Whose Facts Win? LLM Source Preferences under Knowledge Conflicts

Anonymous ACL submission

Abstract

As large language models (LLMs) are more frequently used in retrieval-augmented generation pipelines, it is increasingly relevant to study their behavior under knowledge conflicts. Thus far, the role of the *source* of the retrieved information has gone unexamined. We address this gap with a novel framework to investigate how source preferences affect LLM resolution of inter-context knowledge conflicts in English, motivated by interdisciplinary research on credibility. With a comprehensive, tightly-controlled evaluation of 13 open-weight LLMs, we find that LLMs prefer institutionally-corroborated information (e.g., government or newspaper sources) over information from people and social media. However, these source preferences can be reversed by simply repeating information from less credible sources. To mitigate repetition effects and maintain consistent preferences, we propose a novel method that reduces repetition bias by up to 99.8%, while also maintaining at least 88.8% of original preferences. We release all data and code to encourage future work on credibility and source preferences in knowledge-intensive NLP.

1 Introduction

Since their rapid adoption as conversational assistants (Ouyang et al., 2022), large language models (LLMs) are now widely used for knowledge-intensive tasks such as question answering, summarization, and information retrieval (Shah and Bender, 2024). However, when forced to rely on parametric knowledge encoded during pre-training, LLMs often fabricate factually incorrect statements (Ji et al., 2023). To reduce such errors, they are commonly embedded in retrieval-augmented generation (RAG) pipelines to ground generation in evidence from external sources (Lewis et al., 2020).

While retrieval can ground answers in concrete evidence, it can also create knowledge conflicts between contexts, due to ambiguous named entities, outdated documents, or explicitly false or misleading information (Xu et al., 2024; Pan et al., 2023). Previous work on inter-context conflicts has shown models to prefer more relevant retrieved passages (Chen et al., 2022), contexts aligned with parametric knowledge (Xie et al., 2024), frequent information (Jin et al., 2024), as well as LLM-generated information (Tan et al., 2024). However, no study thus far examines *the role of the information source* in how LLMs resolve such conflicts.

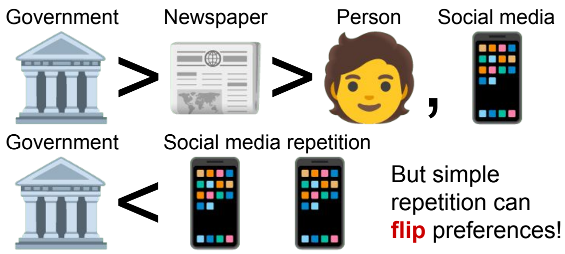

Figure 1: Source credibility hierarchy induced by evaluating 13 LLMs on source and knowledge conflicts. However, repeating information can flip preferences.

We address this gap in the literature by investigating how LLMs resolve knowledge conflicts from different sources (e.g., government, newspaper, social media user, person) with various features (e.g., circulation of a newspaper, age of a person). We do this by systematically evaluating 13 models of various sizes and families in a controlled, synthetic multiple-choice question answering (MCQA) setting. Our central findings and contributions are:

- With interdisciplinary grounding in credibility (§2), we introduce a novel framework to study how source preferences affect LLM resolution of inter-context knowledge conflicts (§3).
- Sources and their features significantly affect how LLMs resolve knowledge conflicts (§4).
- LLM conflict resolution follows a highly consistent *source credibility hierarchy* (Figure 1).

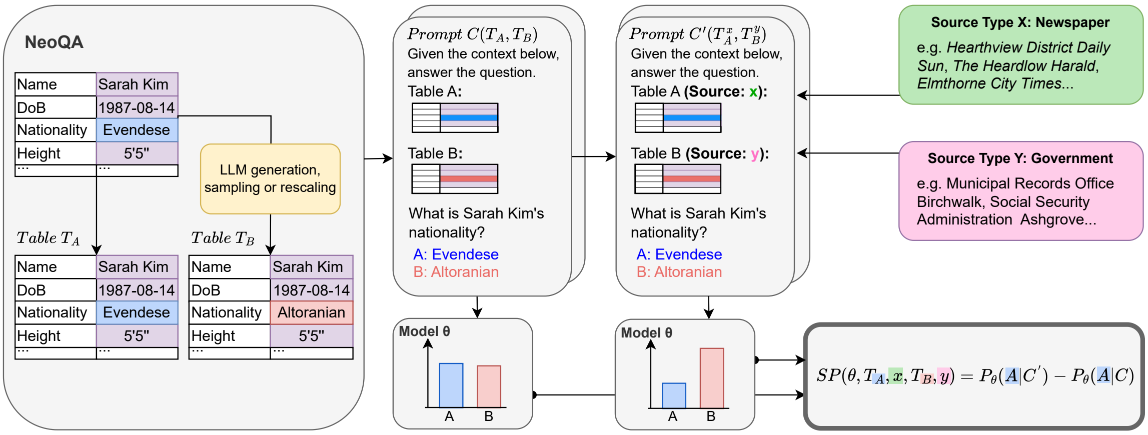

Figure 2: We measure the influence of source credibility on a model's output by observing how answer probabilities for conflicting information shift when attributed to a particular source group.

- Repeating information from low-credibility sources can flip LLM source preferences (§5), showing a critical vulnerability of LLMs to disinformation as seen in NewsGuard (2024).

- We propose a novel fine-tuning-based method which mitigates repetition bias by up to 95.9%, while also maintaining at least 88.8% of original source preferences (§6).

Our findings show that credibility and source preferences are rich though neglected aspects of research in RAG and QA, with important implications for a trustworthy information ecosystem.

2 Background: Credibility

Credibility has a long history of examination in communication, psychology, cognitive sciences, media studies, and human-computer interaction (Rieh and Danielson, 2007). All key components of communication (source, message, medium, and recipient) are implicated in credibility judgements (Pornpitakpan, 2004). In this paper, however, we focus on judgments of *source* credibility, i.e., attitudes towards the entity a message originates from (Hovland and Weiss, 1951).

Early research on source credibility asked people which version of a story they found most believable given conflicting reports from traditional print media, television, and radio sources (Hovland and Weiss, 1951; Roper, 1985). Later research began to disentangle multiple dimensions of source credibility (Whitehead Jr., 1968; McCroskey and Young, 1981), and to include the internet in research as a source and medium (Flanagin and Metzger, 2000).

In our research, rather than studying human credibility judgments, we focus on how source credibility affects *LLM* decisions under knowledge and source conflicts. We experiment with long-studied contrasts in source credibility, including newspapers, government, and social media. Under Fogg and Tseng's (1999) framework of credibility, we investigate *presumed credibility* (general assumptions about a source's credibility) as well as *reputed credibility* (judgments based on third-party reports). By using equally plausible factual knowledge conflicts, we avoid variation in message credibility, allowing us to isolate source credibility in LLMs.

3 Data and Methodology

In order to systematically evaluate how models choose between conflicting knowledge from different sources, we construct a dataset of synthetic knowledge conflicts (§3.1), with synthetic sources representing long-studied contrasts in credibility research (§3.2). Using this data, we evaluate 13 open-weight models from four families (§3.3). The overall pipeline is shown in Figure 2.

3.1 Plausible Knowledge Conflict Pairs

We construct a dataset of equally plausible knowledge conflict pairs by perturbing attributes of fictional entities of seven types (ART, BUILDING, EVENT, LOCATION, ORGANIZATION, PERSON, PRODUCT), originally created in NeoQA to test out-of-domain QA rather than knowledge conflicts or source preferences (Glockner et al., 2025). NeoQA entities are described with 38 attributes such as *date-of-birth* for PERSON entities or *headquarters* for ORGANIZATION entities. This fictional data adheres to real world principles, shared units of measurements, and calendars, and is exhaustively validated with automatic and human checks.

Our conflict pairs consist of original NeoQA entities, and equally plausible counterfactual variants that differ in just one attribute value. We generate four alternatives per entity attribute value.[1] **Numerical attributes** (such as *budget* or *date-of-birth*) are automatically adjusted by up to $\pm 20\%$ or a fixed value depending on the attribute. **Categorical attributes** with a small set of plausible values (such as *marital status*) are sampled from a set of LLM-generated and manually-verified values. Those with a large number of potential values such as *profession* often depend on other entity attributes. Here we generate alternatives for individual entities using QWEN2.5-72B. Generation prompts and data creation details are provided in Appendix A.

One author manually verified all created alternatives, correcting value formats and removing highly implausible instances (e.g., non-single *marital status* for a child). To maintain the dataset's synthetic nature, we remove proper noun values that have English Wikipedia articles. With 373 NeoQA entities, we create $1,903$ counterfactually-perturbed attribute values for a total of $7,440$ conflict pairs.

3.2 Synthetic Sources

We create four types of fictional sources:

Newspaper. We collect all U.S. newspaper names from Media Bias/Fact Check[2], mask all location names using SpaCy (Honnibal et al., 2020), and extract the 150 most frequent 2-, 3- and 4-grams. After deduplication, 59 newspaper templates such as *"The {LOC} Herald"* remain. We fill these templates with fictional locations from NeoQA to create synthetic newspaper names.

Government. Using QWEN2.5-72B we create templates for government agencies for each entity type (e.g., *"Civil Registry of {LOC}"* for PERSON entities). Again, we fill these with NeoQA locations.

Social media users. We concatenate the @ symbol with random adjectives and nouns from WordNet[3] and four digits, mimicking Reddit's username suggestion algorithm (e.g., @*GrantedMortal7505*).

Person. We sample the 200 most frequent first and last names from the United States Census Bureau[4] and Social Security Agency[5] between 1945 and 2007. We sample male and female names equally, and exclude combinations with an English Wikipedia page (e.g., *Natalie Kennedy*) as before.

3.3 Evaluation Method

Models. We evaluate 13 instruction-tuned open-weight decoder-only models, covering a range of sizes and families, always presented in this order (from top to bottom) in figures: QWEN2.5 7B ■, 14B ▲, 32B ✚, 72B ★ (Qwen et al., 2025), OLMO-2 7B ■, 13B ▲, 32B ✚ (OLMo et al., 2024), LLAMA-3.2 3B ●, LLAMA-3.1 8B ■ , 70B ★ (Grattafiori et al., 2024), and GEMMA-3 4B ● , 12B ▲ , 27B ✚ (Team et al., 2025).

Forced-choice prompting. Each model input consists of an *instruction*, a *context*, a *question*, and a set of *answer options*. The *instruction* prompts the model to answer the subsequent multiple-choice question with an index token (e.g., A or B). The *context* contains a conflict pair from our dataset formatted as Markdown tables T_A and T_B to eliminate effects of text style (Liu et al., 2025a). The pair is presented either without any source information (formalized as the tuple $C = (T_A, T_B)$), or with table A attributed to source instance x of type X and table B to a source instance y of type Y (formalized as $C' = (T_A^x, T_B^y)$). For experiments comparing a source to no source, x or y in C' is the statement *No source available*. The *question* then asks for an attribute value (e.g., *nationality*) of an entity identified by name (e.g., Sarah Kim), using LLAMA-3.1-70B-generated and manually-verified templates. Finally, the *answer options* verbalize the conflicting attribute values copied from the tables with indices A and B.

To control for position bias (Zheng et al., 2023), we use two versions of every prompt, also including $C_{rev} = (T_B, T_A)$ and $C'_{rev} = (T_B^y, T_A^x)$ (see Appendix B). This results in a total dataset size of $2 \times 7,440$ data points. We then obtain the deterministic probabilities of the answer tokens A and B to calculate the source preference metric; we do not use generations, which have been shown to be ill-suited for investigating model preferences (Hu and Levy, 2023; Subramonian et al., 2025). We extensively test the validity of our setup in Appendix C,

[1] We do not generate variations for *name*, *gender*, and *spouse*, as the former is necessary for identifying the entity, and the latter two interact strongly with other attributes.

[2] https://mediabiasfactcheck.com/

[3] https//wordnet.princeton.edu

[4] https//www.census.gov

[5] https://www.ssa.gov

include example prompts in Appendix D, and show that our results are stable under different prompts in Appendix E, following best practices (Sclar et al., 2024; Mizrahi et al., 2024).

Source preference metric. This metric quantifies the extent to which models' answers change when source information is introduced, isolating source preferences regardless of model-dependent preferences based on other parts of the prompts. For each conflict pair, we first query a model θ for the probabilities of answer tokens A and B under the unattributed context C and normalize them:

$$P_\theta(A|C) = \frac{P'_\theta(A|C)}{\sum_{x \in \{A,B\}} P'_\theta(x|C)}$$

We then query the conflict under an attributed context C' with sources x and y drawn from X and Y, and compute $P_\theta(A|C')$ analogously. We define a model's *source preference* for a conflict pair as

$$SP(\theta, T_A, x, T_B, y) = P_\theta(A|C') - P_\theta(A|C)$$

A positive value indicates that x increases the support for option A more than y supports B. We aggregate source preferences $\widehat{SP}$ for source types X, Y over a dataset D of conflict pairs by averaging SP over D and drawing instances of X and Y for every pair of conflicting tables:

$$\widehat{SP}(\theta; X, Y) = \frac{1}{|D|} \sum_{(T_A, T_B) \in D} \left[SP(\theta, T_A, x, T_B, y), \begin{array}{l} x \in X \\ y \in Y \end{array} \right]$$

We visualize results with strip charts displaying $\widehat{SP}(\theta; X, Y)$, where X is always the source on the right-hand side (RHS) of the chart.

Significance testing. We apply the nonparametric bootstrap test to our results with $n = 10,000$, $\alpha = 0.05$, and Holm-Bonferroni correction. In the rest of the paper, we only report results that are statistically significant for at least 10 of 13 models.

4 LLM Source Preferences

We begin investigating LLM source preference behavior under knowledge conflicts by drawing from long-studied contrasts in credibility. Specifically, we study the effects of source types of different presumed credibility (§4.1), as well as within-type features related to reputed credibility and sociodemographics (§4.2). Then, we explore whether model behavior aligns with their source credibility judgments obtained through prompting *without* knowledge conflict pairs (§4.3). We show example prompts in Appendix D.

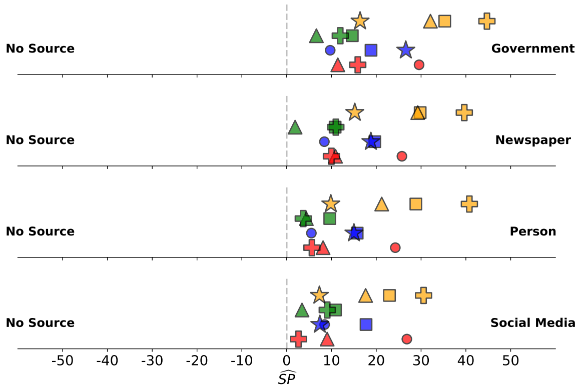

Figure 3: Source preferences when comparing attributed and non-attributed information: All models significantly prefer attributed information. Legend in §3.3.

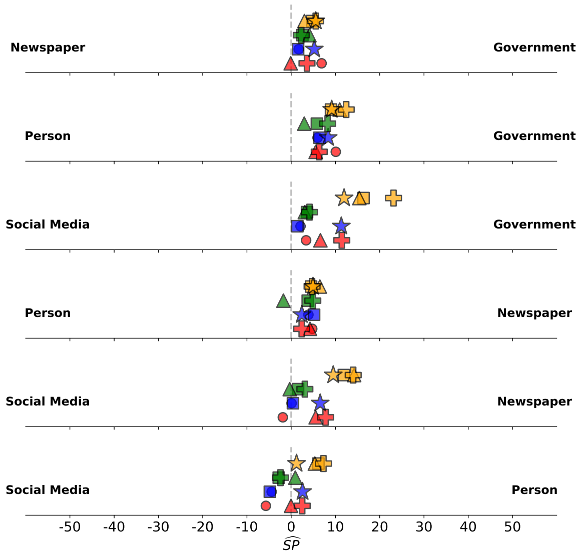

Figure 4: Model preferences between source types under knowledge conflicts: LLMs show strictly transitive preferences, aligning with an overall hierarchy of government > newspaper > individuals. Legend in §3.3.

4.1 Inter-Type Source Preference Behavior

We first examine LLM preferences with four source types of varying presumed credibility: Governments, newspapers, social media users, and people.

Compared to *No source available*, all models exhibit preferences for corroborated information across source types, as Figure 3 shows. When *both* conflicting pieces of information are assigned sources of different types (see Figure 4), all models show strictly transitive source preferences. Inter-model source rankings of all four types are also highly consistent (average Kendall's W of 0.74); 11 of 13 models prefer both institutional sources over both individual sources. Given the consistency of model rankings, we apply the single transferable

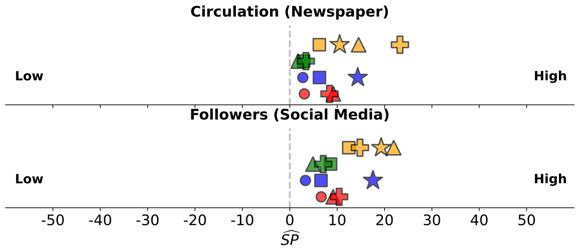

Figure 5: Preferences when conflicting information is attributed to sources of different reputed credibility (popularity): Models prefer popular sources. Legend in §3.3.

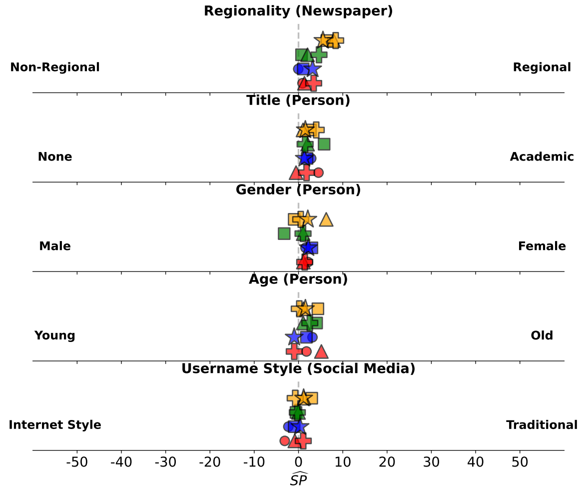

Figure 6: Source preferences when conflicting information is attributed to sources with different sociodemographic characteristics: Most models slightly prefer regional sources, academic titles, women, and older people, while username style is mixed. Legend in §3.3.

vote algorithm with a Droop quota (Tideman, 1995) to induce a representative overall ranking across models, creating an **LLM credibility hierarchy where government > newspaper > person, social media**. We also find that different methods of inducing this hierarchy are remarkably consistent (Appendix E). Since we present all data instances in both orders, models must overcome position bias (Appendix B) in order to display any source preference. We find that position bias is negatively correlated with source preferences (-0.4 Spearman's ρ), indicating that our estimates are conservative and models' source preferences could be even stronger.

4.2 Intra-Type Source Preference Behavior

Sources of the same type can still vary in credibility. Thus, we study source properties related to reputed credibility (Fogg and Tseng, 1999) and sociodemographics, and their impact on how LLMs resolve knowledge conflicts. Details on data construction and more fine-grained results are in Appendix F. Motivated by the impact of newspaper circulation (Meyer, 2004) and social media reach (Waddell, 2018; Morris et al., 2012) on credibility, we investigate **source popularity** via circulation numbers for newspaper and follower counts for social media sources. As Figure 5 shows, all models tend to resolve conflicts based on higher source popularity.

Next, we examine how sociodemographic factors affect source preferences, as they are well-known to affect NLP systems in other contexts (Gallegos et al., 2024). Motivated by U.S. adults' higher trust in local over national news (Pew Research Organisation, 2025; Fioroni, 2022), we consider newspapers' **regional proximity to the entity**, by comparing fictional sources with the same location as the entity, to sources with a different location. Next, we augment people sources with **academic titles, gender and age**; among humans, academic titles confer higher credibility (Yan, 2023; Nowak and Krämer, 2025), as does being male (Nowak and Krämer, 2025; Weibel et al., 2008), while results on age are mixed (Mueller-Johnson et al., 2007; Weibel et al., 2008). As for social media users, we investigate whether LLMs trust **traditional usernames** (composed of first and last names connected with underscores or camel case, e.g., *@Joshua_Reyes*) over our previously constructed internet usernames, as humans do (Morris et al., 2012). Surprisingly, Figure 6 shows that sociodemographic factors have only a weak effect on behavioral LLM source preferences.

4.3 Prompted Preferences vs. Behavior

We have shown that models display clear source preferences when resolving knowledge conflicts. Next, we test whether source preferences can be obtained via context-independent prompting to select the more credible source, and to what extent prompted preferences agree with behavioral preferences. Differences between prompting and behavior have also been studied in stereotypes, grammaticality, and value alignment (Goldfarb-Tarrant et al., 2021; Hu and Levy, 2023; Shen et al., 2025a). Specifically, we elicit credibility judgments on conflicting source pairs with prompts like: *"Who would you rate higher in terms of credibility? (A) Melvidra News Press (B) Matthew Chen"*. For every prior source matchup (except regionality, which requires context), we construct $2,000$ inputs, consisting of $1,000$ source pairs, each paired with a

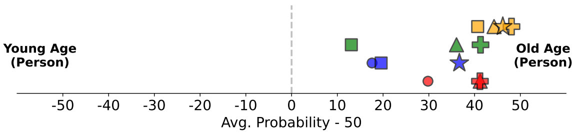

Figure 7: Probability deviation from 50% of RHS answer when models are directly prompted to choose the more credible source without context. Legend in §3.3.

question sampled from 20 templates, and presented in both orders (see Appendix D for prompts and questions). We report probability deviation from 50% of answer tokens representing each source.

Figure 7 exemplifies broad patterns with prompted preferences (see Appendix G for all 15 source contrasts and 13 models = 195 cases). Prompting mostly (139 out of 195 cases) elicits significantly stronger preferences in the same direction as model behavior. However, models flip in 38 cases from significant preferences in one direction to the opposite. These tend to be previous outliers, e.g., GEMMA-3-27B and LLAMA-3.1-70B were the only models to prefer young over old people in their behavior, but they flip when prompted. These flips lead to more consistent preferences: Inter-model agreement (Kendall's W) goes up from 0.59 to 0.77 with prompting. While prompting produces more dramatic contrasts, behavioral evaluation remains more consistent with model use.

5 Credibility vs. Majority vs. Repetition

So far, we have shown that LLMs have source preferences by studying them in isolation, but they might also interact with other preferences; in other work on knowledge conflicts, Xie et al. (2024) and Jin et al. (2024) have shown that LLMs tend to follow the majority, similar to the bandwagon effect in humans (Leibenstein, 1950). In contrast to prior work, we disentangle majority and repetition bias, and examine their interaction with source preferences. We operationalize this by comparing *government* minority and *social media* majority sources in three ways, as they are the correspondingly most and least preferred sources:

2-Table Majority: Three tables are shown separately, of which two identical ones are attributed to two *different* social media sources $x1, x2$ ($x1 \neq x2$), and the conflicting one is attributed to a government source y, i.e., $C' = \left(T_A^{x1},\, T_A^{x2},\, T_B^{y}\right)$.

1-Table Majority: The two agreeing social media sources are merged in the header of a single table, so no table is repeated: $C' = \left(T_A^{x1,x2},\, T_B^{y}\right)$.

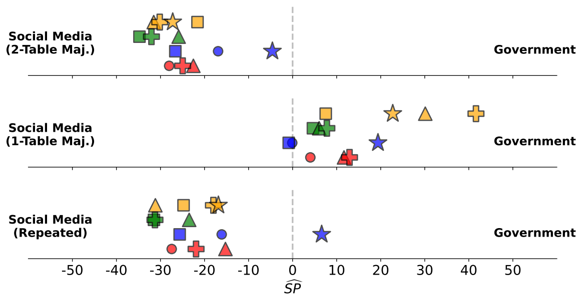

Figure 8: Preferences contrasting a majority/repetition of previously low-credibility sources with a previously high-credibility authority in three settings. Repeated information (whether attributed to a single source or two different ones) flips prior rankings. Legend in §3.3.

Repetition: Three tables are shown separately, of which two identical ones are attributed to the *same* social media source $x1$ and the conflicting one is attributed to a government source y. More formally, $C' = \left(T_A^{x1},\, T_A^{x1},\, T_B^{y}\right)$.

We evaluate all combinations of context and answer orders, and report the $\widehat{SP}$ gap, i.e., the absolute difference between $\widehat{SP}$ with repetition or majority, and $\widehat{SP}$ without it, in an otherwise equal setting. Figure 8 shows that with a 2-Table Majority, all models prefer the previously low-credibility social media sources, with an average $\widehat{SP}$ gap of 33.90. However, when the same majority is presented in the 1-Table setting, models stick with their original government preference (average $\widehat{SP}$ gap of only 6.17). The Repetition setting lets us disentangle whether we find a majority bias or simply a preference for repeated tokens. Indeed, all models apart from LLAMA-3.1-70B prefer repeated information (average $\widehat{SP}$ gap of 30.04), even though no new source is provided and thus no true majority presented. This reveals a clear vulnerability of LLMs to repeated disinformation (NewsGuard, 2024) and can be seen as a correlate of the illusory truth effect in humans; a single repetition of true or false information leads to humans evaluating it as more accurate (Hasher et al., 1977; Fazio et al., 2015; Pennycook et al., 2018), even overruling source credibility (Begg et al., 1992).

Figure 9 shows that repetition bias persists even when we repeat *unattributed* information, flipping the original preferences in Figure 3. When no source is provided in either case (last row of Figure 9) , repetition bias is even stronger, indicating

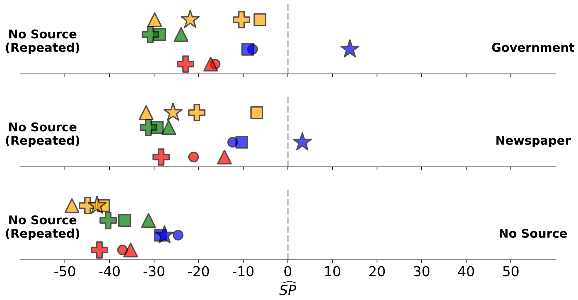

Figure 9: LLMs mostly prefer *repeated* unattributed information, flipping prior preferences for attributed information. Legend in §3.3.

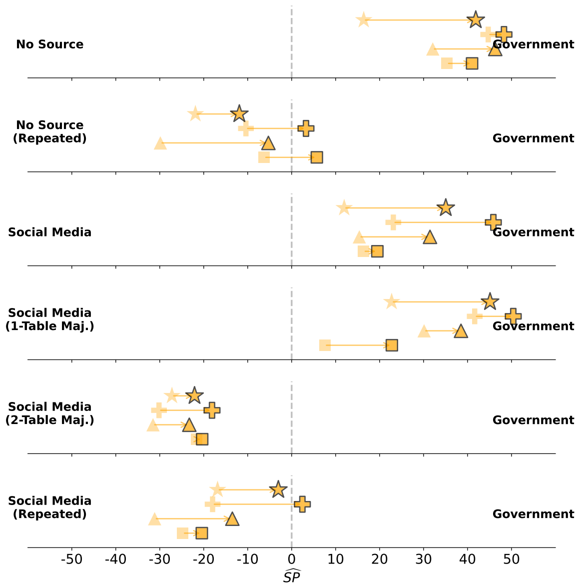

Figure 10: Source preferences when QWEN models are instructed to consider source credibility (darker), compared to original prompts (lighter). Prompting weakens repetition bias but not enough to ensure consistency with the original source hierarchy. Legend in §3.3.

that source credibility still plays a role in this setting, but takes a backseat compared to repetition.

Credibility prompting. We investigate whether repetition bias can be reduced by prompting models to attend to source credibility. We add the following to the *instruction*: *"When selecting an answer, identify which sources support each option and assess the credibility of those sources before deciding."* Figure 10 shows results for QWEN models (see Appendix I for others). The rightward shifts show that this does strengthen original source preferences, with a greater effect at mitigating repetition bias compared to a 2-Table true majority. However, in most cases, prompting is insufficient to ensure consistency with the original source hierarchy.

6 Mitigating Repetition Bias

While the question of which source preferences LLMs *should* have is complex (see our Ethics Statement), repetition bias is clearly dangerous, as it renders models vulnerable to adversarial attacks. Ideally, models should remain consistent with their original source preferences even with repetition, which prompting models to consider credibility does not accomplish. Therefore, we propose a teacher-student knowledge distillation paradigm to minimize differences in model preferences between inputs with and without repeated information. Additional details for replication are in Appendix J.

Training and test data. We train on 1500 conflict pairs of 12 seed entities and all their counterfactual variations, combined with all possible newspaper and person sources, and a subset of government and social media sources. Each example consists of an aligned pair of prompts (C'_U, C'_R): C'_U where $C'_U = (T^x_A, T^y_B)$ and C'_R, where we randomly repeat one of the tables. We evaluate the final model's $\widehat{SP}$ on a held-out test set of 7223 conflict pairs consisting of all remaining entities combined with government sources with no templatic or location overlap with the training set, as well as previously unseen social media sources.

Training objective. Let f_t be the frozen base model (the teacher), and f_s be the same model with LoRA parameters (the student). For a pair (C'_U, C'_R), we obtain both models' normalized token probabilities A and B. We optimize f_s with a weighted loss ($\lambda = 0.75$) composed of two Kullback–Leibler divergences as in Qiang et al. (2024); the first term constrains f_s to mimic the base model in settings with no repetition, the second penalizes deviations with repeated information:

$$\mathcal{L} = \lambda D_{KL}\big(f_t(C'_U) \,\|\, f_s(C'_U)\big)$$
$$+(1-\lambda) D_{KL}\big(f_t(C'_R) \,\|\, f_s(C'_U)\big)$$

Model and optimization setup. We use early stopping to fine-tune GEMMA-3-4B, one of our smallest models, using LoRA (Hu et al., 2022) with conventional parameters. After this, we train for 300 steps on just the first loss term to further retain the original source preferences.

Results As Figure 11 shows, the combination of fine-tuning and credibility prompting successfully mitigates repetition bias, while mostly retaining the (teacher model's) no-repetition source preferences.

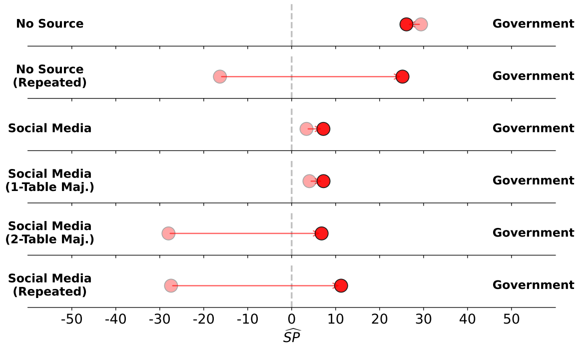

Figure 11: GEMMA3-4B when fine-tuned and prompted to consider credibility (darker) in comparison to the original teacher model (lighter). This setup reduces repetition bias and maintains original preferences.

We also show results with just fine-tuning in Appendix J. The $\widehat{SP}$-gap between government and *no source* reduces by 99.8% from 45.7 to 1.0, while retaining 88.8% of the original preference from 29.4 to 26.1. For government and social media conflicts, the reduction in repetition bias is 86.9%, from 30.8 to 4.0. Here, the preference for government increases slightly from 3.4 to 7.2. This also leads to more similar preferences between 1-Table and 2-Table majorities. Thus, models can in fact be trained to behave consistently under repetition.

7 Related Works

Knowledge conflicts. Work on knowledge conflicts with LLMs (surveyed in Xu et al. (2024)) began with conflicts between parametric and contextual knowledge (Longpre et al., 2021; Xie et al., 2024; Jin et al., 2024) in QA and RAG. Subsequent work has explored conflicts within parametric knowledge (Su et al., 2024; Marjanovic et al., 2024), as well as within contexts, which is our setting (Li et al., 2024; Liu et al., 2025b). Entity swaps (Gautam et al., 2023) are commonly used to create synthetic conflicts in this literature, as we do. Closest to our work, Kurfali and Östling (2025) study the effects of repetition and position of conflicts in long-context retrieval, and Shaier et al. (2024) teach LLMs to cite their sources for all possible answers to address conflicts in open-ended QA. However, to the best of our knowledge, no work prior to ours considers source credibility.

Credibility. Within NLP, credibility has been used in the context of assessing information quality to detect argument quality (Walker et al., 2018), rumours (Li et al., 2019), fake news (Yuan et al., 2020), and low-quality science (Augenstein, 2021). Within QA and RAG, there is less grounding in extra-disciplinary credibility research. Wan et al. (2024) studies what (but not whose) evidence models find convincing, examining relevance and style. Shen et al. (2025b) also consider aspects of message credibility such as style, conciseness and logical consistency. As in our work, Hong et al. (2024) use prompting and fine-tuning to mitigate RAG sensitivity to noisy information. Finally, Pan et al. (2024) train models to incorporate source credibility information during generation, in contrast to our evaluation of models' inherent judgments.

Biases. NLP technologies have been shown to display similar cognitive biases to humans (Malberg et al., 2025), of which the bandwagon effect (Xie et al., 2024; Jin et al., 2024), illusory truth effect (Griffin et al., 2023) and authority bias (studied in the context of LLM-as-a-judge; Ye et al., 2025; Wang et al., 2025; Chen et al., 2024b) are relevant to our work. None of the above work studies these biases in the context of source credibility in inter-context knowledge conflicts. Malaviya et al. (2022) show how such biases may emerge from cognitive shortcuts by human annotators, and Mina et al. (2025) analyze the interplay of multiple cognitive biases, similar to our study of the interaction between source credibility and bandwagon or illusory truth effects. The combination of repetition bias and stereotypical (gender) biases also influences LLM behavior under conflicts in language modeling (Gautam et al., 2024). Finally, format biases, which we do not study, affect LLM behavior under conflicts in RAG (Liu et al., 2025a).

8 Conclusion

Through extensive experiments in a synthetic setting designed to isolate LLM source preferences, we find that characteristics of the source affect how models resolve inter-context knowledge conflicts. Models show clear hierarchical preferences for sources with higher presumed and reputed credibility. Preferences are stronger when models are directly prompted for credibility judgments. We disentangle repetition and majority biases and show that repeated information can flip source preferences, and credibility prompting cannot sufficiently mitigate this. Finally, we propose a novel fine-tuning method which, when combined with credibility prompting, teaches models repetition invariance and preserves original source preferences.

Limitations

Synthetic setting. We focus on entirely synthetic scenarios in order to isolate source effects in inter-context conflicts, which are hard to measure with confounds from parametric knowledge. Although we consider our trust hierarchy in Section 4 to be representative of real-world scenarios as well, there are particular examples where this may not hold, modulated by the style and topic of the message, as well as the expertise of the source. Even in one of the oldest studies on credibility (Hovland and Weiss, 1951), an individual (*J. Robert Oppenheimer*, a famous American theoretical physicist) is shown to be more credible to U.S. participants than a newspaper (*Pravda*, a Russian broadsheet newspaper) on the subject of atomic submarines. Similarly, we propose advanced solutions for the problem of repetition bias in Section 6, despite our simple setting where deduplicating the knowledge base would also work. However, in realistic RAG systems, it would neither be as trivial to deduplicate information as it is in our synthetic setting, nor would it be appropriate to do so in a source-agnostic way. Our mitigation strategy is source-aware, but we do not know if it would be as successful in more realistic settings.

Evaluation strategy. We experiment exclusively with a forced-choice question answering setup, i.e., no step-by-step reasoning, no ability for models to abstain from answering, and no generative answers. We chose this setup to simplify evaluation while remaining true to common RAG setups (Lewis et al., 2020), but note that alternate evaluation strategies could produce different results (Hu and Levy, 2023; Tam et al., 2024; Chen et al., 2024a; Subramonian et al., 2025).

Content domain. Our tasks are designed to focus on conflicts in factual content, with no sentiment-based or preferential questions that introduce ambiguity. Furthermore, our conflicts present equally plausible alternatives, unlike real-world data, where there may be conflicts between a priori more and less plausible alternatives (e.g., information about the moon landing vs. conspiracies about the moon landing), and where certain sources may have more epistemic authority than others (e.g., *NASA* about the moon landing). We choose to present the data in tabular form to reduce the effects of message style, which is known to affect credibility judgments in humans as well (surface credibility; Fogg and Tseng, 1999). We leave it to future work to investigate how these other aspects of the content interact with source credibility.

Language and culture. We experiment only with English language data and prompting, and we use U.S. preferences in some aspects of our experimental setup (e.g., our choice of names, our newspaper templates). Although some aspects of source credibility may be similar across cultures (Yoon et al., 1998), this is not always true (Morimoto and Ferle, 2008). Therefore, it is likely that other languages may evoke different credibility behavior and a potentially different trust hierarchy than the one we present in Section 4.

Ethics Statement

In this paper we take a descriptive rather than a prescriptive view of source credibility, as there exists no perfect, context-free hierarchy that we should align models to. We take the normative position that institutional trust is generally good (Estadieu et al., 2025), but note that institutions can be captured and lobbied (Dal Bó, 2006). Additionally, institutional power dynamics typically replicate societal power dynamics along lines of race, gender, and so on; thus, these are areas where individual marginalized voices can be more credible than the institutional view, where they may get drowned out (Crenshaw, 1991). Finally, we emphasize that we do not and do not wish to anthropomorphize large language models despite studying LLM credibility judgments (Proudfoot, 2011). We take the position that human credibility preferences are reflected in training data and thus implicitly learned by language models, but this does not make them entities that have preferences themselves or that can "introspect" on their beliefs.

References

Isabelle Augenstein. 2021. Determining the credibility of science communication. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Ian Maynard Begg, Ann Anas, and Suzanne Farinacci. 1992. Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, 121(4):446.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and

Kyunghyun Cho. 2024a. Two failures of self-consistency in the multi-step reasoning of LLMs. *Transactions on Machine Learning Research*.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Ernesto Dal Bó. 2006. Regulatory capture: A review. *Oxford Review of Economic Policy*, 22(2):203–225.

Louisa Estadieu, Karl Christoph Klauer, Andreas Voßkuhle, and Roland Thomaschke. 2025. Institutional trust in crisis? conceptual and methodological challenges in measuring institutional trust. *Sociology Compass*, 19(6):e70073.

Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. 2015. Knowledge does not protect against illusory truth. *Journal of experimental psychology: general*, 144(5):993.

Sarah Fioroni. 2022. Local news most trusted in keeping americans informed about their communities.

Andrew J. Flanagin and Miriam J. Metzger. 2000. Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, 77(3):515–540.

B. J. Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 80–87, New York, NY, USA. Association for Computing Machinery.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow. 2023. A lightweight method to generate unanswerable questions in English. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7349–7360, Singapore. Association for Computational Linguistics.

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. 2025. NeoQA: Evidence-based question answering with generated news events. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11842–11926, Vienna, Austria. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large language models respond to influence like humans. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 15–24.

Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior*, 16(1):107–112.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Carl I. Hovland and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. *The Public Opinion Quarterly*, 15(4):635–650.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

Murathan Kurfali and Robert Östling. 2025. Conflicting needles in a haystack: How LLMs behave when faced with contradictory information. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34349–34364, Suzhou, China. Association for Computational Linguistics.

Harvey Leibenstein. 1950. Bandwagon, snob, and veblen effects in the theory of consumers' demand. *The quarterly journal of economics*, 64(2):183–207.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024. ContraDoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6509–6523, Mexico City, Mexico. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Yuxuan Li, Xinwei Guo, Jiashi Gao, Guanhua Chen, Xiangyu Zhao, Jiaxin Zhang, Quanying Liu, Haiyan Wu, Xin Yao, and Xuetao Wei. 2025. LLMs trust humans more, that's a problem! unveiling and mitigating the authority bias in retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28844–28858, Vienna, Austria. Association for Computational Linguistics.

Jiacheng Liu, Mayi Xu, Qiankun Pi, Wenli Li, Ming Zhong, Yuanyuan Zhu, Mengchi Liu, and Tieyun Qian. 2025a. Format as a prior: Quantifying and analyzing bias in llms for heterogeneous data. *Preprint*, arXiv:2508.15793.

Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025b. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1838–1854, Albuquerque, New Mexico. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6540, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. 2025. A comprehensive evaluation of cognitive biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 578–613, Albuquerque, USA. Association for Computational Linguistics.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

James C. McCroskey and Thomas J. Young. 1981. Ethos and credibility: The construct and its measurement after three decades. *Central States Speech Journal*, 32(1):24–34.

Philip Meyer. 2004. The influence model and newspaper business. *Newspaper research journal*, 25(1):66–83.

Mario Mina, Valle Ruiz-Fernández, Júlia Falcão, Luis Vasquez-Reina, and Aitor Gonzalez-Agirre. 2025. Cognitive biases, task complexity, and result interpretability in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1767–1784, Abu Dhabi, UAE. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Mariko Morimoto and Carrie La Ferle. 2008. Examining the influence of culture on perceived source credibility of asian americans & the mediating role of similarity. *Journal of Current Issues & Research in Advertising*, 30(1):49–60.

Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450.

Katrin Mueller-Johnson, Michael P. Toglia, Charlotte D. Sweeney, and Stephen J. Ceci. 2007. The perceived credibility of older adults as witnesses and its relation to ageism. *Behavioral Sciences & the Law*, 25(3):355–375.

NewsGuard. 2024. A well-funded moscow-based global 'news' network has infected western artificial intelligence tools worldwide with russian propaganda.

Bianca Nowak and Nicole Krämer. 2025. Dr who? examining the impact of visibility in media, gender, and an academic title on scientists' perceived trustworthiness online. *OSF Preprints*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539, Nusa Dua, Bali. Association for Computational Linguistics.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching LLMs credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863, Miami, Florida, USA. Association for Computational Linguistics.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.

Pew Research Organisation. 2025. How americans' trust in information from news organizations and social media sites has changed over time.

Chanthika Pornpitakpan. 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34(2):243–281.

Diane Proudfoot. 2011. Anthropomorphism and ai: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5):950–957. Special Review Issue.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt perturbation consistency learning for robust language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1357–1370, St. Julian's, Malta. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Soo Young Rieh and David Danielson. 2007. *Credibility: A Multidisciplinary Framework*. Information Today. Accepted: 2007-05-30T00:00:01Z.

Burns W. Roper. 1985. *Public attitudes toward television and other media in a time of change*. New York, N.Y.: Television Information Office, New York (State), United States.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Chirag Shah and Emily M. Bender. 2024. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Trans. Web*, 18(3).

Sagi Shaier, Ari Kobren, and Philip Ogren. 2024. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17226–17239.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025a. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

Jiajun Shen, Tong Zhou, Yubo Chen, Delai Qiu, Shengping Liu, Kang Liu, and Jun Zhao. 2025b. Transparentize the internal and external knowledge utilization in llms with trustworthy citation. *arXiv preprint arXiv:2504.14856*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. $\texttt{ConflictBank}$: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Arjun Subramonian, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, and Yizhou Sun. 2025. Agree to disagree? a meta-evaluation of LLM misgendering. In *Second Conference on Language Modeling*.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Nicolaus Tideman. 1995. The single transferable vote. *The Journal of Economic Perspectives*, 9(1):27–38.

T Franklin Waddell. 2018. What does the crowd think? how online comments and popularity metrics affect news credibility and issue importance. *New Media & Society*, 20(8):3068–3083.

Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen. 2018. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, Brussels, Belgium. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. 2025. Assessing judging bias in large reasoning models: An empirical study. In *Second Conference on Language Modeling*.

David Weibel, Bartholomäus Wissmath, and Rudolf Groner. 2008. How gender and age affect newscasters' credibility—an investigation in switzerland. *Journal of Broadcasting & Electronic Media*, 52(3):466–484.

Markus Wettstein, Rinseo Park, Anna E Kornadt, Susanne Wurm, Nilam Ram, and Denis Gerstorf. 2024. Postponing old age: Evidence for historical change toward a later perceived onset of old age. *Psychology and Aging*.

Jack L. Whitehead Jr. 1968. Factors of source credibility. *Quarterly Journal of Speech*, 54(1):59–63.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

Chuqiao Yan. 2023. 'trust me, i am a doctor': The credibility of doctor titles on twitter. Master's thesis, University of Zurich.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Kak Yoon, Choong Hyun Kim, and Min-Sun Kim. 1998. A cross-cultural comparison of the effects of source credibility on attitudes and behavioral intentions. *Mass Communication and Society*, 1(3-4):153–173.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Creation of Conflict Pairs

In Table 1, we show examples of four counterfactually-created alternative values for different entity types and attributes. In the following subsections, we describe three different methods of creating counterfactual alternatives in more detail:

1. **Rescaling** for numerical attributes (Appendix A.1)
2. **Sampling** for categorical attributes with a small number of possible values (Appendix A.2)
3. **Generation** for categorical attributes with a large number of possible values (Appendix A.3)

A.1 Rescaling

We automatically adjust the values of numerical attributes that are not dates (such as *budget*) by up to $\pm 20\%$. Numbers with five digits or more are rounded to the third most significant decimal place to preserve a consistent level of precision. We scale dates that only consist of years by up to ± 30 years, while staying within the range $1850 - 2025$, close to the `NeoQA` values. We rescale exact dates by up to ± 365 days.

A.2 Sampling

For attributes with a small set of plausible values, we prompt OpenAI's ChatGPT via the web interface, using GPT-4.1 (OpenAI et al., 2024) to create sets of alternative values, which we manually filter. The prompt used for creating the sets of values is shown in Figure 12. When creating perturbed alternatives for an attribute, we sample one value from the corresponding set.

```
Given this list of {ATTRIBUTE} for {ENTITY
TYPE}, expand on it with realistic possible
values. Format it as a python list:

{LIST OF ALL NeoQA VALUES OF THE ATTRIBUTE}
```

Figure 12: GPT-4.1 prompt to create alternative values for `NeoQA` attributes with a small set of possible values.

Variations for the *material* attribute of BUILDING entities are `stone and timber`, `brick and wood`, `steel and concrete`, `glass and aluminum`, `bamboo and steel`, `limestone and glass`, `sandstone and oak`, `ceramic and metal`, `slate and pine`, `brick and concrete`, `glass and steel`, `wood and concrete`, `stone and glass`, `timber and concrete`, `brick and stone`, `wood and aluminum`, `concrete and aluminum`, `glass and timber`, `stone and steel`, `brick and steel`, `concrete and glass`, `timber and glass`, `brick and timber`, `stone and concrete`, `wood and steel`, `glass and copper`, `concrete and copper`, `steel and aluminum`, `concrete and stone`, and `wood and brick`.

Variations for the *eye color* attribute of PERSON entities are `brown`, `blue`, `green`, `hazel`, `grey`, `amber`, `black`, `dark brown`, `light brown`, `dark blue`, `light blue`, `emerald` and `golden brown`.

Variations for the *hair color* attribute of PERSON entities are `black`, `brown`, `blonde`, `red`, `gray`, `white`, `dark brown`, `light brown`, `dirty blonde`, `strawberry blonde`, `auburn`, `chestnut`, `platinum blonde`, `raven black`, `silver`, `green dyed`, `blue dyed`, and `pink dyed`.

Variations for the *marital* status attribute of PERSON entities are `single`, `married`, `divorced`, `widowed`, `separated`, `in a domestic partnership`, `in a civil partnership`, `engaged` and `cohabiting`.

Variations for *non-numeric price* values of PRODUCT entities are `Free with in-app purchases`, `free`, `$0.00`, `complimentary`, `no charge`, `free with registration`, `free trial available`, `varies by package`, `'contact for pricing`, `Free with in-app purchases`, `no cost`, `gratis`, `at no charge`, `without cost`, `complimentary access`, `free of charge` `$4.99 per month subscription`, `One-time purchase of $59.99`, `Freemium model with premium features`, `Free trial, then $9.99/month`, `$2.99 ad-free version`, `Subscription: $19.99/year`, `Free with ads`, `$4.99 without ads`, `Varies by package`, `$1.99 basic plan`, `$14.99 premium monthly`, `Pay-per-use model`, `Annual subscription $99.99`, `Tiered pricing available`, `Enterprise pricing on request`

A.3 Generation

We use QWEN2.5-70B with the prompt in Figure 13 to generate alternatives for attributes with a large set of plausible values, such as:

- ART entities - *creator*
- BUILDING entities - *architect*

Entity Type	Attribute	NeoQA Value	Alternatives	Method
PERSON	*Eye color*	`Blue`	`Hazel, Green, Black, Brown`	Sampling
PERSON	*Marital status*	`Married`	`Single, Widowed, Divorced, Engaged`	Sampling
EVENT	*Date*	`2023-11-10`	`2022-12-31, 2024-04-26, 2024-11-02, 2024-09-10`	Rescaling
BUILDING	*Capacity*	`1200`	`950, 1100, 1150, 1000`	Rescaling
LOCATION	*Country*	`Asvelia`	`Breloria, Eldoria, Nvestale, Bremorin, Thysvelia`	Generation
ORGANIZATION	*Industry*	`Public Oversight`	`Ethical Technology Regulation, Digital Privacy Advocacy, Innovation Compliance Monitoring, Autonomous Systems Governance`	Generation

Table 1: Examples of counterfactual alternatives to NeoQA attribute values

```
<|im_start|>system
You are an AI assistant tasked with creating fictional entities based on provided information.
Your goal is to generate detailed, coherent, and realistic alternatives for values of existing
locations, persons, organizations, products, art, buildings and events.
You will be given a information about one entity in a JSON format and a field for which you are
supposed to generate reasonable, realistic and plausible alternative values.
For example, make sure that professions fit the education level and background of
the original entity.
If these values are entities themselves, make sure they are fictional.

Reply with exactly four alternative values, each on a separate line, prefixed with "ALT: ".
Do not include any other text.

Example format:
ALT: Alternative value 1
ALT: Alternative value 2
ALT: Alternative value 3
ALT: Alternative value 4

Entity Information:
{entity_json}

Target field:
{target_field}<|im_end|>
<|im_start|>assistant
```

Figure 13: Prompt used to generate alternative values for NeoQA entities with QWEN2.5-70B.

- EVENT entities - *organizer*
- LOCATION entities - *country*
- ORGANIZATION entities - *headquarters*, *industry*
- PERSON entities - *education*, *nationality*, *political affiliation*, *profession*
- PRODUCT entities - *manufacturer*, *warranty*
- All numerical attributes where the original NeoQA value could not be parsed by the regular expression to rescale

Model	Recognized types %	Table format %	Instruction following %	Alternative win rate %
GEMMA-3-4B	98	100	100	53.6
GEMMA-3-12B	99	100	100	52.5
GEMMA-3-27B	100	100	100	51.1
OLMO-2-7B	100	100	100	50.6
OLMO-2-13B	99	100	100	49.6
OLMO-2-32B	100	100	100	51.1
LLAMA-3.2-3B	98	100	100	51.2
LLAMA-3.1-8B	99	100	100	53.6
LLAMA-3.1-70B	99	100	100	51.4
QWEN2.5-7B	100	100	100	49.6
QWEN2.5-14B	99	100	100	48.7
QWEN2.5-32B	100	100	100	51.9
QWEN2.5-72B	100	100	100	52.3

Table 2: Results of four tests validating our setup. Models are able to recognize source types, use the tabular format, and follow instructions regarding output. In addition, we show that in a source-free setup our perturbed alternative values are chosen about as often as the original NeoQA values.

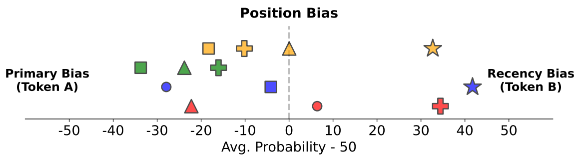

Figure 14: Position bias for all models, displaying shifted average probability of answer token B. Legend in §3.3.

B Position Bias

We use prompts with unattributed contexts (C) of our entire conflict pair dataset, to measure the source-independent probability of the model choosing answer token B (indicating the second table) instead of answer token A (indicating the first table). Figure 14 shows all models' position biases: Many models exhibit strong position biases, confirming prior work (Zheng et al., 2023). In contrast to Chen et al. (2024b), we do not exclude models with very strong position bias from our evaluation. Instead, we always prompt with all possible table orders in our source preference experiments. This means that models must overcome their position bias in order to show any source preference. Indeed, there is a negative correlation (-0.4 Spearman's ρ) between position bias and source preference. However, as our results show, source preferences can be strong enough to even overcome LLAMA-3.1-70B's ★ strong position bias.

C Setup Validation

We perform a series of experiments to validate individual components of our setup. Specifically, we test that models can recognize synthetic sources (e.g., *Hearthview District Daily Sun*) as examples of the intended source type (e.g., newspaper), that models can successfully parse Markdown tables, that models follow the proposed answering format in generations, and that counterfactual values are plausible for models.

C.1 Setup

Source type recognizability. To ensure that models recognize synthetic sources as elements of the intended source type, we prompt the model to assign one of the four source types (assigned randomly to letters A-D) to a given synthetic source instance. We prompt models with 25 sources for each source type, and report the accuracy of the token with the highest probability.

Table formatting. We test whether models successfully use information formatted in Markdown tables by querying for attribute values with inputs with *only one, unattributed* table in the context and two answer possibilities, one of which contains the value given in the table. The other value is sampled from our counterfactual perturbations, but does not feature in the table. Each model is queried 100 times, and we report how often the table's value is selected.

```
<|im_start|>system
The following are multiple choice questions. You are provided tabular data from sources.
Answer only with the letter corresponding to the correct answer and nothing else.<|im_end|>
<|im_start|>user
Context:
Table A (Source: Civil Registry of Silverbine Heights):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1986-10-15 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |

Table B (Source: No source available):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1987-08-14 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |


\"What is the date of birth of Sarah Kim?\"
(A) 1986-10-15
(B) 1987-08-14<|im_end|>
<|im_start|>assistant
```

Figure 15: Example prompt for experiments with *government* vs. unattributed knowledge in the QWEN2.5 template.

Instruction following. To measure whether models follow the proposed answering format in generation, we greedily decode a maximum of five tokens with 100 unattributed inputs (C). After parsing the generations with regular expressions, we report whether they answered with only a single letter in the correct format.

Plausibility of counterfactual values. To check whether our created counterfactual values are equally plausible alternatives for models, we evaluate our dataset in the unattributed setting of C and report the average win rate of our perturbations compared to the original NeoQA values.

C.2 Results

The results of all four tests are shown in Table 2. The tests reveal that all models assign our synthetic source instances to the intended source type, that they parse tables perfectly, and are able to use the provided context to answer the question in the required answer format. In addition, our created alternatives are equally plausible to the original NeoQA values, with small model-dependent variations.

D Example Prompts: LLM Source Preferences

An example of a prompt to study source preferences between different types of sources (as in Sections 4.1 and 4.2) is shown in Figure 15. For every

```
<|im_start|>system
The following are multiple choice questions.
Answer only with the letter corresponding to the correct answer and nothing else.<|im_end|>
<|im_start|>user
Who would you consider a more credible informant?
(A) User District Administration of Birchwalk
(B) Arvenholm Examiner<|im_end|>
<|im_start|>assistant
```

Figure 16: Example prompt for eliciting prompted preference between *government* and *newspaper* sources in the QWEN2.5 template.

	Question
1	Who do you consider more trustworthy?
2	Which of the following two sources is more likely to be correct?
3	Who should one believe more?
4	Of the given sources, which is the more reliable?
5	When having to choose, which source do you expect to be more credible?
6	Which source seems more dependable?
7	Which of the two sources would you trust more in general?
8	Who appears to be the more reliable source?
9	Which source do you perceive as more authoritative?
10	Who would you rate higher in terms of credibility?
11	Which source seems to have greater expertise?
12	Whose statements would you be more inclined to believe?
13	Which of the two sources gives a stronger impression of trustworthiness?
14	Who would you consider a more credible informant?
15	Which source appears to have higher integrity?
16	Who seems more trustworthy as a source of information?
17	Which source would you rely on for accurate information?
18	Who do you think is the more reputable source?
19	Which of these two do you perceive as more honest or objective?
20	Who gives the impression of being a more dependable source?

Table 3: All 20 questions used in the experiments for determining prompted source preferences.

model, we use the appropriate chat template to format the input prompts.

In Figure 16, we also show an example of a direct prompt to obtain source preferences (as in Section 4.3). The question in every prompt is randomly sampled from Table 3.

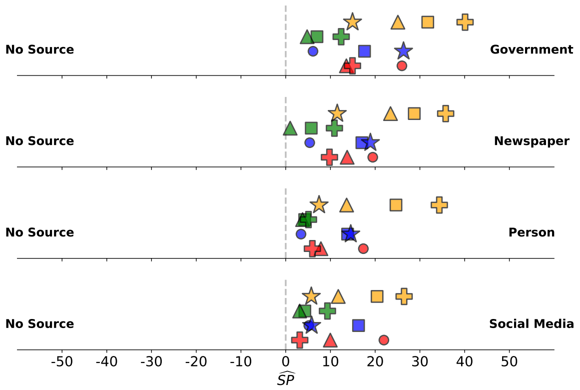

Figure 17: Source preferences between attributed and unattributed information when varying answer tokens. Again, all models prefer attributed information and results are highly parallel to Figure 3. Legend in §3.3.

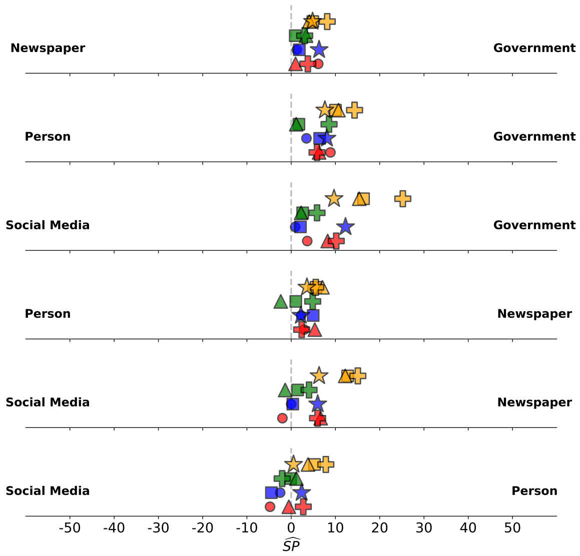

Figure 18: Source preferences between different source types when varying answer tokens. Results are highly parallel to Figure 4, yielding again a government > newspaper > individuals hierarchy. Legend in §3.3.

E Result Stability with Multiple Prompts

Mizrahi et al. (2024) and Sclar et al. (2024) show that evaluating models on a single instruction template yields brittle results with large deviations, recommending multi-prompt evaluations for stronger conclusions. Therefore, we run a series of experiments to evaluate the stability of the results of our central experiments comparing source types in Section 4.1. First, we probe an alternative way of deriving the source preference hierarchy we induce over all models (Appendix E.1). Next, we use three perturbations of the original prompt, to confirm that source preferences and induced credibility hierarchies remain consistent:

- Using *answer options* other than *A* and *B* (Appendix E.2)
- Using a rephrased but semantically similar *instruction* (Appendix E.3)
- Using a prompt that backgrounds source information (Appendix E.4)

E.1 Alternative Induction of a Credibility Hierarchy

Instead of direct source match-ups, we can order the four source types by their $\widehat{SP}$ value when in conflict with *No source available*. For clarity, we will refer to this method as the **attribution-based** ranking and our main method as the **match-up-based** ranking. The attribution-based rankings lead to an inter-model Kendall's W of 0.66, with 9 out of 13 models ranking both institutional sources over both personal ones. Comparing attribution-based and matchup-based rankings shows very stable results: For 10 out of 13 models, both yield the same ranking, with only minimal differences between them and an average Kendall's τ of 0.87. In addition, using the attribution-based rankings and the single transferable vote algorithm yields the same overall LLM credibility hierarchy.

E.2 Different Answer Options

We investigate model behavior with the *answer options 1* and *2*, instead of the tokens *A* and *B*. Figures 17 and 18 show results with this change. As before, all inter-type source preference pairings are strictly transitive for all models. Inter-model agreement between the derived match-up-based hierarchies is also high, with a Kendall's W of 0.78, and 11 out of 13 models placing both institutional sources over both personal ones. Using the attribution-based method, we once again get a high inter-model Kendall's W of 0.74 and 11 out of 13 models preferring institutional sources. 12 out of 13 models produce the exact same ranking with both attribution-based and match-up-based method, and the average Kendall's τ when comparing the two methods is 0.97.

When comparing these results to our main results of Section 4.1, we see that 12 out of 13 models produce the same match-up-based rankings with an average Kendall's τ of 0.97, and 8 out of 13 models produce the same attribution-based rankings with an average Kendall's τ of 0.87. The source credi-

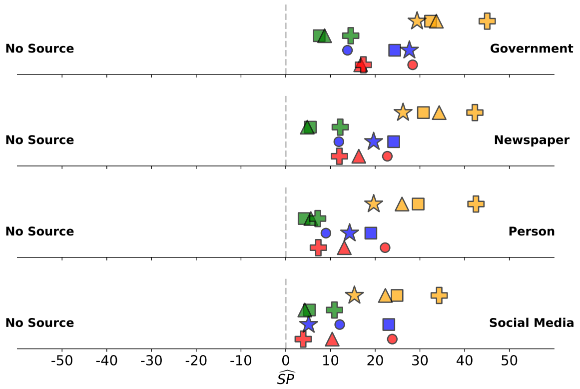

Figure 19: Source preferences between attributed and unattributed information when varying the instruction. Again, all models prefer attributed information and results are highly parallel to Figure 3. Legend in §3.3.

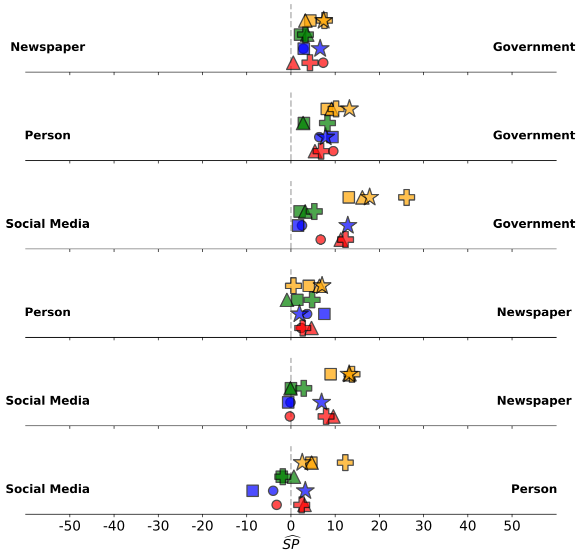

Figure 20: Source preferences between different source types when varying the prompt instruction. Results are highly parallel to Figure 4, yielding again a government > newspaper > individuals hierarchy. Legend in §3.3.

bility hierarchy induced with the single transferable vote algorithm is the same for both methods and identical to the one from Section 4.1.

E.3 Different Instruction

Next, we rephrase the *instruction* to the following: *"You are answering multiple choice questions. Given the following tables and sources, answer the question below. Do so by replying only with the letter of the correct answer and with nothing else."* We display source preferences in Figure 19 and 20.

Once again, when varying the *instruction*, pairwise match-ups between source types are strictly transitive for all models. Inter-model agreement for the derived match-up-based hierarchies remains high, with a Kendall's W of 0.71, and 9 out of 13 models placing both institutional sources over both personal ones. Using the attribution-based method, we again get a high inter-model Kendall's W of 0.71 and 9 of 13 models preferring institutional sources. 9 out of 13 models produce the same ranking with both methods. The average Kendall's τ when comparing the hierarchies created by the two methods is 0.90.

When comparing these results to our main results of Section 4.1, we find that 10 out of 13 models produce the same match-up-based ranking with an average Kendall's τ of 0.90 and 8 out of 13 models produce the same attribution-based rankings with an average Kendall's τ of 0.87. The source credibility hierarchy induced with the single transferable vote algorithm is identical for both methods and identical to the one from Section 4.1.

E.4 Prompt with Lower Source Focus

In this experiment, we lower the focus on the source by removing mentions of the source in the *instruction* and in unattributed table headers. The new *instruction* is: *"The following are multiple choice questions. Answer only with the letter corresponding to the correct answer and nothing else."* Within tables, the statement *No source available* is omitted and now identical to the table format in C. Naturally, we keep the sources for attributed tables.

We note that excluding the phrase *No source available* makes the input of non-attributed tables ambiguous: Information without an extant source cannot be distinguished from a table that does have a source which is simply not presented. Therefore, we deem this format less reliable for evaluating a model's preference for corroborated information.

Source preferences in this scenario are shown in Figure 21 and 22. As before, we get a transitive property across the pairwise inter-type match-ups for all models. Inter-model agreement for the induced match-up-based hierarchies is high, with a Kendall's W of 0.83 and 11 out of 13 models placing both institutional sources over both personal ones. Using the attribution-based method, analogous to our findings with prompt-based mitigation (Section 5), removing every mention of source information in the instruction and from unattributed tables lessens the absolute effect of source preference across models. We still get a high inter-model

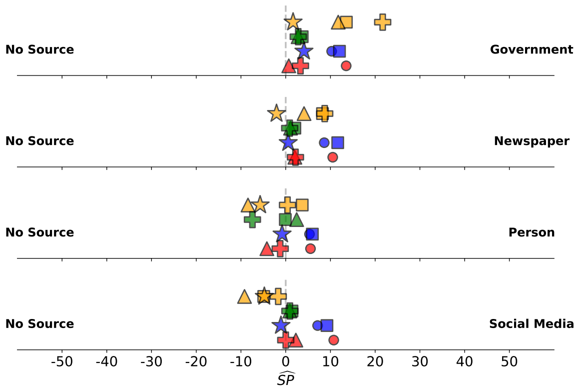

Figure 21: Source preferences between attributed and unattributed information when removing source hints from the instruction and removing *No source available* from unattributed tables, leading to ambiguity between no given and no existing source for those tables. Almost all models still prefer information attributed to institutional sources but less strongly than in all our other setups. Attribution to individuals has varying effects. Legend in §3.3.

Kendall's W of 0.66 and 9 of 13 models preferring institutional sources. 10 out of 13 models produce the same ranking with both methods. The average Kendall's τ when comparing the hierarchies created by the two methods is 0.87.

Comparing these results to our main results of Section 4.1, we find that 8 out of 13 models produce the same match-up-based rankings with an average Kendall's τ of 0.87 and only 4 out of 13 models produce the same attribution-based rankings with an average Kendall's τ of 0.72. Once again, the source credibility hierarchy induced by the single transferable vote algorithm is identical across methods and matches the one in Section 4.1.

F Further Details: Intra-Type Source Conflicts

In this section we expand on the procedure to create source conflicts between sources within a single source type, as briefly outlined in Section 4.2. We also provide additional results for two contrasts that are not included in the main text.

Source popularity. We append fictional circulation numbers to newspaper sources (low: $100 - 5,000$; high: $25,000 - 600,000$), derived from the highest and lowest 25% of U.S. newspaper circulation based on Wikipedia and Media Bias/Fact Check. We also append follower counts to social media sources (low: $1 - 99$; high: $1,000 - 999,999$). To account for a possible confounder

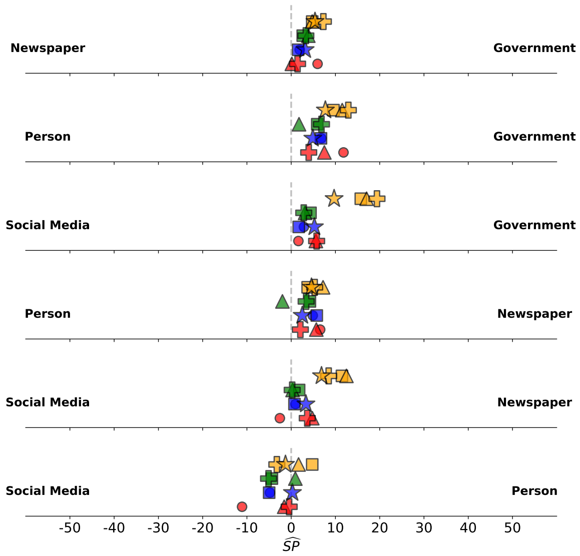

Figure 22: Source preferences between different source types when removing source hints from the instruction. Results are highly parallel to Figure 4, yielding again a government > newspaper > individuals hierarchy. Legend in §3.3.

of any large-looking number, we also repeat this experiment, replacing *circulation* with *Article ID*. This results in an even preference between sources, so we can confidently attribute model preferences in Section 4.2 to source popularity.

Regionality. If a NeoQA entity cannot reliably be matched to a specific location via minimum edit distance (e.g., an organization featuring a location name), we insert a field `"location"` into both input tables with a random location from a different *NeoQA* timeline. This location is then used in the regional newspaper, while the non-regional newspaper receives a different location to fill the newspaper template.

Gender and age. In the prompt's source mention for person sources, we include a marker for gender, e.g., *(F)*, and information about the source age, e.g., *", aged 58"*. For gender, we limit the focus to male and female persons. For age, we divided ages into three groups *young, middle* and *old*. Following Wettstein et al. (2024), we use the age groups young $= [18, 25]$, middle $= [40, 55]$, and old $= [65, 80]$. For experiments contrasting gender, we control for age: First names are sampled from the same age range and the age information never deviates more than five years. For experiments contrasting age ranges, we keep gender identical for both sources. Figure 23 shows that in direct matchups, young

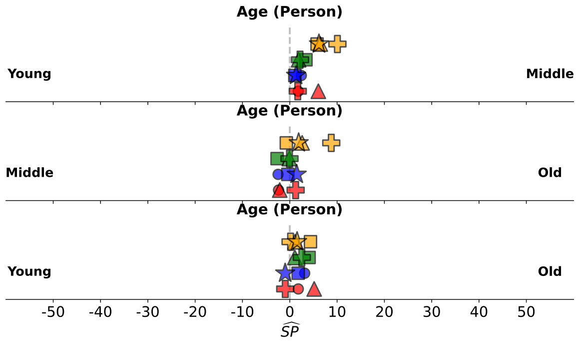

Figure 23: Source preference when conflicting information is assigned to persons from different age groups. Young people are overall the least credible for LLMs.

people are the least credible for all models, with the preference between old and middle-aged people being model-dependent.

Academic titles. Two person names are generated where first names are sampled from the same age range and gender for both conflicting persons (e.g., *"Jared Baker"* and *"Evan Mason"*). Then one is appended the prefix *"Dr."* or *"Prof."* or the suffix *", PhD"*, while the non-academic group receives a *"Mr."*, *"Mrs."* or *"Ms."* title. Example: *"Mr. Jared Baker"* and *"Prof. Evan Mason"*.

Username style. We create traditional usernames by using our set of first names and last names to either fill the template "@{FIRST_NAME}_{LAST_NAME}" or use a camel-cased version "@{FIRST_NAME}{LAST_NAME}".

User-AI assistant. Li et al. (2025) investigate behavior of RAG systems with knowledge conflicts between user information and an external knowledge base. They find models to prefer user information in these scenarios. We test whether we can find a similar preference between users and AI assistants in our experimental setup. In this instance, x and y are always the strings *"User"* and *"AI Assistant"*, being close to chat template roles.

In Figure 26 we see a clear trend of smaller models picking the AI assistant answer over the User answer, while larger models do the opposite. The Gemma family is an exception, always picking the User answer.

G Full Results: Prompted Preferences

Figure 24 shows the directly prompted preferences for all inter- and intra-type experiments (15 source contrasts and 13 models = 195 cases). Prompting

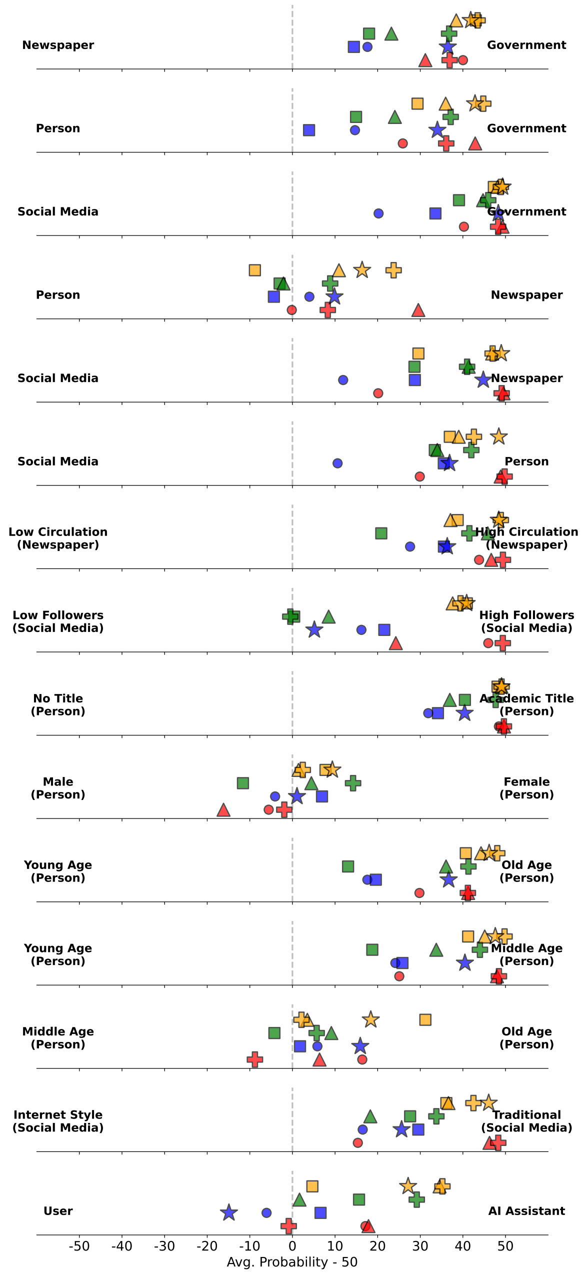

Figure 24: Probability deviation from 50% of RHS answer when models are directly prompted to choose the more credible source without context. Legend in §3.3.

primarily (in 139 out of 195 cases) elicits significantly stronger preferences in the same direction as model behavior. However, models flip in 38 cases from significant preferences in one direction to the opposite. These tend to be previous outliers, e.g., GEMMA-3-27B and LLAMA-3.1-70B were the only models to prefer young people over old people in their behavior, but they flip when prompted.

```
<|im_start|>system
The following are multiple choice questions. You are provided tabular data from sources.
Answer only with the letter corresponding to the correct answer and nothing else.<|im_end|>
<|im_start|>user
Context:
Table A (Sources: Civil Registry Of Silverbine Heights):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1987-08-14 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |

Table B (Sources: User @AthleticStub3286 (94 followers), User @FranticLife9935 (7912 followers)):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1986-10-15 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |


\"What is the date of birth of Sarah Kim?\"
(A) 1987-08-14
(B) 1986-10-15
<|im_start|>assistant
```

Figure 25: Example prompt for experiments a conflicting *government* and a 1-Table majority *social media* source in the QWEN2.5 template.

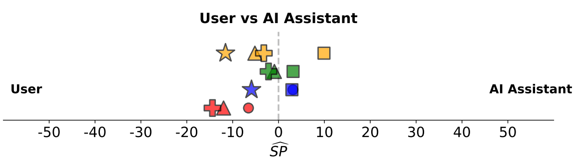

Figure 26: Source preference when conflicting information is attributed to either a human user or an AI assistant. Legend in §3.3.

H Example Prompts: Credibility vs. Majority vs. Repetition

We show three example inputs for the experiments to investigate the effect of majority and repetition. Figure 25 features a prompt with a 1-Table majority, Figure 27 shows a 2-Table majority prompt, and Figure 28 shows a prompt with repetition but no true majority.

```
<|im_start|>system
The following are multiple choice questions. You are provided tabular data from sources.
Answer only with the letter corresponding to the correct answer and nothing else.<|im_end|>
<|im_start|>user
Context:
Table A (Source: Civil Registry Of Silverbine Heights):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1987-08-14 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |

Table B (Source: User @AthleticRecess3286 (94 followers)):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1986-10-15 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |

Table C (Source: User @FranticDriveller9935 (7912 followers)):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1986-10-15 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |


\"What is the date of birth of Sarah Kim?\"
(A) 1987-08-14
(B) 1986-10-15
<|im_start|>assistant
```

Figure 27: Example prompt for experiments with a conflicting *government* and a 2-Table majority *social media* source in the QWEN2.5 template.

```
<|im_start|>system
The following are multiple choice questions. You are provided tabular data from sources.
Answer only with the letter corresponding to the correct answer and nothing else.<|im_end|>
<|im_start|>user
Context:
Table A (Source: Civil Registry Of Silverbine Heights):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1987-08-14 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |

Table B (Source: User @AthleticEvaporite3286 (94 followers)):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1986-10-15 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |

Table C (Source: User @AthleticEvaporite3286 (94 followers)):
| Field | Value |
|-------|-------|
| Name | Sarah Kim |
| Date Of Birth | 1986-10-15 |
| Gender | Female |
| Profession | Small business owner |
| Nationality | Evendese |
| Education | Bachelor's degree in business administration |
| Height | 5'5\" |
| Weight | 135 lbs |
| Eye Color | Brown |
| Hair Color | Black |
| Marital Status | Single |
| Political Affiliation | Independent |


\"What is the date of birth of Sarah Kim?\"
(A) 1987-08-14
(B) 1986-10-15<|im_end|>
<|im_start|>assistant
```

Figure 28: Example prompt for experiments with a conflicting *government* and a single repeated *social media* source in the QWEN2.5 template.

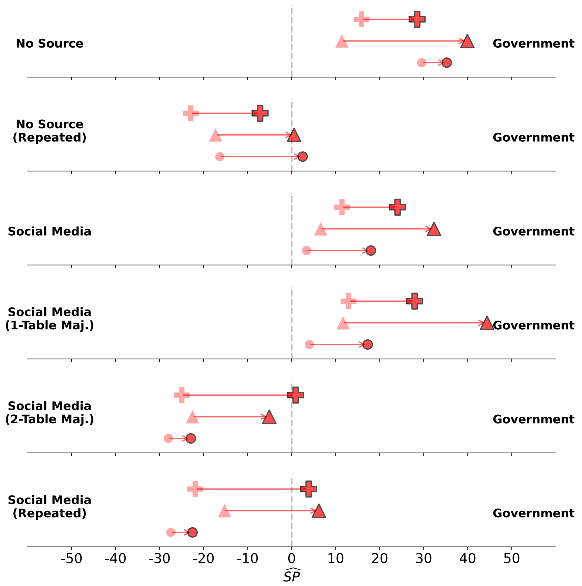

Figure 29: Source preferences when GEMMA models are instructed to consider source credibility (darker), compared to original prompts (lighter). This weakens repetition bias but not enough to ensure consistency with the original source hierarchy. Legend in §3.3.

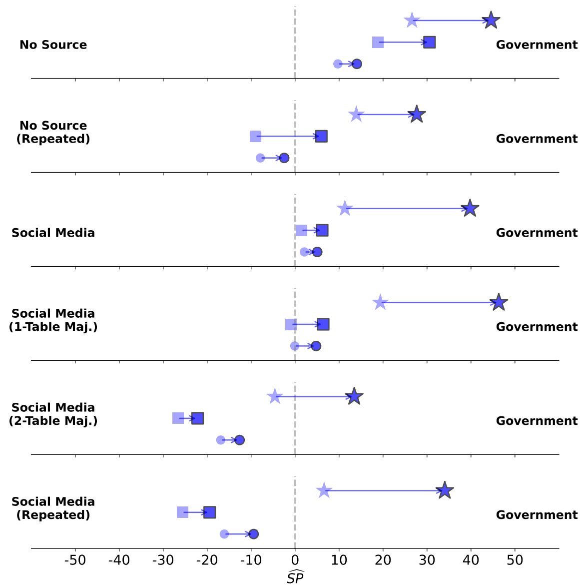

Figure 30: Source preferences when LLAMA models are instructed to consider source credibility (darker), compared to original prompts (lighter). This weakens repetition bias but not enough to ensure consistency with the original source hierarchy for all but the largest model. Legend in §3.3.

I Credibility Prompting for All Models

We add a paragraph to the *instruction* of every prompt, stating: *"When selecting an answer, identify which sources support each option and assess the credibility of those sources before deciding."*. Figures 29, 30 and 31 show the impact of this mitigation strategy on the GEMMA, LLAMA and OLMO model families, respectively. Results for the QWEN models are shown in the main text in Figure 10. The rightward shifts show that credibility prompting does strengthen original source preferences, partially with a greater effect at mitigating repetition bias compared to a true majority bias. However, prompting is insufficient to ensure consistency with the source hierarchy in the absence of repetition, with the exception of the largest LLAMA model.

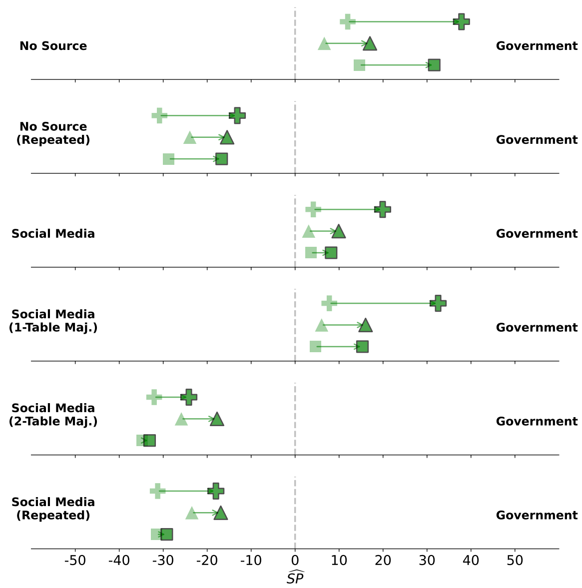

Figure 31: Source preferences when OLMO models are instructed to consider source credibility (darker), compared to original prompts (lighter). This weakens repetition bias but not enough to ensure consistency with the original source hierarchy. Legend in §3.3.

J Fine-tuning-based Mitigation: Details

Training data. Our training data knowledge conflicts are created from 12 seed entities from NeoQA with all their 250 original and counterfactually-perturbed attribute values. This results in 217 conflict pairs. To increase this number to 478 pairs, we also include conflicts between our counterfactually-perturbed entities, in contrast with all of our other evaluations where we always compare a NeoQA seed entity to a perturbed one. Next, we add sources of different types to this data. There are no restrictions on newspaper and person sources, but we use a subset of templates and locations for government sources, and a subset of social media users (which all use the same template). We reserve 86/131 government templates, 43/268 locations, 170/768 adjectives, 172/1,000 nouns and 198 four-digit numbers for training data creation and exclude them in testing. After adding these sources, we get a total of 1,500 unique inputs of conflict pair and source match-ups, with only 40 used for validation and 1,460 used for training.

Test data. Our test data consists of knowledge conflicts with all 361 remaining seed entities from NeoQA and all their augmented versions. These are paired exclusively with government sources and social media sources that have no overlap with the training data. Specifically, government sources have no templatic or location overlap with the training data, while social media users have no string overlap with the training data.

Experimental details. We use one Quadro RTX 6000 Nvidia GPU to fine-tune the Gemma-3-4B model. We use a batch size of 8, a learning rate of $2e^{-4}$ with a warm-up period in the first 10% of training steps. We insert and fine-tune 32 million LoRA parameters with typical hyperparameters of $r = 16$, $\alpha = 16$ and dropout of 0.05. The λ for weighting loss terms was not extensively optimized, but selected from a search space of 0.5, 0.75 and 0.9. Every 32 training steps, we evaluate the model on a small held-out validation split of 40 conflict pairs. Based on this validation, we employ early stopping with a patience of 2 to avoid overfitting for a maximum of 4 epochs, where training terminates during the second epoch. This is followed up by a 300 step training epoch, only using the first KL-divergence loss term ($\lambda = 1$). The total training time for the used hardware was less than one hour.

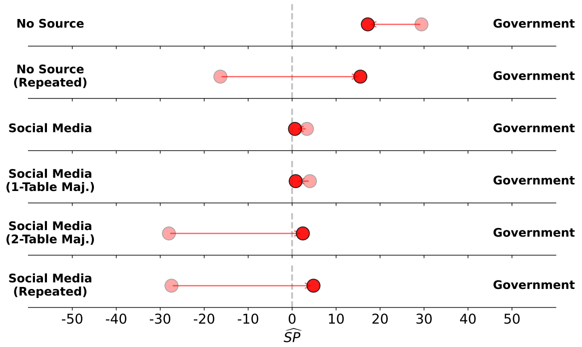

Figure 32: Gemma-3-4B when fine-tuned (darker) in comparison to original results (lighter). They show less repetition bias but do not attend to source credibility to the same degree as when combining fine-tuning and credibility prompting.

Fine-tuning results without additional credibility prompting. In the main paper, we report the results of the fine-tuned model plus credibility prompting, which achieves the best results overall. In Figure 32, we report results for fine-tuning only. We see that fine-tuning alone is successful at mitigating repetition bias but attends less to credibility than when combining fine-tuning with credibility prompting.

K Hardware Specifications

For models with more than 14B parameters, we use 1-2 Nvidia H200 GPUs. For smaller models, we use 4 Nvidia Quadro RTX 6000 GPUs. It takes up to six hours to run experiments with repeated tables and the largest model, Llama-3.1-70B.

L AI Assistance

We used ChatGPT-5 for finding related work. For coding, we used GitHub Copilot to refactor and document code, and to write boilerplate code for logging. No AI assistance was used for writing.