

Using Large Language Models to Measure Impartiality in the Media: ChatGpt-4 and the BBC's Coverage of the Israel-Gaza War as a Case Study

Haran Shani-Narkiss

Acknowledgments

Following the events of October 7th, 2023, and amid controversy over the coverage of the Israel-Hamas war by global news outlets, a group of researchers, lawyers, analysts and assistants formed RIME - Research for Impartial Media - to empirically test for bias in media reporting. They contributed to this work, as well as to other similar attempts, in numerous ways, including data collection, data clearing, data labelling, engaging in discussions, exchanging ideas, and framing the research questions and methodologies needed for rigorous analyses. Their meticulous work and precise scrutiny were invaluable for a comprehensive assessment of this complicated question. These people, marked by diligence and integrity, played a crucial role in highlighting the complexities of quantifying media bias. Their dedication and support have significantly impacted the understanding of this issue, as reflected in the current work.

These are the people of RIME:

Founder and Group Leader - Dr. Haran Shani-Narkiss

Members - Dr. MS, Dr. Oren Tsur, Dr. YKT, Reut Harari-Misgav, Dr. Ariel Goldstein, Khaled Hassan, Dr. MF, Dr. SG.

Contributors – NC, Michael Kan-Tor, Avia S., Nitzan Luxembourg, Sapir Shapira, Eitan Rothenberg, Oded Maor Landaw, Dr. Renana Keydar, Roni Matzkin, Max Cohen, H. Ashkenazi, Gili Rolland, Dr. Jonatan Zalcman, Noa Lubin, Dr. LL, Milka Kaplan, Chen Laster Alon, Liel Binyamin, ID, Tom Marzea, Itamar Aharoni, Nir Hamou, Itay Alon, Guy Niran, Gye kuglovich, Ariel Kirtchuk, IG, Omer Neu, Ouri Levy, IA, IH.

Introduction

The BBC

Some media outlets do not even pretend to be impartial. The public voice of dictatorships is biased, and has limited influence on the public debate, because it is easily dismissed. On the other hand, quality journalism strives for credibility, and so seek to write in an impartial, accurate and balanced manner. The BBC serves as a particularly good subject for studying media-impartiality and media-bias, for several reasons; first, it is legally bound¹ to be impartial and bases its strong reputation on a claim that it complies with that obligation. Second, the BBC is highly influential; having more than a billion monthly visits to its English website alone, it is the most popular news outlet in the world, leading by a wide margin². Lastly, the BBC broadcasts across several languages and formats, enabling experimental research over various settings.

The Importance of Sympathy and Impartial Framing in the Media

Impartial media aims to provide balanced and unbiased reporting, offering comprehensive perspectives on any given issue. Bias can take many forms in media coverage (Rodrigo-Ginés et al., 2023). Still, one of the most insidious is the unequal sympathy extended to the various parties in a conflict, which undermines the principle of journalistic impartiality.

Evoking sympathy for both sides of a conflict does not mean condoning or justifying harmful actions. Instead, it means acknowledging the humanity and experiences of all parties involved. This approach promotes a more profound understanding, fosters empathy, and supports the ethical standards of journalism, ultimately contributing to a more informed and less polarized public.

Creating equal sympathy for different sides in a conflict ensures that all voices and perspectives are represented and acknowledged (Entman, 2007), allowing a nuanced understanding and reducing dehumanization (Haslam, 2006). This is important for illuminating the root causes of conflicts and fostering a dialogue that promotes peaceful resolutions.

Conversely, by promoting a specific agenda or aligning with one group, the media perpetuates negative stereotypes and prejudices that may exist between different groups involved in the conflict. This may promote false narratives and misinformation, which can further escalate tensions and hinder efforts towards resolution, enforcing an "Us vs. Them" mentality (Entman, 2007; Haslam, 2006).

Media which is sympathetic to the different sides of a conflict is also beneficial in and of itself: Impartial and balanced media that reflects sympathy for both sides in a conflict, provides the public with fair, accurate and honest information. As a result, it maintains credibility and public trust

¹ See BBC Guidelines [here](#)

² See ranks and stats [here](#)

(Turcotte et al., 2015). Furthermore, balanced sympathy generally aligns with journalistic ethics, ensuring accuracy and integrity in reporting (Ward, 2009; Galtung, 1969).

Thus, balanced sympathy towards different parties involved in a conflict is essential for fostering an environment of understanding and trust, which is necessary for both the resolution of the conflict and the integrity of journalistic practice.

Measuring Sympathy as a Proxy for Gauging the Degree of Impartiality

Measuring sympathy is an effective way to estimate the degree of impartiality for several complementing reasons:

1. **Direct Indicator of Emotional Tone:** Sympathy is evoked by the emotional tone of the coverage, indicating which side is portrayed more favourably.
2. **Reflects Underlying Bias:** Disparities in sympathetic coverage often reveal underlying biases, whether conscious or unconscious.
3. **Influences Public Perception:** Sympathetic coverage has very 'real' consequences, as it has a great impact on audience attitudes and perceptions, making it a critical factor in assessing actual media influence.

Kwak and An (2014) suggest that news coverage that evokes sympathy can lead to increased compassion for those affected, which can prompt support and collective public action. This is significant due to the media's influential role in shaping public discourse and perceptions of global events (Price & Tewksbury, 1997)

Moreover, measuring the tendency of sympathetic framing given to each side in a conflict encapsulates the essence of the coverage's degree of bias and partiality. By quantifying this aspect, researchers and media analysts can more accurately assess whether a media outlet is maintaining true impartiality or favouring one side, thus providing critical insights into the integrity and balance of conflict reporting.

It is important to note that moderately varying degrees of sympathy and sympathetic framing, in and of itself, are not necessarily a violation of impartiality or evidence for systematic bias. In this report we use systematic framing differences as a proxy. We first use state-of-the-art AI tools to study systematic differences in sympathetic framing. We then compare the framing discrepancies were there should be none, e.g., headlines compared to the main text, stories that cover the same events in different languages (BBC English and Arabic reporting) or different mediums. These systematic discrepancies do demonstrate that the commitment for impartiality is violated and suggest that the BBC actively promotes a heavily biased perspective in a way that misinforms its audience and manipulates public opinion.

The Challenges in Measuring Sympathy in Media Coverage

Measuring sympathetic framing and evocations in media coverage is challenging for several reasons:

1. **Subjectivity of Sympathy:** Sympathy is inherently subjective, making it difficult to define and measure consistently. What one person perceives as sympathetic coverage, another might not. There is no ‘ground truth’.
2. **Nuances of Language:** Language is nuanced and complex. Sympathy can be conveyed through subtle cues such as tone, word choice, and framing, which are not easily quantifiable. These subtleties can vary widely between different contexts and cultures.
3. **Contextual Dependence:** The context in which information is presented can greatly affect its perceived sympathy. The same statement can be seen as sympathetic or neutral depending on the surrounding content and the broader narrative in which it is embedded.
4. **Variability Across Media Type and Genre:** Different media platforms (print, broadcast, online) and formats (news articles, opinion pieces, social media posts) convey sympathy in different ways. This variability complicates the creation of a standardized measurement method.
5. **Interpersonal Differences:** Audience interpretations of sympathy can vary based on personal experiences, biases, and cultural backgrounds (Lee et al., 2024). These differences mean that what resonates as sympathetic to one group may have a different impact on another.
6. **Complexity of Emotions:** Sympathy is often intertwined with other emotions like empathy, compassion, and pity. Distinguishing and measuring these overlapping emotional responses adds another layer of complexity.
7. **Qualitative Nature:** Much of the analysis of sympathy involves qualitative assessment, which lacks the precision and objectivity of quantitative methods. This reliance on qualitative analysis can introduce researcher bias and variability in findings.
8. **Evolving Media Landscapes:** The rapid evolution of media landscapes, including the rise of social media and user-generated content, constantly changes how sympathy is expressed and perceived. Keeping measurement methods up-to-date with these changes is an ongoing challenge.

Due to these factors, developing reliable and valid metrics for measuring sympathy in the media requires a multidisciplinary approach, combining linguistic analysis, psychological principles, and content analysis, while accounting for the dynamic nature of media and audience interpretations. A successful evaluation system should be objective and based on empirical evidence, and for that it should be as automated as possible. Furthermore, due to the complex nature of the task, a successful solution should be one that utilizes large amounts of data. This would average out and overcome the ‘noise’ that is inherent to any kind of measurement, let alone one that puts a complicated concept such as sympathy at its centre.

Automated Media Bias Detection

Media bias detection involves the automated identification of bias within journalistic texts. This process typically combines Natural Language Processing (NLP) and machine learning techniques such as sentiment analysis (Lin et al., 2011), topic modelling (Best et al., 2005), and lexical feature analysis (Hube & Fetahu, 2018). These techniques help identify patterns and features indicating bias, providing a systematic approach to understanding and mitigating media bias.

Media bias detection is a relatively new area in NLP. Initial attempts to automatically detect media bias began in the early 2000s but were limited in accuracy and robustness (Rodrigo-Ginés et al., 2023). Over the last decade, advances in deep learning have significantly impacted NLP, including media bias detection. Specifically, the introduction of recurrent neural networks (RNNs) (Rashkin et al., 2017) and transformer networks (Baly et al., 2020) has enabled the development of sophisticated techniques that outperform traditional methods.

Despite these advances, Large Language Models (LLMs) like ChatGPT have yet to be extensively utilized for media bias detection. LLMs, pretrained on vast amounts of data, have the potential to succeed in this area due to their ability to process context, generate coherent text, and provide reasoning. Their impressive capabilities in natural language understanding, reasoning, and generation could enhance the accuracy and robustness of bias detection, providing deeper insights and more reliable identification of biased content (Hamborg & Donnay, 2018; Wen & Younes, 2024). We believe that the tailored application of state-of-the-art LLMs we developed could be usefully adopted, and no doubt further refined, by media organizations which have a genuine interest in producing balanced, unbiased, high-quality output.

Large Language Models as a Tool to Assess Sympathetic Framing and Media Bias

Large Language Models (LLMs), such as OpenAI's ChatGPT, Google's Gemini and Mistral by MistralAI, among others, may prove to be particularly well-suited for analyzing media-bias for several reasons:

1. **Broad Knowledge Base:** LLMs are *large* pretrained models, often referred to as foundation models. The concept '*large pretrained models*' does not only mean that the number of the model's parameters is large, but also entails that the model-training was done over immense textual corpora containing billions of words spanning almost any domain available. As such, LLMs have the "knowledge" (linguistics statistics) and ability to process and respond to queries (known as *prompts*) across multiple domains without the need for a specific supervised training.
2. **Zero-Shot Abilities:** Consequently, as a result of the size of the model and the sheer volume of the training data, LLMs, and ChatGPT specifically, have demonstrated impressive zero-shot capabilities: the ability to answer specific questions or perform classification tasks without any fine tuning (hence, zero-shot). The zero-shot abilities were demonstrated time and time again in tasks such as sentiment analysis, stance detection, and

hate speech detection – achieving results better or comparable to those of fine-tuned and domain-specific models (Wang et al., 2023).

3. **Emotion and Empathy Understanding:** Specifically, recent research has shown that ChatGPT's ability to detect and express emotions is comparable to that of humans, with an accuracy of approximately 92% for classifying various emotions (Schaaff et al., 2023). In another overwhelming cross-sectional study that was recently published (Ayers et al., 2023), ChatGPT was evaluated for its ability to provide quality and empathetic responses to medical patients' questions from a public social media forum. Evaluators preferred ChatGPT's responses over those of physicians in 78.6% of cases, rating them significantly higher in both quality and empathy.

These attributes suggest that LLMs can effectively analyse and detect sympathy (and therefore sympathy bias) in news media texts. Their ability to 'understand' context and generate nuanced responses makes them a tool with excellent potential for identifying the expression of subtle emotional expressions and providing a basis for comprehensive analysis of how sympathy is expressed in journalistic content.

Furthermore, LLMs' capacity to process vast amounts of data allows them, by subjecting tests to large datasets, to average out signals and filter out noise. This means that while in the process of LLM's classification individual errors might occur, the large-scale patterns and trends it identifies should be reliable. By leveraging its extensive data processing capabilities, using LLMs can provide a consistent and comprehensive picture of media bias, enhancing the robustness and accuracy of bias detection efforts.

In the current study, we present a novel, ChatGPT4-based, method for investigating and quantifying bias in the media. ChatGPT4 capabilities are unprecedented; its advanced contextual understanding, support for multiple languages and ability to perform tasks with minimal examples, were harnessed in the current study for yielding a massive scale of accurate responses.

This allowed for assessments of large datasets, mitigating measurement-related noise that is prominent when the sample size is small. Using ChatGPT4 was also crucial for avoiding potential biases inherent in human subjective judgment.

We focus on the case study of the Israel-Gaza war and its coverage by the BBC in various platforms. The Israel-Gaza war suits well to this end, because it is a polarizing dispute, therefore making achieving impartiality both challenging and important. Also, it has been the most prominent single international story during the course of this study, providing the largest dataset not just of BBC material, but also of comparative material across other global news outlets.

Experiments and Results

In order to empirically confirm impartiality or lack of impartiality thereof, we have conducted a series of seven experiments. In the first three experiments we use LLMs to quantify sympathetic framing towards each party in the conflict. We do this for BBC news stories (Experiment 1), news headlines (Experiment 2), and then repeat these experiments on BBC Arabic (Experiment 3). In the fourth experiment we introduce repetitions and slight prompting variations in order to assess the stability and reliability of the AI model we use in Experiments 1-3. In the fifth experiment we validate our methodology by having the LLM produce reasoning “explaining” its logic in a consistent way and by establishing a comparison to human judgment, based on a sample of items evaluated independently by ten human judges. In the sixth experiment we apply our methodology to a large collection of news stories from hundreds of news outlets from across the globe and the political spectrum. We then align the BBC’s framing tendencies with the framing tendencies of other news outlets that have a clear political orientation. Finally, in the seventh experiment, we move beyond the textual medium and apply our analytic framework to measure framing tendencies in the BBC podcasts, radio, and TV shows. The remainder of this section provides a detailed description of the seven experimental settings and the empirical results.

Experiment #1: Sympathy in Web Articles ’Main Texts

Our dataset for this experiment was initially comprised of 1,529 conflict-relevant items from the BBC website, published within our defined timeframe of interest: four months starting from the day of the Hamas attack on October 7th, 2023 (For the detailed criteria of data inclusion for this and all other experiments, please refer to the Methods section).

To evaluate if the BBC reporting creates greater sympathy towards one side of the conflict compared to the other, we employed a ChatGPT-based analysis: we systematically inputted all conflict-related BBC articles, one by one, providing the same prompt each time, asking the following six questions, hereafter referred to as q1-q6 (The actual prompt is in bold):

“I have the following article.

Please answer only yes or no to the following questions:

1) Does this text create sympathy for Israel?

2) Does this text create sympathy for Gaza?

To enhance the discernment between various aspects of Palestinian or Israeli entities garnering increased sympathy, we introduced two additional pairs of questions:

Questions regarding sympathy towards Israeli and Palestinian civilians:

3) Does this text create sympathy for the Israeli people?

4) Does this text create sympathy for the Palestinian people?

Questions regarding sympathy towards Israeli and Palestinian fighting entities:

5) Does this text create sympathy for IDF/Israel military?

6) Does this text create sympathy for Hamas?"

Note that for the purpose of our analysis and for interpreting the results, questions 3-6 constitute a supplemental extension for questions 1 and 2, whereas q1 and q2 (being more general) are the main questions in our design. For simplicity, when no further resolution is required, or for practical reasons of capacity - we often refer exclusively to q1 and q2.

Experiment#1, Results:

The results of experiment #1 are depicted in Figure 1. Our query yielded 1481/1529 cases for which full responses were received from ChatGPT (See the exact definition in the Methods section). The distribution of answers for q1 and q2 are represented in Figure 1a. ChatGPT classified the majority of articles (75%) as creating sympathy towards one side or both sides. Notably, a significant proportion of the articles (16%) created sympathy towards both sides of the conflict. Concentrating on the difference in sympathy received by each side, we observed that the sympathy expressions towards Gaza were approximately x1.5 more frequent than the sympathy expressions towards Israel (Figure 1a,b).

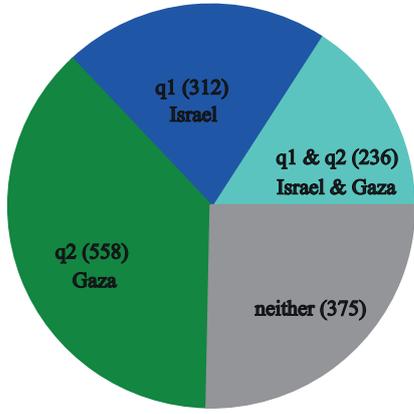
Next, to examine the dynamics of the conflict coverage in time, we grouped all the articles in our dataset into weekly bins, starting from October 7th 2023 (Figure 1c). Analysing the conflict coverage over time unveiled several distinct findings: First, beyond the week of the attack, sympathy level was higher towards the Palestinian side in all weeks but one.

Second, sympathy towards both sides decreased dramatically over time. However, our analysis also revealed a notable decrease in the number of articles over time (Figure 1d). This variability could impact both findings: while an unstable baseline might somewhat obscure the first, the second could be attributed solely to the overall decrease in coverage. To mitigate this inherent variance, we opted to investigate the mean (proportional) expression of sympathy thereafter.

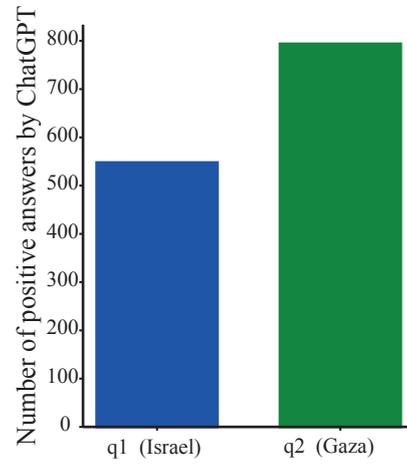
Dividing the number of weekly occurrences of sympathy by the total number of articles per week resulted in a consistent measure, ranging between 0 and 1, and remaining robust despite changes in the number of published articles over time. We used this strategy to evaluate the results yielded for the full query (Questions 1-6, Figure 1e).

In the first two pairs of questions (i.e., q1-q2 and q3-q4), the results demonstrated a clear bias towards showing more sympathy to the Palestinian side, which started after the first week and remained generally stable over time. Questions 3 and 4 qualitatively replicated the results observed for questions 1 and 2. These results not only sharpened our findings, but also validated our methodology, demonstrating that minor alterations in the phrasing of the prompts do not change the results dramatically (compare solid lines to dashed lines in Figure 1e).

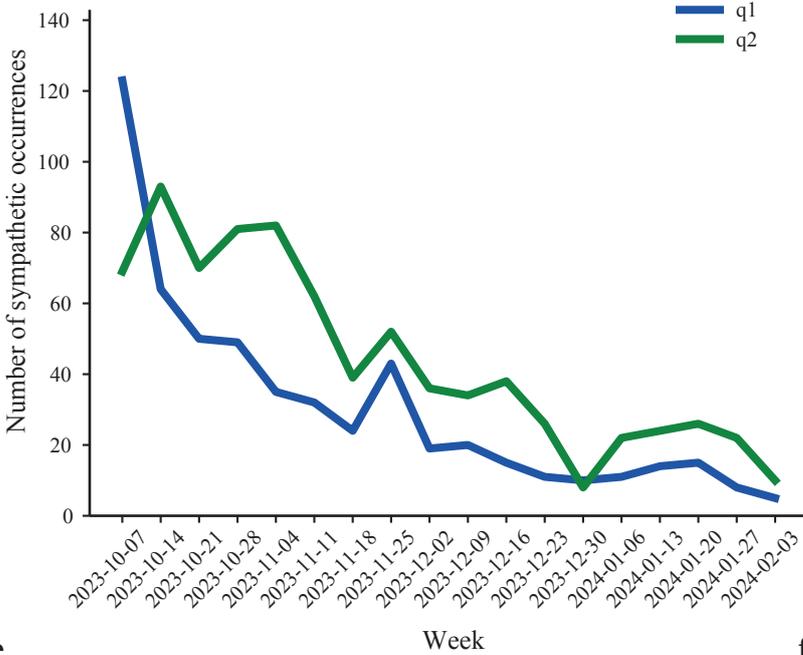
a Distribution of Positive Answers: Main Texts Analysis (N = 1481)



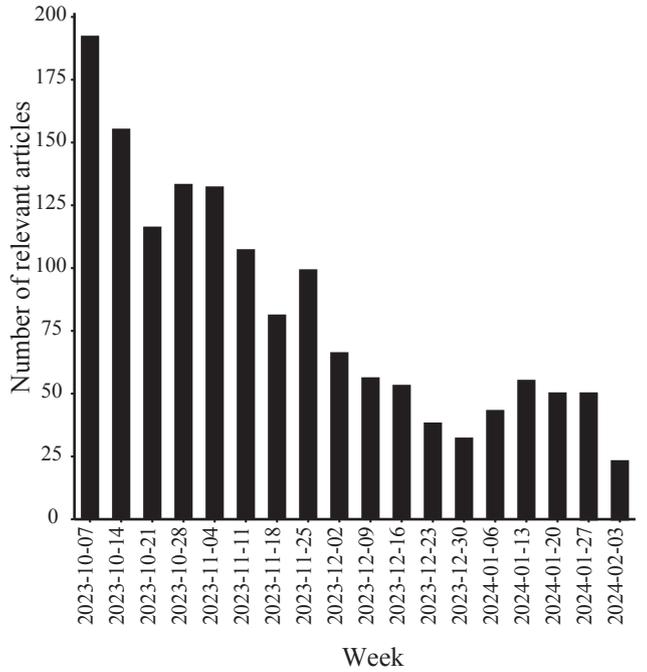
b



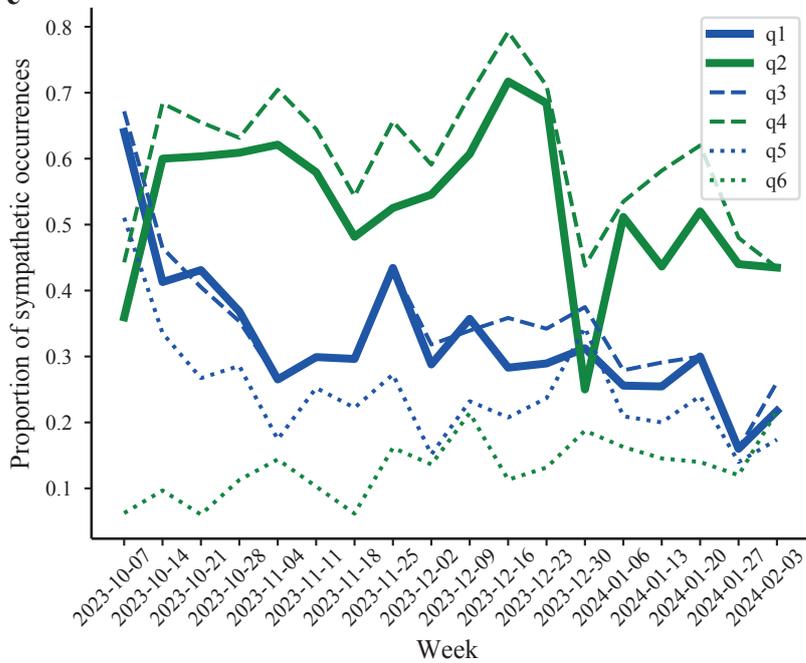
c



d



e



f

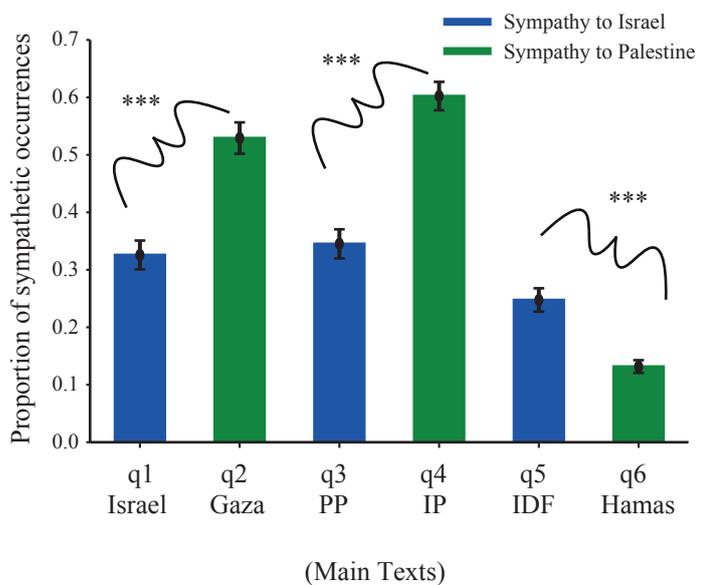


Figure 1. Analysis of 1,481 Main Texts Published on the BBC British Website in the four months following October 7th, 2023. **a**) A pie chart depicting the distribution of sympathetic occurrences towards Israel (q1) and Palestine (q2). **b**) The total number of positive answers for questions 1 and 2 (“Does this text create sympathy to Israel/Palestine”, respectively). **c**) The number of positive responses by ChatGPT to q1 and q2, dynamically aggregated over a week of reporting. **d**) The number of conflict-related articles published on the BBC website weekly. **e**) The proportion of positive responses by ChatGPT to all six questions in our query. Means are dynamically calculated over a week of reporting. Blue lines represent questions regarding sympathy to Israel and green lines represent questions regarding sympathy to Palestine. Dashed lines represent questions regarding civilians and dotted lines represent questions regarding militant entities. **f**) The proportion of positive responses by ChatGPT to all six questions, calculated over the entire 18 weeks within the reporting period. Each bar represents the proportion of positive answers to that particular question. *** indicates $p < 0.0001$, Paired t-test. n.s. – not significant.

In addition to these findings, a contrasting trend emerged for q5 and q6, which regarded sympathy towards the IDF/Hamas, respectively (see the dotted lines in Figure 1e). We observed a pronounced disparity in sympathy favouring the IDF, particularly during the initial weeks of the conflict, when reports of Israeli soldiers being killed or abducted on October 7th were prevalent. Notably, this gap diminished over time (see Figure 1f). This result, combined with the results received for the other pairs of questions, evaluates our methodology again, as it demonstrates ChatGPT's ability to identify subtle (but important) distinctions in its input for yielding separable results. Furthermore, the dissociation between the results received for the pairs of civilian/general entities and those obtained for the pair of military entities, shows that the answer by ChatGPT depends on the precise question, rather than being an outcome of some inherent bias within ChatGPT4 itself towards either Israel or Palestine.

Experiment #2: Sympathy in the Headlines

In an era where attention spans are increasingly fragmented and time is at a premium, headlines offer a quick and efficient way for viewers to grasp the essence of a news story without delving into the entire article (Gabiolkov et al., 2016). However, headlines often reflect sympathy in a binary manner, towards one side of a conflict, due to the need for simplification and clarity, emotional appeal, and the biases of media outlets. This approach resonates with their target audience, maintains narrative consistency, and leverages sensationalism to attract more readers. Additionally, time and space constraints make it easier to highlight one perspective rather than providing a balanced view. These factors collectively contribute to the prevalence of one-sided headlines, despite the potential sacrifice of balance and impartiality.

Moreover, binary sympathy expression in headlines oversimplifies complex issues, fostering a shallow understanding that encourages societal division and polarization. This type of coverage reinforces existing biases and promotes a "black and white" mentality, leading to increased partiality and eroding public trust in media. Additionally, it limits the scope of public discourse by excluding alternative viewpoints and critical analysis, thus preventing constructive dialogue and comprehensive exploration of conflicts.

Recognizing this prominent influence, we turned to evaluate the sympathy levels as reflected exclusively by the headlines. For that goal, we replicated the above-mentioned analysis: we repeated the exact same procedure, presenting ChatGPT with one news item at a time, together with the same six questions, but this time – we changed the first sentence in the prompt to: “**I have the following article headline..”** and then inputted headlines only, instead of the complete articles.

Since headlines are significantly shorter than the body of the article, it was a trivial observation that they include fewer expressions of sympathy than those found in the main text. However, if the headlines were chosen in an unbiased manner, we expected this decrease in proportion to be equal for both sides, so that the results for headlines would broadly reflect the results for the articles themselves. The results of the BBC headlines analysis are depicted in Figure 2.

Experiment #2, Results:

For this experiment to be comparable with experiment #1, our dataset consisted of the same 1481 items for which a complete response was received by ChatGPT when asked about the main text. Our first observation revealed that, as we suspected, a binary sympathy expression characterized the headlines: the portion of items reflecting sympathy for both sides almost completely vanished (Figure 2a; compare the turquoise slice, that now constituted less than 1% of the items, to that in Figure 1a). Focusing on the sympathy ratio for the two sides of the conflict, this analysis revealed a significantly more biased picture: The first two pairs of questions exhibited an even more extreme disparity, with the Palestinian side receiving nearly x3 times sympathetic headlines than the Israeli side (x2.91 times, Figure 2b). This effect was robust and stable across all time points, emerging almost immediately after the first week of coverage (q1 vs. q2 and q3 vs. q4 – see Figure 2c, solid and dashed lines. Note, once again, how similar the answers to these two pairs of questions are).

Examining the full six-question-query results, we have found that, remarkably, in the headlines analysis, Israel did not maintain an advantage even in the comparison between the IDF and Hamas. The slight edge initially observed for the IDF was swiftly overridden by comparable sympathy levels towards Hamas, culminating in a non-significant difference between the two when sympathy was averaged across all weeks (See Figure 2d, last two bars).

Surprised by the last finding, we sought to assess two additional questions. First, we wanted to check how (and to what extent) is the pro-Palestinian tendency, observed at the level of the main text, augmented at the level of the headlines. To answer this question, we plotted the ratio of q1:q2 and compared it to the ratio of q1_headline:q2_headline over time (Figure 2e). Our inquiry yielded a notable finding: Both the main text ratio line (i.e., q1:q2 when the input is the main text, depicted by the solid line in Figure 2e) and the headline ratio line (i.e., q1:q2 when the input is limited to the headline alone, represented by the dashed line in Figure 2e) consistently exceeded one, indicating a consistent bias towards the Palestinian side. However, while the solid line remained relatively stable around two, indicating a typical bias of approximately twice more sympathy towards Palestine, the dashed line skyrocketed, reaching values higher than 10 at certain time points.

The second question we asked regarded the ratios of headlines to main texts showing sympathy for each of the sides (Israel and Gaza) separately. We were interested in quantifying the bias (if any) associated with the stage of headline selection, independently from the unbalanced sympathy in the main texts. To state this explicitly: it is to be expected that headlines will carry less sympathy (and information in general), as they are made of shorter texts. However, if the headlines are chosen in an impartial manner, this reduction in sympathy from main-texts to the headlines level, should be reflected with similar proportions for both sides.

Thus, we examined the ratio of headlines that created sympathy towards Israel to the articles that generated sympathy towards Israel in their main-text, and contrasted the result with the parallel measure for the Palestinian side. For this goal, we plotted ratio of q1_headline:q1 and compared it

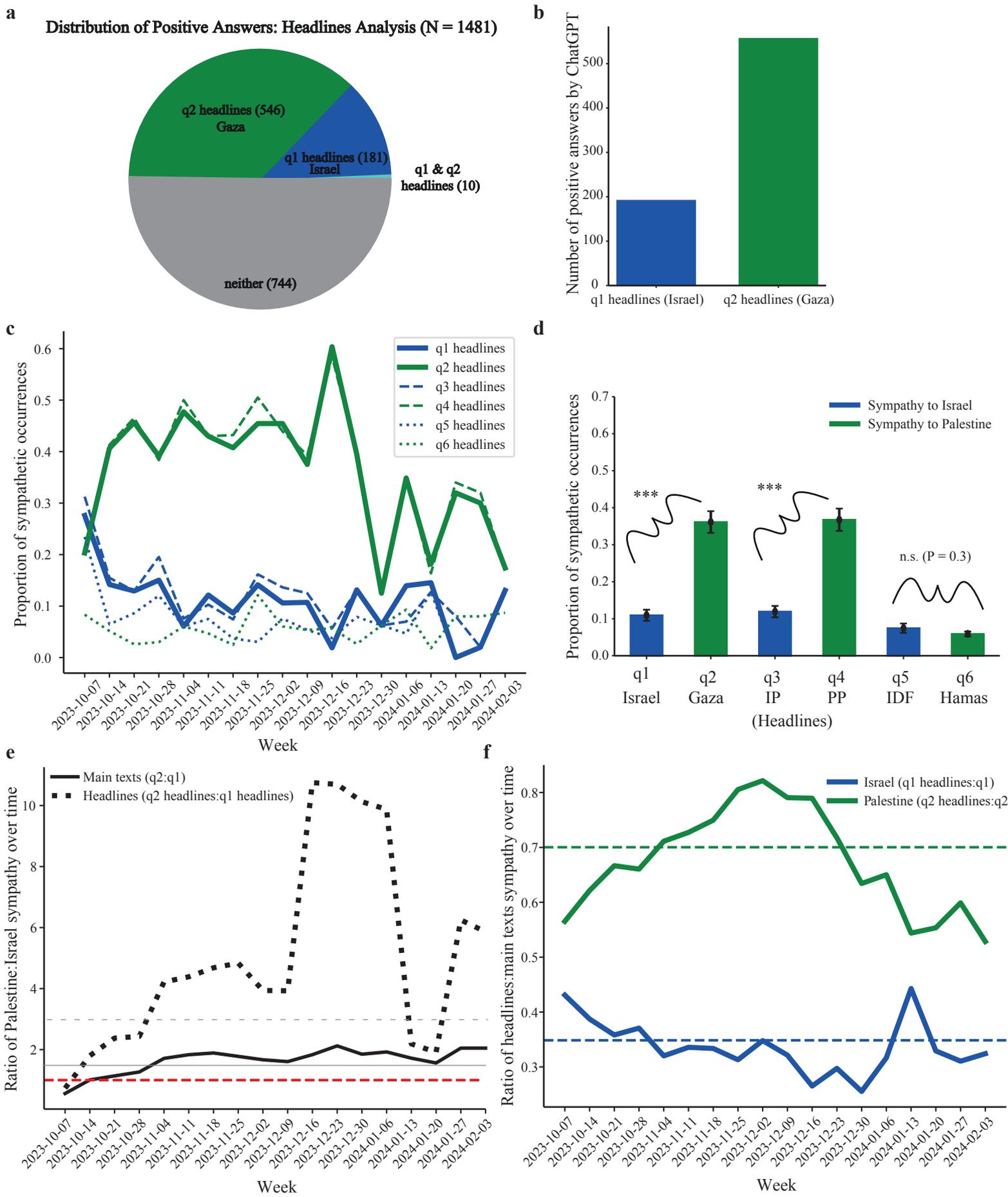


Figure 2. Analysis of 1,481 Headlines Published on the BBC British Website in the four months following October 7th, 2023. **a)** A pie chart depicting the distribution of sympathetic occurrences towards Israel (q1) and Palestine (q2). **b)** The total number of positive answers for questions 1 and 2 when solely headlines were given as an input **c)** The proportion of positive responses by ChatGPT to all six questions in our query. Means are dynamically calculated over a week of reporting. Blue lines represent questions regarding sympathy to Israel and green lines represent questions regarding sympathy to Palestine. Dashed lines represent questions regarding civilians and dotted lines represent questions regarding militant entities. **d)** The proportion of positive responses by ChatGPT to all six questions. Each bar represents the proportion of positive answers to that question, averaged over the entire 18 weeks within the reporting period. *** indicates $p < 0.0001$, Paired t-test. n.s. – not significant. **e)** Sympathy trends, depicted separately for headlines and main texts. The ratio of positive responses by ChatGPT to q1 and q2 when the Main texts are given (solid lines) and when solely articles' headlines are provided as input (dashed line). The flat red dashed line represents a ratio of 1, where a balanced coverage should reside. Faint grey straight lines represent the averages over time. Ratios are calculated over the bins of weekly published articles and then smoothed with a rolling average using a sliding window with size 4. **f)** Headlines vs. main text trends, depicted separately for Israel and Palestine. The graph illustrates the ratio of positive responses by ChatGPT to q1 (for Israel, blue line) and q2 (for Palestine, green line) when only articles' headlines are provided as input vs. when the main texts are given as input. Where an impartial choice of headlines is made, these lines should be generally similar. Dashed lines show the average ratio for each one of the sides. Ratios are calculated over the bins of weekly published articles, and then smoothed with a rolling average using a sliding window with size 4. IP/PP - Israeli/Palestinian People.

to the ratio of q2_headline:q2 over time (Figure 2f). Note that this comparison (q1_headline:q1 vs. q2_headline:q2) goes above and beyond the previously described finding about the different sympathy levels directed at both sides in the main texts, as it controls for this factor.

Once again, our results demonstrate a notable imbalance; When articles' main texts were labelled as sympathetic towards Israel, it was reflected in the headline only in about a third of the cases. For the Palestinian side, this ratio was approximately double (see the dashed lines in Figure 2f, which represent averages over time). Our analysis reveals that, even when considering the lower number of articles showing sympathy to the Israeli vs. the Palestinian side - another notable trend that enhances the imbalance as found in the first experiment - the headlines further change the framing of the story to evoke more sympathy for the Palestinian than the Israeli side. To be clear, when assessing the headlines we observe a trend of further suppressing potential sympathy for the Israeli side, even in cases where the in-depth reporting may suggest otherwise. Recall, the headlines are decided by editors, and not by the authors of the main text, thus suggesting another independent layer of bias.

Experiment #3: BBC in the Arabic Language

The BBC offers news coverage in multiple languages to cater to a global audience. Regardless of the language of reporting, the BBC is committed to adhering to the same high standards of journalism, which include impartiality and accuracy. This commitment ensures that all audiences receive balanced and unbiased news coverage. The BBC's Arabic service, should therefore provide impartial and accurate reporting, as should the BBC do across all platforms and languages, reflecting the corporation's legal obligations. To test the extent that the BBC maintains these standards, we conducted an empirical test, examining the conflict as covered by the BBC Arabic news website.

Experiment #3, Results:

Our evaluation is based on scraping a total of 876 articles published on the BBC Arabic website between October 7th, 2023, and February 7th, 2024. We filtered out 302 items from this dataset, that did not feature at least one of our six designated keywords. The remaining dataset contains 574 items (please refer to the Methods section for the list of these words which were selected to identify coverage of the Gaza War).

Given the potential variance in ChatGPT's performance across languages, with English typically regarded as the most proficient and natural, we initiated our study by translating the articles into English using various methods. Specifically, we employed Google Translate, ChatGPT3.5, ChatGPT4o and ChatGPT4 for this purpose. Subsequently, we enlisted the feedback of several Arabic speakers to evaluate the quality of these translation options. The consensus among them was that the translation produced by ChatGPT4 exhibited superior quality compared to the others, being noted for its high standard (even though not perfect, but see 'Replications and Variations' in the Methods section). Consequently, we utilized the Arabic-to-English translated dataset and

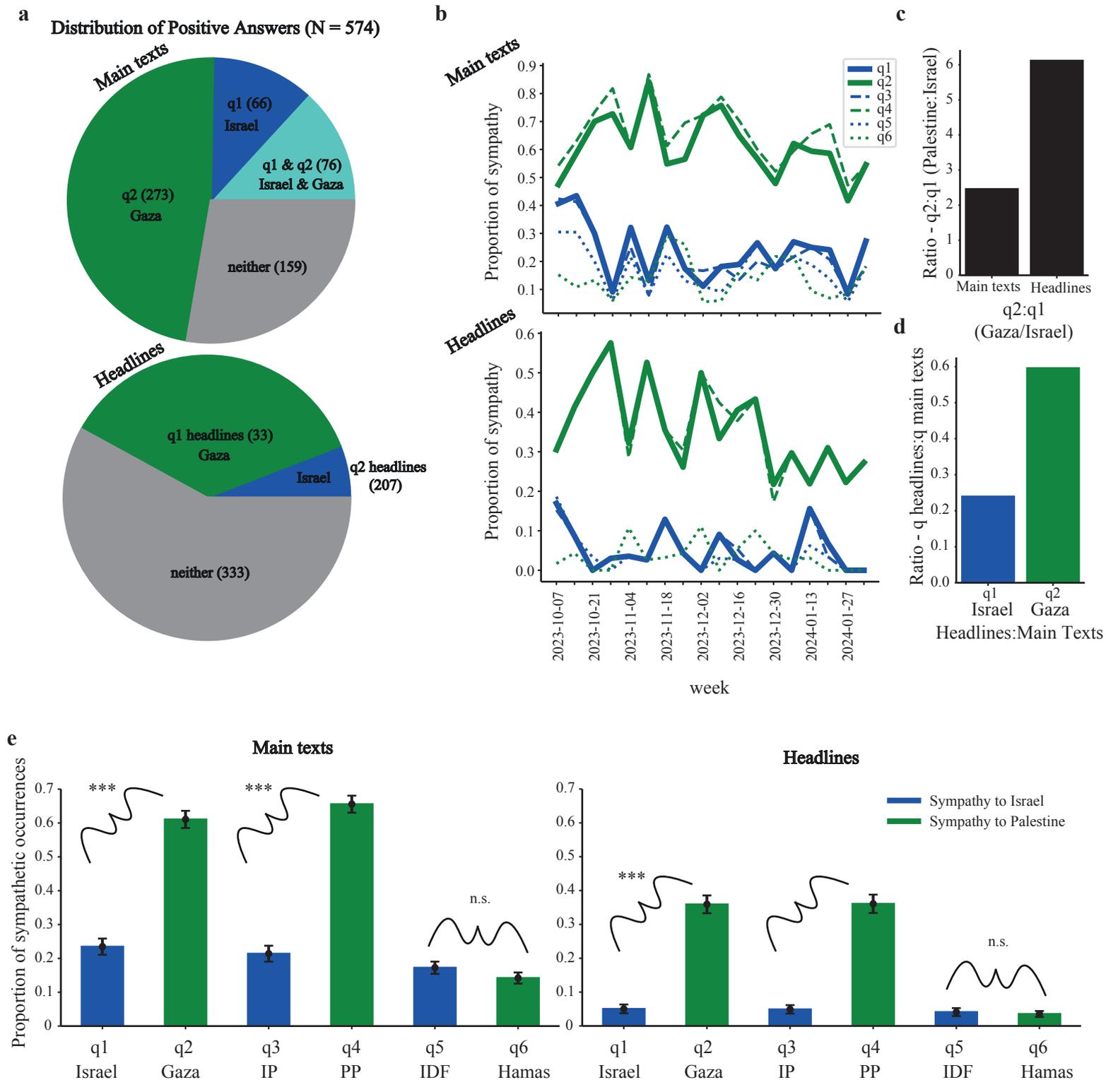


Figure 3. Analysis of 574 Headlines and Main Texts Published on the BBC Arabic Website in the four months following October 7th, 2023. **a)** Pie charts depicting the distribution of sympathetic occurrences towards Israel (q1) and Palestine (q2) in the main texts (**top**) and headlines (**bottom**). **b)** The proportion of positive responses by ChatGPT to all six questions in the main texts (**top**) and headlines (**bottom**) analyses. Means are dynamically calculated over a week of reporting. Blue lines represent questions regarding sympathy to Israel and green lines represent questions regarding sympathy to Palestine. Dashed lines represent questions regarding civilians and dotted lines represent questions regarding militant entities. **c)** Sympathy trends, depicted separately for headlines and main texts. The ratios of positive responses by ChatGPT to q1 and q2 when the main texts are given (**left**) and when solely articles' headlines are provided as input (**right**). **d)** Headlines vs. main text trends, depicted separately for both sides. The total headlines:main-texts sympathy ratios for Israel (q1headlines:q1, blue bar) and for Palestine (q2 headlines:q2, green bar). **e)** The proportion of positive responses by ChatGPT to all six questions for main texts (**left**) and headlines (**right**) analyses. Each bar represents the proportion of positive answers to that question, averaged over the entire 18 weeks within the reporting period. *** indicates p < 0.0001, Paired t-test. n.s. - not significant. IP/PP - Israeli/Palestinian People.

proceeded with the same analyses as those conducted on the dataset obtained from the BBC English website.

This experiment confirmed that the BBC Arabic website exhibits a notably greater degree of imbalance, even when contrasted with that found in the BBC English website (refer to Figure 3 for comparison with Figures 1-2). In BBC Arabic, we observed a greater degree of imbalance in sympathy at the level of both main texts and headlines. This imbalance was evident in each and every time-point, remarkably including even the week of the Hamas attack (see Figure 3b). We note that during that week the shocking details of Hamas' attack on Israel were emerging, weeks before Israel launched its ground campaign. Averaged across all time points, the ratios between the 2 main questions in our query were $\sim x2.5$ fold and $\sim x6$ fold for the main texts and headlines analyses, respectively (Figure 3c). Once again, another independent bias was evident at the choice of headlines, where $\sim 60\%$ of the articles exhibiting sympathy to Palestine reflected this in their headline. Still, only $\sim 25\%$ doing so in the case of Israel (Figure 3d). Considering the full six-questions-query, the difference between the sides was so prominent, such that the sympathy towards Israel and 'The Israeli people' was not comparable to that directed towards Gaza and the Palestinian people, but was comparable to that directed towards Hamas (Figure 3b,e) - a proscribed terrorist organization under the law of the UK, the legal environment in which the BBC is founded and operates.

Experiment #4, Reliability of the Methodology

The output generated by ChatGPT is known to be influenced by subtle variations in the input it receives, primarily the prompt (Salinas & Morstatter, 2024. Errica et al., 2024). Additionally, owing to the inherent stochasticity of the model, even when provided with the exact same prompt, different results may arise across multiple runs. To test the stability of the model and the reliability of our findings, we conducted ten iterations of the same procedure and repeated this for both the main-text and the headlines experiments over the BBC English website dataset. In some of these iterations, we introduced subtle changes to the prompt by altering the order of questions (please refer to Figure 4).

Experiment #4, Results:

The experiment convincingly demonstrated that the conclusions drawn from our results remain stable through multiple runs and despite subtle variations in repetitions with minor perturbations. The results for this experiment, which constitute support for the reliability of our methodology, are depicted in Figure 4 for both the main text (top) and the headlines (bottom) analyses. For each one of these analyses, five identical repetitions are depicted in the upper row and five repetitions where the order of the questions was shuffled are depicted in the lower row.

Experiment #5, Evaluating the Validity of the Methodology against Human Annotation

Reliability and validity are both crucial aspects of measurement tools, but they are not the same and one does not always guarantee the other. Specifically, while reliability refers to the consistency

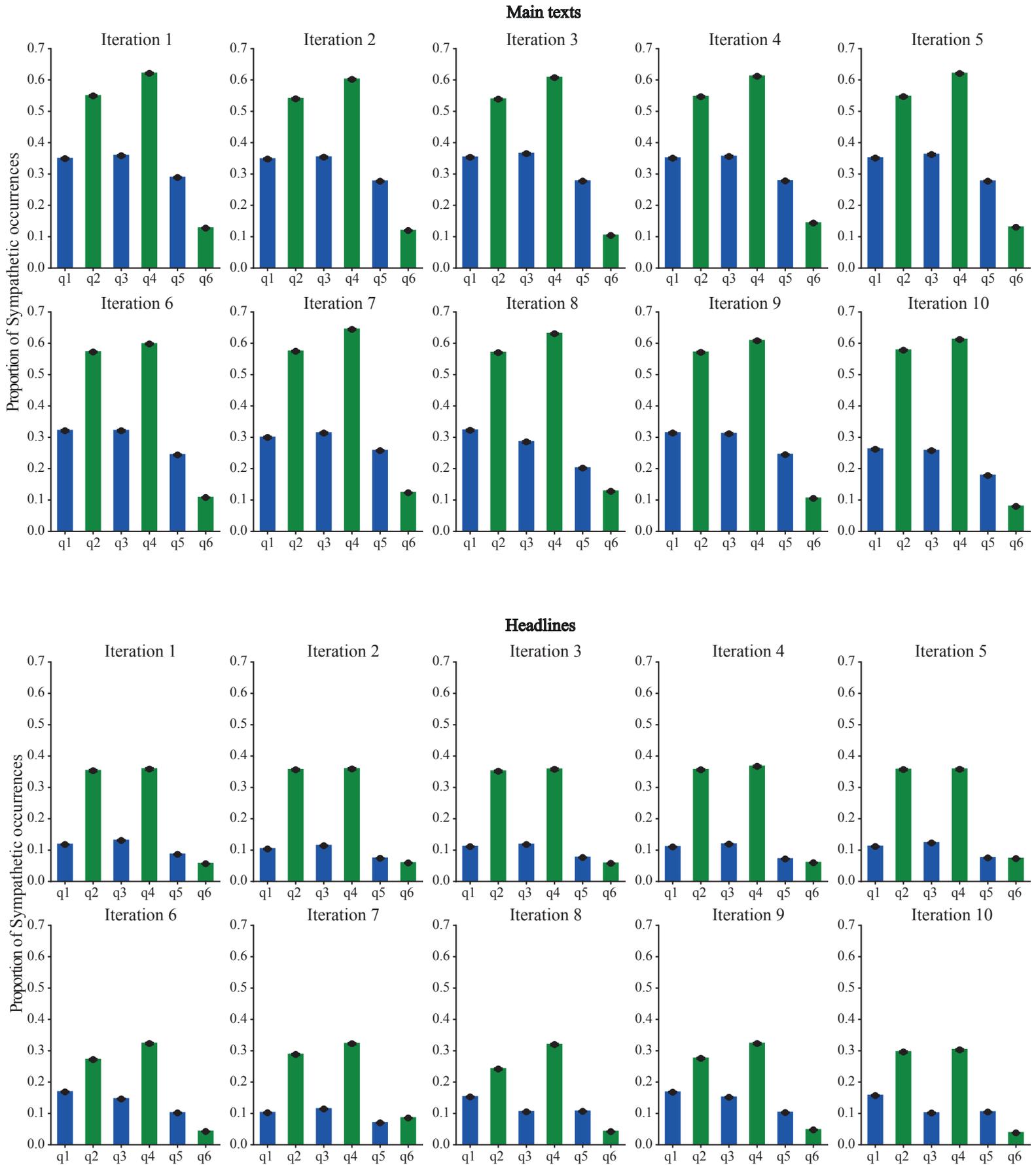


Figure 4. *Stability and Reliability Across Multiple Runs with Minor Variations.* Proportion of positive responses by ChatGPT to each one of the six questions, repeating both the articles' main text (top) and headlines (bottom) experiments ten times. Experiments were done using the data from BBC English (N=1481 articles/headlines). Iterations 1-5 were identical. The questions order for iterations 6-10 was as follows: Iteration 6: 2 1 4 3 6 5, iteration 7: 1 3 5 2 4 6, iteration 8: 2 4 6 1 3 5, iteration 9: 2 1 4 3 6 5, iteration 10: 6 5 4 3 2 1. The high similarity between iterations indicates that the measurement is generally stable and reliable.

or repeatability of a measurement, reliability alone does not ensure validity, which is the accuracy of the measurement in capturing what it is intended to measure. The results we observed gave us several strong reasons to trust the validity of our methodology, e.g. the seemingly sensible locations of an array of reputable news outlets across the axis in Experiment 6 (see below), or the discrepancy between the first week and the following weeks in experiments 1 & 2). To test this more thoroughly, we designed another experiment and started our investigation by probing the ‘black box’.

The ‘black box problem’, in the context of AI based models, refers to the fact that while these models are capable of very powerful computations and impressive performance, understanding how exactly these computations underlie the transformation from the input to the output level remains very difficult (Guidotti et al., 2018; von Eschenbach, 2021; Castelvechi, 2016). To better estimate the validity of our methodology, first, we were interested to know how ChatGPT decides regarding the yes/no answer, after it is given different texts as inputs.

Our first attempt to get closer to that answer and better explain the obtained results, involved utilizing the high-level processing of human language that characterizes ChatGPT: we asked it explicitly to provide reasoning for its answers. Although it may not reflect a rigorous test that implies how the computation is done (e.g. which words/combinations of words are given more weight than others), it may anecdotally hint if the model is ‘directing its attention’ towards the right places. We started by inputting 100 randomly sampled headlines from the BBC English website, asking the six questions just as we did previously, but this time we added the following sentence at the end of the prompt: **“Then, only if the answer is yes, please explain what led you to give this answer?”**.

We were impressed by ChatGPT’s ability to explain and support its decisions with convincing arguments, drawing on the relevant material, reflecting high-level ‘understanding’ of subtle nuances in the language (see Table 1 for several representative examples. The full reasoning for all sampled headlines is provided in Supplementary Document 1).

Finally, we wanted to design an experiment that further evaluates this aspect more rigorously. Rigorous validation of any tool that aims at measuring sympathy cannot be easily established, mostly due to the absence of ground-truth which is usually needed to calibrate AI tools or determine their accuracy. For the current work, our (somewhat philosophical) underlying assumption was that the absence of ground-truth in measuring sympathy necessitates relying on public opinion. Much like a tree falling in the woods making a sound only if there are ears to hear it, the impact of a text in generating sympathy is meaningful only if perceived and felt by its readership as such.

In other words, without an objective standard to measure sympathy, the subjective experiences and emotional reactions of the public should become the primary metric for assessment. Assessing the validity of our tool with this approach ensures that the tool remains relevant and accurate in capturing the genuine emotional impact of texts as perceived by real people. Thus, we next asked

Headline	“Does the following text create sympathy to...”	Explanation by ChatGPT
'I don't eat or sleep' says husband of new mother missing after Hamas attacks	1) Israel? Yes 2) Gaza? No 3) Israeli people? Yes 4) Palestinian people? No 5) Israeli military? No 6) Hamas? No	The text creates sympathy for Israel and its people, specifically for the husband of the new mother because he is the one depicted as suffering due to the situation. 'I don't eat or sleep' implies a state of deep distress and suffering for an Israeli individual, related to the actions of Hamas, thus potentially generating sympathy for the Israeli side. The headline implies that the harm was caused by Hamas, a Palestinian organization. As there is no reference to any suffering by Palestinian people, Gaza, the Palestinian organization Hamas, or the IDF/Israeli military, no sympathy is explicitly created for them in this headline.
Israel Gaza: EU calls for 'corridors and pauses' for humanitarian aid	1) Israel? No 2) Gaza? Yes 3) Israeli people? No 4) Palestinian people? Yes 5) Israeli military? No 6) Hamas? No	The article headline doesn't explicitly create sympathy for Israel, the Israeli people, IDF/Israel military, or Hamas. However, it does imply that those in Gaza potentially are in a situation where they need humanitarian aid, which could create sympathy for Gaza and the Palestinian people. The assertion of the need for 'corridors and pauses' suggests that the current situation may not be allowing for the effective delivery of aid, possibly due to ongoing conflict. Consequently, readers may feel sympathetic to the plight of those affected in Gaza, which includes the Palestinian people.
Rishi Sunak calls for stability after backing Israel against Hamas	1) Israel? Yes 2) Gaza? No 3) Israeli people? No 4) Palestinian people? No 5) Israeli military? Yes 6) Hamas? No	1) The text creates sympathy for Israel since Rishi Sunak is backing Israel, which might suggest to the reader that Israel is in a position needing support. 5) The text also indirectly creates sympathy for the IDF/Israel military since they are the primary force fighting against Hamas. By Sunak backing Israel, he is also indirectly supporting the IDF.
Israel-Gaza: Malaysia will maintain ties with Hamas says PM Anwar Ibrahim	1) Israel? No 2) Gaza? Yes 3) Israeli people? No 4) Palestinian people? Yes 5) Israeli military? No 6) Hamas? No	The text implies a positive action taken by Malaysia towards Hamas, which governs Gaza. Therefore, this text may generate sympathy for Hamas and indirectly for Gaza and the Palestinian people. The act of Malaysia maintaining ties could be viewed as a form of support or approval for Hamas, thus creating sympathy. The text does not mention any specific events, actions, or struggles related to Israel, Israeli people, or IDF/Israel military, so it does not create sympathy for them.
Israel says Hamas stages hit-and-run attacks from tunnels	1) Israel? Yes 2) Gaza? No 3) Israeli people? No 4) Palestinian people? No 5) Israeli military? Yes 6) Hamas? No	The text implies that Hamas, a Palestinian militant group, uses unkempt tactics like initiating hit-and-run attacks on Israel, thus painting them as threats. This can generate sympathy for Israel as it appears as the targeted party and for the Israeli military which has to counteract these attacks. However, the headline does not directly mention or imply anything about the Israeli or Palestinian people

Table 1. Examples for ChatGPT's Explained Answers in Response to Five BBC Headlines. In this (somewhat anecdotal) 'introspection' experiment, the exact prompt given to ChatGPT was as follows:

I have the following article headline. Please answer only yes or no to the following questions:

1) Does this text create sympathy for Israel? 2) Does this text create sympathy for Gaza? 3) Does this text create sympathy for the Israeli people?

4) Does this text create sympathy for the Palestinian people? 5) Does this text create sympathy for IDF/Israel military? 6) Does this text create sympathy for Hamas?

Then, only if the answer is yes, please explain what led you to give this answer?

to what extent ChatGPT reflects the human judgment and designed a straightforward experiment to compare the two.

Experiment #5, Results:

Since humans are limited in their attention span and the amount of text they can practically read, we used only 100 headlines for this experiment (the same randomly sampled items for which ChatGPT produced reasoned responses). We then independently presented these headlines to ten individuals, with the bolded text below, followed by the exact same six questions given to ChatGPT: **“Please read the following news headlines. Try to be as objective as possible and answer only “yes” or “no” to the following questions”**.

The entire data and results for this experiment are presented in Figures 5 and 6, as well as in supplementary Figure 1. Because both the responses of humans and those of ChatGPT may differ between individuals or iterations, respectively, we measure the correspondence in two complimentary ways (please refer to Figure 5a,c): The Majority Vote measure (Figure 5a,c right), as its name suggests, reflects the response received in most repetitions; It gets a “yes” value only if most evaluators answered ‘yes’ for this question.

The Average measure (Figure 5a,c left) reflects the mean “vote” calculated over the 10 iterations/participants (hereafter may be jointly referred to as ‘repetitions’). It ranges from 0 to 1: 0 indicates full agreement of ‘no’ by all ten human annotators; 1 indicates a full agreement of a “yes” response; where 0.5 reflects the lowest level of certainty/agreement, as different repetitions equally yield each type of response. Since the Average measure is the more informative one, encapsulating both the detection of sympathy and the level of consensus for this detection, we use it for further quantification. Both measures are plotted to allow impression of the similar ‘fingerprints’ by humans and by ChatGPT answers for the complete query.

Sorting all headlines in a descending order according to the average difference between q1 and q2 (i.e., q1 minus q2, calculated independently for ChatGPT and Human matrices), we observed generally similar patterns at the matrices depicting ChatGPT and Human responses, as can be seen in Figure 5a,c. However, focusing on specific questions, some differences emerged. Most notably, the almost complete absence of stripes within the Human matrix’s q5 and q6 suggested that the extent to which humans and ChatGPT’s answers are calibrated depends on the exact question (compare the 2 right columns in 5a and those at 5c, either in the Humans or in the Majority Vote matrix).

To further evaluate this, we calculated the average over repetitions for each question separately (Figure 5b,d). Doing this revealed a remarkable similarity between ChatGPT’s and human responses for q1 and q2, as well as for q3 and q4 (compare the four left bars in Figure 5b to those in 5d). For questions 5 and 6, the already low detection rate by ChatGPT became considerably lower in the human responses (two right bars in 5b,d). We conclude that, on average, ChatGPT’s

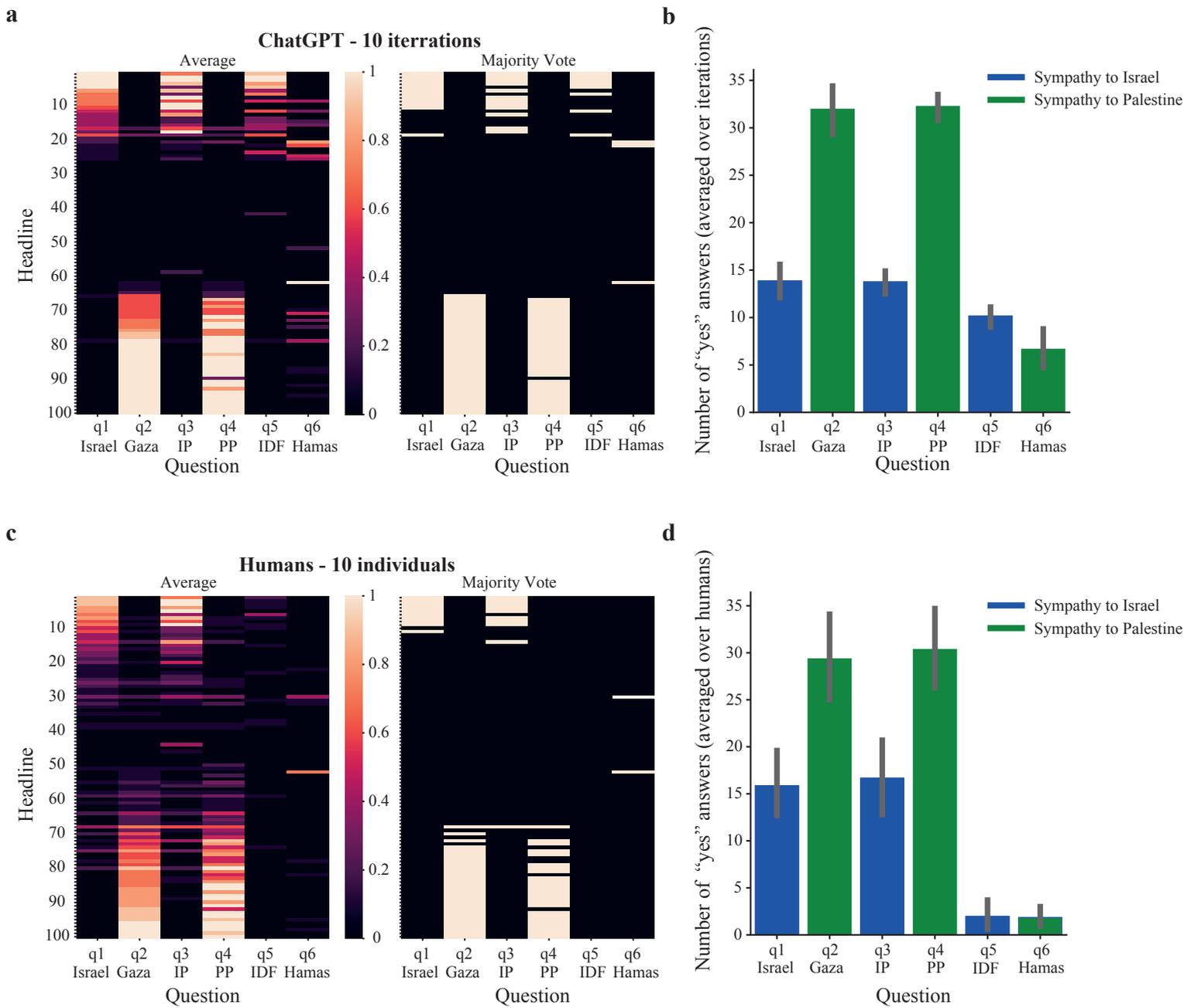


Figure 5. Comparison of Responses by Human Participants and ChatGPT to 100 Headlines Published by the BBC. **a)** Matrices depicting a summary of the “yes” answers provided by ChatGPT across ten iterations for each of the six questions (q1-q6). The vertical y-axis represents the different headlines, while the horizontal x-axis represents the questions. Headlines are sorted by the average value of q1-q2 (i.e. on a pro-Israeli to a Pro Palestinian direction, top to bottom). **Left:** A matrix depicting the average response value. The color coding represents a score in the range 0-1, where 0 means that the headline-question pair yielded negative responses in all 10 iterations, and 1 means the opposite, i.e. a positive answer in all 10 iterations. Note that numbers that are closer to 0.5 reflect lower degree of certainty. **Right:** A binary matrix depicting the “majority vote” measure: headline-question pairs are colored in white if more than half of the iterations yielded a positive answer and remain black otherwise. **b)** The number of positive responses by ChatGPT to all six questions, averaged over 10 iterations. **c-d)** The same as a-b, but 10 human subjects replace 10 ChatGPT iterations. The high similarity between the responses of ChatGPT and human participants across iterations and individuals suggests that both sources produce on average similar measurements of sympathy towards Israel (q1) and Palestine (q2). IP/PP - Israeli/Palestinian People.

judgment regarding sympathy accurately reflects real people's judgment, and that the degree this holds depends on the exact question being asked.

When comparing the responses between humans and ChatGPT so far, we only regarded average trends. Average trends are an important aspect for evaluating our methodology, as it assumes that having relatively few erroneous classifications should not affect the fundamental conclusions, being derived from averages over big numbers. Notably, it is expected that the more samples the averages are based on, the 'cleaner' they are. Thus, we note that while the sample of ten human subjects is very small and should be up-scaled to allow further conclusions, this enabled a surprisingly positive finding; We did not expect that the human average and that of ChatGPT would converge to such high similarity after the evaluation of just as few as ten repetitions having just 100 items each.

Having said that, average trends only provide a coarse measure. To better address our question, we wanted to conduct a direct, fine-grained, comparison between ChatGPT and human judgment, in a paired manner (Figure 6 and supplementary Figure 1). For this goal, we sorted the 100 headlines and corresponding responses according to the average ChatGPT response (specifically, by ChatGPT average value for q1 minus q2, in a descending order), and then sorted the matching human response side by side (See Figure 6. Note that the Humans matrix, now including only q1 and q2 for easier comparison, is not sorted independently as it was in Figure 5a,c but according to the order already dictated by ChatGPT). Finally, we visualized the differences between the matrices by adding a third, subtraction matrix (see Figure 6, right). In this matrix, values range between 1 (bright values – ChatGPT detected sympathy in all iterations, but humans never did) to -1 (dark values – Humans detected sympathy in all iterations, but ChatGPT never did) with 0 (grey) representing complete agreement between the two.

This visualization revealed, once again, the high similarity between the responses, as can be seen by the prevalence of the grey-coloured items. Quantitatively, the responses were highly correlated between humans and ChatGPT for both q1 ($R(100) = 0.62$) and q2 ($R(100) = 0.83$). Further conclusions can be reached by observing the explicitly depicted headlines where discrepancies emerge between humans and ChatGPT (see, for example, q1 for item #3, or q2 for item #20. While examining the findings, it is interesting to refer to Supplementary Document #1 for ChatGPT's explanation regarding each item. The results for the full six-question-query are presented in Supplemental Figure 1).

Experiment #6, comparing the BBC to other global news outlets

The results presented above provide strong evidence of unequal framing and expressions of sympathy towards Israel and Palestine in the BBC coverage of the conflict. It has also demonstrated that the novel methodology we apply for empirical evaluation of the impartiality in sympathetic framing, is robust and valid – reliably detecting sympathy bias in a way that highly resembles human judgments.

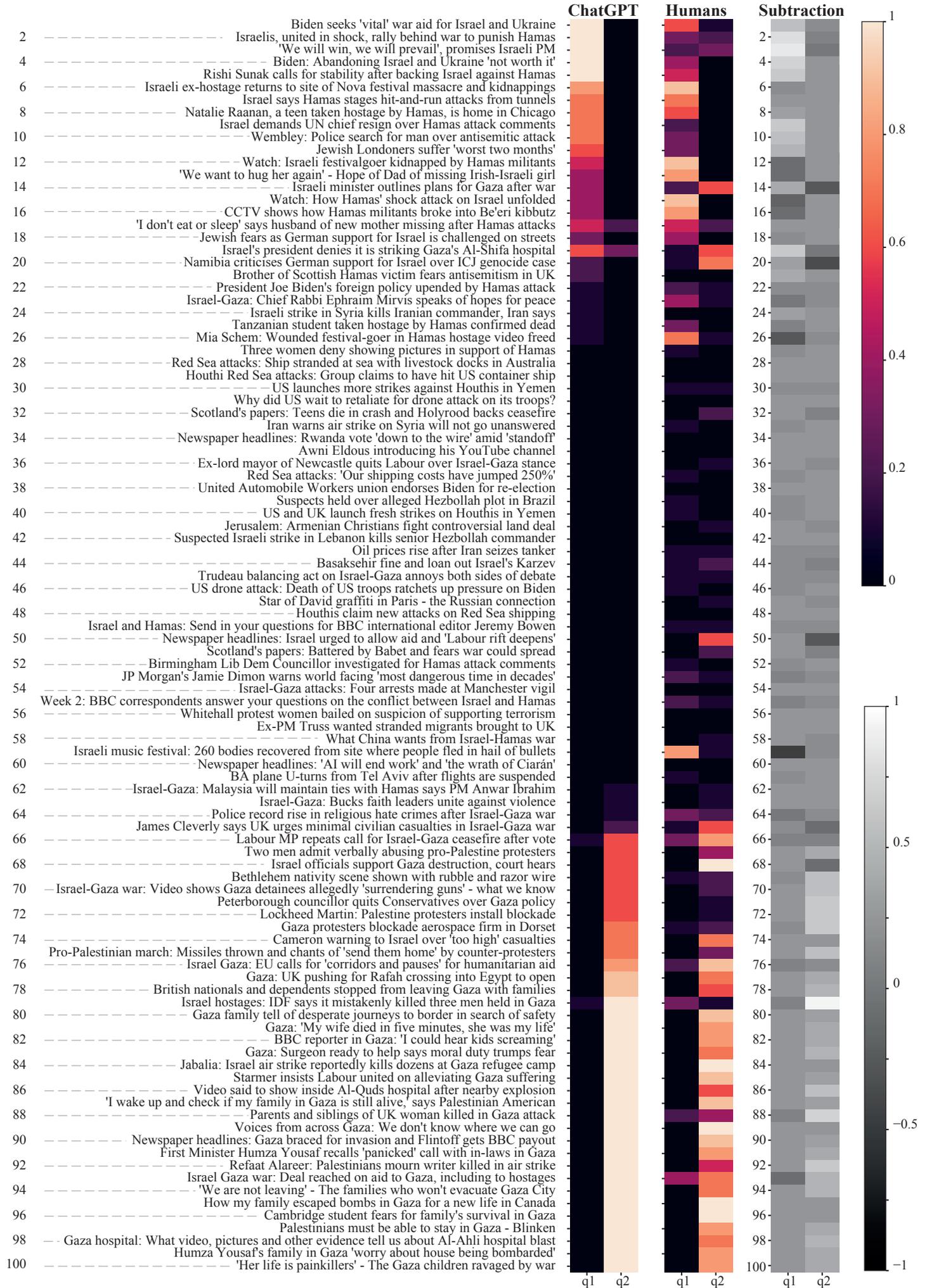


Figure 6. Correlation Between Human Participants and ChatGPT Responses. Responses for q1 & q2 received from ChatGPT for 100 headlines, averaged over 10 iterations and sorted in descending order based on q1-q2 scores. Responses from 10 human subjects are sorted in a matched order, such that the same 100 items could be viewed together with their associated headline. The black & white subtraction matrix represents the difference between the two: darker colors imply higher degrees of sympathy according to human judgment when compared to ChatGPT, brighter colors imply lower degrees of sympathy according to human judgment when compared to ChatGPT and grey colors imply agreement between the responses. The figure highlights the general calibration and disparity between human and AI evaluations of sympathetic occurrences, illustrating the effectiveness and validity of ChatGPT's responses in reflecting human judgments. This agreement could be further appreciated by the high correlation between the human and ChatGPT average values (q1; $R(100) = 0.62$. q2; $R(100) = 0.83$. Complete query; $R(600) = 0.72$), as well as by the fact that average of both subtraction columns is not significantly different than zero (q1; Mean = -0.02 , SD = 0.23 , $p = 0.39$. q2; Mean = 0.026 , SD = 0.24 , $p = 0.29$).

However, it may not be clear to what extent this detected bias, even if accurate, was a result of partial and biased coverage by the BBC, and to what extent the reality unavoidably underlaid this; i.e., that the enormous damage, and especially the high number of casualties in Gaza, naturally and inevitably result in more sympathy towards the side with ‘greater suffering’ - the ‘underdog’.

We note that some of the findings presented thus far already suggest that the latter explanation is implausible, at least in some cases. For example, the discrepancy between the sympathy expressed in the main texts and the headlines. Even if one accepts that the level of imbalance in articles is reasonable, we expect the same level of discrepancy in articles to be reflected in headlines. Yet, the level of imbalance is significantly higher in headlines, compared to the level of imbalance in the main text (compare Figure 1 to Figure 2, or within Figure 3 – compare main-texts’ results to headlines’ results). The same logic is valid for the discrepancy between the BBC English and Arabic websites. The level of imbalance in one language should be reflected in another language. Yet, the data from BBC Arabic website exhibits a consistently higher level of imbalance, compared to the English dataset (compare the results from Figures 1 and 2 to those in Figure 3). Furthermore, the later analysis also demonstrated that higher sympathy levels were always expressed for Gaza, even during the first week following the Hamas attack on October 7th (see Figure 3b).

Nevertheless, we looked for additional ways to dissociate these two alternatives (opinion-induced bias vs. facts-induced). For that purpose, we compare the conflict coverage by the BBC to that of other news outlets worldwide. Here, we utilized the fact that ChatGPT is not a regular ‘reader’, but one capable of consuming and processing vast amounts of data.

Experiment #6, Results:

ChatGPT exhibited a great ability in detecting sympathy within short headline texts. We therefore implemented this strategy to validate our methodology and provide an external reference point for comparison with the BBC, undertaking an assessment of our headline analysis across various other news outlets. For this purpose, we utilized the GDELT dataset (Global Database of Events, Language, and Tone. See [here](#)), an extensive open source that monitors global news media published worldwide. Initially, we filtered the dataset to extract only items from the relevant time frame (the four months following October 7th, 2023), which also contained one or more of the six words used for filtering the BBC data (see Methods for further elaboration). For the robustness of our analysis (which depends on high volumes of data such that the noise inherent to any measurement averages out) we only included news outlets for which we had at least 100 items. This strategy yielded approximately 630,000 items published in about 1,500 news outlets. While the results shown in Figure 7 were filtered for visualization purposes, the full results can be zoomed in and viewed in Supplemental Figure 2

Due to the remarkably similar results obtained from comparing the ratios $q_2:q_1$ and $q_4:q_3$, and to streamline our analysis, we focused solely on quantifying the ratios $q_2:q_1$ across headlines from various news outlets. These ratios were averaged throughout the complete (four months) time-window, to ensure consistency and stability with sparser data points. This process resulted in a single number representing the sympathy ratio for each news outlet in our analysis. All the news outlets were then plotted on the same axis to allow for comparison.

Sympathy ratio: Israel/Palestine (log scale)

- Muslim affiliation
- Jewish affiliation
- Other

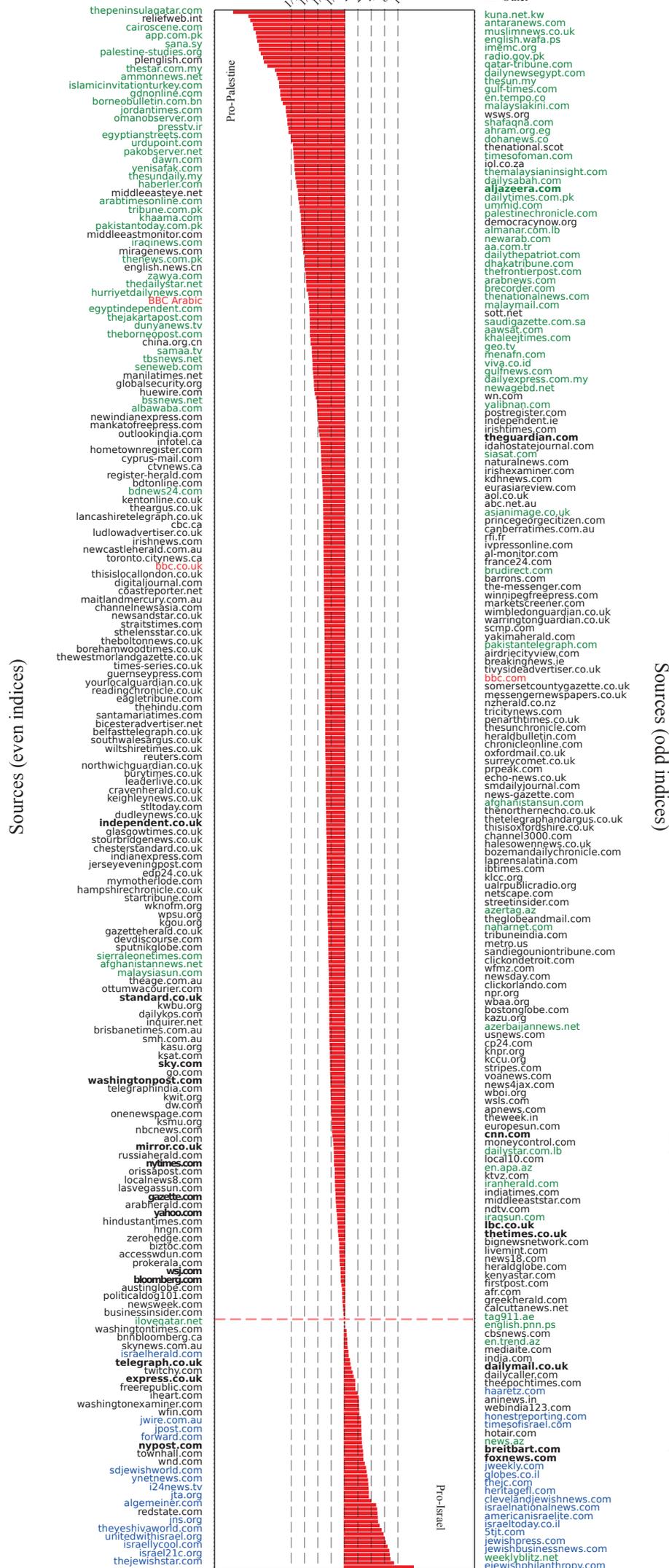


Figure 7. Extensive Comparison of Sympathetic Coverage Between the BBC and Other News Outlets. Distribution of the sympathy ratio (Israel:Palestine, or q1:q2) in headlines taken from 342,559 articles across 376 news outlets, as published in the GDEL dataset. Results for our data (from both the BBC Arabic and BBC English websites) are depicted in red for emphasis. The vertical axis represents different news outlets, accompanied by the number of items they include and sorted in a descending manner according to their total Israel:Palestine sympathy ratio. The horizontal red bars show the sympathy ratio (q1:q2 measured over the complete dataset for every source) in a log scale. Blue and green colored news outlets were classified by ChatGPT as Jewish/Muslim sources, respectively. This classification was done post-hoc, while during the sympathy experiment ChatGPT was blind to anything but the actual headlines. The robust separation between these sources at the two edges of the graph, once again validates our methodology. For this specific graph, Jewish and Muslim outlets were filtered out if they had less than 100 items, whereas all other sources were filtered out if they included less than 500 hundred items. Different parametrization would have led to different number of articles and sources (e.g. setting a threshold of 100 for the inclusion of all sources would yield a 630,000 articles graph, with too many sources to depict in one figure) but the qualitative conclusions of the results would remain similar. The two triangular shapes emerging in the figure demonstrate the different “sides” news outlets take in covering the conflict, where balanced sources are located around a ratio of 1, which stands for equal sympathy to both sides overall, marked with a dashed red line.

We used blue and green hues to colour the extreme regions, where news outlets exhibit a sympathy ratio larger than 2:1 to any direction (i.e. a ratio $<1/2$ for pro-Israeli outlets and >2 for pro-Palestinian outlets). Furthermore, we asked ChatGPT to label both Jewish and Muslim news outlets, as we predicted that they would generally exhibit distinguished bias in the sympathy they create, due to their apparent potential impartiality in relation to the conflict. The results of this experiment both reaffirm the validity of our methodology and demonstrates how extreme is the imbalance exhibited by the BBC in general, and mainly in the Arabic language. Figure 5: See how the Jewish (blue) and Muslim (green) news outlets are grouped at the edges of the graph and note the location of the BBC ENGLISH and BBC ARABIC (both in red) across the wide range of news outlets covered by GDELT). Also note the blue and green regions of extremely unbalanced news outlets, defined as those with a sympathy ratio bigger than 2 towards Israel or Palestine, respectively.

Experiment #7, Expanding the Assessment to Other Platforms

In the experiments undertaken so far, we probed texts from online websites. However, BBC News utilizes a variety of publication channels and formats to reach a global audience, including Television, Radio, Podcasts and Videos. Next, we asked to assess whether our novel methodology could be generalized and used for reliably measuring sympathy across these platforms too. As case studies, we decided to focus on chapters published during our temporal window of interest from the following BBC programs: TV - News at Ten, Newsnight and Breakfast. Radio - BBC Today. Podcasts – by Liz Doucet and Jeremy Bowen. In addition, due to our previous findings regarding the texts published at the BBC Arabic website, we also decided to investigate their published videos (translated to English). Since our methodology receives texts as input while these sources are all based on speech, we were curious to see if the same bias patterns could still be detected (if they exist) even if some of it is likely lost in the translation from speech to text due to, e.g., the absence of tone.

Experiment #7, Results:

We started by converting all the required data from speech to text, using an AI based tool (by a private company named 'Bluskai'). The resulting data was sorted in a table, in which rows containing the texts are divided by utterances for different speakers and accompanied by time stamps. For this analysis, we divided the data into chunks of ~ one minute (see Methods for details). We searched these samples for our 6 keywords and chose our samples of interest if they included one or more of these words or if they were adjacent to such a sample (Figure 8a, 'Analysed minutes'). We then inputted those samples sequentially, just as per our previous analysis, changing only the first sentence in the prompt to fit the new settings: **"I have the following text, taken from a Radio/podcast/Television show..."**

This strategy yielded in total 11,294 (out of 37,176) samples/minutes that were sent to ChatGPT for classification (see division by program in Figure 8a, 'Total Minutes'). These samples included altogether 7,328,260 words, with 197 being the average number of words per sample. The results

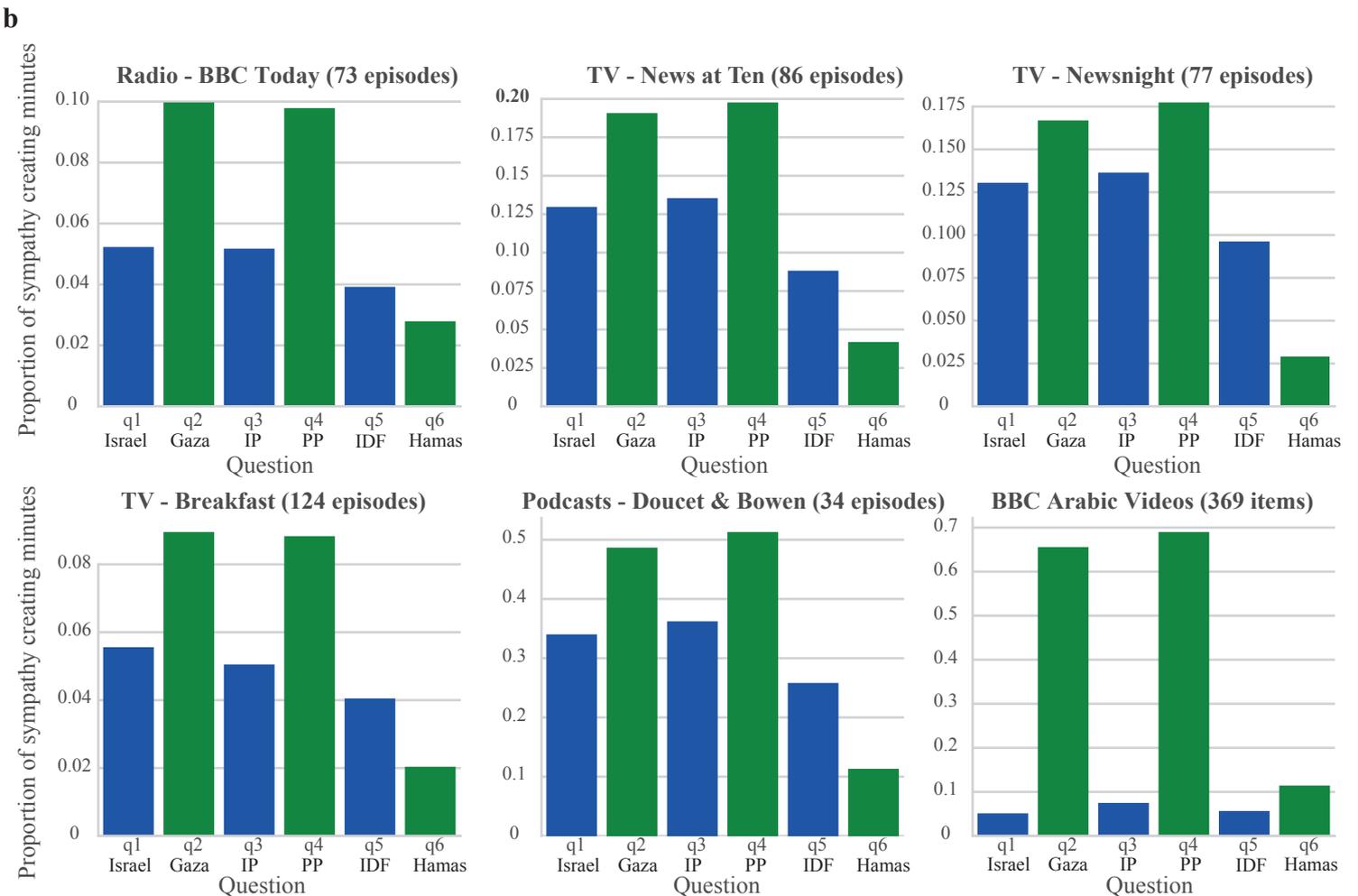
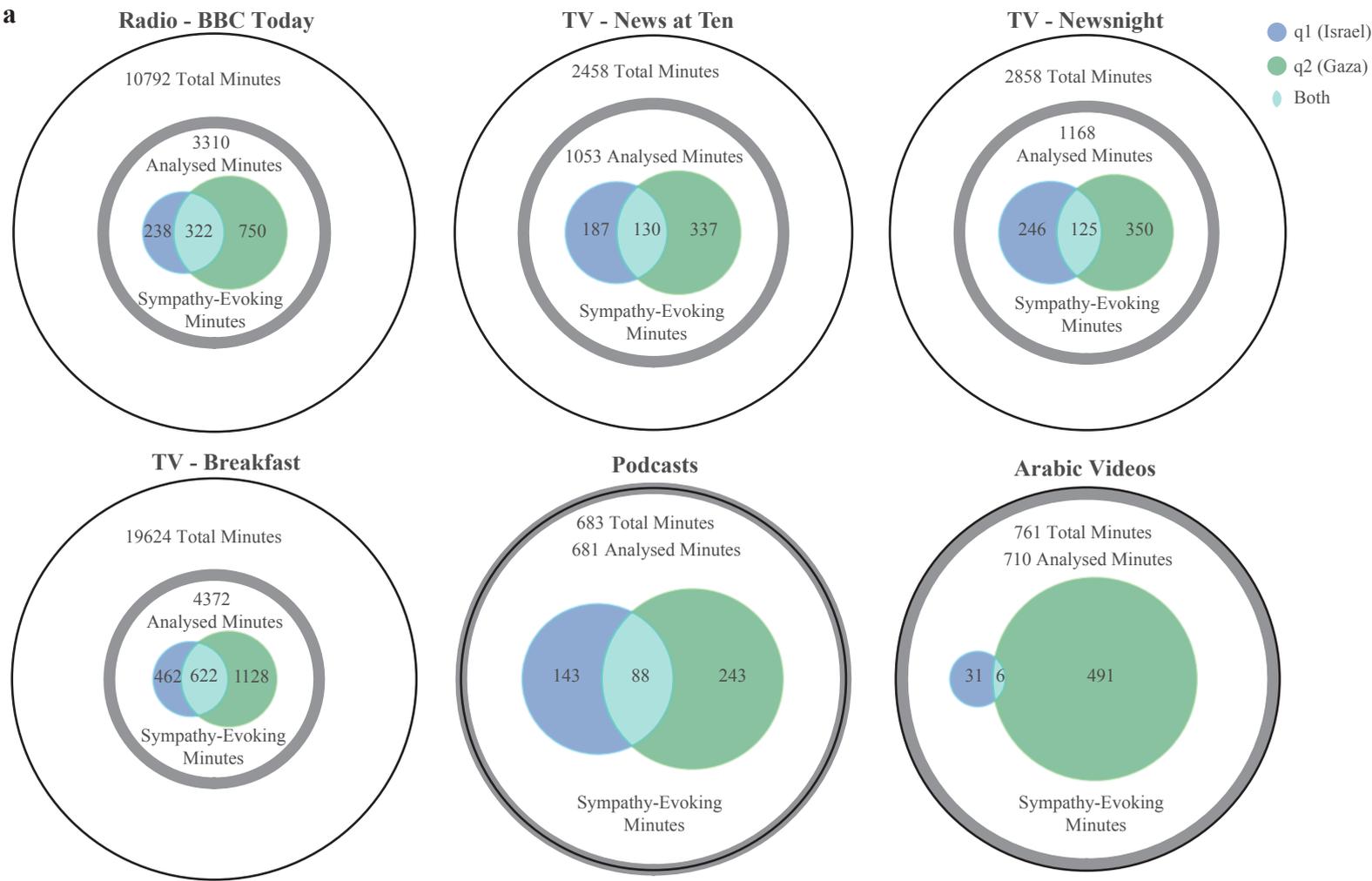


Figure 8. Comprehensive Analysis of Sympathetic Coverage Across Multiple BBC Publication Channels. **a)** The figure depicts the data used in the extensive study regarding sympathy towards Israel and Palestine across various BBC publication channels during the four months following October 7th, 2023. The analysis includes radio programs, TV shows, podcasts, and translated videos. The most sophisticated AI tools were used to transform text to speech for the purpose of this analysis, ensuring accurate and consistent evaluation of the content. The complete datasets, represented by the thin black circles, were divided into chunks, and relevant chunks of text were then chosen based on the presence of one or more of our keywords. The relevant chunks, represented by the thick grey circles, were sent to ChatGPT and yielded the results depicted in the Venn diagrams for the two main questions in our query **b)** The proportion of sympathy-creating minutes is shown for the complete query and for each channel, providing a broad view of how different media formats within the BBC network portray the Israel-Palestine conflict. While all channels exhibit bias in sympathy that favors Palestine, the Arabic videos exhibit this bias on a different scale, showing a much more prominent example of extremely unbalanced coverage. IP/PP - Israeli/Palestinian People.

for this analysis are presented in Figure 8, illustrating how the imbalance of sympathy for Palestinians /Israelis is consistent throughout different BBC platforms, with the content in the Arabic language pushing this trend forward, being extremely biased even when contrasted with the already-material imbalance in the BBC's English content.

Discussion

Contrary to the obligations of impartial reporting, our analysis of sympathy levels in BBC reporting on the Israel-Gaza conflict reveals clear imbalance in the framing of the two sides. The data we analysed consistently indicate a clear bias favouring the Palestinian side, evident across various platforms, formats and languages.

Focusing on articles published online, we showed that the imbalance in coverage was robust in time (Figures 1-3) and especially pronounced in headlines, where sympathy towards Palestine dramatically outweighs that towards Israel, by 3/6-fold on average in the English and Arabic websites, respectively, peaking at tenfold or more at some time-points (see Figure 2 for English and Figure 3 for Arabic headlines analyses). Furthermore, and independently of the main text analysis, we have found that even in the relatively few articles where the main text carries some sympathy towards Israel, this sympathy is rarely reflected in the headline, whereas the same proportion is doubled for Palestine (Figures 2f,3d).

Regarding the BBC Arabic, these tendencies are notably exacerbated: Our analysis demonstrated it to be amongst the most extreme of all the media sources we scrutinized. A thorough investigation unveiled how its bias surpasses that of BBC English across all measured parameters (compare the results in Figures 1,2 to those in Figure 3).

In addition to the conclusions drawn from our analysis, it is noteworthy to highlight the innovative methodology employed in this study. By utilizing ChatGPT-based analysis to assess sympathy levels in big pools of news articles, particularly through the quantifying sympathy occurrences in main-texts, headlines, and their ratios, we introduced a novel approach to studying media bias.

Our method allows for an efficient big-data analysis, via an objective evaluation, thus mitigating potential biases inherent in human subjective judgment and providing insights into the nuances of media coverage. Our innovative, straightforward application of the cutting-edge NLP technology proved to be exceptionally robust; different repetitions yielded similar results, indicating reliability (Figure 4).

Furthermore, our results reflected human judgment remarkably well; responses from human participants and ChatGPT converged to be very similar, even though our sample size was small (Figure 5). This is especially important, as the absence of 'ground-truth' for calibrating our tool makes the average human opinion a very reasonable choice. Moreover, a closer observation, where responses by humans and ChatGPT to the same items were examined side by side, revealed good fits beyond the average level; headlines received similar scores by humans and ChatGPT as reflected in high correlation values, both for the complete 6-questions-dataset, and specifically for

our main questions (Figure 6 and supplementary Figure 1). In the cases where there is disagreement between ChatGPT and humans, reading ChatGPT's 'explanations' provides an interesting (though anecdotal) observation (Table 1, Supplementary Document #1).

After establishing its robustness, we applied our methodology to the assessment of other news outlets. This assessment, based on hundreds of thousands of items, highlighted the distinctiveness of the BBC's standpoint, indicating an exceptionally extreme position even when contrasted with international counterparts, some of which are explicitly hostile to Israel (refer to Figure 7). We pushed the limits further, as we wanted to apply the method on data coming from completely different platforms, all based on speech that was converted to text using advanced AI tools. In this analysis, once again, we revealed a sympathy bias widely spread in various BBC publication channels, with the videos published in the Arabic website being the most biased by far, replicating our previous results again.

Taken together, these findings raise important concerns about journalistic impartiality and the influence of media framing on the public perception of conflict dynamics. While it is essential to recognize the complexities of the Israel-Palestine conflict, media outlets must strive for balanced and fair reporting to foster a nuanced understanding among their audiences.

The novel approach we developed facilitated the systematic analysis of a vast dataset of news articles, offering valuable insights into the portrayal of the Israel-Palestine conflict in the media. Not only does this method enhance the rigor and comprehensiveness of our findings, but it also has the potential to establish a precedent for future research in the realm of media studies. By deepening our understanding of media dynamics, it can contribute to fostering impartial and balanced reporting.

Specifically, the disparity observed in BBC reporting, particularly in the headlines, underscores the need for ongoing scrutiny of media practices and a commitment to transparency and accountability in journalism. Moving forward, efforts to address bias and promote balanced reporting are crucial for fostering informed discourse and facilitating constructive engagement with complex geopolitical issues like the Israel-Palestine conflict.

Methods

Main Dataset (BBC English Website)

Our dataset for experiments 1-3 comprised 1,529 items scraped from the BBC website, published within our defined timeframe of interest (October 7th, 2023, to February 7th, 2024). Items were included only if somehow relevant to the Israel-Gaza conflict., i.e. published in the middle east tab and/or contain at least one of the following six sequences: [israel, gaza, palestin, idf, hamas, antisem]. It is worth noting that we employed abbreviated sequences like 'palestin' to encompass both the terms 'Palestine' and 'Palestinian'. The main dataset in experiments 1-3 was collected by the Asserson legal team, with the result being replicated with a fully automated data collection process (not shown).

After sending the articles sequentially to ChatGPT, including the complete prompt every time, we expected (and accepted) strictly answers in the format 1) yes/no ... 6) yes/no. our query yielded 1481/1529 cases for which responses of this kind were received from ChatGPT. 40 of the 48 missing items did not include any content as their main text, being mostly videos or images. For the other 8 items (about 0.5%), ChatGPT failed to return an answer in the required format.

We replicated the results of this experiment (the results are not shown in the current report) with a dataset we collected with the fully automated scraping system, provided by Bright Data.

ChatGPT API Usage

We wrote a custom Python code, using the Application Programming Interface (API) made available by OpenAI, the creator of ChatGPT. In this and in all other experiments ChatGPT-4 was the model of choice and all default settings (i.e. temperature, context, etc.) were kept.

Transcripts Analysis and Choosing the Right Size for Single Samples

Our methodology is based on quantitative analysis - Multiple complex samples (i.e. texts written in natural language, encoded in high – and due to their different lengths also varying - dimensions) are classified in a binary manner and reduced to a simple yes/no answer. While the basic data units used in our previous analyses were well-defined (main-text/headline of one article), the current analysis posed a new challenge: dividing the data in chunks. First, we needed to choose the basic unit that suits this analysis. While this may sound as a simple, somewhat technical decision, it is in fact an important one and not so trivial.

On one hand, since the power and precision of our methodology does not come from single classifications (which could be noisy), but rather depends on large-scale numbers, the division to samples should provide a high number of samples. For example, if one seeks to apply the methodology on, e.g., a radio show containing 10 chapters, choosing a chapter as the basic unit would result in a far too small a dataset with only 10 samples, reduced to yes/no answers. Furthermore, if the size of these chunks is too big – a question like “does this text create sympathy” may be answered positively in the vast majority of times, reflecting an uninformative query (think about the question: “Is there a person drinking water right now in NYC?”).

On the other hand, choosing chunks that are too small may result in inaccurate results due to the lack of information and context that are needed for the correct classifications of the single samples (think about the question “Is there a person drinking water right now, standing within a certain 1x1 meter area within some unpopulated land?”). Thus, the size of the chunks should be as small as possible to allow the highest number of samples, while still keeping the test informative. Here, we chose to divide the samples using the temporal information (time stamps) associated with them, as this is a reliable measure that should be easy to get with most transcripts. We chose to split the data in chunks of approximately one minute, concatenating all of the utterances within a minute (note that, e.g., if an utterance was longer than a minute, it was sent to ChatGPT as is). Admittedly, while this strategy yielded high numbers of samples, where individual samples contain enough content for informative tests (i.e. including both “yes” and “no” answers to our question, think about the question: “Is there a person drinking water right in an arbitrary group containing, e.g., 200 people?”), it seemed to be somewhat arbitrary.

Future research should regard the ideal size of chunks that would allow the most efficient/precise analysis. It is tempting to hypothesize that the solution would be one that maximizes the entropy of the single tests (i.e., choosing chunks size such that individual tests have ~ 50% chance to result in either yes/no answer, Shani-Narkiss et al., 2020).

Humans vs. ChatGPT

For the human participants essay, we received valid answers for 997 out of the 1,000 questions we asked in total (10 human participants x 100 headlines each). For the three questions that were left blank, all by the same subject, we filled all missing values with “no”.

GDELT Experiment - Labelling the Muslim and the Jewish News Outlets

For the labelling of the Muslim and the Jewish news outlets (Figure 7, green and blue, respectively) we used ChatGPT4o. We note that the labelling was meant to allow a coarse impression, and it may contain some random errors. The exact prompt was as follows:

“I have a list of website sources and would like the following information for each source: [Source, Country, Religion, Global Rank, Local Rank, Ownership].

Here are the specifications:

- Source: The name of the media outlet, exactly as given.*
- Country: The country of origin for this media outlet.*
- Religion: The primary religion associated with the media outlet (or the country of origin, by default).*
- Global Rank: Does this source ranked among the top-100 global news outlets, in terms of readership size (strictly Yes/No/NA).*

- *Local Rank: Does this source ranked among the top-10 local news outlets (country-level), in terms of readership size (strictly Yes/No/NA).*

- *Ownership: Private, Public, NA, Other (strictly one of these options)."*

Replications and Variations

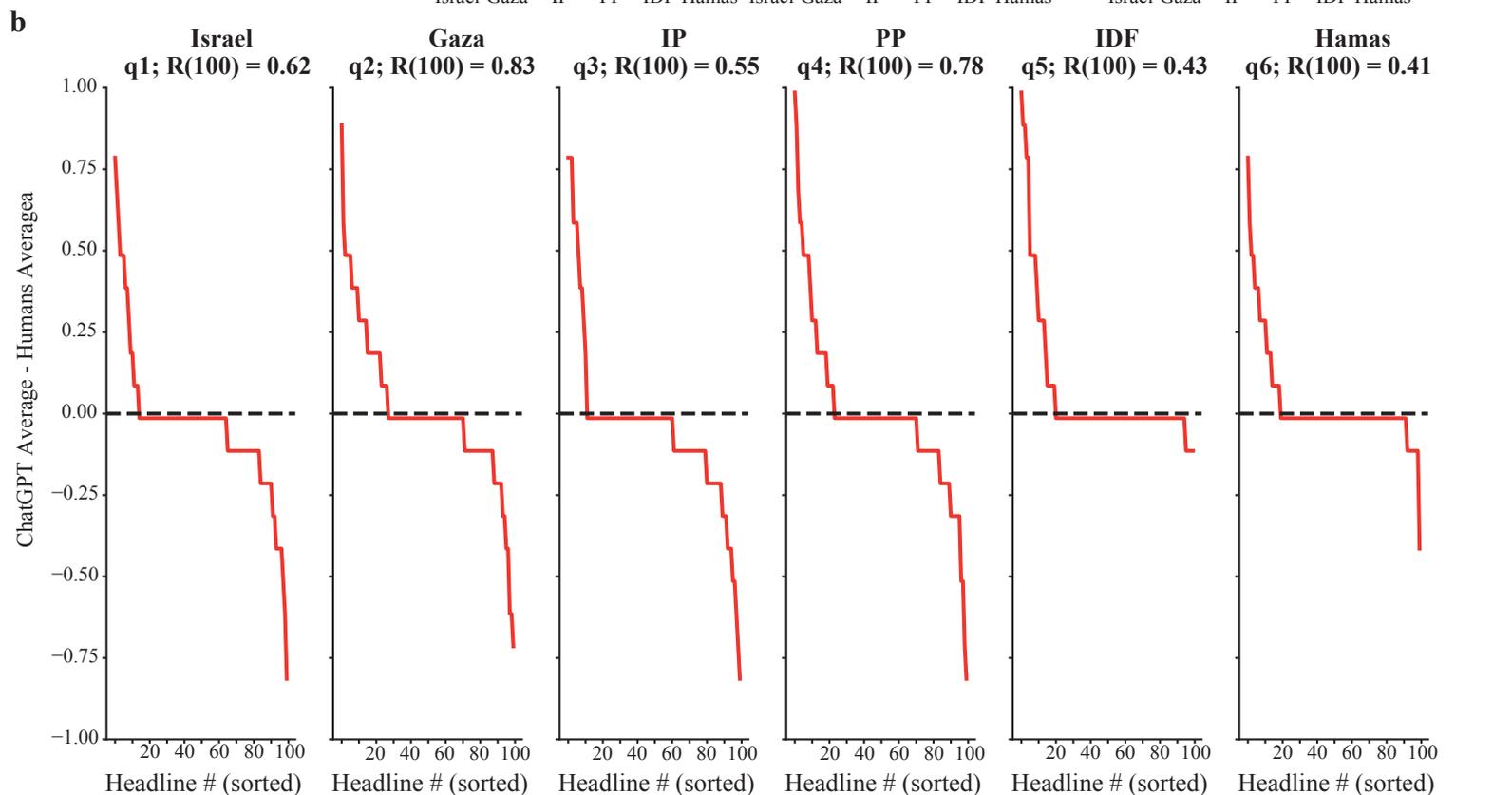
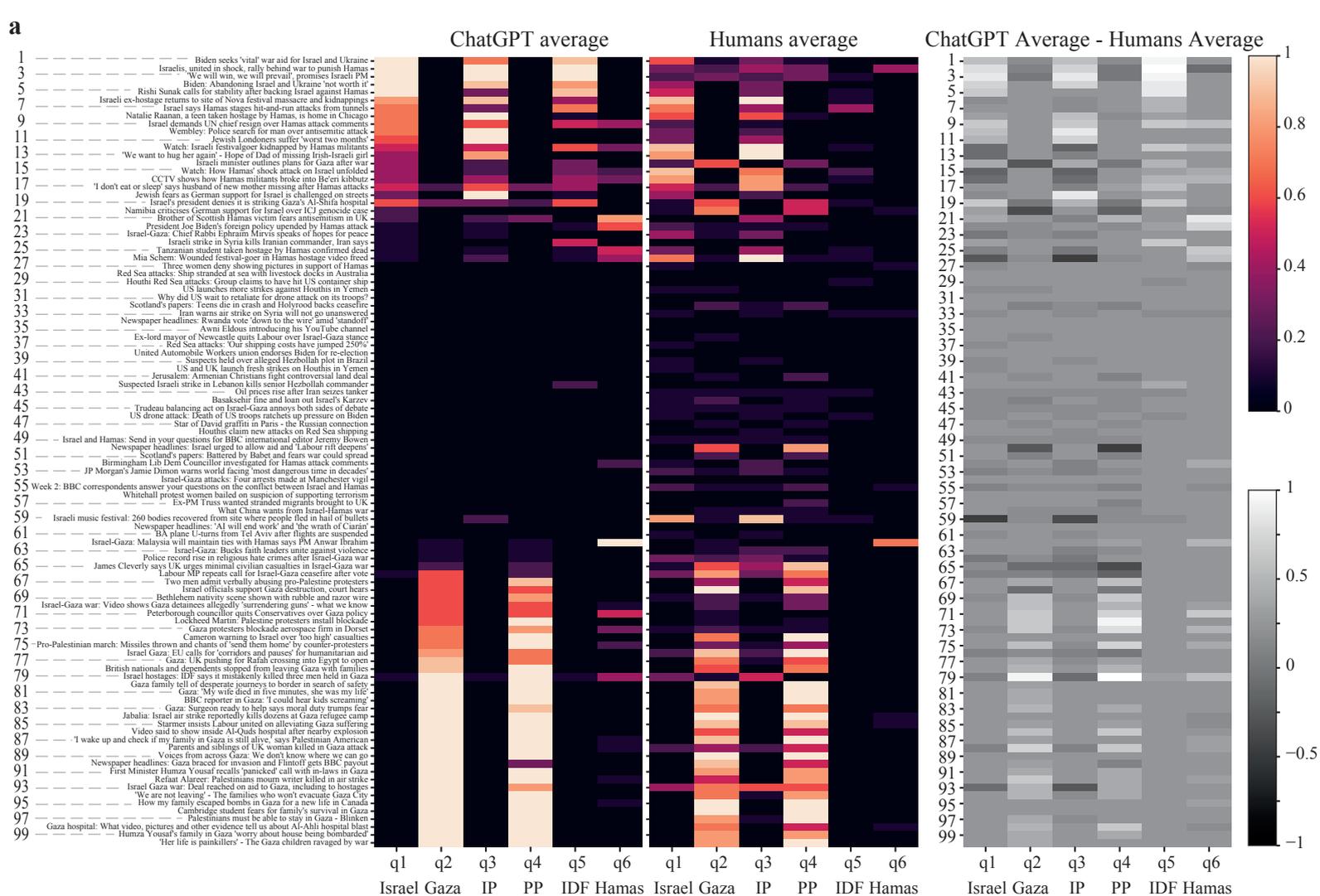
The extensive study described in this report included replications and variations not included in the current document. For example, we analysed another version of the main dataset encompassing the articles from the BBC English website, as we wanted to make sure that a fully automated data collection would yield a similar result to that yielded by the data collected manually by the Asserson legal team. Another example is the experiments we conducted over the BBC Arabic dataset; here the report includes our first attempt, where we used (Arabic to English) translation by ChatGPT4. However, we replicated this experiment using two alternative translations, by ChatGPT4o and by Google Translate.

For clarity and transparency, and as a good scientific practice, we focus exclusively on the original experiments we conducted in this report. However, we note that the trends reported here were consistent, demonstrating the robustness of the findings throughout different settings.

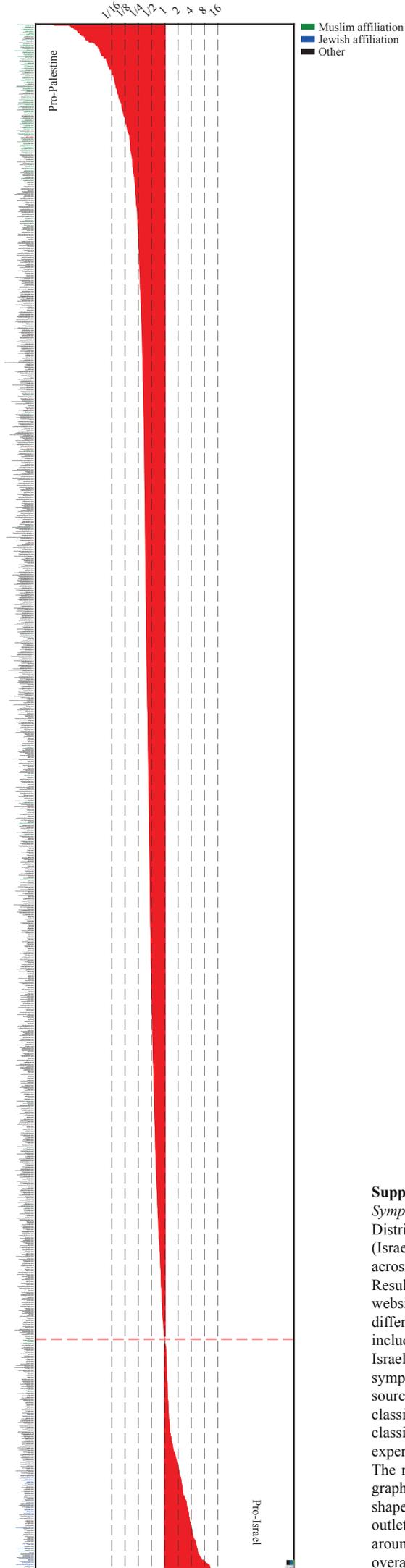
References

- Ayers, J. W., Poliak, A., Dredze, M., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589–596. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Baly, R., Da San Martino, G., Glass, J., & Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 4982–4991).
- Best, C., van der Goot, E., Blackler, K., Garcia, T., & Horby, D. (2005). Europe media monitor. Web intelligence action - technical report EUR221 73 EN, European Commission.
- Castelvecchi, D. (2016, October 5). Can we open the black box of AI? *Nature*. <https://www.nature.com/articles/538020a>
- Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1), 163-173.
- Errica, F., Siracusano, G., Sanvito, D., & Bifulco, R. (2024). What did I do wrong? Quantifying LLMs' sensitivity and consistency to prompt engineering. *arXiv*. <https://arxiv.org/abs/2406.12334>
- Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter? *SIGMETRICS Performance Evaluation Review*, 44(1), 179–192. <https://doi.org/10.1145/2964791.2901462>
- Galtung, J. (1969). Violence, peace, and peace research. *Journal of Peace Research*, 6(3), 167-191.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), Article 93. <https://doi.org/10.1145/3236009>
- Hamborg, F., Donnay, K., & Gipp, B. (2018). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 1-25. <https://api.semanticscholar.org/CorpusID:69309226>
- Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, 10(3), 252-264.
- Hube, C., & Fetahu, B. (2018). Detecting biased statements in Wikipedia. In *Companion Proceedings of the Web Conference 2018* (pp. 1779–1786).
- Kwak, H., & An, J. (2014). A first look at global news coverage of disasters by using the GDELT dataset. In *SocInfo*. Springer. https://doi.org/10.1007/978-3-319-13734-6_22
- Lee, S., Cho, K., Park, J., et al. (2024). Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 NAACL*.
- Lin, Y.-R., Bagrow, J., & Lazer, D. (2011). More voices than ever? Quantifying media bias in networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), 193–200.
- Price, V., & Tewksbury, D. (1997). News values and public opinion: A theoretical account of media priming and framing. *Progress in Communication Sciences*, 173–212.

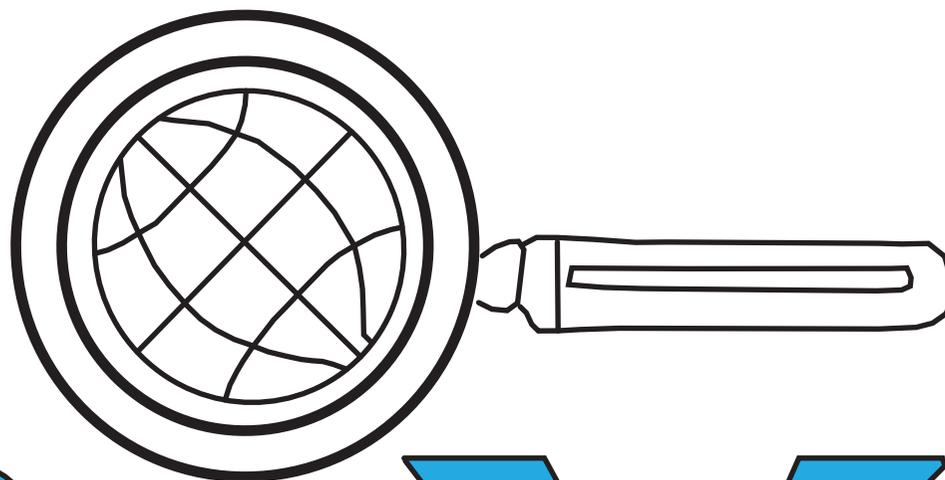
- Rodrigo-Ginés, F. J., Martínez, V., & Rosso, P. (2023). A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237, 121641. <https://doi.org/10.1016/j.eswa.2023.121641>
- Salinas, A., & Morstatter, F. (2024). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv*. <https://arxiv.org/abs/2401.03729>
- Schaaff, K., Reinig, C., & Schlippe, T. (2023). Exploring ChatGPT's empathic abilities. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IU International University of Applied Sciences, Erfurt, Germany.
- Shani-Narkiss, H., Gilday, O., Yayan, N., & Landau, I. (2020). Efficient and practical sample pooling for high-throughput qPCR diagnosis of COVID-19. *medRxiv*. <https://doi.org/10.1101/2020.04.14.20065045>
- Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, 60(3), 556-576.
- Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, 20(5), 520-535.
- Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, 34, 1607–1622. <https://doi.org/10.1007/s13347-021-00477-0>
- Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *School of Computer Science and Engineering, Nanjing University of Science and Technology, China*.
- Ward, S. J. A. (2009). Journalism ethics. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (pp. 295-309). Routledge.
- Wen, Z., & Younes, R. (2023). ChatGPT v.s. media bias: A comparative study of GPT-3.5 and fine-tuned language models. *Applied and Computational Engineering*, 21(1), 249–257. <https://doi.org/10.54254/2755-2721/21/20231153>
- Wolfsfeld, G. (2004). *Media and the path to peace*. Cambridge University Press



Supplemental figure 1. Comparison of Responses Between Human Participants and ChatGPT, the full query (six questions). This figure presents a detailed comparison of the average responses between 10 ChatGPT iterations and 10 human participants for 100 headlines published on the BBC British website. The analysis focuses on all the six different questions (q1-q6) regarding sympathy towards Israel and Palestine. **a**) Exactly like figure 6, but with all six questions depicted. **b**) A quantification of the disagreement between humans and ChatGPT. For each one of the questions, the subtraction between the matrices yields 100 residual values which are then sorted according to size. As can be seen, for questions 1-4, the residuals caused by subtracting the ChatGPT and the Human matrices are centered around zero, reflecting similar assessments between the two on average. For questions 5-6, more positive than negative values are included in the residuals, suggesting that when it comes to sympathy towards militant entities, ChatGPT sensitivity is higher than that of humans. Most residual values are small, reflecting a general agreement between humans and ChatGPT. This could be further appreciated by the correlation values, depicted next to each question. The general correlation value for the full query is $R(600) = 0.72$. IP/PP - Israeli/Palestinian People.



Supplemental Figure 1. Full results, *Extensive Comparison of Sympathetic Coverage Between the BBC and Other News Outlets*. Distribution of the sympathy ratio (Israel:Palestine, or $q1:q2$) in headlines taken from 628,859 articles across 1465 news outlets, as published in the GDEL dataset. Results for our data (from both the BBC Arabic and BBC English websites) are depicted in red for emphasis. The vertical axis represents different news outlets, accompanied by the number of items they include and sorted in a descending manner according to their total Israel:Palestine sympathy ratio. The horizontal red bars show the sympathy ratio ($q1:q2$ measured over the complete dataset for every source) in a log scale. Blue and green colored news outlets were classified by ChatGPT as Jewish/Muslim sources, respectively. This classification was done post-hoc, while during the sympathy experiment ChatGPT was blind to anything but the actual headlines. The robust separation between these sources at the two edges of the graph, once again validates our methodology. The two triangular shapes emerging in the figure demonstrate the different “sides” news outlets take in covering the conflict, where balanced sources are located around a ratio of 1, which stands for equal sympathy to both sides overall, marked with a dashed red line.



RIIME

Research for Impartial Media

