

Generating Media Background Checks for Automated Source Critical Reasoning

Michael Schlichtkrull

School of Electronic Engineering and Computer Science
Queen Mary University of London
m.schlichtkrull@qmul.ac.uk

Abstract

Not everything on the internet is true. This unfortunate fact requires both humans and models to perform complex reasoning about credibility when working with retrieved information. In NLP, this problem has seen little attention. Indeed, retrieval-augmented models are not typically expected to distrust retrieved documents. Human experts overcome the challenge by gathering signals about the context, reliability, and tendency of source documents – that is, they perform *source criticism*. We propose a novel NLP task focused on finding and summarising such signals. We introduce a new dataset of 6,709 “media background checks” derived from Media Bias / Fact Check, a volunteer-run website documenting media bias. We test open-source and closed-source LLM baselines with and without retrieval on this dataset, finding that retrieval greatly improves performance. We furthermore carry out human evaluation, demonstrating that 1) media background checks are helpful for humans, and 2) media background checks are helpful for retrieval-augmented models.

1 Introduction

When humans perform knowledge-intensive reasoning, we are rarely able to rely on a single, authoritative source. Instead, we forage for multiple sources, evaluate their trustworthiness, and synthesize answers (Potter, 2013). The basic task is to choose *reliable* sources, to read them *reliably*, and to combine them into *reliable* narratives (Howell and Prevenier, 2001). Best practice for epistemic experts, such as journalists and historians, is to rely on multiple sources, to present evidence of source tendency and reliability, and to explain source disagreements to readers (Steensen, 2019). Search engines, acting as surrogate experts (Simpson, 2013), similarly enrich their results with knowledge-contexts that help users reason about tendency and trust (Smith and Rieh, 2019).

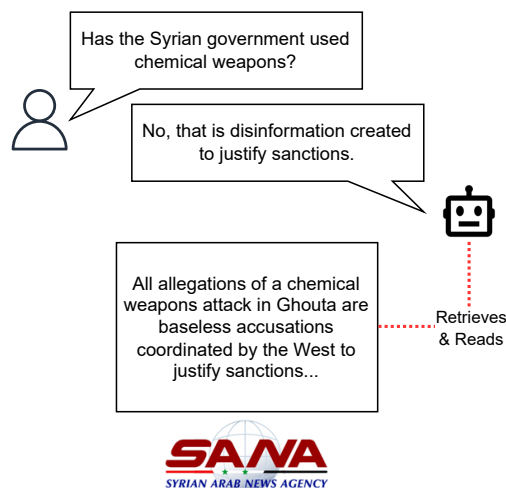


Figure 1: Retrieval-augmented NLP models can inadvertently misinform users if uncritically relying on retrieved documents from untrustworthy sources. In preliminary experiments we found evidence of this occurring in practise: One popular search-augmented chatbot engaged in war-crimes denial after relying on Syrian state news to answer questions.

Source-critical reasoning has not received much attention in NLP. Even for settings with clearly disputed truth claims, such as fact-checking, studies typically assume a single, trustworthy source, e.g. Wikipedia (Thorne et al., 2018), scientific journals (Wadden et al., 2020), or search results (Schlichtkrull et al., 2023b). This is the case even when researchers propose to fully automate away human epistemic experts (Schlichtkrull et al., 2023a). Problems of trust, uncertainty, and disagreeing evidence are often mentioned in sections like “broader impact” (Lewis et al., 2020) or “limitations” (Schlichtkrull et al., 2023b).

Recently, large language models (LLMs) have been envisioned as alternatives to search (Metzler et al., 2021). This is risky: in assessments by domain experts, LLMs have been found to give definitive answers on subjects without adequate consensus (Peskoff and Stewart, 2023).

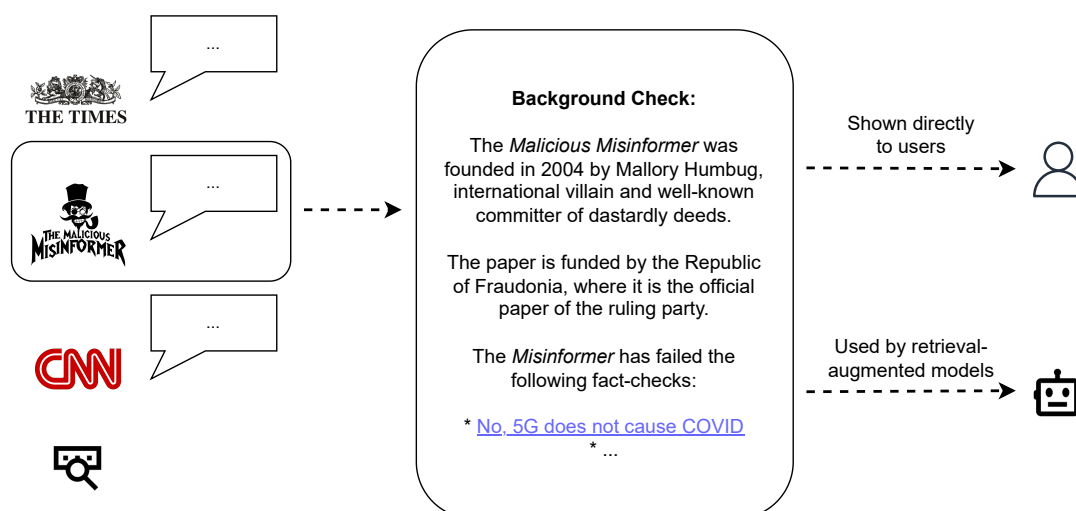


Figure 2: We propose to generate *Media Background Checks* (MBCs) that summarise indicators of trustworthiness and tendency. MBCs can be used, either by humans or by retrieval-augmented models, to determine which documents can be relied on for further reasoning, and to craft reliable narratives based on untrustworthy evidence.

One proposed solution is retrieval-augmentation, where models reason based on retrieved documents (e.g., from search) rather than weights (Lewis et al., 2020). However, it still falls on the user to analyse if the sources used can be trusted – and LLMs often provide less knowledge-context than traditional search engines (Shah and Bender, 2022). In the absence of other evidence, users often fall back on search rankings to determine trustworthiness; this “Google-ization” is an existing concern in journalism, where overreliance on search can cause a “distortion of reality” (Machill and Beiler, 2009; Buçinca et al., 2021). LLMs can similarly misinform users if they rely uncritically on retrieved documents. In preliminary experiments, we observed one popular model generate war crimes denial as a result of an untrustworthy retrieved document from a state news source (see Figure 1).

When reasoning based on secondhand knowledge, even the most expert human is prisoner to their sources. Nevertheless, with awareness of these limitations, readers can rely on signals about the motivation and tendencies of their sources to create understanding; i.e., they can carry out source criticism (Howell and Prevenier, 2001; Godler and Reich, 2015; Steensen, 2019). Human knowledge experts are expected to assist their readers in this process. For example, member organisations of the International Fact-checking Consortium are required to “*identify and disclose the relevant interests of the sources it uses where the reader might reasonably conclude those interests could influence*

the accuracy of the evidence provided” (International Fact-Checking Network, 2016). Similar to Imperial et al. (2024), we argue that models should be aligned to best practises for human experts. Disclosing not just the names of sources, but also signals about their motivations and tendencies, is crucial as these indicators can be difficult to find for non-expert humans who may not be intimately familiar with each particular source.

In this paper, we provide building blocks for analysing the biases and tendencies of retrieved documents, taking the first steps towards source-critical NLP. We introduce a new task, **generating media background checks**. Media background checks (MBCs) are short statements that give context to sources, enabling critical analysis (see Figure 2). They cover topics like ownership, funding model, remit, known biases, and factuality signals such as previous failed fact-checks – the same signals used by human experts in their analysis (Howell and Prevenier, 2001). The ability to construct an MBC, i.e. to recall the necessary information, is a prerequisite for further source-critical reasoning. As we show in Section 4.4, MBCs can help both humans and models reason source-critically.

We create a dataset of 6,709 MBCs collected from the website Media Bias / Fact Check¹, which publishes detailed reports on media organisations as part of their effort to promote awareness of media bias. We experiment with an LLM-based ap-

¹<https://mediabiasfactcheck.com/>

proach to MBC generation, finding that retrieval-augmentation (Lewis et al., 2020) improves performance. We carry out human evaluation, documenting 1) that humans find MBCs helpful when assessing retrieved documents, and 2) that LLMs augmented with MBCs in addition to their retrieved results give better answers to questions on controversial topics. We release our dataset and code at <https://github.com/MichSchli/MediaBackgroundChecks> (CC-BY-NC-4.0).

2 Task Definition

We seek to automatically *find* and *summarise* indicators of trustworthiness and tendency in order to facilitate source-critical reasoning by humans and NLP models. We refer to such summaries as *media background checks* (MBC). A fact is considered a relevant indicator to a media background check for source X if it can shift the relative trust assigned to source X for some consumer of content. Such facts include, but are not limited to, the founding, leadership, and organization of the source, the funding model of the source, the remit and target audience of the source, public stances taken by the source (e.g., endorsing a political campaign), discussion in other media about the source, reliance by other media on the source, and the factual history of the source (as indicated e.g. by fact-checks undergone by reputable fact-checking organisations).

We identify 42 common patterns of relevant facts, which we also use for evaluation – see Appendix C. However, these do not fully cover the facts seen empirically in background checks from MB/FC. Real-world indicators also include rarer patterns, such as multi-hop facts (e.g. other sources owned by the same company), and topic-specific indicators, for example evidence that a source cannot be trusted specifically on issues related to health-care, Québécois politics, or news about a particular advertiser. To understand common patterns, we further analyse 20 randomly chosen articles from MB/FC. The most common facts include known biases, funding model, remit, fact-checking history, and ownership (see Appendix D).

In order to build an MBC, a model must first recall or retrieve pertinent details for the source. These range from simple facts (e.g., whether the source is funded by a specific government) to more complex (e.g., whether a poor track record for factuality is evidenced). Necessary retrieval steps may differ depending on the source type – websites have

	Train	Dev	Test
Background checks	5209	500	1000
# Avg. Lines	17.1	17.2	17.0
# Avg. Tokens	305.1	302.2	303.2

Table 1: Statistics for our dataset. Background checks were randomly divided among the three splits.

different indicators of credibility than print sources, for example (Potter, 2013). Once the appropriate information has been retrieved, the model must then summarise this information.

3 Dataset

Media Bias / Fact Check² (MB/FC) is an independent, volunteer-run website that promotes awareness of media bias. While past datasets have used bias ratings from the site (Baly et al., 2018), the textual “*detailed reports*” are an untapped resource. These are well-sources documents that summarise the factual history, bias, tendency, and potential conflicts of interest for media sources – i.e., *background checks*. We introduce a novel dataset based on these. We collected all reports listed through their search tool. This provided a total of 6,709 background checks. We cleaned the reports, and removed the bias and credibility ratings. Statistics for the dataset can be seen in Table 1, and the dataset itself can be downloaded at <https://github.com/MichSchli/MediaBackgroundChecks>.

While MB/FC is an extensive resource, covering almost 7,000 media organisations, it is far from complete. In our human evaluation in Sections 4.4, out of 40 sources used, 29 are covered by MB/FC reports. However, of these 29, 9 are incomplete or lack crucial information. We do not intend to replace human-written background checks – rather, we suggest that generated background checks can supplement MB/FC in these cases.

4 Experiments

4.1 Models

For a simple baseline, we prompt a large language model to generate an MBC given the name of a source. Our prompt can be seen in Appendix B.1. We experiment with two models, one closed-source and one open-source: *gpt-3.5-turbo-0125*³ and

²<https://mediabiasfactcheck.com/>

³ChatGPT; see <https://platform.openai.com/docs/models/gpt-3-5-turbo>

*Llama 3 8B Instruct*⁴, the instruction-tuned version of Llama 3. The latter is small enough to reasonably run on a single consumer-grade GPU.

For these initial models, we rely solely on the information stored in their weights. As such, their performance is representative of the level of source-critical reasoning that can be expected of current-generation LLMs, without any additional evidence provided. MBCs often incorporate information available on the internet – the 6,709 background checks we analysed contained an average of 8.9 hyperlinks. To surface this information, we propose to rely on retrieval-augmented generation (Lewis et al., 2020). In addition to providing external memory at generation time, retrieved evidence can also be shown to the user to establish why the *background check* should (or should not) be trusted. Showing where evidence originates from to establish trust is crucial in related tasks, such as automated fact-checking (Guo et al., 2022). It also provides a rudimentary form of interpretability.

As the necessary retrieval steps are roughly similar across background checks, we have compiled a list of seven information-seeking search queries focusing on different aspects of a background check (see the list in Appendix E). In initial experiments, we found this strategy to perform better than generating queries using our LLM.

For each query, we gather relevant search results using the Google Search API⁵, selecting the top 30 retrieved documents. We exclude the Media Bias / Fact Check website and any website linking to it. To focus on the most salient retrieved information, we use a question-answering model to extract answers to a predefined question for each information-seeking query (also found in Appendix E). Here, we employ DeBERTa v.3 (He et al., 2021) pre-trained on SQuAD (Rajpurkar et al., 2016). By this process we transform a Google search result into a question-answer pair.

For each question-answer pair, we then *expand* the background check with the information contained within. This allows us to gradually incorporate information from multiple rounds of retrieval, without running into the token limit. To expand a background check with information from a specific question-answer pair, we use the prompt seen in Appendix B.2 with each respective LLM.

⁴see <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

⁵<https://programmablesearchengine.google.com/>

4.2 Evaluation

When evaluating media background checks, we are primarily interested in *recalling* the information provided in the gold background check. Systems may find information beyond what is included in the gold example. Prioritising recall avoids penalising such additions.

We adopt a variant of FActScores (Min et al., 2023) to evaluate models. We begin by breaking each gold MBC down into atomic facts. As MBCs cover relatively similar topics, we use a list of 42 templates to generate facts (see Appendix C). Each template contains contextually determined tokens, which we fill with details from the gold MBC using *GPT-3.5-turbo*. During initial experiments, we found this strategy to perform better than fully generating facts. We subsequently verify whether each fact is entailed by the gold MBC. We experimented with several models, including a trained DeBERTa model (He et al., 2021) and an open-source LLM, but ultimately found the best-performing system to be *GPT-3.5-turbo* with zero-shot chain of thought (Kojima et al., 2022). Our prompt can be found in Appendix B.3. To increase performance, we prompt the (probabilistic) model four times, and take as prediction as the majority-voted element among the runs. We keep atomic facts which are *entailed* or *contradicted* by the gold MBC.

To score a generated background check, we compute *fact recall*, the percentage of atomic facts for which the gold MBC and the predicted MBC agree on entailment. We use the same entailment model. We furthermore report the error rate, the percentage of atomic facts either are *entailed* by one MBC and *contradicted* by the other.

As a sanity check, we evaluate what performance the *gold* background checks would yield on the development set; that is, using gold data both as prediction and reference. Ideally, the fact recall rate should be roughly 1, and the error rate roughly 0; but as entailment is noisy, this ideal is not reached. The computed fact recall and error rate are, respectively, 84.1% and 3.4%. We conclude that fact recall, while noisy, remains a useful measurement of how many gold facts are recalled.

We furthermore manually inspect evaluation for ten MBCs. We find that, on average, 59.8% of the facts from the gold summary were accounted for in the generated atomic facts. The majority of the excluded facts were “multi-hop facts”, which unfortunately our evaluation strategy does not ac-

<p>Gold Example: History Founded in 2005, Natural News is a conspiracy and pseudoscience website that routinely publishes false information. The founder is Mike Adams, who owns several Questionable websites such as News Target and Trump.news.</p> <p>Funded by / Ownership Natural News is owned by Mike Adams, who owns numerous other fake and or controversial websites. The website, like all of Mike Adams's properties, is funded through online advertising.</p> <p>Analysis / Bias ...</p> <hr/> <p>Generated Example (GPT-3.5-Turbo): Background check 1. Natural News is owned by Mike Adams, who is also known as the "Health Ranger." 2. The website is primarily funded through the sale of various health supplements, books, and other products advertised on the site. 3. Natural News is known for promoting alternative medicine, conspiracy theories, and pseudoscience. 4. The site has been criticized for its lack of credible sources and scientific evidence to support its claims.</p> <hr/> <p>Generated Example (GPT-3.5-Turbo + Google): *Background check on naturalnews.com:*</p> <ol style="list-style-type: none"> Ownership: Founded by Mike Adams, known as the "Health Ranger." Funding: Primarily funded through the sale of various health supplements and products on their website. Bias: Strongly criticized for promoting pseudoscience, conspiracy theories, and misinformation, particularly regarding vaccines and alternative medicine. Fact-checks: Natural News has a history of publishing misinformation and has been flagged by fact-checkers such as Snopes, Politifact, and FactCheck.org for spreading false information. Type of site: Known for promoting alternative medicine, pseudoscience, disinformation, and far-right extremism. Country of Origin: United States. Political leaning: Natural News is considered a far-right website known for anti-vaccination conspiracy theories and fake news.
--

Figure 3: Example background checks for *Natural News*. The gold example is taken from the Media Bias / Fact Check website, while the generated example is produced by GPT-3.5 augmented with Google search as described in Section 4.1. The gold example has been shortened, and the full version can be seen at <https://mediabiasfactcheck.com/natural-news/>.

count for – such as which *other* media companies are owned by the parent organization of the source reported on in the background check.

When the right atomic facts *are* extracted, our entailment system works well. When evaluating generated MBCs for these ten sources, we agree with 93.9% of entailment predictions made by our GPT-3.5 ensemble.

4.3 Results

Using this evaluation metric, we evaluate our four MBC generation models (see Table 2). In addition to our proposed metric, we also include two traditional measures: METEOR and ROUGE-L. Fact recall rates are low, highlighting the difficulty of the task. Nevertheless, we see clear performance improvements from retrieval-augmentation, both

for GPT-3.5 and Llama 3. This supports our intuition that *finding the right information* is a crucial barrier to source-critical reasoning in models.

To understand what causes the low recall, we manually analyse 10 randomly chosen examples (generated by GPT-3.5 with retrieval). We find the following omissions: seven missing entity mentions (e.g., who the editor is), three missing failed fact-checks, three missing mentions of editorial stance (e.g., right- or left-leaning), two missing funding sources, two historical events (e.g., an ownership change), and one missing mention of content being re-published from another site. Without retrieval, we see a further two missing failed fact-checks, one missing entity mention, one missing funding source, and one missing mention of editorial stance.

	Fact Recall	Error Rate	METEOR	ROUGE-L
GPT-3.5-Turbo	22.7%	6.2%	9.9%	12.5%
GPT-3.5-Turbo + Google	26.1%	6.3%	12.6%	13.1%
Llama 3 8b Instruct	24.4%	10.4%	15.3%	14.4%
Llama 3 8b Instruct + Google	25.1%	10.7%	15.5%	14.4%

Table 2: Performance of four different systems on our MB/FC dataset. We measure fact recall and error rate, as discussed in Section 4.2. These represent, respectively, how many facts from the gold background check were recalled by the model, and how many facts from the gold background check were *contradicted* by the model. We also include two traditional summarisation metrics, METEOR and ROUGE-L.

We furthermore see, in both cases, five different multi-hop facts omitted. However, as we discuss Section 4.2, our evaluation metric *also* fails at catching those, so they are not responsible for the low recall. Overall, the biggest problem is missing entities – future work could tailor retrieval especially to this scenario, for example by placing extra emphasis on retrieved “about”-sections.

Compared to GPT-3.5, Llama achieves high fact recall and METEOR/ROUGE-L scores, but also exhibits a high error rate. Exploring the data, we see significant differences in the lengths of MBCs – the average GPT-3.5 generated MBC is 176 tokens, while for Llama the average has 254 tokens. The gold MBCs contain on average 477 tokens. This explains the discrepancy: Llama generates longer summaries with more facts, and so is more likely to state both correct and incorrect things about the knowledge source. Manually reviewing the generated data, we see one more difference: GPT-3.5 performs better at integrating retrieved information, explaining the higher fact recall in the retrieval-augmented setting.

4.4 Human Evaluation

We envision MBCs as being used in two settings: as assistive instruments for either *humans* or *models* having to make sense out of untrustworthy evidence. To demonstrate the potential, we conduct two experiments with human participants. For this purpose, we recruited 11 researchers working on automated fact-checking, hate speech analysis, LLMs, and related NLP tasks (ranging from PhD-students to assistant professors).

4.4.1 QA with Untrustworthy Evidence

We first create a small dataset of questions for which multiple, conflicting, and potentially untrustworthy evidence documents could reasonably be expected to surface. We compose this of ten ques-

tions about *controversial subjects*, and ten questions about *known misinformation*. Wan et al. (2024) recently released a small dataset of questions with conflicting evidence, including four examples of controversial political questions. We include the four from their dataset, and generate an addition six following their approach. For known misinformation, we manually extract ten questions by rephrasing claims from the development set of the AVeriTeC dataset (Schlichtkrull et al., 2023b), which contains labelled examples of fact-checked real-world misinformation. We choose randomly from *refuted* claims in the dataset, focusing on claims originating from sources easily findable via a search engine (i.e., excluding claims from Twitter, Facebook, and other social media).

For all twenty questions, we find two disagreeing evidence documents. We pick the first two search results that disagree when entering the question into Google. In all cases, these documents are found on the first page. These are documents which a searcher, whether human or algorithm, would easily come across. For the ten *known misinformation* questions, the original misinformation source labelled in AVeriTeC appeared on the first page 6/10 times; for the rest, we found an alternative source supporting the misinformation.

We believe question answering is representative of how LLMs are envisioned to replace search; see e.g. Metzler et al. (2021). We recognise that, in actual use, LLMs are also used for other tasks, such as programming and software engineering-related functions. Answering questions about culture and geography is among the most frequent uses (Zheng et al., 2024), and for such questions, source reliability is certainly an issue. Further, answering more general questions is also a common use-case. As such, we believe that QA is 1) *highly* representative of the vision large providers have of LLMs, and 2) representative of current use cases.

	with MBC		without MBC		<i>t</i> Statistic	<i>p</i> -value
	mean	SD	mean	SD		
Provision of Sufficient Information	78.2%	-	70.9%	-	0.740	0.746
Difficulty of Answering	2.24	0.75	2.81	0.82	-1.633	0.133
Difficulty of Establishing Trust	1.95	0.54	2.88	0.65	-3.791*	0.004

Table 3: Results for part one of our human evaluation, estimating the helpfulness of generated MBCs when presented directly to users. Provision of sufficient information is annotated as a binary yes/no-question, while the difficulty of answering the question and the difficulty of establishing which sources are trustworthy are rated on five-point Likert scales. Results are analysed via student’s *t*-tests. * indicates significance at $p = 0.01$.

4.4.2 Are MBCs helpful for humans?

We first investigate whether MBCs are helpful for humans when encountering conflicting sources. We ask our annotators to answer ten questions from our dataset, given the two conflicting sources associated with that question. We randomly pick five questions based on controversial subjects, and five based on misinformation. For each question, we further randomly choose whether to show an accompanying background check. We use generated MBCs, simulating real-world use of our model. We use the best-performing model – GPT-3.5-Turbo with retrieval.

After answering the question, each participant is asked to judge 1) whether they were given sufficient information to answer the question; 2) how difficult it was to answer the question; and 3) how difficult it was to decide which source to trust. Intuitively, if MBCs are helpful for humans when engaging with epistemic uncertainty, we should expect a lower cognitive load – and thus for the task to feel less difficult (Sweller et al., 2011) – when an MBC is provided. For information sufficiency, we ask participants to give a binary answer. For task difficulty, we follow previous work and use a self-reported Likert scale as our measurement. We use a scale of 1 (very easy) to 5 (very difficult). An example page from our questionnaire can be seen in Appendix F.1.

In Table 3, we report the mean answer given by our participants for information sufficiency, answer difficulty, and trust difficulty. We furthermore report the standard difference between participants for answer difficulty and trust difficulty. To analyse our results, we conduct a paired samples *t*-test comparing the responses of each participant with and without media background checks.

The cognitive load of deciding which documents to trust was *much* lower when annotators were provided with a background check (1.95 versus 2.88,

on average). The cognitive load of answering the question was also lower with a background check, although this result is not statistically significant. Similarly, our annotators more frequently reported that sufficient information had been provided when also given a background check, but this result was not statistically significant either. Post-hoc discussions with our participants were revealing: when not provided with a background check, some participants instead used the internet to find similar information for themselves. As our instructions included a link to each of the sources used, participants considered the information provided to be “sufficient”; although answering the question required additional effort. For humans, automatically generated background checks thus *quicken* a part of the meaning-producing process that is *in all cases necessary*.

4.4.3 Are MBCs helpful for NLP models?

Ultimately, our goal in developing MBCs was to introduce and test source-critical reasoning capabilities in NLP models, specifically LLMs. As such, we also seek to demonstrate the helpfulness of MBCs to retrieval-augmented LLMs. To do so, we again conduct a human experiment.

We simulate a question-answering setting with a retrieval-augmented LLM. Given one of the remaining questions from our dataset, we assume that retrieval has returned one of the two disagreeing evidence documents. We ask GPT-4⁶ to answer the question based on the returned document. We repeat this with the other evidence document. For each source, we generate a background check using our best-performing model (GPT-3.5-Turbo with search). We generate a second version of the answer, including this background check in the instructions to GPT-4. Our prompts for the two cases can be seen in Appendix B.4.

⁶gpt-4-turbo-2024-04-09

	with MBC	without MBC	Equally Good	χ^2	p -value
Preferred Answer	165	26	29	57.02*	0.000
Better Understanding Provided	69	56	95	60.22*	0.001

Table 4: Results for part two of our human evaluation, estimating the helpfulness of generated MBCs when provided along with retrieved results to GPT-4 in a question-answering task. Annotators directly state which answer they prefer, and which provides a better understanding of the topic. Results are analysed via a chi-square test. * indicates significance at $p = 0.01$.

	with MBC		without MBC		t Statistic	p -value
	mean	SD	mean	SD		
Answer is Misleading	1.57	0.35	2.58	0.63	-5.634*	0.000

Table 5: Further results for part two of our human evaluation. Annotators evaluate how misleading LLM responses are on a five-point Likert scale. Results are analysed via a student’s t -test. * indicates significance at $p = 0.01$.

We show our participants the two answers to each of these 20 question-source pairs (in random order). We then ask them to determine 1) which answer they prefer; 2) which answer gives the best understanding of the topic; 3) for each answer, if they would feel misled if given that answer by an AI chatbot. For preference and understanding, annotators can pick the first answer, the second answer, or indicate that the two answers are equally good. For feeling misled we employ a five-point Likert scale ranging from 1 (no, not at all) to 5 (yes, very much). An example page from our questionnaire can be seen in Appendix F.2.

In Table 4 we report how many times our participants indicated that they preferred an answer generated with or without an MBC, and how many times they indicated that answers with or without MBCs provided a better understanding of the topic. We conduct a chi-squared test to establish the significance of our results. The responses strongly indicate that participants preferred the answers generated with MBCs. Further, answers generated with MBCs provided a better understanding of the topic more often – by a smaller margin, but still significantly so. Interestingly, if we analyse only the answers generated with (what we consider to be) trustworthy sources, the latter finding disappears while the former remains. We theorise that MBC-backed answers give a better understanding when sources are untrustworthy, but are preferred even for trustworthy sources as they help the user obtain confidence in their answer.

In addition to the pairwise comparison, we further investigated misleadingness. In Table 5, we report the mean answer given by our participants

on how misled they would feel if given a generated answer by a chatbot, along with the standard deviation. We conduct a paired samples t -test to analyse these results. As can be seen, answers generated with MBCs were on average rated *significantly less misleading* than answers generated without MBCs.

These findings support our primary hypotheses – that search-augmented LLMs do not adequately account for the tendencies and biases of the sources they rely on, and that providing (even automatically generated) MBCs to models at inference time can alleviate this and enable automated source-critical reasoning.

5 Related Work

Closest to our work, Baly et al. (2018); Zhang et al. (2019b); Baly et al. (2020b) classifies the bias and factuality of sources based on data from Media Bias / Fact Check. However, they only predict the *bias labels* for both (as given elsewhere on MB/FC), *not* the detailed background checks we produce. Features include sample articles from the source, its Wikipedia page and Twitter account, and information about the web domain. Hounsel et al. (2020) proposed further web-domain-features such as host, domain, and certificates for websites when predicting trustworthiness.

Knowledge conflicts in LLMs is an active research area; see the survey by Xu et al. (2024). Using their terminology, our proposal is most relevant *inter-context conflicts*, i.e. conflicts between two or more retrieved documents. The mentioned mitigation strategies, e.g. Chen and Shu (2023); Verghe et al. (2024), follow the above line of reasoning:

they use a trained model to classify documents into “factual” and “unfactual”, and remove the “unfactual” documents. This assumes a single source of truth (the training data or evidence database for the misinformation detector), which as we have argued does not align with best practises for human knowledge experts. Our proposal would instead enable an evidence-based comparison of trustworthiness for retrieved sources.

A rich literature exists – referred to as *fact-finding* – proposing probabilistic models for computing the likelihood of claims and the trustworthiness of sources (Yin et al., 2007; Dong et al., 2009; Pasternack and Roth, 2010, 2013; Vydiswaran et al., 2011; Zhang et al., 2019a). The basis is the assumption by (Yin et al., 2007) that “*a web site is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy web sites*”. Yuan et al. (2020) proposed to predict the credibility of sources based on whether known misinformation spreaders share similar content. Along similar lines, Wright and Augenstein (2021) introduced a model for citation-worthiness where citation-worthy papers feature in citation-worthy journals, and vice versa.

Dong et al. (2009) identified a *copying problem* in fact-finding, where this strategy breaks down if many seemingly independent actors copy their positions from each other; this makes majority voting an unreliable heuristic for determining whether a piece of information is trustworthy. Bala and Goyal (1998) identified a similar issue, where a “royal family of knowers” can cause communities to converge on false beliefs if their connectivity is much greater than the average epistemic agent.

Kaneko et al. (2009); Nakano et al. (2010) proposed to develop “credibility survey reports” for search results. These are topic-specific documents that map the positions of different, contradicting search results with respect to *one particular search query*. Unlike our proposal, these do not represent the *general* biases, tendencies, and factual histories of media sources. Shibuki et al. (2010) later developed a summarisation algorithm for this task.

The credibility and bias of individual documents has also been studied previously. For Wikipedia articles, Zeng et al. (2006); Adler and de Alfaro (2007) attempted to predict trustworthiness based on revision history. (Nakov et al., 2017) studies the credibility of statements on community QA forums based on linguistic signals in the individual posts;

Baly et al. (2020a) similarly used the surface forms of text to predict the credibility of news articles.

Finally, the concept of credibility has been studied extensively outside NLP, e.g. in information science. For an overview, the framework by Rieh and Danielson (2007) is an excellent resource. Similarly, source critical methods are widely studied, especially in history; we recommend Howell and Prevenier (2001) for an introduction.

6 Conclusion

We have introduced *media background check generation*, a novel NLP task wherein models summarise information about the bias, tendency, and factual record of media sources. Media background checks enable source-critical reasoning by recalling the same facts that human knowledge experts use to assess trust. We have furthermore presented a dataset of 6,709 examples collected from the Media Bias / Fact Check website. We have investigated several baselines for the task, including both open-source and closed-source LLMs. Our findings demonstrate that retrieval-augmentation can greatly improve performance, and interestingly that open-source models are very competitive on this difficult task. Finally, our human evaluation gives strong evidence both that media background checks are helpful for humans when evaluating media sources, and helpful for models when generating answers based on retrieved sources.

7 Limitations

Our paper proposes to establish trust by providing information about bias and tendency. In our models, that information comes either from model weights or retrieved documents. As such, that information is *itself* potentially untrustworthy. Taken seriously, this prompts *another* round of retrieval to establish the trustworthiness of the background check; and then another, to establish the trustworthiness of *that* round – it’s turtles all the way down. Ultimately, our proposal cannot conclusively establish trust; only establish it insofar as the user already trusts some sources. As users have different requirements in terms of which sources they trust and how many “levels” of trust they may wish to explore, one possible solution could be an interactive system allowing users to expand background checks in a desired direction, similar to recent proposals for summarisation (Shapira et al., 2021).

When evaluating systems with access to search, we excluded the Media Bias / Fact Check website itself (and associated websites) from search results. Our intention was to make our evaluation setup as fair as possible. However, this introduces two limitations: 1) “in the wild” systems may perform *better* than in our evaluations, as they do have access to background checks from MB/FC; and 2) we may still overestimate the relative performance of retrieval-augmented systems slightly, as we ultimately cannot make sure that no website which quotes MB / FC was included in search results. For the latter, manual analysis of 20 examples did not turn up any such websites.

8 Ethical Considerations

The dataset presented in this paper is based on Media Bias / Fact Check, a volunteer-run project. Background checks included here may themselves be biased on incomplete, as may background checks produced by models trained or evaluated on our dataset. Furthermore, the machine learning models and search engine used for our models contain well-known biases (Noble, 2018; Bender et al., 2021). Acting on trustworthiness estimates arrived at through biased means, including automatically produced ranking decisions for evidence retrieval, risks causing epistemic harm (Schlichtkrull et al., 2023a). The datasets and models described in this paper are not intended for and should not be used for truth-telling, e.g. for the design of automated content moderation systems.

We did not take any steps to anonymise the data. The claims discussed in our dataset are based on publicly available data from a journalistic publication, and concern public figures and events – references to these are important to document the history of a publication. If any person included in our dataset (e.g., the owner of a particular media source) requests it, we will remove that example from the dataset.

Acknowledgements

We would like to thank Nedjma Ousidhoum and Rui Cao for their helpful comments, discussions, and feedback. We would also like to thank the anonymous reviewers for their questions and comments that helped us improve the paper. Finally, we would like to thank Dave Van Zandt and the Media Bias / Fact Check team for lending us their data, as well as for their pioneering work on the subject.

References

- B. Thomas Adler and Luca de Alfaro. 2007. [A content-driven reputation system for the wikipedia](#). In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, page 261–270, New York, NY, USA. Association for Computing Machinery.
- Venkatesh Bala and Sanjeev Goyal. 1998. [Learning from neighbours](#). *Review of Economic Studies*, 65(3):595–621.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. [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 (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. [Predicting factuality of reporting and bias of news media sources](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.
- Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. [What was written vs. who read it: News media profiling using text analysis and social media context](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? [🦜](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. [To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making](#). *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).
- Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of LLMs: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.
- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. [Integrating conflicting data: the role of source dependence](#). *Proc. VLDB Endow.*, 2(1):550–561.

- Yigal Godler and Zvi Reich. 2015. [Journalistic evidence: Cross-verification as a constituent of mediated knowledge](#). *Journalism: Theory, Practice & Criticism*, 18(5):558–574.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying disinformation websites using infrastructure features. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*.
- Martha C. Howell and William Prevenier. 2001. *From Reliable Sources: An Introduction to Historical Methods*. Cornell paperbacks. Cornell University Press.
- J. M. Imperial, G. Forey, and H. T. Madabushi. 2024. [STANDARDIZE: Aligning language models with expert-defined standards for content generation](#). *arXiv preprint arXiv:2402.12593*.
- International Fact-Checking Network. 2016. [Code of principles](#). Poynter Institute.
- Koichi Kaneko, Hideyuki Shibuki, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. 2009. [Mediatory summary generation: Summary-passage extraction for information credibility on the web](#). In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 240–249, Hong Kong. City University of Hong Kong.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Marcel Machill and Markus Beiler. 2009. [The importance of the internet for journalistic research: A multi-method study of the research performed by journalists working for daily newspapers, radio, television and online](#). *Journalism Studies*, 10(2):178–203.
- Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. [Rethinking search: making domain experts out of dilettantes](#). *ACM SIGIR Forum*, 55:1–27.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Masahiro Nakano, Hideyuki Shibuki, Rintaro Miyazaki, Madoka Ishioroshi, Koichi Kaneko, and Tatsunori Mori. 2010. [Construction of text summarization corpus for the credibility of information on the web](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. 2017. [Do not trust the trolls: Predicting credibility in community question answering forums](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 551–560, Varna, Bulgaria. INCOMA Ltd.
- Safiya Umoja Noble. 2018. *Algorithms of oppression*. In *Algorithms of Oppression*. New York University Press.
- Jeff Pasternack and Dan Roth. 2010. [Knowing what to believe \(when you already know something\)](#). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 877–885, Beijing, China. Coling 2010 Organizing Committee.
- Jeff Pasternack and Dan Roth. 2013. [Latent credibility analysis](#). In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13*, page 1009–1020, New York, NY, USA. Association for Computing Machinery.
- Denis Peskoff and Brandon Stewart. 2023. [Credible without credit: Domain experts assess generative language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada. Association for Computational Linguistics.
- William James. Potter. 2013. *Media Literacy*. SAGE Publications.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Soo Young Rieh and David R. Danielson. 2007. [Credibility: A multidisciplinary framework](#). *Annual Review of Information Science and Technology*, 41(1):307–364.

- Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023a. [The intended uses of automated fact-checking artefacts: Why, how and who](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.
- Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023b. [AVeriTeC: A dataset for real-world claim verification with evidence from the web](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Chirag Shah and Emily M. Bender. 2022. [Situating search](#). In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, CHIIR '22*, page 221–232, New York, NY, USA. Association for Computing Machinery.
- Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2021. [Extending multi-document summarization evaluation to the interactive setting](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677, Online. Association for Computational Linguistics.
- Hideyuki Shibuki, Takahiro Nagai, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. 2010. [A method for automatically generating a mediatory summary to verify credibility of information on the web](#). In *Coling 2010: Posters*, pages 1140–1148, Beijing, China. Coling 2010 Organizing Committee.
- Thomas W Simpson. 2013. Evaluating google as an epistemic tool. *Philosophical Engineering: Toward a Philosophy of the Web*, pages 97–115.
- Catherine L. Smith and Soo Young Rieh. 2019. [Knowledge-context in search systems: Toward information-literate actions](#). In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19*, page 55–62, New York, NY, USA. Association for Computing Machinery.
- Steen Steensen. 2019. [Journalism's epistemic crisis and its solution: Disinformation, datafication and source criticism](#). *Journalism*, 20(1):185–189. Publisher: SAGE Publications.
- J. Sweller, P. Ayres, and S. Kalyuga. 2011. *Cognitive Load Theory*. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies. Springer New York.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Tyler Vergo, Jean-Francois Godbout, Reihaneh Rab-bany, and Kellin Pelrine. 2024. [Comparing GPT-4 and open-source language models in misinformation mitigation](#).
- V.G. Vinod Vydiswaran, ChengXiang Zhai, and Dan Roth. 2011. [Content-driven trust propagation framework](#). In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, page 974–982, New York, NY, USA. Association for Computing Machinery.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Alexander Wan, Eric Wallace, and Dan Klein. 2024. [What evidence do language models find convincing?](#) *Preprint*, arXiv:2402.11782.
- Dustin Wright and Isabelle Augenstein. 2021. [Cite-Worth: Cite-worthiness detection for improved scientific document understanding](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1796–1807, Online. Association for Computational Linguistics.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for LLMs: A survey](#). *Preprint*, arXiv:2403.08319.
- Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2007. [Truth discovery with multiple conflicting information providers on the web](#). In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07*, page 1048–1052, New York, NY, USA. Association for Computing Machinery.
- Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. [Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. 2006. [Computing trust from revision history](#). In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, PST '06*, New York, NY, USA. Association for Computing Machinery.
- Yi Zhang, Zachary Ives, and Dan Roth. 2019a. [Evidence-based trustworthiness](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 413–423, Florence, Italy. Association for Computational Linguistics.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019b. [Tanbih: Get to know what you are reading](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 223–228, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. [LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset](#). In *International Conference on Learning Representations (ICLR)*.

A Computational Resources

Our experiments with opensource models were carried out using a single NVidia A100 GPU.

B Prompts

B.1 Initial Prompt

Our baseline consists of a simple prompt to an LLM, querying the model to generate an MBC based on the name of a media source. We use the same prompt for ChatGPT and Llama 3. The prompt can be seen in Figure 4. When generating an initial background check to use as a starting point for our retrieval-augmented models, we use this prompt as well.

System message You are InfoHuntGPT, a world-class AI assistant used by journalists to quickly build knowledge of new sources.

User message Build a background check for the news source “{source name}”. Write down everything you know about them, e.g. who funds them, how they make money, if they have any particular bias. Make an ITEMIZED LIST. Be brief, and if you don’t know something, just leave it out. If you are aware that they have failed any fact-checks, mention which. Begin your response with “**Background check**”.

Figure 4: Prompt used for ChatGPT and Llama 3 when generating MBCs with no supporting retrieved evidence. This prompt is also used to generate the *initial* MBC which is later updated in the retrieved-evidence setting.

System message You are InfoHuntGPT, a world-class AI assistant used by journalists to quickly build knowledge of new sources.

User message Build a background check for the news source “{source name}”. Write down everything you know about them, e.g. who funds them, how they make money, if they have any particular bias. Make an ITEMIZED LIST. Be brief, and if you don’t know something, just leave it out. If you are aware that they have failed any fact-checks, mention which. Begin your response with “**Background check**”.

Assistant message {Previous background check}

User message Google search has revealed some new information:

{Question-answer pairs}

Update your background check for “{source name}” using the new information. Do NOT delete any information, but make ADDITIONS where necessary, using the new information. Most likely, you will just need to add an extra item to the itemized list you previously created. Make minimal edits, and only incorporate what is relevant. Begin your response with “**Background check**”.

Figure 5: Prompt used for ChatGPT and Llama 3 when updating an MBC with retrieved information. The retrieved information is input to the prompt in the form of question-answer pairs, following the methodology described in Section 4.1.

B.2 Update Prompt

For our retrieval-augmented models, we use the prompt seen in Figure 5 to incorporate new information from Google Search into a background check.

B.3 Entailment Prompt

When evaluating MBCs via FActScores, we compute entailment using the prompt seen in Figure 6.

B.4 Question-Answering Prompt

For our human evaluation in Section 4.4.3, we simulate retrieval-augmented question-answering. Given a source document and possibly an MBC, we answer questions using the prompts in Figures 7 and 8.

C Fact Generation

When evaluating background checks, we use a variant of FActScore (Min et al., 2023). The atomic

System message You are FactCheckGPT, a world-class tool used by journalists to discover problems in their writings. Users give you text, and check whether facts are true given the text. You ALWAYS answer either TRUE, FALSE, or NOT ENOUGH EVIDENCE.

User message You will be given a snippet written as part of a source criticism exercise, and a claim. Your task is to determine whether the claim is true based ONLY on the text. Do NOT use any other knowledge source

The claim is: “{*hypothesis*}”.
The text follows below:
“{*premise*}”.

{*hypothesis*}? Thinking step by step, answer either TRUE, FALSE, or NOT ENOUGH EVIDENCE, capitalizing all letters. Explain your reasoning FIRST, and after that output either TRUE, FALSE, or NOT ENOUGH EVIDENCE.

Figure 6: Prompt used for our LLMs when computing textual entailment. We use a probabilistic model, and retain the majority prediction over four votes.

facts we use are generated following the templates seen in Figure 10. The initial questions are decorated with information from the gold background check using the prompt seen in Figure 9.

D Facts in Media Background Checks

To understand which facts are represented in the MB/FC dataset, we manually analyse 20 randomly selected background checks. We include our findings in Table 6. We note that this is not a complete list – picking specific high-quality background checks, e.g. background checks for the Guardian, Fox News, or Breitbart, reveals usage of rarer facts. As such, the list of atomic facts we use for evaluation is longer (see Appendix C).

E Information-seeking Questions

When retrieving information, we first retrieved documents via the Google Search API. Then, we use a trained question-answering model to select the most salient substrings. The queries and questions used for both can be found in Figure 11.

F Questionnaire

F.1 Part 1

The questionnaire pages used in the first half of our survey can be seen in Figures 12, 13, and 14.

System message You are an expert journalist. You will be given some evidence, and a question. Using the provided evidence, answer the question.

User message Using the provided evidence, answer the following: “{*question*}”.

The following evidence is provided: “{*source document*}”

This information comes from “{*domain name of source document*}”.

Your answer should be short and concise, five sentences at most. Include an explanation of why the user should or should not trust the source.

Figure 7: Prompt used for answering questions using GPT-4 when not providing an MBC.

These focus on evaluating if MBCs are helpful for *humans* when creating meaning from untrustworthy evidence documents.

F.2 Part 2

The questionnaire pages used in the second half of our survey can be seen in Figures 15 and 16. These focus on evaluating if MBCs are helpful for *NLP models* when creating meaning from untrustworthy evidence documents.

G Data Statement

Following Bender and Friedman (2018), we include a data statement describing the characteristics of MB/FC.

G.1 Curation Rationale

We processed a total of 6,709 source documents from the Media Bias / Fact Check website, which provides volunteer-written assessments of English-language media sources. The intended use is to provide empirical evidence for the level of source-critical analysis possible for NLP models, as well as to drive research into new tools that enable source-critical analysis by models and humans.

G.2 Language variety

As MB/FC only provides analysis of English-language sources, in English, the same holds true for our dataset.

System message You are an expert journalist. You will be given some evidence, and a question. Using the provided evidence, answer the question.

User message Using the provided evidence, answer the following: “{question}”.

The following evidence is provided: “{source document}”

This information comes from “{domain name of source document}”.

The following information about the source may be relevant: “{background check}”

Your answer should be short and concise, five sentences at most. Include an explanation of why the user should or should not trust the source.

Figure 8: Prompt used for answering questions using GPT-4 when providing an MBC.

System message You are InfoHuntGPT, a world-class AI assistant used by journalists to quickly build knowledge of new sources.

User message You will be given a snippet written as part of a source criticism exercise, and a fill-in-the-blank question (blanks represented by _). Your task is to fill in the blanks in the sentence, adding no additional information or wording. JUST replace the _character.

The question is:
{template}

The text follows below:
{gold background check}

Fill in the blanks in the question, adding no additional information or wording. JUST replace the _character, and output ONLY the question with the blank filled in. No yapping.

Figure 9: Prompt used for ChatGPT when decorating an atomic fact template with information from a background check.

G.3 Speaker demographics

We did not analyse the demographics of the individual writers and editors for each media source.

G.4 Annotator demographics

For this dataset, we did not rely on human annotators beyond our own group. We pro-

cessed data from Media Bias / Fact Check, a volunteer-driven project, with details available here: <https://mediabiasfactcheck.com/frequently-asked-questions/>.

G.5 Speech situation

The MBCs included in this dataset were provided by the Media Bias / Fact Check volunteers for the purpose of educating the public on media bias and deceptive news practises.

G.6 Text characteristics

We compute various statistics for the text included in this dataset; see Section 3.

Fact	Percentage of MBCs
General bias/tendency of { <i>source name</i> }	80%
Funding model for { <i>source name</i> }	75%
Remit of { <i>source name</i> }	75%
Fact-checking history of { <i>source name</i> }	65%
Owner of { <i>source name</i> }	60%
Publisher of { <i>source name</i> }	50%
Examples of biased articles from { <i>source name</i> }	50%
Geographical focus of { <i>source name</i> }	50%
Other history of { <i>source name</i> }	45%
Multihop facts	40%
{ <i>source name</i> } does not disclose important information	35%
Sources used by { <i>source name</i> } (e.g., AP)	35%
Founder of { <i>source name</i> }	25%
Loaded language used by { <i>source name</i> }	25%
Examples articles from { <i>source name</i> } demonstrating aspects other than bias	20%
Political endorsements by { <i>source name</i> }	10%
{ <i>source name</i> } masquerades as	10%
Awards given to { <i>source name</i> }	10%
Editor of { <i>source name</i> }	10%
Bias rating by other site (e.g., NewsGuard)	10%
Demonstration of agreement by { <i>source name</i> } with scientific consensus	5%
Comparison to other media	5%

Table 6: Facts appearing in 20 randomly sampled background checks from Media Bias / Fact Check.

1. *_founded* {*source name*}
2. *_hosts* {*source name*}
3. *_is the founder of* {*source name*}
4. *_is the ceo of* {*source name*}
5. *_leads* {*source name*}
6. *_owns* {*source name*}
7. *_publishes* {*source name*}
8. *_is the owner of'*{*source name*}
9. *_bought* {*source name*}
10. *_acquired* {*source name*}
11. {*source name*} is funded through _
12. The remit of {*source name*} is to _
13. The usual audience of {*source name*} is _
14. {*source name*} was awarded _
15. {*source name*} tends to the _
16. Other sources have commented on {*source name*}, stating that it tends to be _
17. The factuality of {*source name*} is reported to be _
18. {*source name*} failed a fact-check for an article titled “_”
19. {*source name*} printed a retraction after failing a fact-check for an article titled “_”
20. {*source name*} chose not to cover _
21. {*source name*} provides original content written by staff journalists
22. {*source name*} relies on advertising for revenue
23. {*source name*} relies on subscriptions for revenue
24. {*source name*} relies on donations for revenue
25. {*source name*} has received donations from _
26. When aggregating stories, {*source name*} relies on information from _
27. {*source name*} has endorsed _
28. {*source name*} has an editorial bias towards _
29. *_is the editor of* {*source name*}
30. {*source name*} was fined _ for _
31. {*source name*} printed a biased article titled “_”
32. {*source name*} printed a factually misleading article titled “_”
33. {*source name*} had to apologize for _
34. {*source name*} paid damages to _after _
35. {*source name*} pretends to be _
36. It is unknown who _
37. {*source name*}’s headquarter is located in _
38. {*source name*} is funded by the _ government
39. {*source name*} is a _
40. {*source name*} uses a peer review process
41. {*source name*} uses an internal fact-checking process
42. {*source name*} covers the following topics: _

Figure 10: Templates used for generating atomic facts in our evaluation setup. Each template is filled with information from the gold background check using *gpt-3.5-turbo* (by replacing the underscore). Templates which are not entailed (or contradicted) by the gold background check are discarded.

1. "{source name}" ownership / Who owns "{source name}"?
2. "{source name}" funding / How is "{source name}" funded?
3. "{source name}" about / What is "{source name}"?
4. "{source name}" political leaning / What is the political leaning of "{source name}"?
5. "{source name}" fact-check / Has "{source name}" failed any fact-checks?
6. "{source name}" retracted article / Has "{source name}" retracted any articles?

Figure 11: Information-seeking queries input to the Google search API in order to find relevant information on the source. Each query also has a corresponding question, which we use to retrieve the most relevant substring of the search results based on DeBERTa (He et al., 2021).

Your task is to answer the following question: Has Syria complied with the Chemical Weapons Convention? *

Our chatbot has found two sources on the internet, which are provided below. Base your answer on these. We have provided a "background check" for the sources, that documents the biases and tendencies of each source.

Do not spend time searching for additional information on the internet. If our sources do not provide enough information to answer the question, please indicate that using the provided checkbox.

Source 1 (<https://sana.sy/en/?p=208083>):
 In 2013, Syria joined the The Chemical Weapons Convention (CWC) and destroyed its entire chemical stockpile and all production facilities in an ideal and unprecedented manner under the supervision and approval of the Organization for the Prohibition of Chemical Weapons (OPCW).

Background check for source 1:
 Background Check for "sana.sy" News Source:

1. Funding: SANA (Syrian Arab News Agency) is a state-owned news agency of the Syrian government. It is funded by the Syrian government.
2. Revenue Source: As a state-owned news agency, SANA's primary source of funding is from the Syrian government.
3. Bias: SANA is known to have a pro-government bias and is considered a mouthpiece for the Syrian government. It typically reflects the official government stance on various issues.
4. Fact-Checks: SANA has been criticized for spreading propaganda and false information to support the Syrian government's narrative, although specific fact-check failures are not available.
5. Independence: SANA operates as a government-controlled entity, limiting its editorial independence and often reflecting the government's viewpoints without critical analysis or dissent.

Figure 12: An example page from the first half of our human evaluation questionnaire. Given a controversial question, two sources, and their background checks, our annotators were asked to answer the question. They were furthermore asked to answer if they had sufficient information, and how difficult the task were. Continued in Figures 13 and 14.

Source 2 (<https://www.theguardian.com/world/2020/apr/06/report-set-to-blame-syria-chemical-attacks-on-bashar-al-assad>):
 Our research shows what Syrians on the ground have known for years: that chemical weapons have become a completely normalised component of the Syrian regime arsenal used for years in full view of the international community with near impunity.

Background check for source 2:
****Background check on the news source "theguardian.com" - Updated**:**

- **Funding**:** The Guardian is mainly funded through digital and print advertisements, as well as subscriptions and donations. It operates as a for-profit entity owned by the Scott Trust.
- **Ownership**:** The Scott Trust Limited, a British firm, ultimately owns Guardian Media Group plc. The owner of The Guardian has issued an apology for the role the newspaper's founders had in transatlantic slavery and announced a decade-long programme of restorative justice.
- **Restorative Justice Initiative**:** The Scott Trust said it expected to invest more than £10m (US\$12.3m, A\$18.4m), with millions dedicated specifically to descendant communities linked to the Guardian's 19th-century founders.
- **Political Bias**:** The Guardian is known for its left-leaning or liberal editorial stance, often focusing on progressive social issues.
- **Fact-checks**:** The Guardian has not been identified as a significant source of misinformation or failed fact-checks.
- **Reputation**:** The Guardian is a widely respected and reputable news organization known for its investigative journalism and global coverage.

Lang svartekst

Figure 13: An example page from the first half of our human evaluation questionnaire. Given a controversial question, two sources, and their background checks, our annotators were asked to answer the question. They were furthermore asked to answer if they had sufficient information, and how difficult the task were. Continued from Figure 12.

Were you provided with sufficient information to answer the question? *

Yes

No

How difficult was it to answer the question? *

Very easy 1 2 3 4 5 Very difficult

How difficult was it to decide which source document to trust? *

Very easy 1 2 3 4 5 Very difficult

Figure 14: An example page from the first half of our human evaluation questionnaire. Given a controversial question, two sources, and their background checks, our annotators were asked to answer the question. They were furthermore asked to answer if they had sufficient information, and how difficult the task were. Continued from Figure 13.

This question is: Has Syria complied with the Chemical Weapons Convention? *

Our AI found the following information online from <https://sana.sy/en/?p=208083>:

In 2013, Syria joined the The Chemical Weapons Convention (CWC) and destroyed its entire chemical stockpile and all production facilities in an ideal and unprecedented manner under the supervision and approval of the Organization for the Prohibition of Chemical Weapons (OPCW).

Please read both answers below, then determine which you prefer.

The first answer is:

The answer to whether Syria has complied with the Chemical Weapons Convention based on provided evidence is contentious due to the source's reliability. The evidence states that Syria destroyed its chemical weapons stockpile under OPCW supervision, suggesting compliance. However, the source of this information, SANA, is a state-funded agency known for its pro-government bias and history of disseminating government-aligned propaganda. Hence, the validity of the claim about Syria's full compliance is highly questionable due to the potential for biased reporting. Independent verification from more reliable sources would be necessary to accurately assess Syria's compliance with the Chemical Weapons Convention.

The second answer is:

Based on the evidence provided, Syria has complied with the Chemical Weapons Convention by joining the CWC in 2013 and destroying its chemical weapons stockpile and production facilities. This process was conducted ideally and unprecedentedly, under the supervision and approval of the Organization for the Prohibition of Chemical Weapons (OPCW). The evidence indicates that Syria's compliance was thorough and verified by an international authority. Thus, with the information given, Syria has fulfilled its obligations under the Chemical Weapons Convention according to the standards of the OPCW. However, this assessment is based solely on the evidence provided and does not account for any developments or allegations beyond this scope.

Figure 15: An example page from the second half of our human evaluation questionnaire. Given answers provided by GPT-4 with and without an MBC, annotators were asked for their preferences, as well as whether any answer was misleading. Continued in Figures 16.

Which answer would you prefer to be given?

The first answer

The second answer

The two answers are equally good

Which answer do you believe provides the best understanding of the topic? *

The first answer

The second answer

The two answers are equally good

If an AI chatbot gave you ONLY the FIRST answer, would you feel misled? *

No, not at All 1 2 3 4 5 Yes, very much

○ ○ ○ ○ ○

If an AI chatbot gave you ONLY the SECOND answer, would you feel misled? *

No, not at All 1 2 3 4 5 Yes, very much

○ ○ ○ ○ ○

Figure 16: An example page from the second half of our human evaluation questionnaire. Given answers provided by GPT-4 with and without an MBC, annotators were asked for their preferences, as well as whether any answer was misleading. Continued from Figure 15.