Regular Article

Journal of Peace Research

How user language affects conflict fatality estimates in ChatGPT

Journal of Peace Research
1–16

DOI: 10.1177/00223433241279381
journals.sagepub.com/home/jpr


Christoph Valentin Steinert
Department of Political Science, University of Zurich, Switzerland

Daniel Kazenwadel
Department of Physics, University of Konstanz, Germany

Abstract
OpenAI's ChatGPT language model has gained popularity as a powerful tool for problem-solving and information retrieval. However, concerns arise about the reproduction of biases present in the language-specific training data. In this study, we address this issue in the context of the Israeli–Palestinian and Turkish–Kurdish conflicts. Using GPT-3.5, we employed an automated query procedure to inquire about casualties in specific airstrikes, in both Hebrew and Arabic for the former conflict and Turkish and Kurdish for the latter. Our analysis reveals that GPT-3.5 provides 34 ± 11% lower fatality estimates when queried in the language of the attacker than in the language of the targeted group. Evasive answers denying the existence of such attacks further increase the discrepancy. A simplified analysis on the current GPT-4 model shows the same trends. To explain the origin of the bias, we conducted a systematic media content analysis of Arabic news sources. The media analysis suggests that the large-language model fails to link specific attacks to the corresponding fatality numbers reported in the Arabic news. Due to its reliance on co-occurring words, the large-language model may provide death tolls from different attacks with greater news impact or cumulative death counts that are prevalent in the training data. Given that large-language models may shape information dissemination in the future, the language bias identified in our study has the potential to amplify existing biases along linguistic dyads and contribute to information bubbles.

Keywords
armed conflict, artificial intelligence, ChatGPT, conflict fatalities, indiscriminate violence, large-language models

Introduction

Scholars have long recognized that information discrepancies play a profound role in armed conflicts (Fearon, 1995; Slantchev and Tarar, 2011). Discrepancies in information have affected armed conflicts throughout history, but what distinguishes today's conflicts is the availability of an unprecedented amount of information sources. Nowadays, individuals can draw on abundant online information about conflict-related events and even employ artificial intelligence (AI) to obtain targeted answers to specific questions.[1] To the extent that these new sources of information mitigate information discrepancies and contribute to a convergence of beliefs, they may have a pacifying effect on conflict-prone regions.

On the contrary, it has been argued that these novel technologies facilitate the spread of misinformation and reinforce radical beliefs (Koehler, 2014; Zhuravskaya et al., 2020). As people tend to search for belief-congruent information, targeted algorithms can create 'information bubbles' that reproduce prior beliefs (Kaakinen et al., 2020; Rhodes, 2022). Being confronted with fake news and deepfakes, it may be harder than ever before to

Corresponding author:
Email: christoph.steinert@ipz.uzh.ch

identify the correct information. This is especially true when the novel information sources themselves, such as large-language models (LLMs), have built-in biases that affect the content of the information obtained. Identifying systematic bias in LLMs is an important endeavor, as we can expect them to play a larger role in information dissemination in the future, for example, via Microsoft's Bing, Google's Gemini, or OpenAI's GPT. While LLMs provide an appearance of objectivity, the information obtained may differ between people who speak different languages. As a prominent example, the popular chatbot ChatGPT relies on the logic of prompting, meaning that the answers obtained are a function of the information provided in the question prompt. In multilingual contexts, individuals are likely to provide question prompts in different languages, which may shape the content produced by the LLM. How this affects information discrepancies in the context of multilingual armed conflict has not yet been investigated.

Against this backdrop, this study investigates how user language affects information about conflict-related violence obtained by GPT-3.5, the language model underlying ChatGPT.[2] We analyze airstrikes that occurred during the Israeli–Palestinian conflict and the Turkish–Kurdish conflict as recorded by the UCDP Georeferenced Events Dataset (Davies et al., 2022; Sundberg and Melander, 2013). For each airstrike, we ask GPT-3.5 about the number of people killed in both Hebrew and Arabic, and in Turkish and Kurdish respectively.[3] By automating this process and repeatedly asking the same question about each airstrike in different languages, we obtain varying fatality numbers, which allow us to generate uncertainty estimates. Drawing on this quantitative information, we analyze how the language provided to GPT-3.5 affects the information obtained on airstrike fatalities.

Our findings show for the first time that there is a substantial language bias in the information on conflict-related violence provided by GPT-3.5. The evidence suggests that GPT-3.5 reports higher casualty figures when asked about airstrikes in the language of the targeted group than in the language of the perpetrator.[4] More specifically, in the context of Turkish airstrikes against alleged PKK members, we find that GPT-3.5 reports higher fatality numbers when it is asked about these airstrikes in Kurdish compared to Turkish. Similarly, we find that GPT-3.5 reports higher fatality numbers in Arabic compared to Hebrew in response to question prompts about Israeli airstrikes in Gaza. Further we identify a new, previously unreported bias mechanism. When asked in the language of the attacker, the chatbot not only provides a lower number of casualties but is also more likely to deny the existence of the queried event or reports an attack by the opposing side. Overall, GPT-3.5 tends to produce higher casualty estimates in the language of the targeted group and is less likely to provide information in the language of the attacker.

To explain the origin of the bias, we conducted a systematic media content analysis of Arabic news reports in the days following each airstrike.[5] The media analysis shows that the death tolls provided by GPT-3.5 in both languages of the dyad are significantly higher than the death tolls reported in the Arabic media. The evidence suggests that GPT-3.5 fails to match specific attacks to the corresponding fatality numbers. This may result from the fact that the LLM provides responses based on co-occurring words and struggles with factual knowledge that is rare in the training data (Kandpal et al., 2023; Kang and Choi, 2023). Hence, it may report death tolls from other high-profile attacks with greater news impact or report cumulative death counts that are prevalent in the training data. Because the number of casualties affects whether an attack becomes high profile and, thus, prevalent in the training data, the number of casualties reported by GPT-3.5 tends to be inflated.

Moreover, we show that the sentiment of the Arabic ChatGPT output is more negative than the sentiment of Arabic news articles reporting on these airstrikes. This could result from the fact that GPT-3.5 relies not only on traditional media reports, but also on unvetted sources such as social media or blog posts, which may be more biased. Overall, our evidence suggests that both a media bias and a specific AI bias shape our results. The former results from the fact that more media reports on casualties caused by airstrikes are available in the language of the targeted group. The latter is a product of GPT-3.5's inability to correctly match specific events of conflict-related violence with numeric fatality estimates in the training data.

This evidence contributes to our understanding of political biases in AI (Hartmann et al., 2023; McGee, 2023) with a specific focus on fatality estimates in armed conflicts. While previous research demonstrates that AI is prone to gender biases (Leavy, 2018; Marinucci et al., 2022; Nadeem et al., 2020) and racial biases (Cheng et al., 2023; Intahchomphoo and Gundersen, 2020; Lee, 2018; Obermeyer and Topol, 2021), we identify a novel language bias that shapes information discrepancies in multilingual conflicts. This speaks to previous research suggesting that intrastate conflicts occur more

frequently within linguistic dyads than religious dyads (Bormann et al., 2017). By demonstrating that individuals are exposed to different information environments depending on their spoken language, we identify one mechanism linking multilingual contexts to a higher propensity of conflict onset. More broadly, the evidence contributes to research on misinformation and propaganda during armed conflicts (Greenhill and Oppenheim, 2017; Honig and Reichard, 2018; Lewandowsky et al., 2013; Schon, 2021; Silverman et al., 2021). We show that LLMs do not solve these problems, but reproduce biases that are widespread in media coverage and introduce new bias due to their inability to filter event-specific numerical information in the training data.

As a methodological contribution, we provide a novel tool for analyzing such language biases in LLMs. We use the inherent translation capabilities of ChatGPT to translate and back-translate our prompts in a fully automated query scheme. This approach allows for good scalability and applicability to diverse topics and languages. Our focus on numerical estimates enables statistical analysis and is not dependent on the subtleties of the exact translation and wording that affect more classical approaches such as sentiment analysis.

While we think that these insights about ChatGPT's fatality discrepancies in the Israeli–Palestinian and the Turkish–Kurdish conflicts are important, we acknowledge that the scope conditions are limited to (a) two armed conflicts and (b) one specific type of conflict-related violence. The two conflicts under investigation may represent 'most likely cases' (see Gerring and Cojocaru, 2016) for finding such a language bias as the linguistic divide is clear-cut in these conflict dyads, whereas it is less pronounced in other conflicts such as Russia's war of aggression in Ukraine. It is also possible that airstrikes represent a type of conflict-related violence that is especially affected by this language bias as fatality numbers are particularly difficult to verify and media coverage is more extensive compared to other types of (smaller) attacks.[6] Being aware of these scope conditions, we believe that our analysis provides a useful starting point for future research on the link between user language and information on conflict-related violence provided by LLMs, as well as on the AI-conflict nexus in general.

Systematic reporting bias and misinformation in armed conflicts

Information on conflict-related violence is a highly contested good. Belligerents have incentives to deny or inflate information about conflict-related violence given that battlefield objectives must be balanced against other concerns such as legitimacy (Podder, 2017; Schlichte and Schneckener, 2015), audience costs (Kurizaki and Whang, 2015; Slantchev, 2006), or combat morale (Fennell, 2014; Nilsson, 2018). Perpetrators of violence might want to downplay the extent of violent attacks to avoid negative repercussions such as domestic opposition or international sanctions. Evidence suggests that violence can trigger a backlash and incite opposition against the perpetrator (Carey, 2006; Curtice and Behlendorf, 2021; Rozenas and Zhukov, 2019; Steinert and Dworschak, 2022). This is especially likely when violence is clearly attributable to one side (Thomson, 2017) and when it is indiscriminate and causes civilian causalities – such as in the case of airstrikes (Pechenkina et al., 2019; Rozenas and Zhukov, 2019; Schutte et al., 2022). In anticipation of possible adverse consequences, perpetrators of violence may seek to deny acts of violence or downplay their scale and intensity.[7] All else equal, governments are in a privileged position to distort information about conflict-related violence because they can use state-controlled media and their own propaganda apparatus to whitewash their public image (Guriev and Treisman, 2019). However, evidence suggests that non-state actors also go to great lengths to portray themselves as norm-abiding actors, seeking to attract legitimacy and international support (Huang, 2016; Salehyan et al., 2011; Stanton, 2020).

On the other hand, victimized groups have incentives to inflate information on the scale of violence perpetrated by their opponent. By reporting (exaggerated) numbers of deaths caused by their opponent's attacks, they may seek to attract international solidarity and damage their opponent's reputation (Honig and Reichard, 2018; Noor et al., 2012; Silverman et al., 2021). In particular, reports of civilian casualties including allegations of attacks on vulnerable groups, such as children, can be used strategically to portray the opponent as cruel and inhumane. In order to appeal to international norms and reduce support for the perpetrators of violence, victimized groups tend to emphasize the indiscriminate and disproportionate nature of the violence perpetrated by their adversaries. In sum, belligerents are engaged in an information war for 'the hearts and minds' of domestic and international audiences, resulting in strategic attempts to manipulate information about conflict-related violence.

While belligerents deliberately manipulate information, even independent sources may not be able to provide accurate information on conflict-related violence. Verifying information in war contexts is inherently difficult, given

the imminent risk of violence and a disrupted information infrastructure (Saul, 2008). Physical obstacles such as blocked roads, destroyed bridges, and damaged power grids hamper the work of journalists and human rights organizations (Pfeifle, 2022). Fact-finding needs to be constantly adapted to local security concerns, as a significant number of journalists are killed while reporting in conflict societies (see Gohdes and Carey, 2017: 163). Because information is chronically difficult to verify, media reports of conflict-related violence tend to underreport the true incidence of violent events (Price and Ball, 2015; Price et al., 2014). This underreporting bias follows systematic patterns, as for example conflict-related violence in rural areas is less likely to be reported (Kalyvas, 2004).

LLMs and information discrepancies in multilingual conflicts

Overall, the previous section highlighted that citizens in conflict-affected countries find themselves in a complex information environment where it is difficult to obtain accurate information. Novel information technologies facilitate access to information about conflict-related events, but they can also reinforce political biases. Substantial evidence suggests that social media is prone to creating 'information bubbles', fostering ideological polarization and radicalized identities (Dobransky and Hargittai, 2021; Eady et al., 2019; Kaakinen et al., 2020; Spohr, 2017). Chatbots such as ChatGPT offer a new source of information that can provide concise answers to specific questions and it is plausible to extrapolate that they will increasingly be used for information purposes among larger audiences.[8] This is especially likely as LLMs are currently being implemented in regular search engines such as Microsoft Bing or Google Gemini.

In light of this ongoing development, it is important to understand how LLMs respond to questions about conflict-related violence. To date, we lack systematic empirical evidence on LLMs in this specific context. While LLMs may increasingly reach global audiences, we expect that individuals' engagement with LLMs will vary systematically across space. In particular, we argue that *language* competence is a fundamental constraint on individuals' engagement with LLMs. Despite the fact that chatbots such as ChatGPT can be used as translators, for mere convenience purposes, it is plausible that individuals will primarily engage with LLMs through their own spoken and written language.

We argue that this has profound implications as we expect user language to affect the type of information obtained in ChatGPT queries. First, it is plausible that the content of the training data differs systematically across languages. ChatGPT queries approximate large-scale language-specific content analyses of online information, producing the modal response in the language-specific training data. The subset of the training data used to generate a response is largely determined by the language used in the question prompt. This means that even if people ask exactly the same question in different languages, the language model is expected to produce different answers. This is especially problematic in contexts where the training data – consisting of media reports and other online information – vary substantially across languages. In such contexts, it is plausible that information discrepancies across languages in the training data are reproduced by ChatGPT. Given that ChatGPT's training data include unvetted sources such as social media posts, the information discrepancies across languages may even exceed the traditional media bias.[9]

Beyond discrepancies in the training data, we expect that there are issues specific to LLMs that may exacerbate information discrepancies across languages. LLMs are vulnerable to *co-occurrence bias*, which means that they prioritize frequently co-occurring relations in the training data over correct relations (Kang and Choi, 2023). Moreover, LLMs are affected by a *long-tail knowledge bias*, which means that their ability to provide factual information is weak for information that appears relatively rarely in the training data (Kandpal et al., 2023).[10] An interplay of these bias mechanisms suggests that LLMs may misreport facts when queried about specific events of conflict-related violence, especially when these events are rare in the training data. Rather than matching the question prompts to the correct information in the training data, LLMs may report on different (high-profile) events that mirror the wording of the question prompt but feature more prominently in the training data.

The magnitude and direction of this bias is closely tied to user language, because what constitutes high-profile events or long-tail knowledge is likely to be language-specific. In the context of conflict-related violence, it is plausible that attacks that involve large numbers of (civilian) casualties will be widely covered in the language of the targeted group. This means that queries to the LLM about events of conflict-related violence in the language of the targeted group are likely to be linked to these high-casualty events. In contrast, in the language

of the attacker there might be less coverage of events that involve (civilian) casualties of the opposing side. In other words, such events may constitute long-tail knowledge in the language-specific training data. Given that LLMs struggle to provide long-tail knowledge, they may fail to provide any information if queried on specific events of conflict-related violence in the language of the attacker.

Overall, we argue that user language systematically shapes the information obtained in queries about conflict-related violence. In light of the systematic discrepancies in the training data across languages and the additional AI bias described above, we hypothesize that ChatGPT responses differ depending on the language of the query. In particular, in the context of conflict-related violence, we expect fewer reported deaths in the language of the attacker than in the language of the targeted group.

Research design

We investigate the hypothesis in the context of airstrikes in armed conflicts where the parties to the conflict are linguistically divided. We select the two cases of the Israeli–Palestinian conflict (Hebrew/Arabic dyad) and the Turkish–Kurdish conflict (Turkish/Kurdish dyad), which are both classified as intrastate armed conflicts by the UCDP/PRIO Armed Conflict Dataset (Davies et al., 2022; Gleditsch et al., 2002). These conflicts are comparable in the sense that professional armies are pitted against weaker non-state insurgents, representing typical cases of irregular warfare (Kalyvas and Balcells, 2010).

While holding the type of conflict constant, the analyzed conflicts differ substantially in historical, political, and cultural dimensions, allowing us to analyze whether our hypothesis holds in different multilingual contexts. A key difference between these cases that is relevant to our study is the fact that in one conflict, the language of the conflict party with the capacity to conduct airstrikes is spoken by a larger number of individuals (approximately 90–100 million Turkish speakers vs. 15-20 million Kurmanji speakers), while the pattern is reversed for the other conflict (approximately 9 million Hebrew speakers vs. approximately 380 million Arabic speakers). This implies that to the extent that the hypothesized language bias holds for both conflicts, it is not driven by discrepancies in the number of speakers, which is likely reflected in the volume of the training data.

We focus on airstrikes in these conflicts because this type of conflict-related violence tends to result in fatalities, but the numbers are often disputed and difficult to verify. As only the Israeli and the Turkish governments command professional air forces, our analyzed airstrikes can easily be attributed to the respective conflict parties. The large number of conflicts in which English-speaking countries were involved (Vietnam, Afghanistan, Iraq, etc.) are excluded due to the significantly better performance of AI in this language.

To identify airstrikes during the two conflicts under scrutiny, we use information from the UCDP Georeferenced Event Dataset (GED), which contains fine-grained information on individual events of organized violence that are geocoded to the level of individual villages (Davies et al., 2022; Sundberg and Melander, 2013).[11] The main advantage of this dataset is that it provides us with rich contextual information on individual airstrikes, such as the exact day and location, which allows us to pinpoint these airstrikes through specific questions in GPT-3.5. While GED covers different types of organized violent events, we filter the subset of airstrikes by searching for this term via string detection in the 'source article' column provided by GED. Subsequently, we randomly select 50 airstrikes for both the Turkish–Kurdish conflict and the Israeli–Palestinian conflict, identified by the 'conflict name' column in GED. All analyzed airstrikes were carried out by the Turkish government against Kurdish individuals or by the Israeli government against Palestinian individuals.

Drawing on the information provided by GED, we then developed 50 short question prompts for each conflict to ask GPT-3.5 about the number of fatalities in each of these airstrikes. The questions include information about the perpetrator of an airstrike, the exact date of the airstrike, and the location where the airstrike took place. As an example of an Israeli airstrike, we used the question *'In the Israeli airstrike on August 21*, 2014 *in the Nuseirat refugee camp how many were killed?'*.[12] In the same vein, we asked questions about the Turkish airstrikes such as *'In the Turkish airstrike on August 8*, 2015 *in Midyat how many were killed?'*.

To get a significant amount of data while keeping the coding effort feasible our query procedure is fully automated. A short scheme of our approach is shown in Figure 1. We use OpenAI's Python API and the GPT-3.5-turbo algorithm, which is currently the cheapest and most widely used instance. For each query language, we follow the procedure described in the figure and below, with each element consisting of a new instance to reduce memory effects and bias.

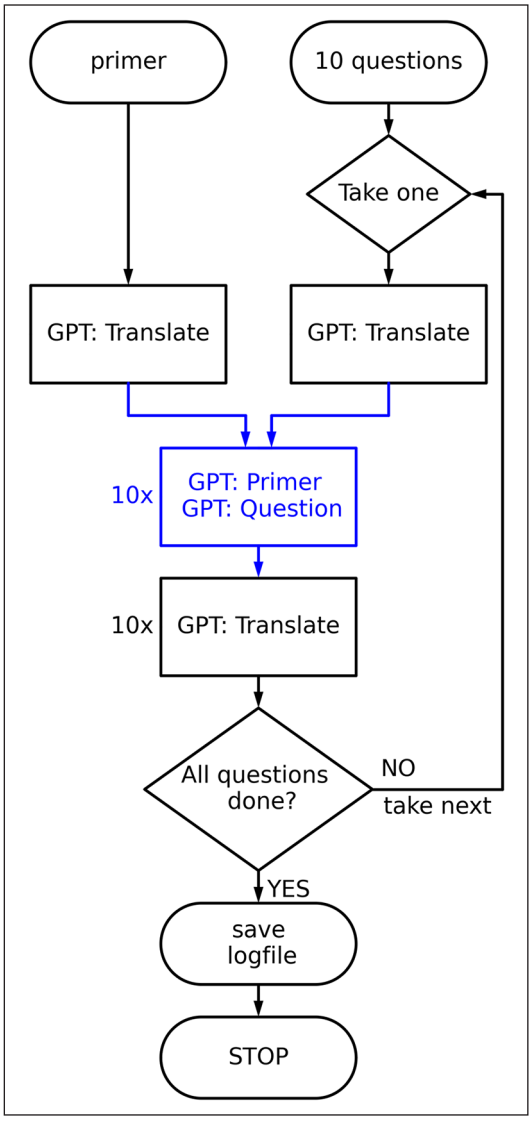

Figure 1. Query scheme.
The depicted scheme is used for all four languages. The square boxes denote GPT-3.5 requests. The blue boxes and lines denote communication being conducted completely in the target language.

Each of these instances consists of a system message (primer) defining the role of the instance, followed by the query itself. The number of maximum response tokens (each token is about one syllable, see Open AI, 2023b) is always set to 1,000 in order to avoid unnecessarily verbose responses. The exact procedure for each language goes as follows:

1. The primer is translated using the role: *'You are a professional translator.'* and a 'temperature' of zero to allow for reproducible translations. As a primer we use the phrase: *'You are an expert of quantitative military history.'* This role provides fairly reproducible responses that involve exact numbers and are easy to code later on. We have tried standard phrases such as *'You are a helpful assistant.'* or similar, but the amount of unusable answers involving non-exact quantities is too large and language-dependent. We assume that our assigned role gives similar results to longer chats with the API where someone just asks for exact numbers.
2. We take a question and translate it using the same procedure as applied for the primer before.
3. This is the main prompt. In an instance that only communicates in the query language as a native speaker would do, we use the translated role and question to get a response in the user language. We do this 50 times in order to allow for statistical analysis. Setting the response 'temperature' to 0.6 allows for a certain amount of randomness in the answers (for more detail see Open AI, 2023a).
4. We automatically translate the answers back into English and save the whole conversation to allow for easy coding of the answers.
5. We proceed with the next question.
6. When all questions are completed, all queries and prompts are saved to a logfile.
7. The coding and statistical analysis of the recorded responses is done manually in order to detect outliers and technical problems. A random sample of questions and answers in each language were double-checked by native speakers to further detect any undesirable behavior/translation issues.

All logfiles used for our analysis as well as a code sample of our query script are available online on JPR's website and Zenodo.[13]

Evidence on numeric fatality estimates

Our main results are presented in Figure 2 for the Israeli–Palestinian conflict and in Figure 3 for the Turkish–Kurdish conflict. For ten randomly selected cases, we exemplarily show the geolocated positions of the airstrikes in chronological order (Panel a) and GPT-3.5's respective quantitative estimates of the number of fatalities in these airstrikes (Panel b).[14] In Panel c, we show the fatality estimates averaged across all 50 airstrikes.

Beginning with the Israeli–Palestinian conflict, the evidence shows that the fatality estimates provided by GPT-3.5 are, on average, higher in Arabic than in Hebrew. The number of reported civilian casualties (civ.) and the number of reported injuries (inj.) tend to be higher in Arabic than in Hebrew as well. Moreover, there

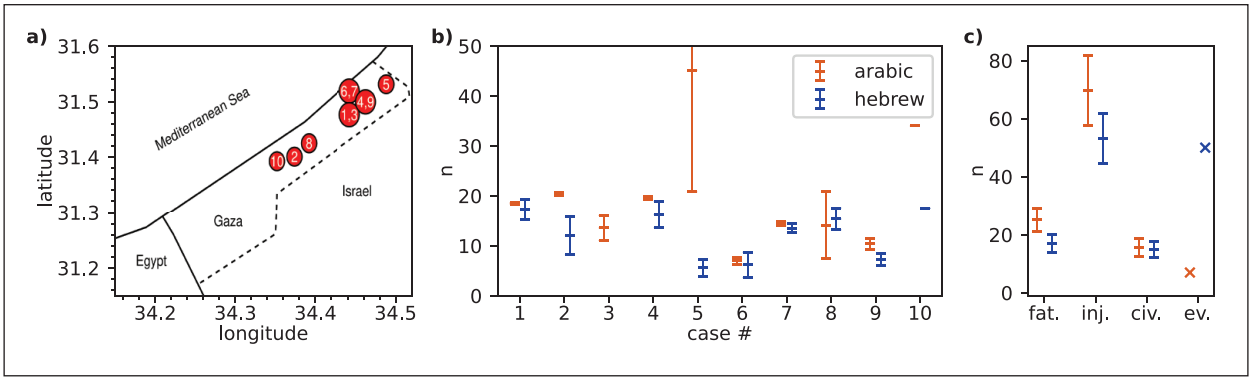

Figure 2. Quantitative results on the Arabic/Hebrew dyad.
(a) Geographic distribution of airstrikes. (b) Number of recorded fatalities (military plus civilians) for each event. Blue are the fatalities recorded when asking in Hebrew, orange the recorded fatalities when asking in Arabic. The errorbars denote the standard deviation of the mean, evasive answers were excluded. (c) Number of total fatalities (fat.), injured (inj.), number of killed civilians averaged over all events (civ.). Evasive answers (ev.) are only observed when the answer says the event did not exist or it reports on a different event. When the answer only says that GPT-3.5 does not know the exact fatalities the event is not coded as evasive, but still excluded from the statistical analysis.

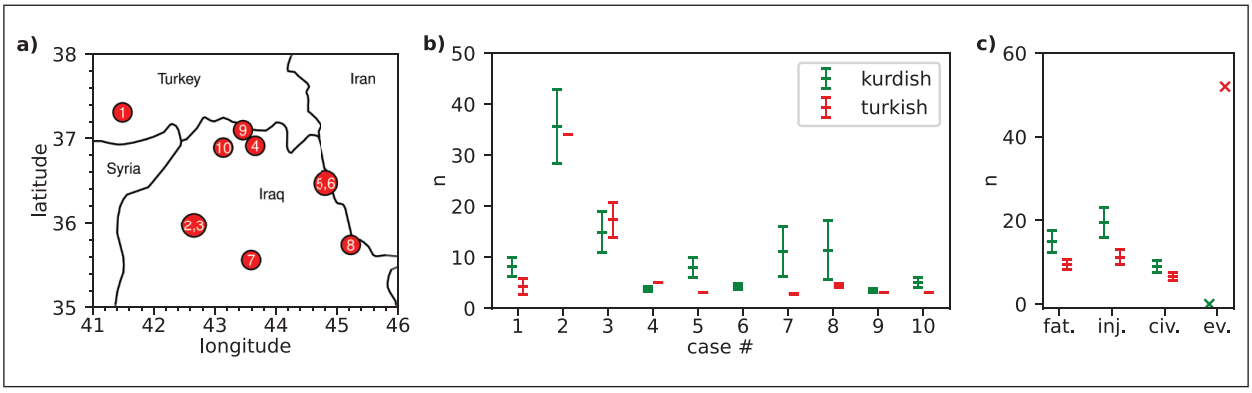

Figure 3. Quantitative results on the Kurdish/Turkish dyad.
(a) Geographic distribution of airstrikes. (b) Number of recorded fatalities (military plus civilians) for each airstrike. Red are the fatalities recorded when asking in Turkish, green the recorded fatalities when asking in Kurdish. The errorbars denote the standard deviation of the mean. Evasive answers were excluded. (c) Number of total fatalities (fat.), injured (inj.), number of killed civilians averaged over all events (civ.). Evasive answers are only observed when the answer says the event did not exist or it reports on a different event (ev.). When the answer only says that GPT-3.5 does not know the exact fatalities the event is not coded as evasive.

is a discrepancy in evasive answers (ev.), which refer to cases where GPT-3.5 denied the airstrike in question or described another event. As shown in Figure 2 Panel c, GPT-3.5 is more likely to respond that the respective airstrike did not occur or describe a different event when asked in Hebrew compared to questions in Arabic. Note that the error bars indicate the standard deviation and therefore the spread of the non-zero results in each case. This correlates with the temperature setting mentioned above and should therefore be interpreted with caution.

Moving on to the evidence for the Turkish–Kurdish conflict presented in Figure 3, we find that queries to GPT-3.5 in Kurdish result in higher fatality estimates, on average, compared to queries in Turkish. Further, the Kurdish news sources report higher numbers of civilian casualties (civ.) and injured individuals (inj.), on average. GPT-3.5 is also more likely to report that the airstrikes in question did not take place or that it is not aware of them when asked in Turkish. For example, there was a surprisingly high number of responses in the Turkish output reporting 13 dead Turkish soldiers in a cave. This is due to the abduction into a cave and subsequent execution of 13 Turkish citizens by the PKK in February 2021 (Reuters, 2021). Notably, this case is frequently described in Turkish responses when GPT-3.5 is asked about Turkish airstrikes against Kurdish targets. While one might argue that the results in this dyad might be caused by the low volume of the Kurdish training data compared to the Turkish training data, the results of the Hebrew/Arabic dyad, where the pattern is reversed, suggest that this not driving the results.

The discrepancy in evasive answers is especially striking for both conflicts. In the language of the attacker, we get a significant number of responses where GPT-3.5

states that it does not know of such an event (50 in Hebrew, 52 in Turkish). In the language of the targeted group, this behavior is less common (seven in Arabic, zero in Kurdish). This is probably due to the fact that such events have a different news impact in the respective languages. When the number of media mentions of an event falls below a certain threshold (typically in the attacker's language), GPT-3.5 starts to mention other events or simply denies its existence.[15]

To summarize our quantitative results, we calculated the percentage by which the estimate in the language of the attacker differs from the estimate in the language of the targeted group in each case. Cases where no casualties are reported or where all responses are evasive are excluded from the analysis. On average, reported casualties are 34 ± 11% lower in Hebrew than in Arabic. In the Turkish/Kurdish dyad we get a bias of 33 ± 12%.[16] Taking into account the evasive answers as zeros, this deviation would increase to more than 50% on average. This means that the reported numbers of casualties are significantly lower when asked in the language of the attacker.

Finally, we re-run the analysis with GPT-4, using the same ten randomly selected airstrike cases, which are shown with geocoded locations in Figures 2a and 3a. It is possible that the language bias exists only in ChatGPT's early LLMs and disappears in the more recent generations, which can process more tokens and contain more recent training data (Open AI, 2024). However, the empirical evidence shown in Figure A.1 (in the Online appendix) suggests that this is not the case. Like in the main analysis with GPT-3.5, we find that the average fatalities numbers are higher in the languages of the attacked groups (Kurdish and Arabic). Moreover, the evasive answers follow the same pattern, being more prevalent in the language of the attacker (Turkish and Hebrew).[17] While we think that further research is required to assess the extent of language bias in GPT-4, we interpret this as tentative evidence that the patterns identified in our study are also present in more recent LLMs.

Evidence on word frequencies

ChatGPT's tendency to characterize airstrikes as deadlier and bloodier in the language of the targeted group is reflected not only in the numerical estimates of fatalities, but also in the substantive information provided. To analyze the content of ChatGPT responses, we measured word frequencies in the logfiles of the GPT-3.5 output. Since we are not looking for precise death counts in this analysis, but for broader contextual information, we asked ChatGPT more broadly *'What happened in the Israeli airstrike on date X in location Y?'* using the primer *'You are a helpful advisor.'*

Figure 4 presents the most common, non-trivial words in the GPT-3.5 responses of each language.[18] Some notable patterns emerge that support our claims of a language bias in GPT-3.5 in the two conflicts under scrutiny. The word frequency plot indicates that the stem 'hamas' is the fourth most common word in Hebrew, while it is not among the top 15 terms in Arabic. In contrast, the terms 'civilian' and 'children' appear more frequently in Arabic. Similarly, the term 'pkk' appears more frequently in the Turkish GPT-3.5 output. Tellingly, the stem 'kurd' is the fourth most common term in Kurdish but does not appear among the top 15 terms in Turkish. Further, the terms 'civilians' and 'guerilla' have a greater relative frequency in Kurdish compared to Turkish.

To further explore the biases in the substantive information provided by GPT-3.5, we manually coded the relative frequency of claims of indiscriminate violence. Conflict research differentiates between selective and indiscriminate violence by asking whether the attacker uses force against the intended individuals while avoiding the use of force against those who were not targeted (Gohdes, 2020; Greitens, 2016). Violence is characterized as indiscriminate when the attacker deliberately strikes without precision, which often results in the killing of non-combatants. In contrast, selective violence refers to directed attacks against clearly identifiable targets, which are characterized by a higher degree of precision. We conceptualize this important dimension of violence by manually coding statements about killed civilians and non-combatants such as children. As indiscriminate violence violates international humanitarian law, we further searched for references to the United Nations and human rights in the responses.

Civilian casualties are mentioned more than twice as often in the Arabic responses as in the Hebrew responses. Killed children are mentioned six times more often and female victims three times more often in the Arabic version. GPT-3.5's responses in Arabic are also more likely to emphasize that these airstrikes violated international law. The United Nations is mentioned 13 times in the Arabic responses, while it is never mentioned in the Hebrew output. Moreover, the propensity that it was highlighted that these airstrikes were condemned by the international community differs by a factor of 11. In contrast, the term 'terrorist' is mentioned more than six times more frequently in the Hebrew responses compared to the Arabic responses.

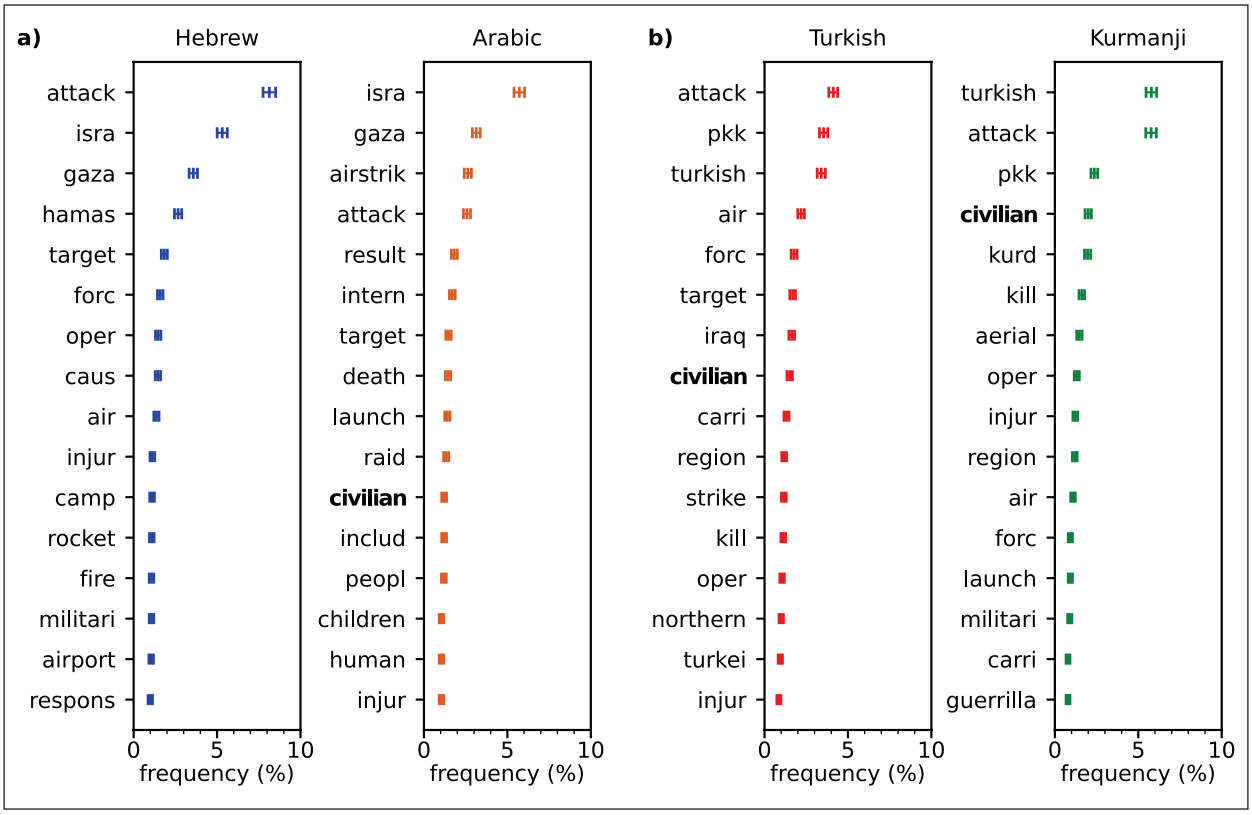

Figure 4. Word frequency analysis.
Top 15-word stems in each language after removal of stopwords, with their respective word frequencies. The uncertainties are based on the shot noise limit. (a) Hebrew/Arabic dyad. (b) Turkish/Kurdish dyad.

With regard to the Turkish–Kurdish conflict, we find that civilian casualties are mentioned about 50% more often in Kurdish than in Turkish. Killed children appear ten times more often in the Kurdish responses compared to the Turkish ones. Notably, the term 'innocent' appears only in the GPT-3.5 output in Kurdish. Furthermore, the term 'human rights' is mentioned 33% more often in Kurdish than in Turkish. Overall, terms related to indiscriminate victimization or international condemnation appear more frequently in GPT-3.5 responses in Kurdish. In contrast, the term 'terrorist' is mentioned 8% more often in the Turkish text compared to the Kurdish output. See Figure A.2 in the Online appendix for a visual summary of these findings.

Tracing the source of the bias

While we have shown that user language shapes the information about conflict-related violence provided by GPT-3.5, the evidence presented so far remains silent on the source of the bias. The question arises whether GPT-3.5 simply reproduces a media bias in the training data or whether there is an additional bias mechanism specific to LLMs.

To investigate this question, we conducted a systematic media content analysis of the airstrikes under scrutiny. We use the online platform LexisNexis, which provides a large corpus of media sources in different languages and has been frequently used by conflict researchers to systematically track media coverage of conflict-related phenomena (Baum and Zhukov, 2015; Carey et al., 2022). Ideally, we could conduct the media analysis in the four languages under investigation. However, LexisNexis only covers two (Arabic and Turkish) of these four languages.[19] While this precludes the possibility of comparing the media bias within conflict dyads, the media analysis allows us to gain a better understanding of the relationship between the information in the news reports that feeds into the training data and the output provided by GPT-3.5.

We employ the following procedure in LexisNexis to systematically track media reports in Arabic (and Turkish) on the 50 airstrikes under investigation. We use the Arabic (respectively Turkish) term for 'airstrike' and the location of the airstrike (such as al-Shati refugee camp or Beit Lahiya) as search operators and retrieve all news reports that contain these terms on the day of a given attack and the following three days.[20] We translate these news sources

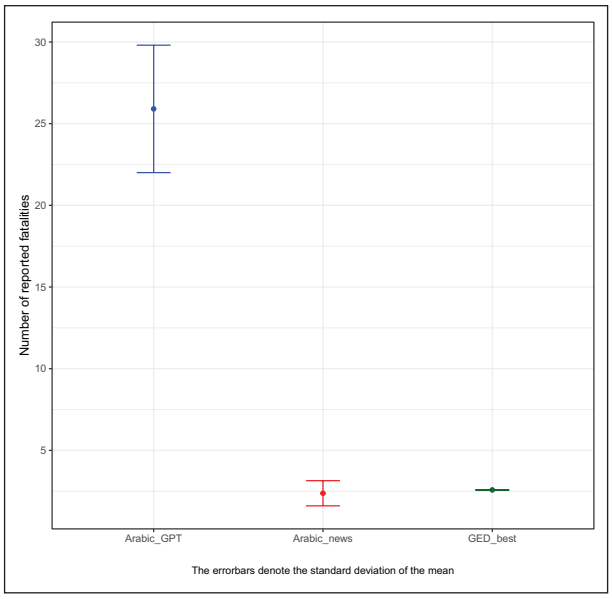

Figure 5. Media analysis.
This figure compares the fatality estimates derived from the media analysis with the death counts provided by GPT-3.5 and by the GED dataset. The label *Arabic news* refers to the average number of fatalities in the Arabic media analysis. *Arabic GPT* provides the average number of fatalities in the Arabic GPT-3.5 output for the subset of airstrikes that are covered by the media analysis. *GED best* captures the average number of fatalities for these airstrikes as reported by the GED dataset based on international news reports.

into English using the built-in Google translator and manually filter the subset of news sources that relate to airstrikes occurring in the respective conflicts. Subsequently, we manually study these news reports and retrieve the reported death counts of the airstrikes under investigation. To ensure that the fatality numbers correspond to these specific attacks, we use additional contextual information about these airstrikes such as the identity of the victims or the time of the day, which we match with the event descriptions provided by the GED. We manually exclude all cumulative death counts, which aggregate fatality numbers from multiple events.[21]

Figure 5 displays the results of the media analysis of the Arabic news sources. The evidence shows that the fatality numbers provided by GPT-3.5 (in the Arabic version) are significantly higher than the fatality numbers reported in the Arabic media. This suggests that GPT-3.5 is unable to link the correct fatality numbers to the corresponding events. Since GPT-3.5 lacks the contextual granularity to identify these specific attacks, it misreports the fatality numbers and may provide death tolls from other high-profile events or cumulative death counts that are prevalent in training data. Hence, the misreported fatality numbers are likely linked to the co-occurrence bias of LLMs.

Another notable pattern in Figure 5 is that the death counts provided in the Arabic news sources match with the fatality numbers in the GED, which are largely based on international media sources (Högbladh, 2023; Sundberg and Melander, 2013). If the media bias were the main driver of our results, we would expect a significant discrepancy between the death tolls reported in the Arabic media compared to the international media reports captured in the GED. However, the media analysis shows that this is not the case. Hence, the evidence suggests that the bias results from the inability of GPT-3.5 to link the queried airstrikes to the corresponding fatality numbers, rather than from misreporting in Arabic news sources.

Subsequently, we run the media analysis on Turkish news sources, using the same procedure in LexisNexis. However, we found only for one among the 50 Turkish airstrikes a Turkish news source that provided any fatality numbers.[22] This suggests that the Turkish media tends to refrain from providing information about casualties caused by Turkish airstrikes against Kurdish targets. This media bias may explain the tendency of GPT-3.5 to provide evasive answers in the language of the state that conducts the airstrikes. To the extent that no or only few fitting sources to the prompts are available in the training data, GPT-3.5 may describe different events or deny the attacks due to the long-tail knowledge bias.

To analyze whether GPT-3.5 makes the wording more one-sided compared to the news articles, we conducted sentiment analyses. For five randomly selected airstrikes in the Israeli–Palestinian conflict, we ran sentiment analyses on the full text of all news articles identified on LexisNexis that report on a given airstrike and on the ChatGPT output related to that airstrike (both in the Arabic version). The results of the sentiment analyses are shown in Figure 6. The sentiment scores are consistently more negative in the Arabic ChatGPT output than in the Arabic news sources. This may reflect the fact that the corpus of the LLM includes not only official news reports but also other, potentially more one-sided sources such as social media or blog posts.

Overall, these additional analyses suggest that two distinct sources of bias drive our results. First, there is a specific AI bias caused by GPT-3.5's inability to match specific airstrikes to corresponding fatality numbers and by its tendency to describe these attacks in more negative terms compared to official media reports. Second, there is a media bias with regard to the reduced propensity to report on airstrike fatalities in the language of the attacker, resulting in more evasive answers by GPT-3.5.

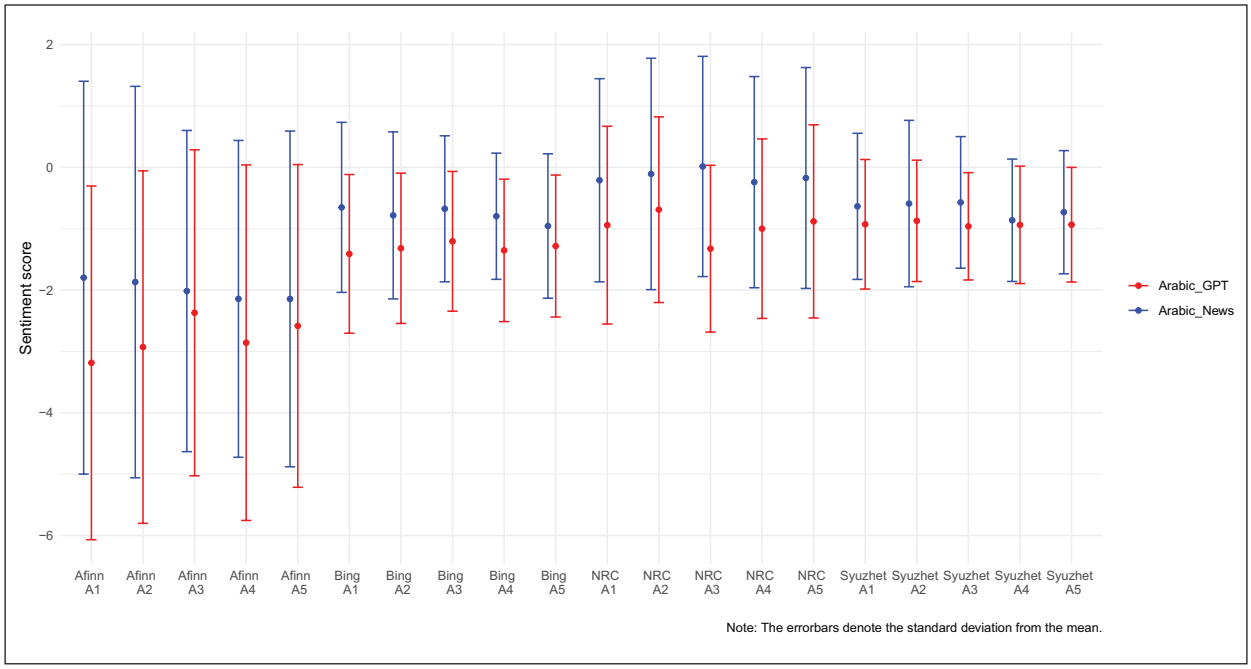

Figure 6. Sentiment analyses.
The figure compares the sentiment scores of the Arabic GPT-3.5 outputs with the sentiment scores of the Arabic news reports for five randomly selected airstrikes (both translated into English). *Arabic GPT* is based on the full text of the GPT-3.5 output to the question *'What happened in the Israeli airstrike on date X in location Y?'. Arabic news* includes the full text of all news articles traced on LexisNexis that report on a given airstrike on the same or the following day. The different sentiment scores are calculated with the Syuzhet-package in R. The labels refer to the following airstrikes: A1: 21 August 2014; A2: 4 August 2014; A3: 12 November 2019; A4: 27 June 2014; A5: 18 August 2011.

Discussion and conclusions

This study demonstrates that information on conflict-related violence in the Israeli–Palestinian and the Turkish–Kurdish conflicts generated by ChatGPT is affected by a substantial language bias. We show that the language model tends to produce higher airstrike fatality estimates when queried in the language of the targeted group compared to the language of the attacker. In the context of Turkish airstrikes against Kurdish targets, we find that GPT-3.5 produces higher fatality estimates if it is enquired in Kurdish compared to Turkish. In the same vein, GPT-3.5 reports higher fatality estimates in Arabic compared to Hebrew when asked about Israeli airstrikes in the Gaza Strip. Moreover, we show that the attacks are described as more indiscriminate in the language of the targeted group, which is reflected in discrepancies in information about civilian casualties, killed children, and female victims. While it is well established that LLMs can generate misinformation (Buchanan et al., 2021; Solaiman et al., 2019) and that they are linked to ethical and social risks (Bahrini et al., 2023; Weidinger et al., 2021), our study provides the first evidence of language bias in the context of conflict-related violence.

It is important to consider these findings in light of the underlying data generation process. By generating responses from a multilingual corpus of online sources, queries to ChatGPT approximate a large-scale language-specific content analysis of online information. Hence, although we specifically identify biases in the responses generated by the LLM, they partially reproduce broader biases that are present in the training data. The evidence shows that airstrikes are described in a different tone depending on the origin of the online sources, and they might not be described at all in the media sources of the state that conducts these attacks. While this media bias is problematic in itself, ChatGPT makes it especially difficult for citizens to identify this bias. Critical consumers may be able to distinguish between high-quality and low-quality news sources, but they are less likely to understand the origins of the biases produced by the LLM.

On top of this media bias, we show a novel source of bias that has no equivalent in classical search engines. GPT-3.5 retrieves casualty numbers from the language-specific training data but fails to correctly match them to the corresponding events. Due to its reliance on co-occurring words, GPT-3.5 may report casualty counts from different attacks with greater news impact, or provide cumulative casualty counts that are prevalent in the news media. Because such high-profile events are systematically different from low-profile events – for example, involving more deaths and more civilian casualties – the

reported death counts tend to be inflated. However, what constitutes a high-profile event is likely to be language-specific, leading to systematic differences across languages. If there are no high-profile events to link to in a particular user language, GPT-3.5 may provide evasive answers due to its weaker performance with long-tail knowledge (Kandpal et al., 2023).

Overall, this language bias could have important implications for multilingual armed conflicts. Public opinion plays a crucial role in armed conflicts as governments tend to rely on loyal troops and public support to wage war (Feinstein, 2022; Tomz and Weeks, 2013; Voeten and Brewer, 2006). Our findings suggest that citizens of states that have conducted airstrikes may underestimate the human toll of these attacks based on the information obtained through LLMs. In contrast, citizens of attacked groups may perceive these airstrikes as especially brutal and indiscriminate based on the information available in their language. These antithetical perceptions may contribute to radicalized identities and intensify dynamics of mutual blaming (Hameleers and Brosius, 2022). In so doing, information discrepancies may nurture grievances and ultimately reinforce conflicts within linguistic dyads.

Our findings have implications beyond the specific context of armed conflict. It is possible that similar language biases affect information generated by LLMs in other topic areas, especially where the training data is likewise heterogeneous and differs across languages. This is likely to be the case for other areas of contested information such as sensitive political issues, religious beliefs, or cultural identities. Future research could explore to what extent language biases in LLMs are present in other topic areas and which languages are particularly susceptible to these biases.

To conclude, our study shows that user language systematically shapes conflict fatality estimates produced by ChatGPT. We further provide a novel method for quantitatively analyzing bias in LLMs, offering a more robust quantitative alternative to classical sentiment analysis approaches. Using this new method, we show significant discrepancies in information on conflict-related violence between different user languages in the Israeli–Palestinian and the Turkish–Kurdish conflict. To the extent that LLMs are increasingly used for information purposes, potentially through their integration in search engines such as Microsoft Bing or Google Gemini, this bias could promote information bubbles in the future.

Authors' note

Both authors contributed equally.

Replication data

The dataset and scripts for the empirical analysis in this article, along with the Online appendix, are available at https://www.prio.org/jpr/datasets/ and https://doi.org/10.5281/zenodo.8181226. The analyses were conducted using Python and R.

Acknowledgements

The authors would like to thank J Holder and J Weisser for many helpful comments and discussions, our native speakers K Mahtouch, R Rauchwerger and R Tadik for the translation checks, and H Zentner for providing their contacts. We thank Prof. Tina Freyburg for her great support and Lionel Perruchoud for his excellent research assistance. We thank the 'Konstanzer WG', especially P Gebauer, S Fonseca, L Heyden, F Boehringer, and H Zentner for providing a highly conducive and stimulating environment for this research project.

Funding

We acknowledge funding from Evangelisches Studienwerk e.V. (Daniel Kazenwadel) and from the International Postdoctoral Fellowship of the University of St Gallen (Christoph Steinert).

ORCID iDs

Christoph Valentin Steinert https://orcid.org/0000-0003-1760-3214
Daniel Kazenwadel https://orcid.org/0000-0003-0225-5317

Notes

1. This may not apply to citizens of states where the internet is completely censored. However, there are only a small number of states where internet censorship is comprehensive, and individuals tend to be savvy in circumventing these restrictions (e.g. King et al., 2013: 328).
2. As a robustness test, we also cover GPT-4.0. The results align for both versions of OpenAI's LLM.
3. We use Northern Kurdish, also known as Kurmanji, which is a widely spoken form of Kurdish. Kurmanji is predominantly used in south-eastern Turkey and the Kurdistan region of Iraq, which are the regions covered in our article.
4. We do not take a political stance on the question of who is the aggressor in the overall conflicts but use the terms

'perpetrator' and 'attacker' only in relation to specific airstrikes carried out by one side.

5. Kurdish and Hebrew are not available on platforms such as LexisNexis and Factiva, which precludes a systematic media analysis in these languages. While Turkish is available, we found hardly any Turkish news reports that provide information on casualties caused by Turkish airstrikes against Kurdish targets. This lack of coverage may explain the tendency of GPT-3.5 to provide evasive answers in the language of the state that conducts these attacks.
6. We have explored other types of conflict-related violence such as arrests during these conflicts, but the numerical information provided by ChatGPT was too sparse to rerun the same analysis.
7. In some contexts, perpetrators of violence may wish to exaggerate the scale of their attacks in order to spread fear and signal their resolve. Such exaggeration of violence is particularly common in the case of terrorist groups that seek to maximize fear and attention (Blankenship, 2018; Braithwaite, 2013).
8. Recent evidence from web tracking and survey data suggests that higher levels of education and younger age are important predictors of ChatGPT usage (Kacperski et al., 2023). Notably, political knowledge is positively associated with ChatGPT usage even when controlling for socio-demographic factors. This suggests that LLMs are currently disproportionately used by individuals who may be more attentive to political bias. To the extent that LLMs are increasingly implemented in standard search engines, this may no longer be the case.
9. Moreover, there might be discrepancies across languages due to different volumes of training data. The training data – a mixture of different sources, including a copy of open access internet data (Common Crawl), an overview of open source books, and Wikipedia – is heavily biased towards the English language (over 50%) (Artetxe et al., 2022; Crawl Archives, 2023; Dodge et al., 2021). Since the performance of language models depends on the amount of training data, this results in a significantly worse performance for languages with less training data (Fang et al., 2023; Lai et al., 2023).
10. Kandpal et al. (2023) show that retrieval-augmented generation is a promising way to improve the performance of LLMs on questions where few relevant documents are available in the pre-training dataset.
11. The GED relies on international media reporting, which implies that only airstrikes that have been covered by the international media make it into our sample. Given that this may imply that these are the most likely cases of events that ChatGPT would 'know' about, it is interesting that we observe substantial numbers of 'evasive answers' in the language of the attacker. This suggests that despite international media coverage, there is no or only very limited reporting on these events in the language of the attacker.
12. We had tried different question wordings and chose this question based on the highest propensity to provide numerical estimates of fatalities in GPT-3.5.
13. See: https://doi.org/10.5281/zenodo.8181226 (Kazenwadel, 2024).
14. We highlight these ten airstrikes as these cases are also used in the analysis with GPT-4, which is discussed below.
15. Moreover, ChatGPT is more likely to remain vague or not to provide any answer in the language of the attacker. We consider evasive answers as a special case of the instances where ChatGPT does not provide any fatality numbers.
16. The exact formula used to calculate this difference is

$$\delta = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{x_{\mathrm{i}},\mathbf{arab} - x_{\mathrm{i}},\mathbf{heb}}{0.5(x_{\mathrm{i}},\mathbf{arab} - x_{\mathrm{i}},\mathbf{heb})}$$

where ($X_p$heb / $X_p$arab) denotes the average fatalities in a single case reported in the respective language and N is the number of cases where fatalities were reported in both languages. The normalization over the mean of both cases gives better numerical stability to outliers. The uncertainty was calculated via the standard deviation of the mean. The formula used on the Kurdish/Turkish dyad is equivalent.

17. Evasive answers are less common with GPT-4 compared to GPT-3.5 and, overall, there are fewer responses with precise fatality numbers. This may reflect that the model is fine-tuned to avoid hallucinations and to refuse to provide facts when the training data is sparse (Open AI, 2024). Hence, there is a trade-off between the risk of hallucinations and the 'laziness' of the model.
18. We applied stemming (with the tidytext-package in *R*) and removed stopwords.
19. We had also tried different platforms that have been used for systematic media analyses such as Factiva but could not find any that covers Kurdish or Hebrew.
20. See Online appendix A.3 for an overview of the Arabic news sources from which we derived the information.
21. We provide the sources for all coding decisions in the Online supplementary material.
22. The news source reported only one killing for the airstrike in Sinjar on 15 August 2018, whereas the GED reported five killings.

References

Artetxe M, Aldabe I, Agerri R, et al. (2022) Does corpus quality really matter for low-resource languages? In: *Proceedings of the 2022 conference on empirical methods in natural language processing* (eds Goldberg Y, Kozareva Z and Zhang Y), Abu Dhabi, UAE, 7–11 December 2022, pp. 7383–7390. Abu Dhabi, UAE: Association for Computational Linguistics.

Bahrini A, Mohammadsadra K, Hossein A, et al. (2023) ChatGPT: Applications, opportunities, and threats. In:

2023 systems and information engineering design symposium, Charlottesville, pp. 274–279. New York, NY: IEEE.

Baum MA and Zhukov YM (2015) Filtering revolution: Reporting bias in international newspaper coverage of the Libyan civil war. *Journal of Peace Research* 52(3): 384–400.

Blankenship B (2018) When do states take the bait? State capacity and the provocation logic of terrorism. *Journal of Conflict Resolution* 62(2): 381–409.

Bormann NC, Cederman LE and Vogt M (2017) Language, religion, and ethnic civil war. *Journal of Conflict Resolution* 61(4): 744–771.

Braithwaite A (2013) The logic of public fear in terrorism and counter-terrorism. *Journal of Police and Criminal Psychology* 28(2): 95–101.

Buchanan B, Lohn A, Musser M, et al. (2021) Truth, lies, and automation. How language models could change disinformation. Report, Center for Security and Emerging Technology, University of Georgetown, May. Available at: https://cset.georgetown.edu/publication/truth-lies-and-automation/ (accessed 31 July 2024).

Carey SC (2006) The dynamic relationship between protest and repression. *Political Research Quarterly* 59(1): 1–11.

Carey SC, Mitchell NJ and Paula K (2022) The life, death and diversity of pro-government militias: The fully revised pro-government militias database version 2.0. *Research & Politics* 9(1): DOI: 10.1177/ 20531680211062772.

Cheng M, Durmus E and Jurafsky D (2023) Marked Personas: Using natural language prompts to measure stereotypes in language models. arXiv preprint, 29 May. DOI: 10.48550/arXiv.2305.18189.

Crawl Archives (2023) Statistics of common crawl monthly archives. Available at: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages (accessed 31 July 2024).

Curtice TB and Behlendorf B (2021) Street-level repression: Protest, policing, and dissent in Uganda. *Journal of Conflict Resolution* 65(1): 166–194.

Davies S, Pettersson T and Oberg M (2022) Organized violence 1989–2021 and drone warfare. *Journal of Peace Research* 59(4): 593–610.

Dobransky K and Hargittai E (2021) Piercing the pandemic social bubble: Disability and social media use about COVID-19. *American Behavioral Scientist* 65(12): 1698–1720.

Dodge J, Sap M, Marasović A, et al. (2021) Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint, 30 September. DOI: 10.48550/arXiv.2104.08758.

Eady G, Nagler J, Guess A, et al. (2019) How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *SAGE Open* 9(1): DOI: 10.1177/ 215824401983270.

Fang C, Ling J, Zhou J, et al. (2023) How does ChatGPT-4 perform on non-English national medical licensing examination? An evaluation in Chinese language. *PLOS Digital Health* 2(12): e0000397.

Fearon JD (1995) Rationalist explanations for war. *International Organization* 49(3): 379–414.

Feinstein Y (2022) *Rally 'Round the Flag: The Search for National Honor and Respect in Times of Crisis*. Oxford: Oxford University Press.

Fennell J (2014) Morale and combat performance: An introduction. *Journal of Strategic Studies* 37(7): 796–798.

Gerring J and Cojocaru L (2016) Selecting cases for intensive analysis: A diversity of goals and methods. *Sociological Methods & Research* 45(3): 392–423.

Gleditsch NP, Wallensteen P, Eriksson M, et al. (2002) Armed conflict 1946–2001: A new dataset. *Journal of Peace Research* 39(5): 615–637.

Gohdes A (2020) Repression technology: Internet accessibility and state violence. *American Journal of Political Science* 64(3): 488–503.

Gohdes A and Carey S (2017) Canaries in a coal-mine? What the killings of journalists tell us about future repression. *Journal of Peace Research* 54(2): 157–174.

Greenhill KM and Oppenheim B (2017) Rumor has it: The adoption of unverified information in conflict zones. *International Studies Quarterly* 61(3): 660–676.

Greitens SC (2016) *Dictators and Their Secret Police: Coercive Institutions and State Violence*. Cambridge: Cambridge University Press.

Guriev S and Treisman D (2019) Informational autocrats. *Journal of Economic Perspectives* 33(4): 100–127.

Hameleers M and Brosius A (2022) You are wrong because I am right! The perceived causes and ideological biases of misinformation beliefs. *International Journal of Public Opinion Research* 34(1): edab028.

Hartmann J, Schwenzow J and Witte M (2023) The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint*, 5 January. DOI: 10.48550/arXiv.2301.01768.

Högbladh S (2023) UCDP GED Codebook version 23.1. Available at: https://ucdp.uu.se/downloads/ged/ged231.pdf (accessed 31 July 2024).

Honig O and Reichard A (2018) Evidence-fabricating in asymmetric conflicts: How weak actors prove false propaganda narratives. *Studies in Conflict & Terrorism* 41(4): 297–318.

Huang R (2016) Rebel diplomacy in civil war. *International Security* 40(4): 89–126.

Intahchomphoo C and Gundersen OE (2020) Artificial intelligence and race: A systematic review. *Legal Information Management* 20(2): 74–84.

Kaakinen M, Sirola A, Savolainen I, et al. (2020) Shared identity and shared information in social media: Development and validation of the identity bubble reinforcement scale. *Media Psychology* 23(1): 25–51.

Kacperski C, Roberto U, Bonnay C, et al. (2023) Who are the users of ChatGPT? Implications for the digital divide from web tracking data. *arXiv preprint*, 5 September. DOI: 10.48550/arXiv.2309.02142.

Kalyvas S (2004) The urban bias in research on civil wars. *Security Studies* 13(3): 160–190.

Kalyvas S and Balcells L (2010) International system and technologies of rebellion: How the end of the Cold War shaped internal conflict. *American Political Science Review* 104(3): 415–429.

Kandpal N, Deng H, Roberts A, et al. (2023) Large language models struggle to learn long-tail knowledge. In: *Proceedings of the 40th international conference on machine learning*, vol. 202 (eds Krause A, Brunskill E, Kyunghyun C, et al.), Honolulu, HI, 23–29 July, 15696–15707. Honolulu, HI: PMLR.

Kang C and Choi J (2023) Impact of co-occurrence on factual knowledge of large language models. *arXiv preprint*, 12 October. DOI: 10.48550/arXiv.2310.08256.

Kazenwadel D (2024) Replication data: How user language affects conflict fatality estimates in ChatGPT. *Zenodo*. DOI: 10.5281/zenodo.8181226.

King G, Pan J and Roberts ME (2013) How censorship in China allows government criticism but silences collective expression. *American Political Science Review* 107(2): 326–343.

Koehler D (2014) The radical online: Individual radicalization processes and the role of the Internet. *Journal for Deradicalization* 1(Winter): 116–134.

Kurizaki S and Whang T (2015) Detecting audience costs in international disputes. *International Organization* 69(4): 949–980.

Lai VD, Ngo NT, Veyseh AP, et al. (2023) ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint*, 12 April. DOI: 10.48550/arXiv.2304.05613.

Leavy S (2018) Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In: *Proceedings of the 1st international workshop on gender equality in software engineering,* Gothenburg, 27 May–3 June, 14–16. New York, NY: Association for Computing Machinery.

Lee NT (2018) Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society* 16(3): 252–260.

Lewandowsky S, Stritzke W, Freund A, et al. (2013) Misinformation, disinformation, and violent conflict: From Iraq and the war on terror to future threats to peace. *American Psychologist* 68(7): 487–501.

Marinucci L, Mazzuca C and Gangemi A (2022) Exposing implicit biases and stereotypes in human and artificial intelligence: State of the art and challenges with a focus on gender. *AI & Society* 38(2): 747–761.

McGee RW (2023) Is ChatGPT biased against conservatives? An empirical study. *SSRN*, 15 February. DOI: 10.2139/ssrn.4359405.

Nadeem A, Babak A and Marjanovic O (2020) Gender bias in AI: A review of contributing factors and mitigating strategies. In: *27th ACIS 2020 proceedings, Wellington, New Zealand*, 1–4 December, 1–12. Wellington, New Zealand: AIS Electronic Library.

Nilsson M (2018) Primary unit cohesion among the Peshmerga and Hezbollah. *Armed Forces & Society* 44(4): 647–665.

Noor M, Shnabel N, Halabi S, et al. (2012) When suffering begets suffering: The psychology of competitive victimhood between adversarial groups in violent conflicts. *Personality and Social Psychology Review* 16(4): 351–374.

Obermeyer Z and Topol E (2021) Artificial intelligence, bias, and patients' perspectives. *The Lancet* 397(10289): 2038.

Open AI (2023a) Chat completions. Available at: https://platform.openai.com/docs/guides/chat (accessed 31 July 2024).

Open AI (2023b) What are tokens and how to count them? Available at: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (accessed 31 July 2024).

Open AI (2024) Models. Available at: https://platform.openai.com/docs/models/overview (accessed 31 July 2024).

Pechenkina AO, Bausch AW and Skinner KK (2019) How do civilians attribute blame for state indiscriminate violence? *Journal of Peace Research* 56(4): 545–558.

Pfeifle B (2022) Detecting armed conflict damages in satellite imagery using deep learning. PhD thesis, Universität Konstanz.

Podder S (2017) Understanding the legitimacy of armed groups: A relational perspective. *Small Wars & Insurgencies* 28(4): 686–708.

Price M and Ball P (2015) Selection bias and the statistical patterns of mortality in conflict. *Statistical Journal of the IAOS* 31(2): 263–272.

Price M, Gohdes A and Ball P (2014) Updated statistical analysis of documentation of killings in the Syrian Arab Republic. Report, Human Rights Data Analysis Group. Available at: https://www.ohchr.org/sites/default/files/Documents/Countries/SY/HRDAGUpdatedReportAug2014.pdf (accessed 31 July 2024).

Reuters (2021) Turkey says militants executed 13, including soldiers, police, in Iraq. *Reuters*, 14 February. Available at: https://www.reuters.com/article/us-turkey-iraq-security-idUSKBN2AE050 (accessed 31 July 2024).

Rhodes SC (2022) Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Political Communication* 39(1): 1–22.

Rozenas A and Zhukov YM (2019) Mass repression and political loyalty: Evidence from Stalin's 'terror by hunger'. *American Political Science Review* 113(2): 569–583.

Salehyan I, Gleditsch KS and Cunningham DE (2011) Explaining external support for insurgent groups. *International Organization* 65(4): 709–744.

Saul B (2008) The international protection of journalists in armed conflict and other violent situations. *Australian Journal of Human Rights* 14(1): 99–140.

Schlichte K and Schneckener U (2015) Armed groups and the politics of legitimacy. *Civil Wars* 17(4): 409–424.

Schon J (2021) How narratives and evidence influence rumor belief in conflict zones: Evidence from Syria. *Perspectives on Politics* 19(2): 539–552.

Schutte S, Ruhe C and Linke AM (2022) How indiscriminate violence fuels conflicts between groups: Evidence from Kenya. *Social Science Research* 103: 102653.

Silverman D, Kaltenthaler K and Dagher M (2021) Seeing is disbelieving: The depths and limits of factual misinformation in war. *International Studies Quarterly* 65(3): 798–810.

Slantchev BL (2006) Politicians, the media, and domestic audience costs. *International Studies Quarterly* 50(2): 445–477.

Slantchev BL and Ahmer Tarar (2011) Mutual optimism as a rationalist explanation of war. *American Journal of Political Science* 55(1): 135–148.

Solaiman I, Brundage M, Clark J, et al. (2019) Release strategies and the social impacts of language models. *arXiv preprint*, 24 August. DOI: 10.48550/arXiv.1908.09203.

Spohr D (2017) Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34(3): 150–160.

Stanton JA (2020) Rebel groups, international humanitarian law, and civil war outcomes in the post-Cold War era. *International Organization* 74(3): 523–559.

Steinert CV and Dworschak C (2022) Political imprisonment and protest mobilization: Evidence from the GDR. *Journal of Conflict Resolution* 67(7–8): 1564–1591.

Sundberg R and Melander E (2013) Introducing the UCDP georeferenced event dataset. *Journal of Peace Research* 50(4): 523–532.

Thomson H (2017) Grievance attribution, mobilization and mass opposition to authoritarian regimes: Evidence from June 1953 in the GDR. *Comparative Political Studies* 51(12): 1594–1627.

Tomz MR and Weeks JLP (2013) Public opinion and the democratic peace. *American Political Science Review* 107(4): 849–865.

Voeten E and Brewer PR (2006) Public opinion, the war in Iraq, and presidential accountability. *Journal of Conflict Resolution* 50(6): 809–830.

Weidinger L, Mellor J, Rauh M, et al. (2021) Ethical and social risks of harm from language models. *arXiv preprint* 8 December. DOI: 10.48550/arXiv.2112.04359.

Zhuravskaya E, Petrova M and Enikolopov R (2020) Political effects of the internet and social media. *Annual Review of Economics* 12: 415–438.

CHRISTOPH VALENTIN STEINERT, b. 1991, PhD in Political Science (University of Mannheim, 2022); Postdoctoral Research Fellow, University of Zurich (2024–present); current main interests: human rights, artificial intelligence, and international organizations.

DANIEL KAZENWADEL, b. 1995, MSc in Physics (University of Konstanz, 2020); PhD student (University of Konstanz, 2020–present); current main interest: ultrafast transmission electron microscopy.