated using Equation (4), this represents the positive influence that *cost* has on the development of policy, technology, market and supply chain.

EnergyEval correctly predicts the increase in positive opinions as cost falls, representative data essential for accurate results. Evidenced in Figure 15a, both models accurately detect the trends in LCOE reduction. Both EnergyEval and the input energy literature reflect this increase in positive sentiment towards declining costs. For instance, the 2008 literature mentions that ‘capital costs have more than doubled over the last five years’, which EnergyEval correctly identifies as negative. Hence, the good agreement between predicted cost trends and LCOE are a result of the accuracy of EnergyEval and the representative cost-related opinions of the input data. A significant decrease in the positive sentiment associated with cost in 2010 is also evident in Figure 15a. The reasoning for this is that the 2010 document, *Great Expectations: The Cost of Offshore Wind in UK Waters*, is primarily focused on cost. By introducing an in-depth assessment on cost, the information gives rise to an equal distribution of positive and negative opinions. Hence, this leads to the lowest proportion of positive sentiment. This issue results from the input data itself not being representative enough for that year.

EnergyEval’s positive-objective sentiment for cost gives more accurate results to the true LCOE data, compared to only-positive. Observed in Figure 15a, the positive-objective measurement exhibits less rapid fluctuations in sentiment throughout 2008–2020. This is because the additional objective layer filters out the data that is less direct and factual. Hence, the sentiment is more likely to be reliable and true to the real data since these opinions are more probable to happen or have already happened.

Cost has a strong influence on all sectors, especially on the Market. As expected, Figure 15b shows that Market gives the highest proportion of positive sentiment from cost. This is anticipated since the demand for offshore wind is intrinsically linked to fluctuations in cost. In addition, all topics follow similar trends with time—this indicates the strong influence that cost has on all sectors. It should also be noted that the high frequency of cost in 2010 is a result of the cost-skewed input data.

6.4.2. Capacity

Figure 16a displays the comparison between *capacity*-associated positive and positive-objective sentiments with the offshore wind capacity installed each year. This is calculated using Equation (3) for *capacity*. To visualise the impact of *capacity* on offshore wind, Equation (4) is used on each topic to determine which sectors were influenced the most by increases in installed capacity—shown in Figure 16b.

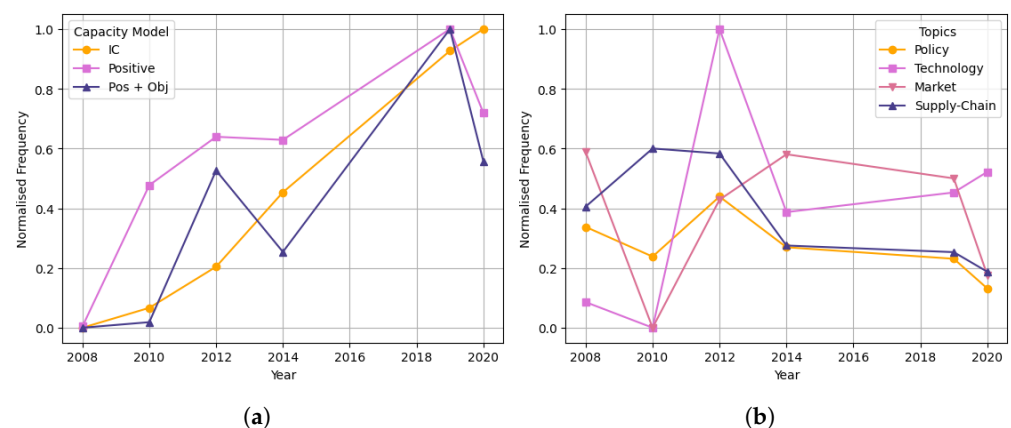


Figure 16. EnergyEval: Capacity-associated positive sentiment (a) and frequency of topics (b).

EnergyEval correctly detects the expected increasing trend for installed capacity. However, whilst the general predicted trend follows this increase correctly, certain fluctuations are evident in Figure 16a. Specifically, results from 2020 imply a reduction in capacity—which is incorrect. The slight inaccuracy of this model is not that the literature itself is suggesting negative sentiment towards capacity progress; it is the frequency-based measurement that causes small changes in opinion to greatly affect the overall assessment. This result arose from the frequency-based calculation of capacity-associated positive sentiment. By taking the count of positive sentiment as a fraction of the total sentiment, we are not measuring the global sentiment of a certain topic, only the number of times a report mentions a positive attribute of offshore wind for each sentence. The significant reductions seen in 2020 are attributed to the lower frequency of positive sentiment in the text compared to the previous reports in each year. To improve on this measurement metric, a way to quantify global sentiment of ‘offshore wind’ needs to be calculated independent of the size of the report. A likewise observation can be seen in Figure 15a for cost-associated sentiment.

EnergyEval's positive-objective sentiment for capacity gives more accurate results to the true data, compared to only-positive. A key observation from Figure 16 is that, like cost, the positive-objective measurement more closely aligns with real installed capacity data, which is more realistic and less hopefully optimistic than only-positive as expected. Similar to Figure 15a, the positive-objective model appears to reduce irrelevant opinions contributing to the prediction. Compared to cost, Figure 16a displays a larger difference between these two predictions. From observation of the labelled data, this is found to arise from the large quantity of unopinionated sentences in the only-positive sentiment. EnergyEval correctly filters these out in the subjectivity classification, responsible for the gap in predictions.

Installed Capacity has a positive influence on Technology. Figure 16b indicates that the increase in capacity has positively influenced developments in technology. As required, this parallels how greater power outputs have been enabled through larger turbines. This signifies the capability of EnergyEval to correctly interpret key trends from the literature.

6.4.3. Turbine Size

Figure 17a illustrates the comparison between *turbine-size-associated* positive (purple line) and positive-objective (blue) sentiments with the largest turbine installed offshore (yellow line). This is calculated using Equation (3) for the *turbine size* keyword. For the impact of *turbine size* on offshore wind, Equation (4) is used to identify which sectors are influenced the most by increases in turbine size—results are shown in Figure 17b.

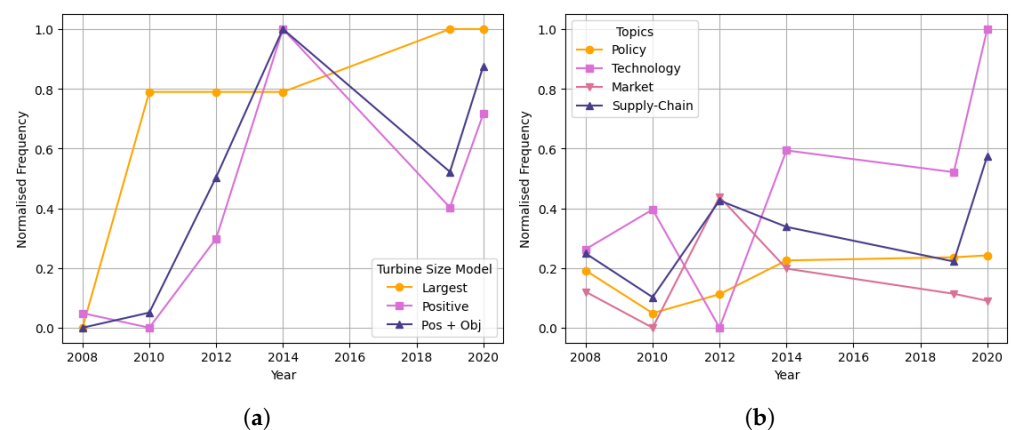


Figure 17. EnergyEval: Turbine-size-associated positive sentiment (a) and frequency of topics (b).

EnergyEval's positive-objective sentiment for turbine size gives more accurate results to the true data. Consistent with the observations from Figures 15–17 denotes that the positive-objective model more closely approaches the true turbine size data. What can be further noticed is that the gap between positive and positive-objective measurements is smaller compared to the cost and capacity predictions. The reasoning for this is that turbine-related sentences tend to be factual and unopinionated due to its technical descriptions. This inherent factualness explains why the difference between objective and non-objective predictions is so minimal.

Turbine size strongly influences developments in technology. EnergyEval accurately captures the increasing trend in technology. This is expected due to turbine size being fundamental to its development, from rotor diameter to turbine height. The discontinuity in 2012 is a result of the limited dataset size that falsely influences the output with its lack of representation.

7. Discussion

This section discusses the future potential of EnergyEval on further net-zero technologies, as well as the limitations that should be addressed prior to these practical applications.

7.1. *EnergyEval* in Practice

7.1.1. Extension to Net-Zero Technologies

In order to apply EnergyEval on further low-carbon technologies, redesigning the heterogeneous role prompts would be the only adaptation required. This is because these are the components specialised to the energy technology—defining the area of expertise of the model. Advantageously, this requires significantly less time to specialise the model to its particular use case compared to time-consuming alternatives like fine-tuning or training deep learning models. With labelled data for energy scarcity, LLMs are an essential tool for sentiment analysis, as they eliminate the need for training datasets. Therefore, EnergyEval holds strong potential for further application.

7.1.2. Extension to Multi-Modal LLMs

In the next frontier of large language models, multi-modal LLMs offer a host of greater capabilities beyond traditional LLMs. For the purposes of EnergyEval, integration with multimedia inputs could enable the analysis of energy-related podcasts, documentaries or images. This would improve dataset representation and could even lead to more realistic sentiment predictions.

7.1.3. Comparison to Existent Methods

As part of our preliminary investigations, we evaluate a semantic lexicon-based approach. However, this method performs poorly in our domain-specific context. The main limitation is that generic sentiment lexicons lack coverage of offshore wind-related terminology, especially nuanced indicators of cost, capacity, and public perception. As a result, many relevant positive or negative signals are missed, leading to significantly lower accuracy compared to our LLM-based approach. A more detailed comparison with existent models would be beneficial as part of future work for contextualising performance.

7.2. *Limitations of EnergyEval*

7.2.1. Gathering Reliable Energy Data

A key limitation noticed in the offshore wind results is the influence of energy data on predictions. The lack of representation in the dataset leads to unexpected fluctuations in the forecasted trends. Hence, it is critical to gather a variety of diverse literature for the model to deduce informed sentiments. This means acquiring data from a range of different sources and years, along with ensuring each year contains sufficient data on the examinable features.

It should be also repeated here that while our goal is indeed to evaluate and simulate expert-level predictions using LLMs, in the current investigation, we incorporate LLMs in the ground truth construction process primarily for practical reasons—namely, to reduce the time and cost associated with manual expert annotation. Obtaining a sufficiently large volume of expert-labelled data is both resource intensive and time-consuming. To address this, we adopt a hybrid labelling strategy: LLMs are first used to generate initial annotations, which are then reviewed, refined, or corrected by human annotators. This approach enables us to benefit from the efficiency of LLMs while ensuring quality and reliability through human oversight. A different approach might be required in future studies.

7.2.2. Extracting Broader Sentiment Information

Another limitation observed is the sensitivity in using frequency measurements to quantify global sentiment opinions from sentence-level analysis. Here, the 2020 results in Figure 16a show that minor differences in the frequency of capacity-related positive sentiment lead to large fluctuations when scaled against the other years. This issue of this is that sentence-level sentiment analysis does not capture the overall sentiment of aspects in the document, and document-level analysis is not possible due to context limitations. Possibly, an additional LLM layer could be used to extract this broader sentiment information from the sentence-level analysis. This approach would integrate and prioritise multiple opinions within the document to evaluate a global sentiment. Overall, further research into document-level sentiment analysis using LLMs is needed for more advanced results.

7.2.3. Increasing the Number of Debate Rounds

In attempt to improve the polarity debate accuracy, increasing the number of debate rounds could be a way to enhance debater capability. Due to the time constraints of this project, the effect of further debate rounds is not investigated. Recent MAD studies [30,32] indicate performance improvement with increased rounds. Therefore, a promising research direction involves exploring the influence of greater debate rounds on the ability of GPT-3.5 to reach accurate conclusions during debate.

7.2.4. Using Smaller, Energy-Efficient LLMs

Whilst EnergyEval utilises the smaller, open-source Mixtral model, an interesting research question would be to determine whether GPT-3.5 could be replaced by a smaller language model. Motivated by the environmental and financial costs of exceedingly large LLMs, leveraging performance improvements from MAD could allow smaller models to compete with larger ones at a lower cost. Future research could investigate the effect of different SLMs on EnergyEval.

8. Conclusions

In this study, we present a novel multi-agent large language model (LLM) debate framework—EnergyEval—designed to evaluate polarity, subjectivity, and thematic trends within technical energy literature. The novelty of our approach lies in optimising multi-agent debate mechanisms for energy-specific data, leveraging heterogeneous LLMs with specialised role prompts. We employ one-shot chain-of-thought and least-to-most prompting strategies to enhance reasoning accuracy and mitigate hallucinations. Among the evaluated models, Mixtral and GPT-3.5 emerge as the most effective generator and debater, respectively, offering a strong balance between accuracy and computational efficiency.

We apply EnergyEval to offshore wind literature, where sentiment analysis revealed that the positive and objective sentiment trends identified by the model broadly aligned with real-world offshore wind indicators such as the Levelised Cost of Energy and installed capacity. Furthermore, EnergyEval effectively identifies key thematic drivers—cost, capacity, and turbine size—and capture how domains such as policy, technology, market, and supply chain evolved in response to these factors. These findings highlight the value of robust, diverse, and large-scale energy datasets for accurate time-series predictions. They also suggest that aggregating broader sentiment information can address the limitations of sentence-level sentiment analysis, improving our understanding of high-level sectoral trends.

While our results are encouraging, it is important to recognize that data quality and potential model hallucinations remain ongoing challenges in LLM applications. As previously discussed, these limitations can affect output reliability. However, our proposed multi-agent debate (MAD) framework mitigates such risks through dynamic cross-verification among agents. By allowing agents to critically evaluate and challenge one another's reasoning, the framework reduces the propagation of erroneous or low-quality information, improving robustness and interpretability.

In conclusion, this study demonstrates the promise of EnergyEval for sentiment analysis within energy research, offering insights that extend across net-zero technologies. Future research should investigate whether increasing the number of debate rounds enhances GPT-3.5's reasoning effectiveness, and whether integrating an additional LLM aggregation layer can refine document-level sentiment predictions. Exploring EnergyEval's generalizability across other energy sectors and regions also presents a valuable direction, given its modular design and adaptability through prompt and evaluation customisation.

Author Contributions: Conceptualisation, H.F., K.V. and S.R.; Data curation, H.F.; Formal analysis, H.F., K.V. and S.R.; Methodology, H.F., K.V. and S.R.; Software, H.F.; Writing—original draft, H.F.; Writing—review and editing, K.V. and S.R. All authors have read and agreed to the published version of the manuscript.

Funding: K.V. would like to acknowledge the financial support of the UK Engineering and Physical Science Research Council through grant EP/Y004930/1.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompt Engineering

Prompt engineering is the process of structuring inputs [46], a critical technique in optimising the accuracy of LLMs. Below details a brief overview of the core methods used in prompt engineering.

1. **Comprehensive Description:** A detailed definition of the desired task is necessary in prompting. This is because a basic instruction lends itself to a number of possible options, whereas more descriptive prompts generate more precise results [46]. In the latent concept theory, this could be due to the fact that, by increasing the number of in-context concepts, the probability of a certain output given the prompt increases, narrowing to a more relevant answer.
2. **Few-shot Prompting:** This is when a number of examples are given in the prompt, alongside a description of the task. How many demonstrations are used has been found to vary from task-to-task. For simpler tasks, one-shot prompting is typically sufficient, whereas few-shot prompts tend to improve performance for more invented tasks [46]. The reasoning behind this depends on whether few-shot learns or recalls latent concepts. Despite evidence suggesting few-shot potential [10], zero-shot prompts have been shown to outperform few-shots in certain scenarios [47].
3. **Chain of Thought:** CoT works by prompting the model to provide intermediate reasoning steps in order to guide the response [46]. This approach has demonstrated highly significant improvement to accuracy, such as an increase from 17% to 78.7% for zero-shot arithmetic tasks [48].
 - (a) **Zero-shot CoT:** This is an instance of CoT where the task instruction is followed by 'Let's think step by step', with 'zero-shot' referring to there being no examples of reasoning steps.

- (b) **Few-shot CoT:** Likewise to few-shot prompting, few-shot CoT contains a number of manual demonstrations, along with reasoning included in the answer for each example. The disadvantage is the complexity in implementing these examples, requiring correct manual reasoning.
4. **Self-Consistency:** The self-consistency prompting method aims to ensure the model's responses are consistent with each other by replacing the naive greedy decoding used in chain-of-thought [49]. This follows a 3-stage process. The first step prompts the language model using chain of thought; the second samples multiple, diverse responses to the same prompt. The most consistent answer is found using the majority vote over all reasoning paths.
 5. **Generated Knowledge:** This approach prompts the model to generate potentially useful information prior to the task instruction [46], leveraging the pre-trained knowledge within LLMs to provide additional context. This method could be interpreted to increase the output probability by adding more latent concepts to the context from its pre-trained 'subconscious'.
 6. **Least-to-Most:** Least-to-most prompting consists of a series of prompts that build upon each other, where each represent a fundamental step to the desired task—reducing 'cognitive overload'.
 7. **Tree of Thoughts:** ToT follows a problem-solving process: an initial input prompts the model to outline the steps to solve the problem, and the subsequent steps allow the model to automatically approach the problem step-by-step. Unlike least-to-most, this methodology is based on the thought-process prior to the task—hence, more versatile and less-specific to a certain problem.
 8. **Retrieval Augmentation:** This incorporates up-to-date external information into the prompt, treated as foundational knowledge for the model [46]. The aim of this method is to reduce hallucinations.
 9. **Role Prompting:** This method involves assigning the model a specific role to play, which has been shown to be effective in aligning to a desired output [46].

References

1. Poynting, M. Climate Change: Is the UK on Track to Meet Its Net Zero Targets? May 2024. Available online: <https://www.bbc.co.uk/news/58160547> (accessed on 3 April 2024).
2. McNally, P. An Efficient Energy Transition: Lessons from the UK's Offshore Wind Rollout. February 2022. Available online: <https://www.institute.global/insights/climate-and-energy/efficient-energy-transition-lessons-uks-offshore-wind-rollout> (accessed on 1 January 2024).
3. OpenAI. GPT-3.5 Turbo. Available online: <https://chatgpt.com/pricing> (accessed on 14 February 2024).
4. Bloomberg. Introducing BloombergGPT. March 2023. Available online: <https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/> (accessed on 4 April 2024).
5. Research, G. Med-PaLM. March 2023. Available online: <https://sites.research.google/med-palm/> (accessed on 3 April 2024).
6. DeepMind, G. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* **2024**, arXiv:2312.11805.
7. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Rozière, N.; Goyal, N.; Hambro, E.; Azhar, A.; Rodriguez, A.J.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
8. Luk, M. Generative AI: Overview, Economic Impact, and Applications in Asset Management; SSRN. September 2023. Available online: <https://ssrn.com/abstract=4574814> (accessed on 1 December 2023).
9. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv* **2023**, arXiv:2205.11916. [CrossRef]
10. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]
11. Zhang, Z.; Liu, M.; Sun, M.; Deng, R.; Cheng, P.; Niyato, D.; Chow, M.Y.; Chen, J. Vulnerability of Machine Learning Approaches Applied in IoT-based Smart Grid: A Review. *arXiv* **2023**, arXiv:2308.15736v3. [CrossRef]

12. Zhang, Z.; Yang, Z.; Yau, D.K.Y.; Tian, Y.; Ma, J. Data security of machine learning applied in low-carbon smart grid: A formal model for the physics-constrained robustness. *Appl. Energy* **2023**, *347*, 121405. [CrossRef]
13. Poynting, M. What are fossil fuels? Where does the UK get its energy from? *BBC News*, 12 March 2024.
14. Department for Energy Security & Net Zero. *UK Energy in Brief 2023*; National Statistics: Newport, UK, 2023.
15. Department for Energy Security & Net Zero. The Ten Point Plan for a Green Industrial Revolution. November 2020. Available online: <https://www.gov.uk/government/publications/the-ten-point-plan-for-a-green-industrial-revolution/title> (accessed on 29 April 2024).
16. Ploszajski, A. The UK's Future Energy Mix: What Will It Look Like If We're to Reach Net Zero? Available online: <https://www.imperial.ac.uk/Stories/future-energy> (accessed on 29 April 2024).
17. UK Offshore Wind History. Available online: <https://guidetoanoffshorewindfarm.com/uk-offshore-wind-history/> (accessed on 1 May 2024).
18. *Developments in Wind Power*; Houses of Parliament Parliamentary Office of Science and Technology: London, UK, 2019.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
20. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Technical Report, OpenAI. 2018. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 December 2023).
21. Xie, S.M.; Raghunathan, A.; Liang, P.; Ma, T. An Explanation of In-context Learning as Implicit Bayesian Inference. *arXiv* **2022**, arXiv:2111.02080. [CrossRef]
22. von Oswald, J.; Niklasson, E.; Randazzo, E.; Sacramento, J.; Mordvintsev, A.; Zhmoginov, A.; Vladymyrov, M. Transformers Learn In-Context by Gradient Descent. *arXiv* **2023**, arXiv:2212.07677. [CrossRef]
23. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual, 3–10 March 2021; pp. 610–623. [CrossRef]
24. Kalai, A.T.; Vempala, S.; Nachum, O.; Zhang, E. Why Language Models Hallucinate. *arXiv* **2025**, arXiv:2509.04664. [PubMed]
25. Bailey, N. The Carbon Footprint of LLMs—A Disaster in Waiting, September 2024. Available online: <https://nathanbaileyw.medium.com/the-carbon-footprint-of-llms-a-disaster-in-waiting-6fc666235cd0> (accessed on 27 September 2024).
26. Liu, B. *Sentiment Analysis and Opinion Mining*; Springer: Berlin/Heidelberg, Germany, 2012.
27. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.J.; Bing, L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. *arXiv* **2023**, arXiv:2305.15005. [CrossRef]
28. Fatemi, S.; Hu, Y. A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis. *arXiv* **2023**, arXiv:2312.08725.
29. Xing, F. Designing Heterogeneous LLM Agents for Financial Sentiment Analysis. *arXiv* **2024**, arXiv:2401.05799. [CrossRef]
30. Xing, F. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv* **2023**, arXiv:2305.14325. [CrossRef]
31. Chan, C.M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; Liu, Z. ChatEval: Towards Better LLM-Based Evaluators Through Multi-Agent Debate. *arXiv* **2023**, arXiv:2308.07201.
32. Zhang, J.; Xu, X.; Zhang, N.; Liu, R.; Hooi, B.; Deng, S. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. *arXiv* **2024**, arXiv:2310.02124. [CrossRef]
33. Sun, X.; Li, X.; Zhang, S.; Wang, S.; Wu, F.; Li, J.; Zhang, T.; Wang, G. Sentiment Analysis through LLM Negotiations. *arXiv* **2023**, arXiv:2311.01876. [CrossRef]
34. Jennings, T.; Tipper, H.; Daglish, J.; Grubb, M.; Drummond, P. *Policy, Innovation and Cost Reduction in UK Offshore Wind*; Technical Report; The Carbon Trust: London, UK, 2020.
35. Whitmarsh, M.; Canning, C.; Ellson, T.; Sinclair, V.; Thorogood, M. The UK Offshore Wind Industry: Supply Chain Review. 2019. Available online: https://safety4sea.com/wp-content/uploads/2019/02/OWIC-The-UK-Offshore-Wind-Industry-Supply-Chain-Review-2019_02.pdf (accessed on 27 September 2024).
36. Trewby, J.; Kemp, R.; Green, R.; Harrison, G.; Gross, R.; Smith, R.; Williamson, S. *Wind Energy: Implications of Large-Scale Deployment on the GB Electricity System*; Royal Academy of Engineering: London, UK, 2014.
37. Spencer, M.; Tipper, W.A.; Coats, E. *UK Offshore Wind in the 2020s: Creating the Conditions for Cost Effective Decarbonisation*; Green Alliance: London, UK, 2014.
38. Chinn, M. *The UK Offshore Wind Supply Chain: A Review of Opportunities and Barriers*; Offshore Wind Industry Council: London, UK, 2014.
39. *Wind Energy in the UK: State of Industry Report*; Renewable UK: London, UK, 2012.
40. Greenacre, P.; Gross, R.; Heptonstall, P. *Great Expectations: The Cost of Offshore Wind in UK Waters—Understanding the Past and Projecting the Future*; UK Energy Research Centre: London, UK, 2010.

41. Jennings, T.; Delay, T. *Offshore Wind Power: Big Challenge, Big Opportunity*; The Carbon Trust: London, UK, 2008.
42. OpenAI. GPT-4 Turbo. Available online: <https://chatgpt.com/pricing> (accessed on 14 February 2024).
43. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [[CrossRef](#)]
44. Jiang, A.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Hanna, E.B.; Bressand, F.; et al. Mixtral-8x7B-Instruct. 2023. Available online: <https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1> (accessed on 14 February 2024).
45. Falcon. Falcon-7b-Instruct. Available online: <https://huggingface.co/tiiuae/falcon-7b-instruct> (accessed on 14 February 2024).
46. Chen, F.; Zhang, Z. Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review. *arXiv* **2023**, arXiv:2310.14735. [[CrossRef](#)]
47. Reynolds, L.; McDonell, K. Prompt programming for large language models: Beyond the few-shot paradigm. In Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–7.
48. Mu, J. Stanford CS224N | 2023 | Lecture 10—Prompting, Reinforcement Learning from Human Feedback. 2023. Available online: https://www.youtube.com/watch?v=SXpJ9EmG3s4&ab_channel=StanfordOnline (accessed on 29 January 2024).
49. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv* **2023**, arXiv:2203.11171. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.