

# SPACE: Source Preferences Affected by varying Credibility of Expressions

Florenis Fejzaj, Carolin Flath, Luca Fuhrmann, Cilian Kerskens, Ole Lorbacher

{fejzaj, flath, fuhrmann, kerskens, lorbacher}@cl.uni-heidelberg.de

University of Heidelberg

Department of Computational Linguistics

## Abstract

Given a knowledge conflict, prior work has shown that large language models (LLMs) favor certain sources over others, thereby forming a source hierarchy when the information from those sources appears similarly credible. Until now, the impact of *message credibility*, the plausibility of the content of a message, in these scenarios has not been studied. By introducing three levels of non-credibility into the knowledge conflicts, we examine their effect on the source hierarchy. We find that introducing non-credible elements into the knowledge conflicts has a much larger effect than the source hierarchy. Our findings also indicate that the strength of this effect depends on the particular magnitude of non-credibility, suggesting a message credibility preference.

## 1 Introduction

Large language models (LLMs) are being used increasingly for question answering and subsequently as an alternative to search engines (Pew Research Center, 2025; Activate, 2024). In this regard, recent research findings suggest that LLMs exhibit preferences for different source types when confronted with conflicting and *equally credible* information (Schuster et al., 2026), indicating that some information is given greater weight based on its source. However, it has not yet been studied how they resolve knowledge conflicts containing at least *one statement with non-credible information*, which may be crucial and relevant when answering questions.

Following the work of Schuster et al. (2026), we aim to evaluate how question answering under such knowledge conflicts is influenced by reducing the message credibility (Appelman and Sundar, 2016) in different gradations. We do this by gradually diminishing the credibility of single attributes in the contexts provided to the LLMs.

Our key findings and contributions are:

- We introduce a novel, structured framework for assessing the credibility of messages (ranging from *credible* to *nonsensical*), thus a dataset for inter-contextual knowledge conflicts arising from messages with varying levels of credibility. (Section 2).
- We find that LLMs show a strong affinity for message credibility, overriding the latent source preferences, while at the same time being sensitive to certain degrees of message credibility, suggesting a *message credibility preference* (Section 6, Section 6.5.4).

## 2 Data

To ensure comparability to Schuster et al. (2026), we build on their dataset, which is based on NeoQA (Glockner et al., 2025), including its source types (government, newspaper, social media user, person, NO SOURCE), entities (ART, BUILDING, EVENT, LOCATION, ORGANIZATION, PERSON, PRODUCT) and their associated attributes (e.g., capacity for BUILDING entities or number\_of\_employees for ORGANIZATION entities). Although the data is synthetically generated, it is designed to reflect real-world principles.

### 2.1 Categories of non-credibility

We consider all data from Schuster et al. (2026) to be credible. In addition to credible messages, we introduce three levels of non-credible messages by gradually reducing their credibility: improbable, implausible and nonsense.

Improbable messages are consistent with real-world regularities, but statistically slightly rare, (e.g. a body height of 7.5 feet). Since both credible and improbable messages remain somewhat possible, we expect a strong preference for credible over improbable messages. Implausible messages exceed the physically realizable value for the given

attribute (e.g. a body height of 20 feet). Here, we expect a *categorical* rather than a graded preference for credible messages, which should be preferred not just mostly, but *always*.

Finally, messages that are nonsense commit a categorical error: the assigned value belongs to a semantic domain incompatible with the respective attribute, hence nonsensical (e.g. a body height of “lettuce”). We therefore expect the following message credibility preference:  $\text{credible} > \text{improbable} > \text{implausible} > \text{nonsense}$

All categories satisfy the condition of analytical truth (Quine, 1951), meaning that every message has to be not self-contradictory in the sense that it does not assert and deny the same proposition simultaneously. By ensuring this, we prevent obviously contradictory information, which rarely (if ever) occurs in real human communication.

## 2.2 Data Adjustments

Based on the original data, we define rules to generate five non-credible values for each category for every attribute.<sup>1</sup> A detailed list of these rules can be found in Appendix A.1. Where needed, we alter the dataset to fit the project specific needs, e.g. by removing credibility conflicts in the base data. A detailed description of the adjustments can be found in Appendix A.2. Additionally, we remove the entity type MISCELLANEOUS. No other entities are removed.

## 3 Metrics

To both comprehensively interpret our results and experiments and compare those to the earlier findings of source hierarchy (Schuster et al., 2026), we choose to formulate a diverse set of metric calculations. When talking about any probabilities of decisions or choices, we refer to the probability distribution created from selected logits (corresponding to the answer option tokens) from the models’ outputs.

### 3.1 Accuracy

We define the accuracy as the rate at which the credible value is preferred over the non-credible value. This way we have a straight forward option to analyze if the message credibility in general has an impact and whether our understanding of credibility is displayed in this result.

<sup>1</sup>Not all categories can be adapted to each attribute. See Appendix A.1 for more information.

For the abstention studies we measure the amount of times the answer token for “Can’t be answered” has the highest probability.

## 3.2 Direction and Boost Scores

Our direction score quantizes the probability a source gets when compared to another source for all three scenarios of credibility: 1) both sources having only credible values. 2) a non-credible value is attributed to the first source 3) or to the second source.

For the boost score the direction score of the scenario “credible vs. non-credible” is then normalized with the direction score of the scenario where both sources are attributed with a credible value to display the boost a source gets when itself has a credible value attributed and the other one a non-credible value.

The anti-boost score shows how much other sources (with a credible value) are boosted when a source has a non-credible value.

### 3.2.1 Direction Score

Let  $ST = \{\text{Government, Newspaper, Social Media, Person}\}$   
 $CC = \{\text{credible, improbable, implausible, nonsense}\}$

$$d(S1, S2, \text{conf}) = \frac{(p_{\text{conf}} - p_{\text{cred}}) + (p_{\text{conf}}^{\text{rev}} - p_{\text{cred}}^{\text{rev}})}{2}$$

$$D_{\text{c\_conf}}(S1, S2, \text{conf}) = \frac{\sum^N d(S1, S2, \text{conf})}{N} = -D_{\text{conf\_c}}(S2, S1, \text{conf})$$

$$D_x^{\text{all}}(S1, S2) = \frac{1}{|CC|} \sum_{c \in CC} D_x(S1, S2, c), \quad x \in \{\text{c\_conf, conf\_c}\}$$

with

- Source Pair  $S1, S2 \in ST, S1 \neq S2$
- $p$  being the score of the answer

When talking about the Direction score, we refer to the  $D^{\text{all}}$  scores, if not specified otherwise.

### 3.2.2 Boost Score

$$b(S1, S2, \text{conf}) = |D_{\text{c\_conf}}(S1, S2, \text{conf}) - D_{\text{c\_conf}}(S1, S2, \text{credible})|$$

$$b_{\text{anti}}(S1, S2, \text{conf}) = |D_{\text{conf\_c}}(S1, S2, \text{conf}) - D_{\text{conf\_c}}(S1, S2, \text{credible})|$$

$$B_{\{\text{anti}\}}(S, \text{conf}) = \frac{1}{|ST|} \sum_{S' \in ST \setminus S} \sum \frac{b_{\{\text{anti}\}}(S, S', \text{conf})}{N}$$

$$\text{Boost}_{\{\text{anti}\}}(S) = \frac{1}{|CC|} \sum_{c \in CC} B_{\{\text{anti}\}}(S, c)$$

## 3.3 Confidence Metrics

We implement confidence metrics to further investigate the model’s confidence and position bias, regardless of any answer being a correct one.

### 3.3.1 Consistency

The consistency of a model depicts how consistent the pattern of answer tokens of a model are for all permutations of a prompt (10 prompts for one entity with one attribute and one credibility category). E. g. the pattern is always “B B”, therefore the choices are always the same, regardless of the permutation.

### 3.3.2 Positional Coherence

If the answer for a normal and a reversed prompt is the same (pattern “A B” or “B A”) the positional coherence is high and the positional bias low.

### 3.3.3 Inter-Model Agreement

For comparing the models we use the inter-model agreement which measures how often two models make the same prediction.

### 3.3.4 Setup Validation

To see whether the models are actually confident in their decision and the interpretation of the logits works correctly with our framework, we evaluate how confident it is in picking “A”, “B” (or “C” for abstention studies), meaning that the answer token is also the token with the highest probability.

## 4 Models

We select models out of the model families used by Schuster et al. (2026) for comparability of the source hierarchy.

- QWEN2.5-7B (Qwen-Team, 2024) ■
- LLAMA-3.1-8B (Patterson et al., 2022) ■
- LLAMA-3.2-3B (Liu et al., 2025) ●
- GEMMA-3-4B (Gemma-Team et al., 2025) ●
- OLMo-2-7B (OLMo-Team et al., 2024) ■

## 5 Experiments

We perform various experiments to investigate the effects of message credibility in different settings. The first setup is to examine whether a variation of message credibility has an impact in question answering under a knowledge conflict without introducing any sources. The prompts (Appendix B.1) for this experiment type consist of the instruction, two tables (without sources) of which one has a non-credible value for one attribute, and a question about the conflicting attribute with the two answer tokens “A” and “B” each assigned to one value of the tables.

We then change the prompts for the next experiment type while introducing sources to all the tables (Appendix B.2). Everything else stays the same. We aim to compare the results of these experiments with the results from Schuster et al. (2026) in order to find out whether the source preferences and source hierarchy are affected by a varied message credibility of the table assigned to a source.

To take a closer look on the impact of the message credibility of the whole table and not just the conflicting attribute which is also asked about (and therefore in the focus of the prompt) we perform the first two experiment types with a different question. We now ask for an attribute where the two possible values are credible in both tables and another attribute is made non-credible for one table (Appendix B.4) (“query experiments”).

Like Schuster et al. (2026), we also implement abstention studies (Madhusudhan et al., 2025) to analyze whether models select the answer option “Can’t be answered” when both tables have a non-credible value for the questioned attribute (Appendix B.5). We conduct this experiment with and without sources. Like this we want to investigate the models’ ability to acknowledge uncertainty (Yang et al., 2023).

## 6 Evaluation

The significancies of our results are layed out in Appendix D.

### 6.1 Experiments without Sources

To evaluate the experiments without sources we mainly focus on the results of the accuracy, as this shows whether models can actually differentiate knowledge conflicts without any influence from information about sources.

Across all credibility categories the accuracy reaches 74.5%. This result aligns with our understanding of message credibility as humans would as well rather choose a credible value over a non-credible value.

| Category    | Accuracy     |
|-------------|--------------|
| improbable  | 0.593        |
| implausible | 0.704        |
| nonsense    | <b>0.902</b> |

Table 1: Mean Accuracy grouped by credibility category in the standard experiment without sources. Mean = 0.745.

Furthermore it is visible that a credibility category lower in the hierarchy (Section 2.1) shows a higher accuracy meaning models detect low credibility more reliably.

## 6.2 Experiments with Sources

To further investigate the models’ behavior in knowledge conflicts and the interaction between source and message credibility, source information is added to the prompts.

Regarding the accuracy in this scenario, it is noticeable that adding the source information does not have a notable impact on the models’ decisions (Figure 2). Additionally, this credibility hierarchy is further underlined by the positional coherence (Figure 3). From these two metrics we can conclude that models can more easily distinguish absurd information than (rather) credible one, which confirms our previously proposed message credibility hierarchy (Section 2.1).

Further, when looking at the metrics regarding the models confidence, models seem to share a substantial agreement between each other in their decisions as well as overall confidence in their decisions (Table 9, Figure 4).

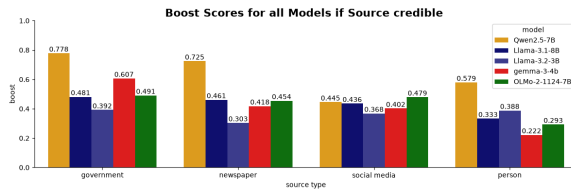


Figure 1: Boost scores for all models grouped by source type of the credible source. Models prefer institutional sources.

When further investigating the source preference, a vague hierarchy can be distinguished from the models’ predictions using the boost score, even though most singular comparisons are not significant. Here it is visible (Figure 1), that all models prefer governmental sources above others, while the scores from the anti-boost metric show a more equal distribution throughout the source types (Figure 5). When combining source-preference and message-credibility, a visible hierarchy in source types and credibility categories can be described (Table 6), which also aligns with the source preference established in Schuster et al. (2026).

## 6.3 Differing Query Attribute

When changing the attribute the model is asked about in the prompt, a decrease in performance ( $\approx 15\%$ ) in the accuracy is directly noticeable.

| Category    | Accuracy     |
|-------------|--------------|
| improbable  | 0.538        |
| implausible | 0.559        |
| nonsense    | <b>0.623</b> |

Table 2: Mean Accuracy grouped by credibility category in the query experiment. Mean = 0.577.

This raises the question whether models actually pay substantial attention to the whole tables presented in the prompt or focus more on the question at the end of the prompt. For the latter case, this would impose major concerns about the actual considered context windows of the models.

Yet, the source preference shows similar values to the standard experiment with sources and does not indicate any changes in hierarchy (Figure 12, Figure 13). Also, the models show confidence in the pattern of credibility hierarchy (Figure 14) and consistency (Figure 15). Additionally, a slight decrease in the inter-model agreement is notable (Table 10).

## 6.4 Abstention Studies

Following the trend in the positional coherence (Figure 22), the accuracy (Table 3) still resembles the same hierarchy as defined in Section 2.1, although indicating a decrease in performance.

Judging from the evaluation of these accuracies, the models do not reliably abstain from decisions even when the option is explicitly presented in the instruction, although remaining confident in their decisions (Figure 23). Still, the models keep their hierarchy in sources similar to those in the standard experiment (Figure 24, Figure 25).

| Category    | Accuracy     |
|-------------|--------------|
| improbable  | 0.227        |
| implausible | 0.291        |
| nonsense    | <b>0.502</b> |

Table 3: Mean Accuracy grouped by credibility category in the abstention experiment. Mean = 0.350.

## 6.5 Linear Mixed Effects Model

To determine whether the effects captured by the direction and boost scores (Section 3.2) hold

once other factors are taken into account, we fit a linear mixed-effects (LME) model that jointly estimates the contributions of message credibility, source pairing, attribute position, and entity class to question-answering behavior. We therefore fit a LME that jointly estimates the contribution of these factors to question-answering behavior.

### 6.5.1 Dependent Variable

The outcome variable is the mean credible-answer probability

$$\bar{p}_{\text{credible}} = \frac{p_{\text{cred}} + p_{\text{cred}}^{\text{rev}}}{2}$$

where  $p_{\text{cred}}$  and  $p_{\text{cred}}^{\text{rev}}$  are the probabilities assigned to the credible answer token in the normal and reversed prompt, respectively (cf. Section 3.2.1), averaged to remove positional presentation bias.

### 6.5.2 Fixed Effects

We include four fixed effects: credibility category CC paired with source pair type – the order-agnostic combination of the two source types  $\text{ST} \times \text{ST}$  (reference level:  $\text{no\_source} \times \text{no\_source}$ ,  $\text{credible}$ ) – attribute position, and entity class. Attribute position is treated as a categorical predictor with one indicator per position rather than a continuous linear term, since preliminary fits showed a non-monotonic relationship between position and  $\bar{p}_{\text{credible}}$  that a single linear slope would misrepresent.

### 6.5.3 Random Effects

Entities serve as the grouping factor, since each entity is observed under multiple combinations of credibility category, source pairing, and position. We include a random intercept together with random slopes for credibility category, allowing entities to differ not only in their baseline credible-answer probability but in how strongly that probability shifts as the conflicting information becomes less credible.

### 6.5.4 Hypothesis testing

To investigate what accounts for the source type boost and anti-boost scores, we turn to the LME and compute significance tests with the hypotheses  $H_0$ : the two source-pairing conditions being compared produce no difference in the model’s probability of selecting the credible answer, once credibility category, conflict-attribute position, and entity class are controlled for;  $H_1$ : they produce a nonzero difference. This is computed

using a pairwise contrast test against the baseline of  $\text{no\_source} \rightarrow \text{no\_source}$  at the given credibility level (Appendix E.6).

To test whether source pairing matter, we look at each pairing individually and run a omnibus Wald test per model with the hypotheses  $H_0$ : source pairing has no effect on the probability of selecting the credible answer at all, every coefficient involving  $\text{source\_pair\_type}$  (main effects and interactions with credibility category) is simultaneously zero;  $H_1$ : at least one of those coefficients is nonzero.

The heat map (Figure 32) shows the same preferences for government sources that the boost score depicted with directional asymmetry with a increase when credible and a decrease when non credible. This pattern is broadly consistent with the source credibility hierarchy (Schuster et al., 2026) induced from pairwise matchups across 13 models:  $\text{government} > \text{newspaper} > \text{person} > \text{social media}$ , with person and social media left unranked against each other. The size and significance of these source-pairing effects shrink as message credibility decreases.

## 7 Conclusion

Our experiments show, that when faced with a knowledge conflict, LLMs generally choose answers with higher message credibility. They also follow the order of categories we proposed ( $\text{credible} > \text{improbable} > \text{implausible} > \text{nonsense}$ ).

LLMs are especially sensitive to non-credibility when it affects the queried attribute itself, rather than the source presenting it. We therefore conclude that LLMs weigh the answer’s credibility itself much higher than the source’s credibility it comes from. This matches our observation of vanishing source preferences when the credibility changes.

Given two equally non-credible answers, models still tend to pick one rather than mark the question as unanswerable, even when that option is available. This shows that the introduction of non-credibility does not change the underlying tendency not to abstain.

## Limitations

**Computational Resources.** We were limited to smaller or the smallest models for each model family, due to limited compute resources. Smaller models have, in general, less pronounced source hierarchies (Schuster et al., 2026). Repeating our experiments with larger and with more models would strengthen the findings we made.

**Data Language and format.** Our prompts and our data are exclusively in english. Additionally, only 3 of our chosen attributes are textual (hair\_color, eye\_color, material), the rest is numerical data.

**Table formatting.** While good for our task and very representative for the application with RAG architectures, formatting our data in tables instead of continuous text limits the applicability of our results to ChatBot-like scenarios where users supply their own sources.

**Multiple Choice Questioning.** Our experimental setup doesn't allow for continuous text answers from the prompted models but restricts them to a multiple choice scenario. This, obviously, makes it more suitable for comparing it to other tasks and some RAG applications but isn't representative for chat bot like usage. Based on the results of our abstinence studies, it is likely that the models would give more fine grained or differentiated answers in an unrestricted answering setup.

**Instruction.** Similar to the instruction tuning performed in Schuster et al. (2026), we could have altered our instruction more, in order to make the focus on the assignment of message credibility from sources more explicit.

**Abstention Study - Multiple Choice Answering.** As visible in Appendix B.5, the abstention option is placed at the third position for both normal direction and reversed prompts. This could also impose a form of position bias for the model and could be solved by permutating all combinations of the three options.

## References

- Activate. 2024. [Artificial intelligence-powered search usage among adults in the United States in 2024](#), by search type [Graph]. *Statista*.
- Alyssa Appelman and S. Shyam Sundar. 2016. [Measuring Message Credibility: Construction and Validation of an Exclusive Scale](#). *Journalism & Mass Communication Quarterly* 93(1):59–79. visited on 2026-04-29.

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, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Naveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Naveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan

- Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 Technical Report](#). arXiv: 2503.19786 [cs.CL].
- 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.
- Rom Himmelstein, Amit LeVi, Yonatan Belinkov, and Avi Mendelson. 2025. [Silent Tokens, Loud Effects: Padding in LLMs](#). arXiv: 2510.01238 [cs.CL].
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025. [SpinQuant: LLM quantization with learned rotations](#). arXiv: 2405.16406 [cs.LG].
- Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. [Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345, Abu Dhabi, UAE.
- OLMo-Team, 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, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. [2 OLMo 2 Furious](#). arXiv: 2501.00656 [cs.CL].
- David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. [The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink](#). arXiv: 2204.05149 [cs.LG].
- Pew Research Center. 2025. [Share of adults in the United States who used ChatGPT as of March 2025, by activity \[Graph\]](#). *Statista*.
- W. V. Quine. 1951. [Main Trends in Recent Philosophy: Two Dogmas of Empiricism](#). *The Philosophical Review* 60(1):20–43. visited on 2026-05-17.
- Qwen-Team. 2024. [Qwen2.5: A Party of Foundation Models](#).
- Jakob Schuster, Vagrant Gautam, and Katja Markert. 2026. [Whose Facts Win? LLM Source Preferences under Knowledge Conflicts](#). arXiv: 2601.03746 [cs.CL].
- Sarah Catherine Walpole, David Prieto-Merino, Phil Edwards, John Cleland, Gretchen Stevens, and Ian Roberts. 2012. [The weight of nations: an estimation of adult human biomass](#). *BMC Public Health* 12(1):439.
- Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. 2023. [Improving the Reliability of Large Language Models by Leveraging Uncertainty-Aware In-Context Learning](#). arXiv: 2310.04782 [cs.CL].

# Appendix

## A Data Generation

### A.1 Rules

For all of the attributes that were altered to have non-credible values, we defined rules to generate the non-credible values. All rules have been discussed thoroughly in our research group. Whenever possible and suitable, these rules were based on real world data, statistics or scientific theories.

For the values in the nonsense category, we defined a list of words that fit for all possible attributes (Appendix F).

Not all categories can be adapted to each attribute. For example, there is no way to alter the values of the attribute `hair_color` in such a way that these values are improbable, because, biologically spoken, rather a certain hair color is *or* is not natural and therefore implausible or not implausible.

#### A.1.1 person\_weight

- improbable: 300 - 500 kg
- implausible: 800 - 2000 kg
- Source: Walpole et al. (2012)

#### A.1.2 product\_weight

- improbable: 0.5 - 1 g and 50 - 70 tons
- implausible: -550 - -10 kg
- Source: A piece of paper weight around 5g, cars and trucks weigh up to 44 tons in Europe. Based on that, we picked the ranges mentioned above

#### A.1.3 floors

- improbable: 250 - 1000
- implausible: floats between -60 and 8
- Source: 163 floors for the highest building<sup>2</sup>, almost no buildings with more than 100 floors<sup>3</sup>

#### A.1.4 budget

- improbable: 100,000,000 - 1,000,000,000
- implausible: -120,000,000 - -1

#### A.1.5 hair\_color

- implausible:
  - Combination of

- Natural light
- Natural dark
- Natural

▸ and

- green
- yellow
- blue
- purple
- orange

- Sources: <https://www.haut.de/wie-bilden-sich-die-natuerlichen-haarfarben/> and <https://www.ulprospector.com/knowledge/10385/pcc-hair-pigmentation/>

#### A.1.6 number\_of\_participants

- improbable: 10 million - 70 million
- implausible: 6 billion - 10 billion
- Sources:
  - Most participants for single day events: 76 million,<sup>4</sup>
  - Number of people alive: around 8 billion

#### A.1.7 number\_of\_employees

- improbable: 2.5 million - 6 million
- implausible: 67 million - 10 billion
- Source: Existing maximum is 3 million, runner ups are 2.3 million and 1.61 million<sup>5</sup>

#### A.1.8 annual\_revenue

- improbable: one of
  - 900 billion - 1.1 trillion
  - -420 billion - -10 billion
- implausible: one of
  - 5 trillion - 10 trillion
  - -100 trillion - -6 trillion
- Source: Highest real life value is 716 billion<sup>6</sup>

#### A.1.9 date

- implausible:
  - year: -5000 - 5000, excluding 500 - 2500
  - month: 1 - 12
  - day: 1 - 31

#### A.1.10 release\_date

- implausible:
  - year: -5000 - 5000, excluding 500 - 2500
  - month: 1 - 12

<sup>2</sup>[https://en.wikipedia.org/wiki/Burj\\_Khalifa](https://en.wikipedia.org/wiki/Burj_Khalifa)

<sup>3</sup><https://www.skyscrapercenter.com/criteria#:~:text=Height%20Relative%20to,the%20urban%20norm>

<sup>4</sup>[https://en.wikipedia.org/wiki/List\\_of\\_most-attended\\_concerts](https://en.wikipedia.org/wiki/List_of_most-attended_concerts)

<sup>5</sup>[https://en.wikipedia.org/wiki/List\\_of\\_largest\\_employers](https://en.wikipedia.org/wiki/List_of_largest_employers)

<sup>6</sup>[https://en.wikipedia.org/wiki/List\\_of\\_largest\\_companies\\_by\\_revenue](https://en.wikipedia.org/wiki/List_of_largest_companies_by_revenue)

- day: 1 - 31

#### A.1.11 price

- improbable: 1 million - 30 billion
- implausible: 30 trillion - 500 quadrillion

#### A.1.12 warranty

- implausible: -500 - 500 years, excluding 0 - 100 years, given in number of days

#### A.1.13 duration

- improbable: 2 - 20 days
- implausible: 300 - 15,000 days

#### A.1.14 elevation

- improbable: -3500 - 8800 meters, excluding -50 - 4000 meters
- implausible: -15,000 - 15,000 meters, excluding -11,000 - 9000 meters

#### A.1.15 material

- implausible:
  - Combination of:
    - Sugar
    - Flour
    - Salt
    - Silk
    - Wool
    - Cotton
    - Polyester
    - Leather
  - the word “and” and
    - water
    - sunflower oil
    - apple juice
    - perfume
    - cardboard
    - paper
    - napkins
    - bandages

#### A.1.16 year\_built

- implausible: 2300 - 10,000

#### A.1.17 date\_of\_birth

- improbable:
  - year: 2300 - 10,000
  - month: 1 - 12
  - day: 1 - 29
- implausible:
  - year: 0 - 2022

- month: 13 - 100
- day: 32 - 100

#### A.1.18 eye\_color

- implausible: One of:
  - Purple
  - Violet
  - Pink
  - Orange
  - Yellow
  - Black
  - Light orange
  - Dark orange
- Source: Possible range of eye colors<sup>7</sup>

#### A.1.19 area

- improbable:
  - If entity has the “country” attribute: one of:
    - 20 million - 40 million km<sup>2</sup>
    - 0.1 - 0.4 km<sup>2</sup>
  - Else: one of:
    - 15,000 - 600,000 km<sup>2</sup>
    - 1 - 10 km<sup>2</sup>
- implausible:
  - One of:
    - 0.00001 - 0.0005 km<sup>2</sup>
    - 2 billion - 3 billion km<sup>2</sup>
- Sources:
  - smallest country: 0.44 km<sup>28</sup>
  - biggest country: 17,098,246 km<sup>29</sup>
  - smallest city: 13 km<sup>210</sup>
  - Different ways of calculating for the largest city in the world<sup>111213</sup>

#### A.1.20 founded

- improbable: one of:
  - 0 - 578 AD
  - 1 - 298,000 BC
- implausible: one of:
  - 10 billion - 298,001 BC
  - 2300 - 10 billion
- Sources:

<sup>7</sup>[https://en.wikipedia.org/wiki/Eye\\_color#Eye\\_color\\_range](https://en.wikipedia.org/wiki/Eye_color#Eye_color_range)

<sup>8</sup><https://whc.unesco.org/en/list/286/maps/>

<sup>9</sup><https://en.wikipedia.org/wiki/Russia>

<sup>10</sup>[https://en.wikipedia.org/wiki/Hum,\\_Istria\\_County](https://en.wikipedia.org/wiki/Hum,_Istria_County)

<sup>11</sup><http://www.demographia.com/db-worldua.pdf>

<sup>12</sup><https://worldstats.com/global-stats/largest-cities-by-area/>

<sup>13</sup><https://worldpopulationreview.com/world-city-rankings/largest-city-in-the-world-by-area>

- ▶ oldest documented company: 578 AD<sup>14</sup>
- ▶ Appearance of Homo Sapiens: 298,000 BC<sup>15</sup>

### A.1.21 population

- improbable: one of:
  - ▶ 0 - 34
  - ▶ 1,417,492,001 - 8,295,655,181
- implausible: one of:
  - ▶ 10 billion - -1
  - ▶ 8,295,655,182 - 100 billion
- Sources:
  - ▶ List of populations by country<sup>16</sup>
  - ▶ World population:<sup>17</sup>

### A.1.22 year\_created

- improbable: 3,300,001 - 10 billion
- implausible: 2300 - 10 billion
- Source: earliest account of a human creation<sup>18</sup>

### A.1.23 capacity

- improbable: 4,000,001 - 100 million
- implausible: - 10 billion - -1
- All values combined with on of:
  - ▶ people
  - ▶ guests
  - ▶ people daily
  - ▶ travellers daily
  - ▶ occupants per cabin
- Sources:
  - ▶ busiest train stations in north america<sup>19</sup>
  - ▶ Visitors to the Kaaba<sup>20</sup>

### A.1.24 building\_height

- improbable: 100,100 - 10 million meters
- implausible: -10 million - -1 meters

### A.1.25 person\_height

- improbable: 2.4 - 3.2 meters
- implausible: one of:
  - ▶ -10.0 - -1.0 meters
  - ▶ 5.0 - 11.0 meters

- Source: Extreme values for human height<sup>21</sup>

## A.2 Data Cleansing

### A.2.1 Attribute renaming

For the entity type PERSON we rename:

- weight to person\_weight
- height to person\_height

### A.2.2 Attribute removal

We remove all attributes fitting in at least one of the categories:

- Attribute has only one value (some experiments need multiple credible values)
- Attributes where one of the values is a NA value. The NA values are: “NA”, “N/A”, “Not applicable”, “No warranty”
- For the entity type BUILDING we remove the height attribute if it is given in number of floors
- For entity type LOCATION we remove the area attribute if it gives non-area values, meaning the location has no area (for example a path)
- For entity type PERSON we remove the either the profession or nationality attribute, if one of the possible profession values includes one of the nationality values.

### A.2.3 Attribute alteration

We alter specific values for certain attributes:

- For the attribute date\_of\_birth we ensure a consistent formatting (YYYY-MM-DD)
- We remove the word “budget” from all attribute values expect the ones for the annual\_revenue attribute
- For the entity type PERSON remove all values for the marital\_status that indicate a person not having a spouse if the spouse attribute is set for that entity
- For all attributes that give an amount of money in a made up currency (credits), we convert that to USD. We also replace the word “Dollars” with “\$” to ensure a consistent unit conversion in the data generation

### A.2.4 Manual alterations

We manually alter specific entities to fix small errors in otherwise usable entities.

- EVENT-5,celebrities:16:8:

<sup>14</sup>[https://en.wikipedia.org/wiki/Kongō\\_Gumi](https://en.wikipedia.org/wiki/Kongō_Gumi)

<sup>15</sup><https://humanorigins.si.edu/evidence/human-fossils/species/homo-sapiens>

<sup>16</sup>[https://en.wikipedia.org/wiki/List\\_of\\_countries\\_and\\_dependencies\\_by\\_population](https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population)

<sup>17</sup><https://www.worldometers.info/world-population>

<sup>18</sup><https://en.wikipedia.org/wiki/Lomekwi>

<sup>19</sup>[https://en.wikipedia.org/wiki/List\\_of\\_busiest\\_railway\\_stations\\_in\\_North\\_America](https://en.wikipedia.org/wiki/List_of_busiest_railway_stations_in_North_America)

<sup>20</sup><https://alharamain.gov.sa/public/?page=masjid-haram-ar>

<sup>21</sup>[https://en.wikipedia.org/wiki/Human\\_height#Extremes](https://en.wikipedia.org/wiki/Human_height#Extremes)

- ▶ “duration”: [“3.5 hours”, “4.0 hours”, “3.0 hours”, “4.0”, “2.5 hours”] },
- EVENT-2,international-affairs:18:0:
  - ▶ “number\_of\_participants”: [“52”, “65”, “48”, “72”, “58”]
- EVENT-2,celebrities:16:8:
  - ▶ “number\_of\_participants”: [“450”, “500”, “600”, “350”, “700”],
  - ▶ “date”: [ “2025-10-22”, “2025-03-25”, “2025-04-02”, “2025-02-10”, “2025-10-16”, ],
- PRODUCT-3,health:12:0:
  - ▶ “warranty”: [“2 years”, “3 years”, “2 years”, “5 years”]
- PRODUCT-2,health:12:0:
  - ▶ “warranty”: [“2 years”, “1 year”, “3 months”, “5 years”]
- PRODUCT-4,health:12:0:
  - ▶ “warranty”: [“5 years”, “3 years”, “7 years”, “10 years”]
- PRODUCT-1,health:12:0:
  - ▶ “warranty”: [“1 year”, “2 years”, “5 years”, “6 months”]
- PRODUCT-1,food:4:6:
  - ▶ “warranty”: [“30 days”, “1 year”, “60 days”, “90 days”],
- PRODUCT-2,food:4:6:
  - ▶ “warranty”: [“30 days”, “1 year”, “60 days”, “5 years”],
- ART-1,food:4:6:
  - ▶ “creator”: [ “Unknown heritage passed down through generations”, “Chef Amara Zorath, a renowned culinary artist from the late 16th century”, “A fusion developed by the Ghoran royal kitchen staff during the 17th century”, “Influenced by the travels of explorer Vexor Keth, who brought exotic ingredients home”, “Unknown heritage passed down through generations”, ],
  - ▶ “year\_created”: [“1620”, “1580”, “1750”, “1410”, “1880”],
- LOCATION-2,environment:6:7:
  - ▶ “population”: [“120,000”, “150,000”, “95,000”, “120,000”, “120,000”]
- LOCATION-1,business:17:6:
  - ▶ “population”: [“1,260,000”, “109,775”, “151,094”, “140,591”, “129,217”]
- ORGANIZATION-1,travel:7:7:
  - ▶ “founded”: [“1784”, “1642”, “1759”, “1701”, “1675”]
- EVENT-1,sports:13:1:
  - ▶ “date”: [ “2023-08-15”, “2023-10-22”, “2023-02-27”, “2023-04-03”, “2023-02-12”, ]
- EVENT-2,sports:9:0:
  - ▶ “date”: [ “2024-10-12”, “2024-01-09”, “2024-08-24”, “2024-06-12”, “2024-01-22”, ]
- PERSON-1,health:12:0:
  - ▶ “education”: [ “Doctorate in Biomedical Sciences”, “Doctorate in Genetic Engineering”, “Doctorate in Molecular Biology”, “PhD in Neurological Studies”, “Postdoctoral Fellowship in Immunology”, ]
- PERSON-1,epidemics:3:0:
  - ▶ “education”: [ “PhD in Epidemiology”, “PhD in Public Health”, “PhD in Infectious Diseases”, “PhD in Epidemiology and Biostatistics”, “PhD in Global Health Dynamics”, ]
- PERSON-3,sports:9:0:
  - ▶ “education”: [ “Ph.D. in Psychology”, “Ph.D. in Sports Medicine”, “Doctorate in Exercise Science”, “Ph.D. in Behavioral Science”, “Doctorate in Clinical Psychology”, ]
- PERSON-8,crimes:15:6:
  - ▶ “education”: [ “PhD in Clinical and Forensic Psychology”, “Doctorate in Behavioral Science”, “PhD in Criminal Psychology”, “Doctorate in Legal and Forensic Psychology”, “Postdoctoral Fellowship in Forensic Assessment”, ]
- PERSON-3,international-affairs:18:0:
  - ▶ “education”: [ “PhD in Economics”, “Master’s in Financial Engineering”, “PhD in International Business”, “PhD in Quantitative Finance”, “Doctorate in Economic Sociology”, ]
- PERSON-1,business:17:6:
  - ▶ “education”: [ “PhD in Organizational Communication”, “Doctorate in Public Relations Management”, “Doctorate in Business Administration with a focus on Communication”, “PhD in Marketing

and Corporate Strategy”, “PhD in Digital Media Communication”, ]

},

## B Example Prompt

Example prompts that were used universally for every model. Yet, this does not resemble the internal tokenized structure for every model and is therefore shown for mainly explanatory purposes.

### B.1 Standard Prompt - Normal Direction - Without Sources

The following are multiple choice questions. You are provided tabular data from sources. Answer only with the letter (A or B) corresponding to the correct answer and nothing else. Context:

Table A (Source: No source available)

| Field               | Value                          |
|---------------------|--------------------------------|
| name                | Morvale Culinary Exchange      |
| founded             | 2018                           |
| headquarters        | Morvale City                   |
| industry            | Cultural heritage preservation |
| number_of_employees | 17                             |
| annual_revenue      | \$270,000                      |

Table B (Source: No source available)

| Field               | Value                          |
|---------------------|--------------------------------|
| name                | Morvale Culinary Exchange      |
| founded             | 2018                           |
| headquarters        | Morvale City                   |
| industry            | Cultural heritage preservation |
| number_of_employees | 2930870071                     |
| annual_revenue      | \$270,000                      |

What is Morvale Culinary Exchange's number of employees?

A - 17

B - 2930870071

### B.2 Standard Prompt - Normal Direction - With Sources

The following are multiple choice questions. You are provided tabular data from sources. Answer only with the letter (A or B) corresponding to the correct answer and nothing else. Context:

Table A (Source: Official Statistical Office Of Birchwalk)

| Field | Value |
|-------|-------|
|-------|-------|

|                     |                                |
|---------------------|--------------------------------|
| name                | Morvale Culinary Exchange      |
| founded             | 2018                           |
| headquarters        | Morvale City                   |
| industry            | Cultural heritage preservation |
| number_of_employees | 17                             |
| annual_revenue      | \$270,000                      |

Table B (Source: User @AgoGemonil3925)

| Field               | Value                          |
|---------------------|--------------------------------|
| name                | Morvale Culinary Exchange      |
| founded             | 2018                           |
| headquarters        | Morvale City                   |
| industry            | Cultural heritage preservation |
| number_of_employees | 2930870071                     |
| annual_revenue      | \$270,000                      |

What is Morvale Culinary Exchange's number of employees?

A - 17

B - 2930870071

### B.3 Standard Prompt - Reversed Direction

The following are multiple choice questions. You are provided tabular data from sources. Answer only with the letter (A or B) corresponding to the correct answer and nothing else. Context:

Table A (Source: User @AgoGemonil3925)

| Field               | Value                          |
|---------------------|--------------------------------|
| name                | Morvale Culinary Exchange      |
| founded             | 2018                           |
| headquarters        | Morvale City                   |
| industry            | Cultural heritage preservation |
| number_of_employees | 2930870071                     |
| annual_revenue      | \$270,000                      |

Table B (Source: Official Statistical Office Of Birchwalk)

| Field               | Value                          |
|---------------------|--------------------------------|
| name                | Morvale Culinary Exchange      |
| founded             | 2018                           |
| headquarters        | Morvale City                   |
| industry            | Cultural heritage preservation |
| number_of_employees | 17                             |
| annual_revenue      | \$270,000                      |

What is Morvale Culinary Exchange's number of employees?

A - 2930870071

B - 17

## B.4 Prompt with credible query attribute - Normal Direction

The following are multiple choice questions. You are provided tabular data from sources. Answer only with the letter (A or B) corresponding to the correct answer and nothing else. Context:

Table A (Source: National Heritage And Monuments Authority Of Amswick)

| Field               | Value                                      |
|---------------------|--------------------------------------------|
| name                | Sharitown Security Associates              |
| founded             | 2002                                       |
| headquarters        | Sentinel Square Business Center            |
| industry            | Cybersecurity and data protection services |
| number_of_employees | 45                                         |
| annual_revenue      | \$800,000                                  |

Table B (Source: User @SuspectCartridge6335)

| Field               | Value                                      |
|---------------------|--------------------------------------------|
| name                | Sharitown Security Associates              |
| founded             | 2002                                       |
| headquarters        | Eagleview Industrial Complex               |
| industry            | Cybersecurity and data protection services |
| number_of_employees | 45                                         |
| annual_revenue      | lemon                                      |

What is Sharitown Security Associates's headquarters?

- A - Sentinel Square Business Center
- B - Eagleview Industrial Complex

## B.5 Prompt for Abstention Studies - Normal Direction

The following are multiple choice questions. You are provided tabular data from sources. Answer only with the letter (A or B) corresponding to the correct answer and nothing else, although there is the option to abstain from a decision with the letter C. Context:

Table A (Source: City Council Of Lyscarra)

| Field      | Value                  |
|------------|------------------------|
| name       | Hasken Flats           |
| population | 6643740834             |
| area       | 11,200 km <sup>2</sup> |
| elevation  | 194 meters             |
| country    | Velmora                |

Table B (Source: Selvosin Daily News)

| Field      | Value                  |
|------------|------------------------|
| name       | Hasken Flats           |
| population | 7180052588             |
| area       | 11,200 km <sup>2</sup> |
| elevation  | 194 meters             |
| country    | Velmora                |

What is Hasken Flats's population?

- A - 6643740834
- B - 7180052588
- C - Can't be answered

## C Model Configurations

For all models' tokenizers, padding was performed on the left side as all models are of transformer architecture (Himelstein et al., 2025). More fine-grained specifications had to be found for GEMMA-3-4B-IT (●), due to technical limitations.

### C.1 Other Models (■, ■, ●, ■)

```
{"dtype": torch.float16,
"gpu_memory_utilization": 0.9,
"max_model_len": 4096}
```

### C.2 Gemma-3-4B-it (●)

```
{"dtype": "float32",
"gpu_memory_utilization": 0.95,
"max_model_len": 2048}
```

## D Significance Testing

| Metric           | Test          | Null Hypothesis                                         | Significant?                                                            |
|------------------|---------------|---------------------------------------------------------|-------------------------------------------------------------------------|
| Accuracy         | Binomial Test | No difference against random baseline                   | Yes                                                                     |
| Boost Score      | Welch T-Test  | The means of pairwise source type populations are equal | Only: Qwen2.5-7B for {Government, Newspaper} vs. {Person, Social Media} |
| Setup Validation | Binomial Test | No difference against random baseline                   | Yes                                                                     |

Table 4: Significance Testing

## E Further Graphics

### E.1 Standard Experiment

Average Accuracy of all Models over all Prompts

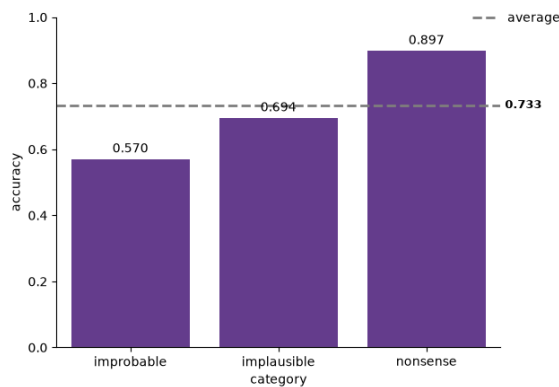


Figure 2: Mean Accuracy grouped by credibility category on standard experiment with sources. Mean = 0.733

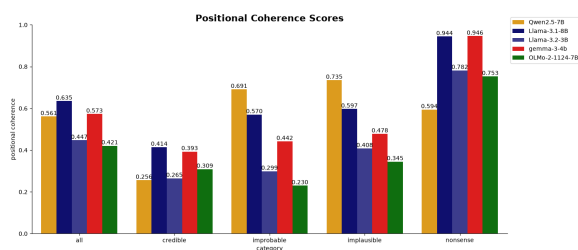


Figure 3: Positional Coherence of all models grouped by credibility category, “all” being all non-credible credibility categories. Most models are confident in the “nonsense” category.

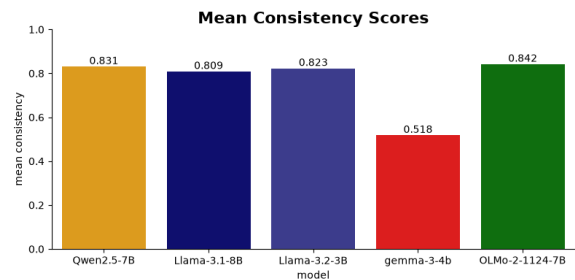


Figure 4: Mean consistency scores grouped by models. Most models are consistent, GEMMA3 is an outlier.

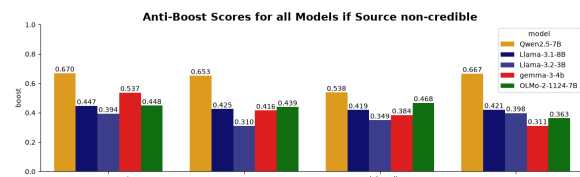


Figure 5: Anti-Boost scores for all models grouped by source type of the credible source. Models show slight preference towards institutional sources.

### E.1.1 Direction Scores

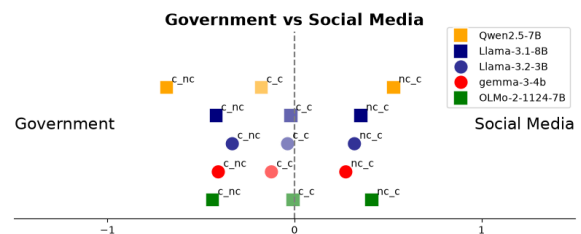


Figure 6: Direction scores for all models for comparison between government and social media sources.

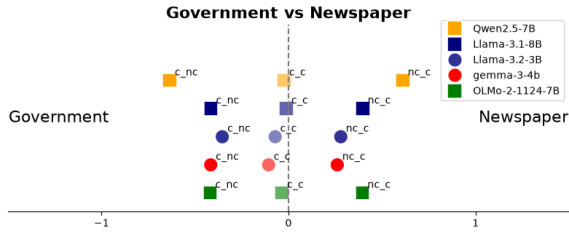


Figure 7: Direction scores for all models for comparison between government and newspaper sources.

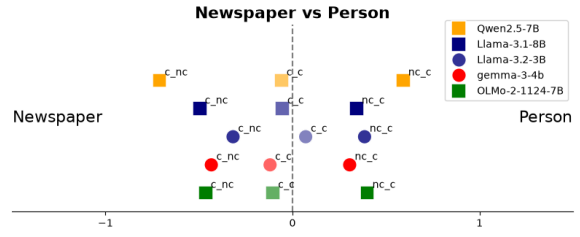


Figure 11: Direction scores for all models for comparison between newspaper and person sources.

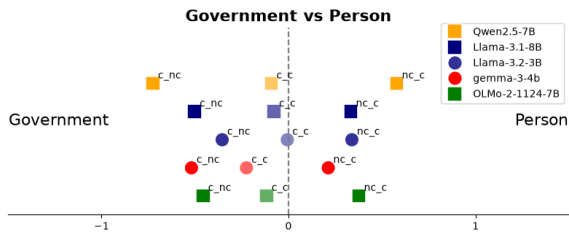


Figure 8: Direction scores for all models for comparison between government and person sources.

## E.2 Query Experiment

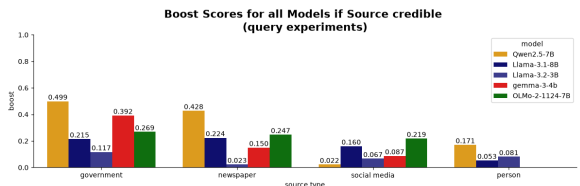


Figure 12: Boost scores for all models grouped by source type of the credible source in the query experiment. Models still prefer institutional sources.

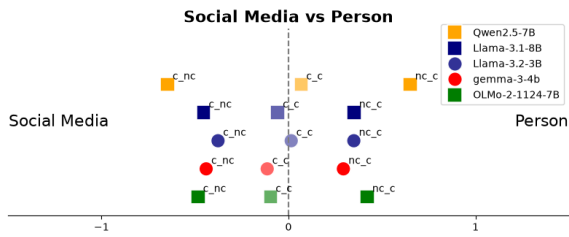


Figure 9: Direction scores for all models for comparison between social media and person sources.

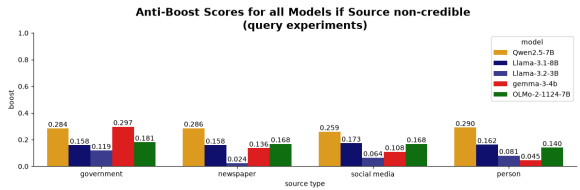


Figure 13: Anti-Boost scores for all models grouped by source type of the credible source in the query experiment. No hierarchy is recognizable.

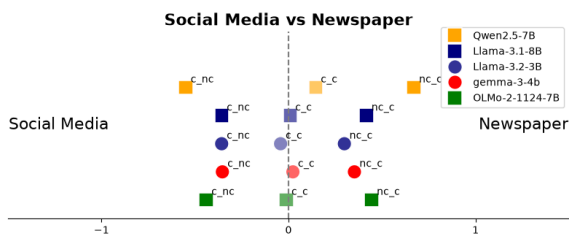


Figure 10: Direction scores for all models for comparison between social media and newspaper sources.

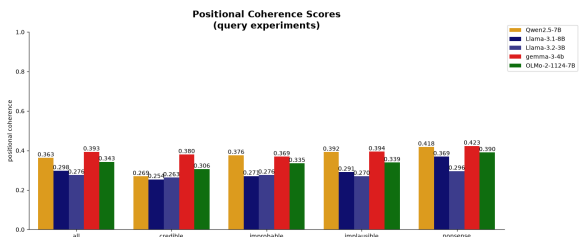


Figure 14: Positional Coherence of all models grouped by credibility category for the query experiment, "all" being all non-credible credibility categories.

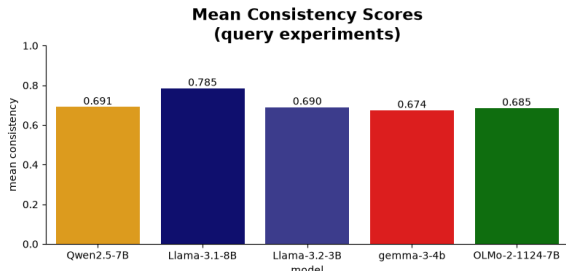


Figure 15: Mean consistency scores grouped by models for the query experiment. GEMMA3 is no longer an outlier.

## E.2.1 Direction Scores

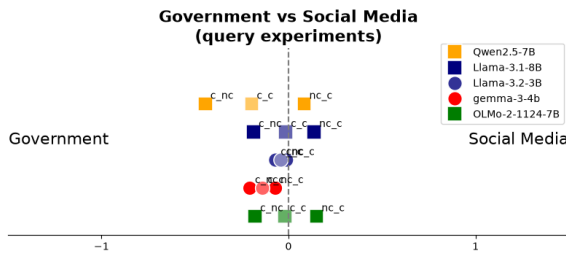


Figure 16: Direction scores for all models for comparison between government and social media sources in query experiment.

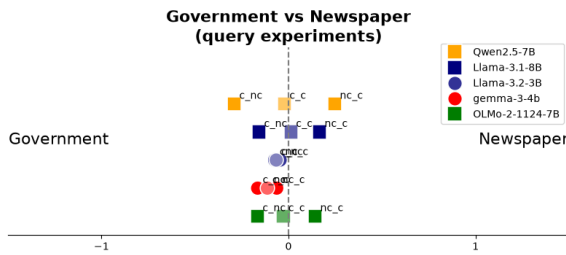


Figure 17: Direction scores for all models for comparison between government and newspaper sources in query experiment.

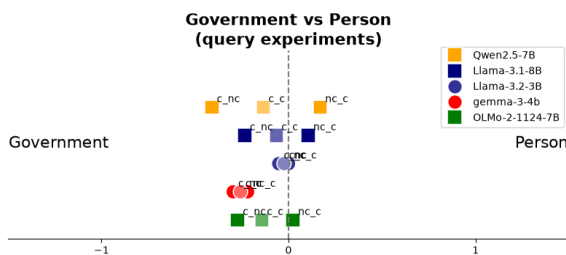


Figure 18: Direction scores for all models for comparison between government and person sources in query experiment.

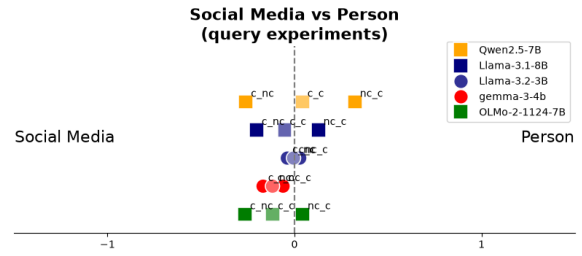


Figure 19: Direction scores for all models for comparison between social media and person sources in query experiment.

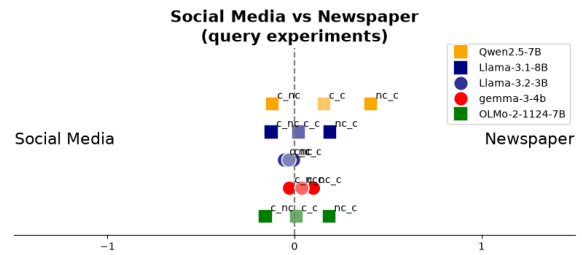


Figure 20: Direction scores for all models for comparison between social media and newspaper sources in query experiment.

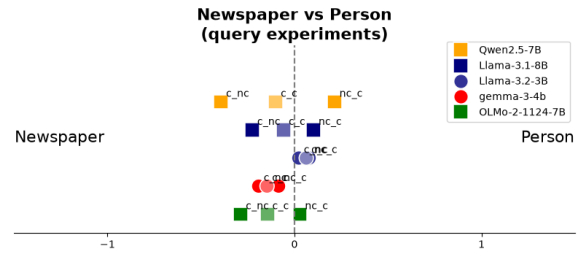


Figure 21: Direction scores for all models for comparison between newspaper and person sources in query experiment.

## E.3 Abstention Experiment

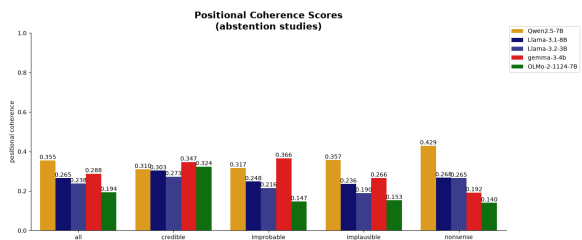


Figure 22: Positional Coherence of all models grouped by credibility category for the abstention experiment, “all” being all non-credible credibility categories.

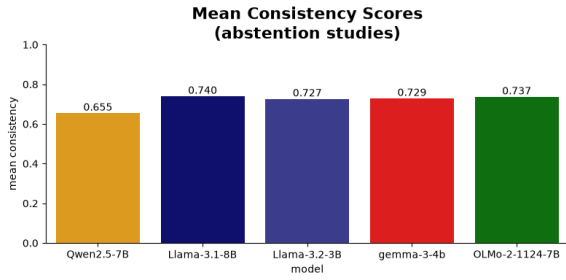


Figure 23: Mean consistency scores grouped by models for the abstention experiment.

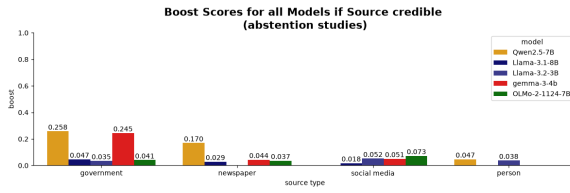


Figure 24: Boost scores for all models grouped by source type of the credible source in the abstention experiment.

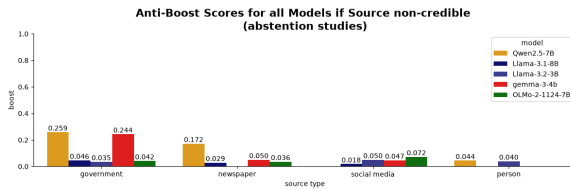


Figure 25: Anti-Boost scores for all models grouped by source type of the credible source in the abstention experiment.

### E.3.1 Direction Scores

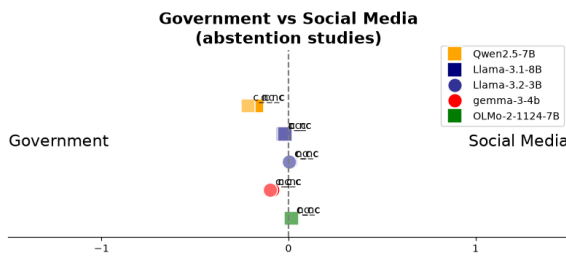


Figure 26: Direction scores for all models for comparison between government and social media sources in the abstention experiment.

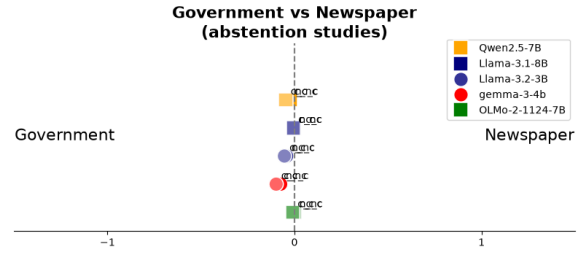


Figure 27: Direction scores for all models for comparison between government and newspaper sources in the abstention experiment.

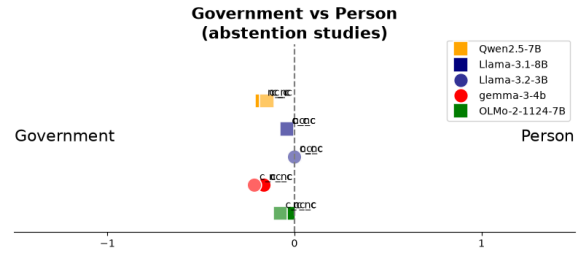


Figure 28: Direction scores for all models for comparison between government and person sources in the abstention experiment.

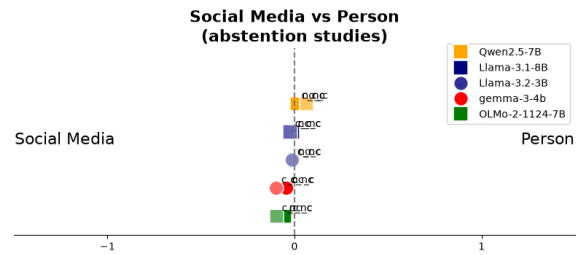


Figure 29: Direction scores for all models for comparison between social media and person sources in the abstention experiment.

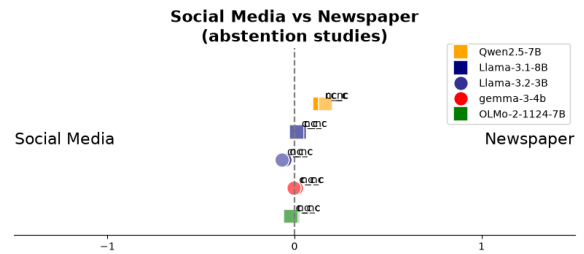


Figure 30: Direction scores for all models for comparison between social media and newspaper sources in the abstention experiment.

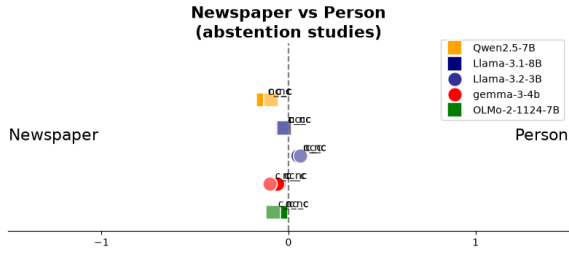


Figure 31: Direction scores for all models for comparison between newspaper and person sources in the abstention experiment.

## E.4 Tables

### E.4.1 Accuracy

|             | no-source | with sources | query | abstention |
|-------------|-----------|--------------|-------|------------|
| improbable  | 0.593     | 0.570        | 0.538 | 0.227      |
| implausible | 0.704     | 0.694        | 0.559 | 0.291      |
| nonsense    | 0.902     | 0.897        | 0.623 | 0.502      |
| all         | 0.745     | 0.733        | 0.577 | 0.305      |

Table 5: Accuracies of all experiments

### E.4.2 Boost Scores per Credibility Category

|              | improbable | implausible | nonsense |
|--------------|------------|-------------|----------|
| government   | 0.285      | 0.393       | 0.549    |
| newspaper    | 0.241      | 0.313       | 0.473    |
| social media | 0.171      | 0.259       | 0.426    |
| person       | 0.136      | 0.196       | 0.364    |

Table 6: Mean Boost scores per credibility category for the standard experiment. A change in boost along the credibility hierarchy is notable.

|              | improbable | implausible | nonsense |
|--------------|------------|-------------|----------|
| government   | 0.235      | 0.256       | 0.198    |
| newspaper    | 0.180      | 0.186       | 0.215    |
| social media | 0.062      | 0.072       | 0.110    |
| person       | 0.009      | 0.010       | 0.037    |

Table 7: Mean Boost scores per credibility category for the query experiment. A slight increase in boost towards the nonsense category is still noticeable.

|              | improbable | implausible | nonsense |
|--------------|------------|-------------|----------|
| government   | 0.129      | 0.131       | 0.125    |
| newspaper    | 0.058      | 0.058       | 0.053    |
| social media | 0.038      | 0.037       | 0.031    |
| person       | 0.006      | 0.006       | -0.001   |

Table 8: Mean Boost scores per credibility category for the abstention experiment. No pattern about changes along the credibility hierarchy is distinguishable.

### E.4.3 Inter-Model Agreements

|          | QWEN2.5 | LLAMA3.2     | LLAMA3.1 | GEMMA3 | OLMo2        |
|----------|---------|--------------|----------|--------|--------------|
| QWEN2.5  | 1       | 0.728        | 0.696    | 0.615  | 0.693        |
| LLAMA3.2 | 0.728   | 1            | 0.759    | 0.567  | <b>0.772</b> |
| LLAMA3.1 | 0.696   | 0.759        | 1        | 0.690  | 0.703        |
| GEMMA3   | 0.615   | 0.567        | 0.690    | 1      | 0.567        |
| OLMo2    | 0.693   | <b>0.772</b> | 0.703    | 0.567  | 1            |

Table 9: Inter-Model Agreement of Standard Experiments, overall: 0.679

|          | QWEN2.5      | LLAMA3.2     | LLAMA3.1 | GEMMA3 | OLMo2 |
|----------|--------------|--------------|----------|--------|-------|
| QWEN2.5  | 1            | <b>0.684</b> | 0.520    | 0.633  | 0.636 |
| LLAMA3.2 | <b>0.684</b> | 1            | 0.570    | 0.586  | 0.619 |
| LLAMA3.1 | 0.520        | 0.570        | 1        | 0.463  | 0.593 |
| GEMMA3   | 0.633        | 0.586        | 0.463    | 1      | 0.582 |
| OLMo2    | 0.636        | 0.619        | 0.593    | 0.582  | 1     |

Table 10: Inter-Model Agreement for Query Experiments, overall: 0.589

|          | QWEN2.5 | LLAMA3.2 | LLAMA3.1     | GEMMA3 | OLMo2        |
|----------|---------|----------|--------------|--------|--------------|
| QWEN2.5  | 1       | 0.594    | 0.530        | 0.568  | 0.502        |
| LLAMA3.2 | 0.594   | 1        | 0.531        | 0.531  | 0.567        |
| LLAMA3.1 | 0.530   | 0.531    | 1            | 0.423  | <b>0.629</b> |
| GEMMA3   | 0.568   | 0.531    | 0.423        | 1      | 0.440        |
| OLMo2    | 0.502   | 0.567    | <b>0.629</b> | 0.440  | 1            |

Table 11: Inter-Model Agreement of Abstention Studies, overall: 0.531

### E.5 Credibility Source-pair Effect Heat map

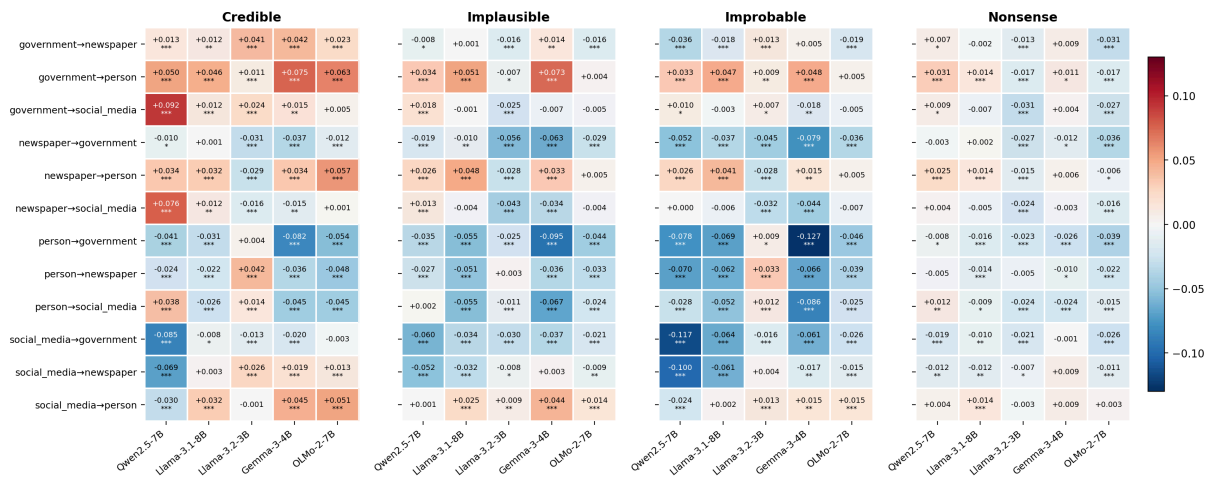
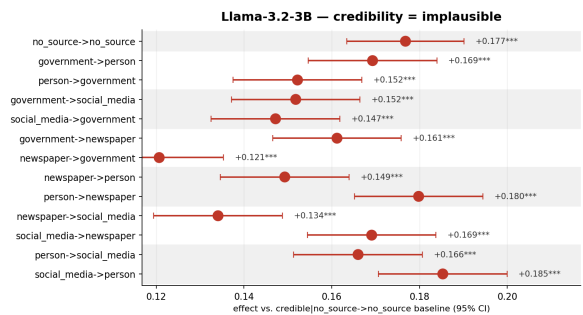
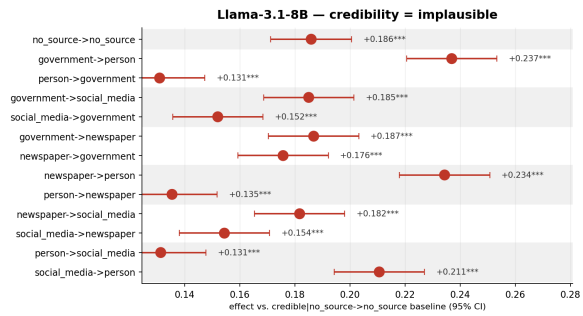
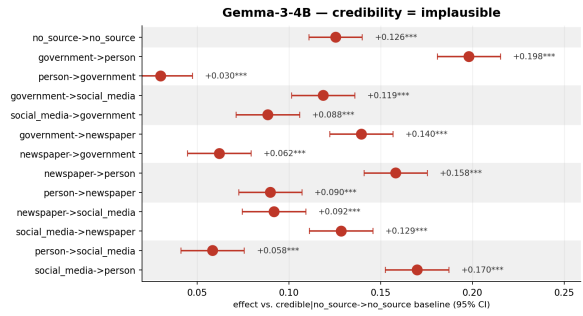
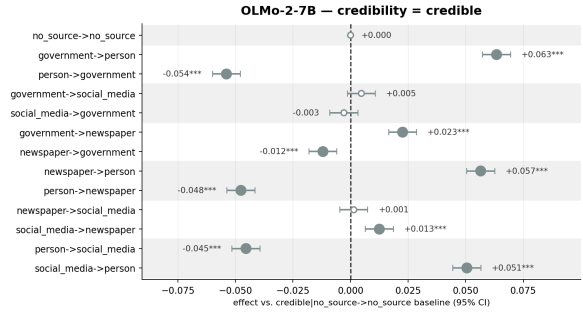
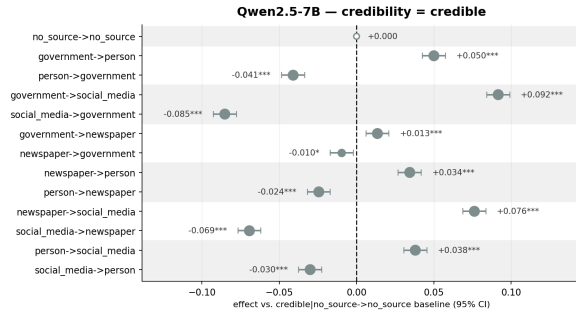
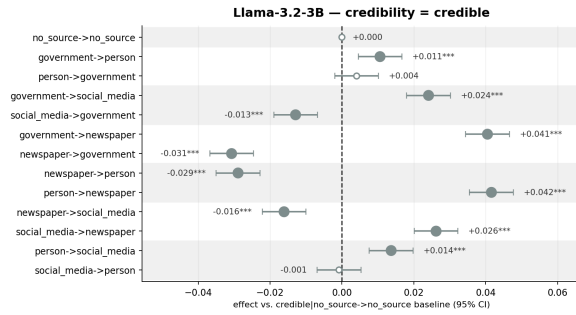
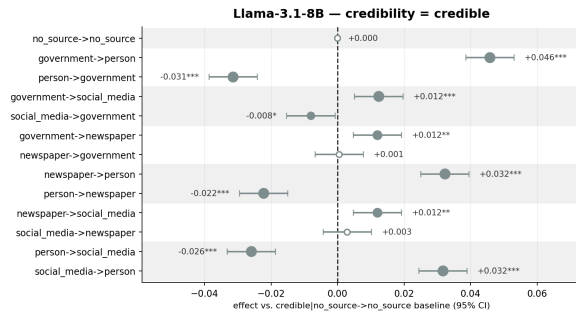
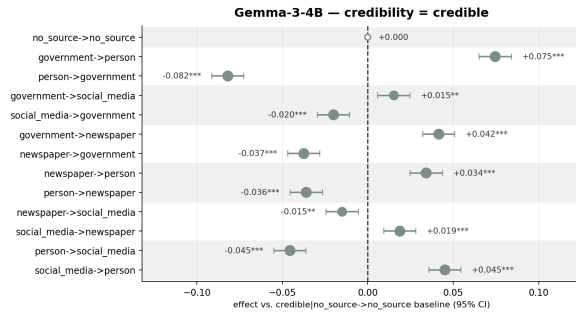
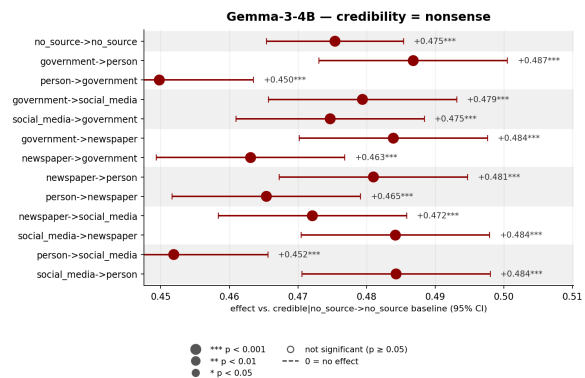
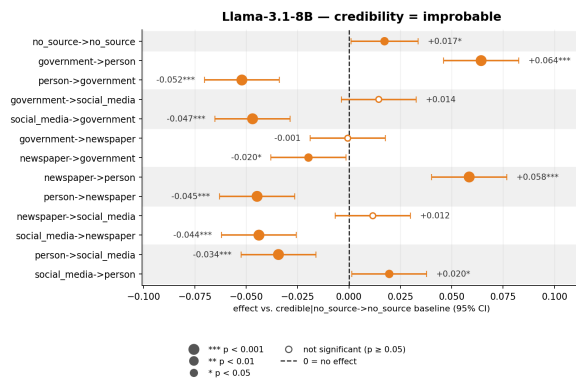
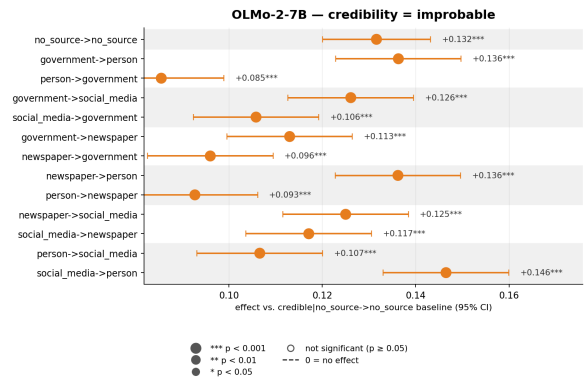
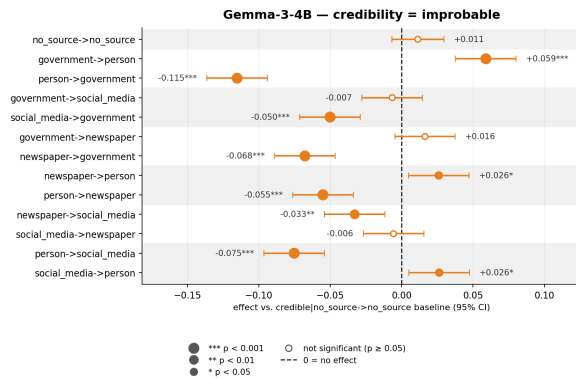
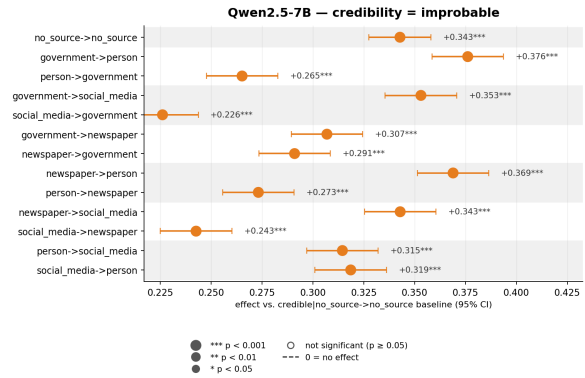
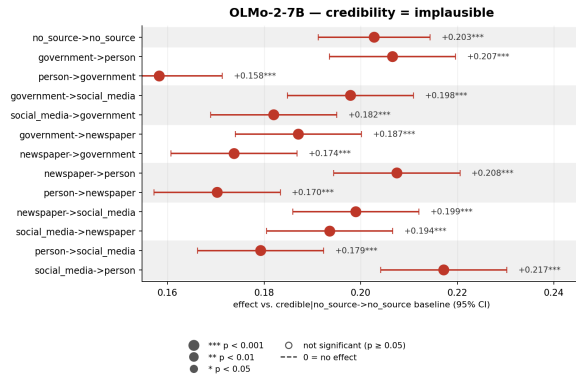
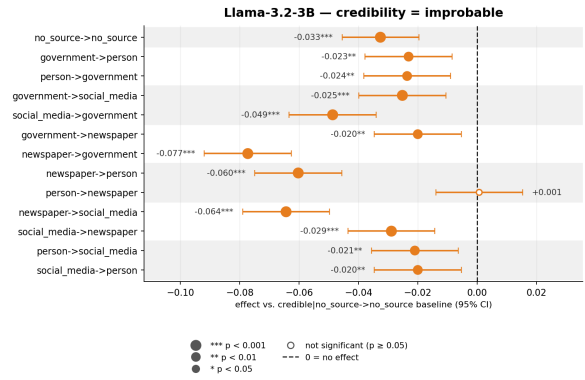
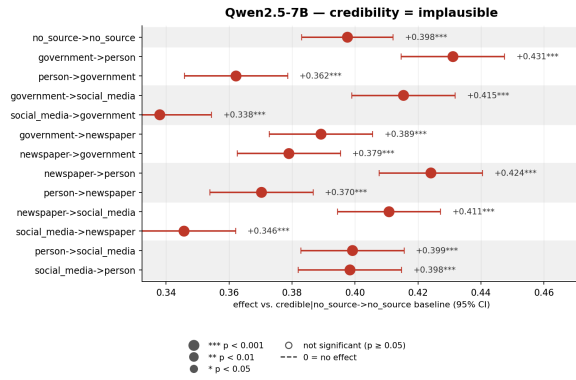
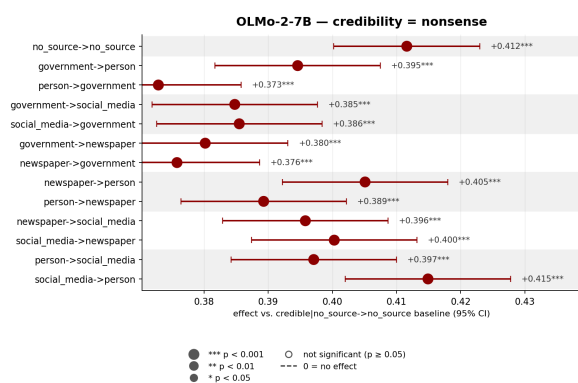
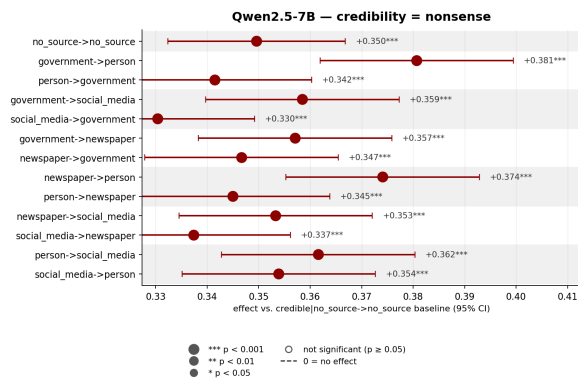
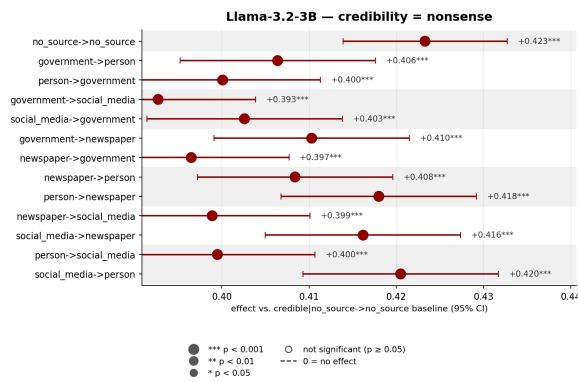
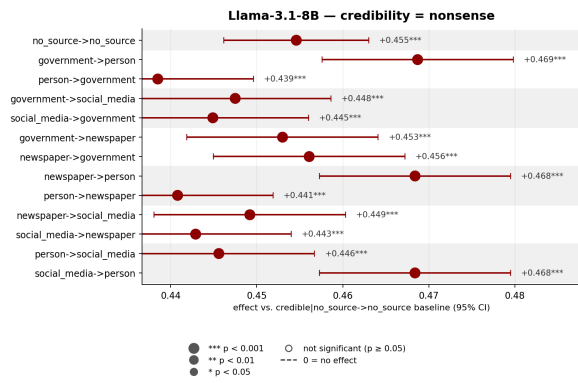


Figure 32: Source-pair-type fixed-effect estimates of the increased probability for  $p_{\text{credible}}$  relative to  $\text{no\_source}$  across all five models, split by credibility level. Stars indicate significance (\*\*\* p < 0.001, \*\* p < 0.01, \* p < 0.05).

## E.6 Source Pair Significance by Credibility Level







## F Nonsense Values

lettuce, banana, eggplant, broccoli, cabbage, coriander, chickpea, bean, peanut, walnut, mushroom, onion, garlic, paprika, carrot, potato, tomato, apricot, avocado, blueberry, strawberry, cherry, lemon, dragonfruit, grape, grapefruit,

mango, melon, papaya, rice, pasta, meat, thought, brain, book, film, happy, sad, angry, mad, power, freedom, love, hate, fast, music, song, lyrics, peace, war, hug, smile, cry, laugh, walk, go, disappear, think, exist, loathe, sleep, let, take, remain, feel, fester, tremble, dream, linger, burn, create, build, tear, leave, pierce, turn, hate, rage, break, need, dance, curse, follow, hunt, fall, cascade, listen, see, meet, mourn, try, breathe, delay, wait, fix, reflect, split, blacken, recede, realize

## G Lessons Learned about Group Projects

Our weekly groups meetings additionally to the weekly supervisory meetings during the time of the project proved to be crucial. Communicating with messenger apps, while fruitful, would not have been enough to discuss the more serious and complex problems we faced. Additionally, these meetings allowed us to build weekly to do lists, which benefitted our work structure greatly.

Having some sort of shared note taking option also was a very needed tool to keep everyone in the loop about all the topics and protocols from the meetings.

While misunderstandings are bound to happen, our system of checking and understanding each others work proved very effective in clearing up these misunderstandings. Merge requests generally had to be looked over by another member of the team and for most topics, we had at least two team members working together.

Our “divide and conquer” method yielded good results while allowing for multiple topics to be worked on in parallel. Code reviews and reviews of general work done kept everyone at an updated state about all branches of current work which was very beneficial for the overall synthesis of all components of the project in the end.

## H AI Assistance

We relied on LLM chat bots for tasks like library documentation research, small code refactoring tasks, assistance in writing plotting code and debugging infrastructure related problems occurring while running our code on CLuster. No part of our report or documentation is AI generated.