Language Bias under Conflicting Information in Multilingual LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have been shown to contain biases in the process of integrating conflicting information when answering questions. Here we ask whether such biases also exist with respect to which language is used for each conflicting piece of information. To answer this question, we extend the conflicting needles in a haystack paradigm to a multilingual setting and perform a comprehensive set of evaluations with naturalistic news domain data in five different languages, for a range of multilingual LLMs of different sizes. We find that all LLMs tested, including GPT-5.2, ignore the conflict and confidently assert only one of the possible answers in the large majority of cases. Furthermore, there is a consistent bias across models and prompting languages in which languages are preferred, with a general bias against Russian and, for the longest context lengths, in favor of Chinese. The language preferences are consistent between models trained inside and outside of mainland China, though somewhat stronger in the former category. There is also a general tendency among models to prioritize information that matches the language used for prompting. We hope to make users and developers of multilingual LLMs aware of this category of biases, to spur further research on their causes and possible mitigation.

1 Introduction

Given the enormous impact and real-world deployment of large language models (LLMs) in recent years, it is essential to evaluate how these models deal with the complexity of information that they are used to process. One particular such complexity is contradictory information, which is abundant in today's information landscape. A very common application for LLMs is to summarize information from multiple sources, and if sources disagree on some points, it is of critical importance to know how these disagreements are reflected in the output. Previous work (Kurfali and Östling, 2025) has investigated this question with respect to factors such as repetition and positioning of information in the context window given to models, by using a *conflicting needles in a haystack* paradigm where conflicting pieces of information (*needles*) are inserted into large amounts of unrelated textual information (*haystacks*). We expand on this line of research by systematically varying the *languages* of the needles and haystacks in order to test whether the language information is presented in affects its chances of making it to the final output.

Specifically, we ask the following research questions:

- RQ1: How do current multilingual LLMs behave when faced with conflicting information in a naturalistic information retrieval setting?
- RQ2: Are current multilingual LLMs biased with respect to which language they prioritize when extracting contradictory information?
- RQ3: Are any languages consistently favored or disfavored across LLMs?
- RQ4: Does the country of origin of the LLMs influence their handling of contradictory information?

We identify four major contributions of this paper, both methodological (1) and empirical on the (in)ability to deal with conflicting information (2) as well as a detailed look at how the biases made apparent by this failure are structured (3, 4).

1. We introduce the *multilingual conflicting needles in a haystack* paradigm for investigating language biases in multilingual LLMs, using manually constructed naturalistic templates that simulate a real-world information retrieval scenario.

2. Contrary to our expectation, we find that even the state-of-the-art models, such as GPT-5.2, consistently fail to detect conflicting information, even under the easiest conditions in our experiments.

3. We find that there is a clear language bias as to which of two contradictory pieces of information is chosen by a model, and that this bias is surprisingly consistent across models.

4. We find that the patterns of bias are similar between models trained in the West (North America and Europe) as compared to the East (PRC), but that there seems to be a slight bias against English in PRC-trained models only.

2 Related Work

The Needle In A Haystack (NIAH) paradigm, originally introduced by Kamradt (2023), has been used in LLM research to test information retrieval under controlled circumstances. The key idea is to build a collection of text ("haystack") into which one or more pieces of information ("needles") are inserted. An LLM is then asked to retrieve the information, and the rate of success indicates how well the LLM under test is able to deal with this particular configuration of haystack and needles. Hsieh et al. (2024) compared retrieval performance with contexts of different lengths, with the goal of mapping how usable context size compares to the architectural context size limit for common LLMs. Given their finding that many models show a substantial drop in retrieval performance with longer contexts, we systematically vary context length in our experiments. Wang et al. (2024) extended the NIAH paradigm to multimodal data, using combined textual (English) and visual haystacks, finding that the accuracy for retrieving, counting, and reasoning from visual information was considerably lower than from information in text.

Kurfali and Östling (2025) extended the NIAH paradigm to *conflicting* needles, where contradictory information is inserted into a haystack of text. They found that LLMs in general do not identify the conflict, and that in choosing which needle to retrieve, there are both model-specific biases and biases that are consistent across models. In particular, models tend to rely more on repeated needles. Schuster et al. (2026) extended this research to study how LLMs take source credibility into account when faced with conflicting information. They find that repetition of information can even make LLMs prioritize information from unknown or unreliable sources over high-credibility sources. While they were unable to mitigate this repetition bias with prompting-based approaches alone, they propose a fine-tuning method which reduces repetition bias and increases reliance on source credibility. Other mitigation approaches include training models to explicitly report sources for their claims (Shaier et al., 2024).

Lovering et al. (2025) also studied how LLMs deal with conflicting information, both in a conflicting NIAH scenario with inserted needles, and by identifying naturally occurring contradictions within documents. They found that top models can identify about 60% of contradictory needles and achieve a precision of about 50% when identifying naturally occurring contradictions. Similar results have been found in other studies focusing on direct detection of contradictions (Li et al., 2024; Tan et al., 2026). In general, this setting leads to the identification of a larger share of contradictions than the other studies referred to above (Kurfali and Östling, 2025; Schuster et al., 2026), but their focus is on using state-of-the-art models with chain-of-thought (CoT), and prompting with the explicit goal of looking for contradictions. This contrasts with our focus, which is to test how contradictions are handled during "mundane" tasks like information retrieval.

Given the different political systems, level of information control, and general cultural differences between mainland China (PRC) and what could be loosely referred to as the "west" (represented in our work by the US, Canada, and EU), it is reasonable to ask to what extent LLMs from these two areas are different. Previous work has shown that, for instance, the dominating political ideology in a country has an effect on the sentiments expressed by LLMs about prominent people (Buyl et al., 2026). We are interested in whether there are differences with respect to how different languages are treated. Wen-Yi et al. (2025) found that in spite of what one may expect given the political and cultural differences, performance of PRC-trained LLMs is mostly in line with that of "western" models on a broad sample of languages including minority languages of the PRC, Mandarin, and several major languages from Asia and Europe. We test (RQ4) whether this result also holds for the more subtle biases introduced by conflicting multilingual information.

3 Method

We follow Kurfali and Östling (2025) in inserting multiple conflicting *needles*, artificially generated pseudo-facts, into *haystacks* of authentic factual news text. Here we extend previous work by using *multilingual* haystacks, which allows us to investigate how multilingual LLMs select information depending on which language it is presented in.

A *haystack* in this work consists of up to $N = 34$ news articles of approximately 25 000 English words in total, originally in English [eng] but machine-translated[1] into Chinese [cmn], German [deu], Russian [rus] and Turkish [tur]. Simplified characters are used for Chinese, Cyrillic letters for Russian, and Latin letters for the remaining languages. In order to simulate a realistic information retrieval scenario from a multilingual collection of news articles, we construct our haystacks using articles from established news outlets. To facilitate reproducibility, we only source articles from media outlets[2] that publish content under Creative Commons licenses.

In our work, a specific haystack contains two languages, L_1 and L_2 (typically different). The two articles into which the needles are inserted are tied to L_1 and L_2, in that order, but the languages for each of the remaining $N - 2$ articles are chosen independently and pseudo-randomly from a uniform (50/50) distribution over L_1 and L_2. The permutation of the N articles is also pseudo-randomly chosen. In both cases, the pseudo-random seed is based on the needle to be inserted, which ensures that language assignments and permutations are comparable across languages for a given needle. Table 1 summarizes the variables used to determine a haystack.

We define four categories of needles. Each category has two selected articles, chosen manually to topically fit with the theme of the needle. For instance, the articles selected for category 1 needles are both about the entertainment industry. For this purpose, we only use articles with a permissive license (CC-BY) which permits rewriting the article. Figure 1 gives an example in context, and Table 2 lists the English versions of the eight needles we use. Data including translations into all languages are available in the Supplementary Materials, except for CC-NC-ND licensed articles where translation scripts are provided as translations could be considered derivative work.

Variables	Values
Category	1, 2, 3, 4
L_1, L_2	cmn, deu, eng, rus, tur
X_1, X_2	Delcroft, Quellman, Pikehart
Y	Cinderfax, Noiseweld, Motelvine, Brovencia, Clevantra, Teraluxis

Table 1: Configuration for a given haystack. The name of the entity Y as well as the language assignment (of the non-needle articles) and order of articles is chosen deterministically in a pseudo-random manner, conditioned only on the question category and pair of surnames. This implies that each pair L_1, L_2 of languages has a matching haystack configuration in all the other language pairs. Names were generated by GPT-5 and manually verified by web searches to be very rare or unused.

An LLM is given one haystack as context, and inside the haystack we insert two sentences at different locations with the same value of Y but (typically) distinct X_1 and X_2. By querying an LLM for the lead vocalist/editor/CEO/chairman of Y, we can see which of X_1 and/or X_2 is retrieved. The main novelty in this work is our use of *multilingual* haystacks, which allows us to investigate to what extent the language of the needle and the language of the prompt influence which X is retrieved. In particular, if swapping the languages that X_1 and X_2 are presented in results in the opposite X being retrieved, we interpret that as indicating a bias towards the language that the retrieved X was presented in. Figure 1 illustrates one pair of English/Chinese contrastive haystacks. If haystack (a) yields "Ashwren" while haystack (b) yields "Delcroft" when a given LLM is queried for the name of the original lead vocalist of Cinderfax, this indicates a possible bias towards Chinese in this model.

We automatically generate 900 haystacks for each model, by producing all combinations of pseudo-fact category (4), ordered pairs of surnames ($3^2 = 9$, of which 6 are conflicting), ordered pairs of languages ($6 \cdot 5/2 = 15$, of which 5 are monolingual and 10 are bilingual). Each haystack is queried using each of the five prompt languages, for a total of 4500 LLM queries per model. The 900 haystacks can be categorized as follows:

[1] We initially used the translation model of Cui et al. (2025), but found the translation quality was low for some of the languages, so used Google Translate for the results presented in this work.

[2] We downloaded articles from Wikinews (CC-BY) and ProPublica (CC-NC-ND), as well as transcripts from Democracy Now! (CC-NC-ND)

Cat.	Article	Template
1	wn250819/3	John SURNAME, the original lead vocalist of BANDNAME, praised Rondell's work on the album cover picture.
1	wn250924/4	The original lead vocalist of BANDNAME, John SURNAME, called the early death of Brett James a tragic loss for country music in the United States.
2	wn250830/1	Paul SURNAME, the editor of PAPERNAME, called the events a tragic story.
2	wn250806/5	The editor of PAPERNAME, Paul SURNAME, called the accident highly entertaining.
3	wn250918/7	The long-time CEO of COMPANYNAME, George SURNAME, disagreed with the court's decision.
3	wn250910/5	United States businessman George SURNAME, CEO of COMPANYNAME, witnessed the incident and said he is shocked by it.
4	wn250724/5	Richard SURNAME, chairman of the ORGNAME organization, expects Mitchell to resign from office soon.
4	wn250723/5	The chairman of the ORGNAME organization, Richard SURNAME, supported the government's plan.

Table 2: English needle templates used in this work, along with their category and the article/paragraph number in which they are inserted. The same eight needles have also been translated into the other four languages.

- 480 bilingual conflicting haystacks (4 questions times 6 conflicting needles times 10 language pairs times two language pair directions) which form the basis of our analysis, after dividing them into 240 contrastive pairs.
- 120 monolingual conflicting haystacks ($4 \cdot 6 \cdot 5$) which we only use as part of the analysis of conflict detection.
- 240 bilingual non-conflicting haystacks ($4 \cdot 6 \cdot 10 \cdot 2$) which we do not analyze further in this work.
- 60 monolingual non-conflicting haystacks ($4 \cdot 3 \cdot 5$) which we only use as a control to verify that models are able to perform the retrieval task under optimal conditions.

To answer the question of whether a given language is preferred over another by a model, we treat the 240 contrastive pairs as independent experiments with four possible outcomes:

1. The surname X presented in language L_1 is retrieved in both haystacks.
2. The surname X presented in language L_2 is retrieved in both haystacks.
3. The same surname X is retrieved in both haystacks of the pair, that is, the choice was based on factors other than language.
4. Retrieval failure: neither of X_1 or X_2 is retrieved in one or both of the haystacks.

Discarding the last two cases, we count the number of instances of (1) and (2) and use a binomial null model with $p = 0.5$, to account for the possibility of switching languages sometimes introducing non-systematic changes of which language's needle being retrieved, as well as inherent non-deterministic behavior during model inference. If (1) or (2) occurs significantly more often than the null model predicts, we consider this evidence of L_1 being preferred over L_2 (1), or vice versa (2).

4 Experimental Setting

We performed the experiments on a set of 12 LLMs (listed in Table 3 and described in more detail in Appendix A). We used greedy decoding or a sampling temperature of 0 when applicable, aiming at getting as close to a deterministic results as possible given existing infrastructure. For the OpenAI models, we used their API, and for all other models we ran the experiments on a single H100 GPU. In total, we used approximately 100 GPU hours and USD 331.53 of API credits. For chain-of-thought models, we used the lowest setting of "thinking effort". For chat models, we constructed a single long "user" message consisting of the haystack, a blank line, and the question. For non-chat models, we used the simple template "Question: [question text]", followed by a blank line and "Answer:".

Code and full results data files are provided as Supplementary Materials.

[...] It took a long time to persuade Ronnie to stand exactly as I wanted but in the end he was very brave and it was a perfect composition." John **Delcroft**, the original lead vocalist of **Cinderfax**, praised Rondell's work on the album cover picture. He sustained numerous injuries throughout his career, including but not limited to concussions and broken ribs.
[...]
Jesus Take the Wheel 荣获格莱美"最佳乡村歌曲"奖。它还被提名为"年度歌曲"。 **Cinderfax**的原主唱John **Ashwren**称Brett James的早逝是美国乡村音乐的悲剧性损失。 [...]

(a) L_1 = English, L_2 = Chinese

[...] 我花了很长时间才说服罗尼完全按照我想要的方式站着，但最终他非常勇敢，这是一张完美的构图。" **Cinderfax**乐队原主唱John **Delcroft**对Rondell的专辑封面图片创作工作表示赞赏。他在整个职业生涯中遭受过多次伤害，包括但不限于脑震荡和肋骨骨折。
[...]
Jesus Take the Wheel won a Grammy award for "Best Country Song." It was also nominated for "Song of the Year." The original lead vocalist of **Cinderfax**, John **Ashwren**, called the early death of Brett James a tragic loss for country music in the United States. [...]

(b) L_1 = Chinese, L_2 = English

Figure 1: Excerpts from a pair of contrastive haystacks with Y = Cinderfax, X_1 = Delcroft, X_2 = Ashwren. Each haystack contains up to about 25 000 words. The contents of (a) and (b) are identical except that the languages are swapped for the two needle sentences and the article they are integrated to. In the excerpts, only parts of two articles are shown. The remaining articles are in the same order and language across the two haystacks. Light gray indicates authentic news text that is part of the context (most of which is left out for brevity), while the text in black are generated from our manually constructed needle templates described in Table 2.

5 Results and Discussion

5.1 RQ1: handling of conflicting information

Starting with our first research question, we begin by looking at how the LLMs we study are able to deal with conflicting information overall. Table 3 shows that all LLMs are unable to reliably identify conflicting information in the haystacks.[3] For small haystacks (approximately 1 000 words) almost all queries result in a single answer without mentioning the alternative. With large haystacks (approximately 25 000 words) the main qualitative change is that, with some models, a larger proportion of queries fail to give *any* of the two correct answers. Results for monolingual haystacks are found in Appendix B (Table 10), and are qualitatively very similar.

Our prompt specifies the expected output format by adding the sentence "Answer with only the full name." This decision was made to simulate a typical prompting-based information retrieval scenario, where the user does not want any additional information besides the name. In order to test whether this sentence restricts the model too much, we also repeated our experiments for some settings with the sentence removed, leaving the task more open. Although the number of detected contradictions increases, as expected, we still observe the same qualitative pattern: a large majority of conflicting needles go undetected. A limited set of experiments with this reduced prompt are summarized in Appendix D, but apart from this we refrain from further prompt exploration due to the associated computational cost.

5.2 RQ2: presence of language bias

Given that our results for RQ1 indicate that models overwhelmingly ignore contradicting information and present a single answer, we now turn to RQ2 which asks whether or not the choice of which information to report depends on the language it is presented in. We begin to answer this by looking at how often, over all prompting languages and contrastive haystack pairs, information in one language is chosen over another (full results are found in Appendix C, with additional haystack sizes in Appendix F). Since we work with contrastive pairs of haystacks where all other parameters apart from the languages of the needles (and their immediate context) are equal, we assume that without systematic impact of the language order we would see a binomial distribution for a given pair of languages.

For most of the models we see that information in Russian is ignored significantly above chance level (column *rus* of each sub-table). Apart from this, several models display a preference for Chinese (row *cmn* for GEMMA-3-27B, LLAMA-3.1-8B, GPT-5-NANO, MINISTRAL-8B, GLM-4-9B, YI-1.5-9B). There are also considerable differences in the overall level of language bias, which

[3] Due to the large number of queries, we classify the outputs automatically using the heuristic that if both valid answers are found in the output, this likely represents a refusal due to identifying the conflict. While in theory it is possible for a model to correctly mention the existence of a conflict without specifying the alternatives, during our manual inspection of LLM outputs we have so far not found any exceptions to this heuristic.

Model	1 000 words			25 000 words		
	Both	None	One	Both	None	One
GEMMA-3-27B-IT	14	0	2386	0	178	2222
GEMMA-3-4B-IT	22	22	2356	0	676	1724
LLAMA-3.1-8B-INSTRUCT	9	49	2342	0	41	2359
GPT-5.2-2025-12-11	0	0	2400	13	2	2385
GPT-5-MINI-2025-08-07	23	0	2377	12	0	2388
GPT-5-NANO-2025-08-07	1	2	2397	0	6	2394
C4AI-COMMAND-R7B-12-2024	4	16	2380	0	52	2348
MINISTRAL-8B-INSTRUCT-2410	0	22	2378	0	70	2330
MISTRAL-NEMO-INSTRUCT-2407	0	28	2372	0	1061	1339
GLM-4-9B-0414	0	9	2391	2	773	1625
QWEN3-4B	9	75	2316	0	203	2197
YI-1.5-9B-32K	3	32	2365	0	1316	1084

Table 3: Summary of outcomes from all conflicting multilingual haystack retrievals. We report the number of times that the model correctly identifies **both** answers, or retrieval fails so that **none** of the answers is identified, or reports only **one** of the two possible answers without mentioning the other.

is most clearly seen in the sums in the rightmost column, where C4AI-COMMAND-R7B has an almost flat distribution over languages whereas for instance GPT-5-MINI has 221 instances of English being picked over other languages, compared to just 38 for Russian.

Note that there are a total of 240 contrastive pairs times 5 prompt languages, which makes the maximum sum for a language pair $240 \times 5 = 1200$. We observe that in all cases, counts are considerably lower than that, with a maximum of $93 + 4 = 97$ for English/Russian with the GPT-5-MINI model, and only somewhat higher in some of the reduced-haystack complementary experiments reported in Appendix F. This implies that factors other than language seem to determine which needle is chosen in at least 90% of cases. Previous research indicates that these factors include for example the form of (novel) names, since models often prefer the form of certain words over others when answering questions on the basis of conflicting information (Kurfali and Östling, 2025). The significance of our results here is that in the subset of cases where language is the decisive factor, there are clear biases towards certain languages.

While we are mainly interested in the effect of the language that *information* is presented in, we also need to take the language used for *prompting* into account. For several of the models we see a very clear bias towards preferring information in the prompting language. This is particularly clear in models like C4AI-COMMAND-R7B-12-2024, which has a low overall level of language bias when averaged over prompt languages, but a substantial bias towards the prompting language. These results show that language biases are even more severe when conditioned on the prompting language. For instance, of the cases when GPT-5-NANO-2025-08-07 chooses the Russian needle over other languages, this is nearly always when prompting in Russian. A full table is available as Table 12 in Appendix C.

5.3 RQ3: language bias across models

To test whether some languages are favored over the other languages of our sample, in a way that is consistent across different LLMs, we fit a hierarchical statistical model using Stan (Stan Development Team, 2026).

$$
\begin{aligned}
s &\sim \text{half-Normal}(0, 100) \\
b_{l_1,l_2} &\sim \text{Normal}(0, 100) \\
b_{l_1,l_2,l_p,m} &\sim \text{Normal}(b_{l_1,l_2}, s) \\
w_{l_1,l_2,l_p,m} &\sim \text{Bernoulli}(\sigma(b_{l_1,l_2,l_p,m}))
\end{aligned}
$$

where l_1 and l_2 are two different languages that make up a subset of contrastive pairs of bilingual haystacks, l_p is the language used for prompting the model m, $w_{l_1,l_2,l_p,m}$ is the binary outcome (win for l_1 or for l_2) for a given contrastive pair when model m is prompted with language l_p. Our main interest is in the parameter b_{l_1,l_2}. This represents the log-odds overall bias across models and prompting languages for choosing information from l_1 over l_2. We have chosen to treat models as independent, even though some of them are from the same

Language pair l_1, l_2	$P(>0)$	95% CI
Chinese vs German	99.0%	$[0.11, 1.52]$
Chinese vs English	96.6%	$[-0.04, 1.28]$
Chinese vs Russian	100.0%	$[2.54, 4.17]$
Chinese vs Turkish	88.4%	$[-0.27, 1.16]$
German vs English	58.7%	$[-0.62, 0.79]$
German vs Russian	100.0%	$[1.74, 3.36]$
Turkish vs German	64.1%	$[-0.60, 0.88]$
English vs Russian	100.0%	$[2.01, 3.64]$
Turkish vs English	77.3%	$[-0.44, 0.97]$
Turkish vs Russian	100.0%	$[2.39, 4.12]$

Table 4: Estimates of the log-odds bias parameter b_{l_1,l_2}, representing the overall bias (over all prompting languages and models) of language l_1 over language l_2. Haystack size is 25 000. CI = Bayesian credibility interval.

Language pair l_1, l_2	$P(>0)$	95% CI
Chinese vs German	87.5%	$[-0.27, 1.05]$
English vs Chinese	52.8%	$[-0.61, 0.66]$
Chinese vs Russian	100.0%	$[2.53, 3.96]$
Chinese vs Turkish	93.7%	$[-0.14, 1.16]$
English vs German	100.0%	$[0.91, 2.33]$
German vs Russian	100.0%	$[2.43, 3.81]$
German vs Turkish	96.2%	$[-0.07, 1.24]$
English vs Russian	100.0%	$[2.68, 4.16]$
English vs Turkish	100.0%	$[0.67, 1.98]$
Turkish vs Russian	100.0%	$[2.33, 3.72]$

Table 5: Estimates of the log-odds bias parameter b_{l_1,l_2}, representing the overall bias (over all prompting languages and models) of language l_1 over language l_2. Haystack size is 1000. CI = Bayesian credibility interval.

family (we have two Llama models and three GPT models). We feel that this is justified, given that we have a relatively small sample of model and that previous work has found that models from the same family can behave very differently in this task (Kurfali and Östling, 2025).

We fit one model for each haystack size. Table 4 summarizes the posterior of b_{l_1,l_2} for large (up to 25 000 words) haystacks, and Table 5 for small (up to 1000 words) haystacks. Tables for intermediate sizes can be found in Appendix E. In our sample of languages, we see that information in Russian is strongly dispreferred in comparison to all of the remaining languages, and that this applies to all haystack sizes. For the longest haystacks, which is the most challenging version of the task, there is also strong evidence that Chinese is preferred over the other languages, with the possible exception of Turkish.

5.4 RQ4: language bias across geographic areas

Finally, we ask whether the origin of each LLM has an effect on its language bias. To answer this, we extend the model to include a variable for the origin o_m of a particular model, and add the origin o as an additional index to the bias parameter b_{o,l_1,l_2}.

$$
\begin{aligned}
s &\sim \text{half-Normal}(0, 100) \\
b_{o,l_1,l_2} &\sim \text{Normal}(0, 100) \\
b_{l_1,l_2,l_p,m} &\sim \text{Normal}(b_{o_m,l_1,l_2}, s) \\
w_{l_1,l_2,l_p,m} &\sim \text{Bernoulli}(\sigma(b_{l_1,l_2,l_p,m}))
\end{aligned}
$$

For ease of analysis and due to the limited number of models from each origin, we make a coarse classification only consider two origins: West (USA, Canada, France) and East (PRC).

The language biases are shown separately in Table 7 (eastern models) and Table 8 (western models). In addition, the difference between these tables are shows in Table 6. Overall, the general patterns are similar, with a bias against Russian for models from both origins, and a tendency of a bias towards Chinese. In both cases, the pattern is stronger for the eastern models than for the western. One difference to note is that the eastern models have a weak overall bias against English, that is not found in the western models. There is not a single case of Russian being favored over Turkish in any of the eastern models, and combined with our weak prior this leads to a very high estimated bias of Turkish over Russian.

Altogether, there is some evidence of models favoring the dominant language (Chinese and English, respectively) of their origin. On the other hand, there also seems to be a common (in our sample) bias towards Chinese, and a common bias against Russian.

6 Discussion and conclusions

We found that all LLMs under study were surprisingly consistent in their behavior of asserting a single answer in the presence of conflicting information (RQ1). Although there are some differences between conditions, the same general behavior is observed for both monolingual and multilingual haystacks, across all five languages and 12 models, with different prompts and widely different context lengths. This is a surprising and worrying mode of

	cmn	deu	eng	rus	tur
cmn		0.9	1.1	**2.6**	0.7
deu	-0.9		1.3	2.0	0.1
eng	-1.1	-1.3		0.5	-1.5
rus	**-2.6**	-2.0	-0.5		**-77.0**
tur	-0.7	-0.1	1.5	**77.0**	

Table 6: Difference in bias among eastern models over western models for choosing l_1 (row) over l_2 (column). A *row* with high values indicates a preference for the language of that row by eastern models, while a *column* with high values indicates a preference of the column language among western models. Text weight indicates model confidence in the sign of the difference: **Bold**: $p > 0.99$, black: $p > 0.90$, gray: $p \leq 0.90$. Values presented are posterior medians. Haystack size is 25 000.

	cmn	deu	eng	rus	tur
cmn		1.5	1.4	**5.5**	1.0
deu	-1.5		1.1	**4.3**	-0.0
eng	-1.4	-1.1		**3.2**	-1.4
rus	**-5.5**	**-4.3**	**-3.2**		**-79.9**
tur	-1.0	0.0	1.4	**79.9**	

Table 7: Bias among eastern models for choosing l_1 (row) over l_2 (column). All other details are identical to Table 6.

	cmn	deu	eng	rus	tur
cmn		0.6	0.3	**2.9**	0.3
deu	-0.6		-0.3	**2.2**	-0.2
eng	-0.3	0.3		**2.7**	0.1
rus	**-2.9**	**-2.2**	**-2.7**		**-2.8**
tur	-0.3	0.2	-0.1	**2.8**	

Table 8: Bias among western models for choosing l_1 (row) over l_2 (column). All other details identical to Table 6.

failure, in a task of great practical importance that intuitively would seem trivial compared to many of the more complex tasks that LLMs excel at. It falls beyond the scope of this study to identify the mechanisms behind this behavior, but we identify this as an important area of future work.

We proceeded to ask whether the language that a piece of information is presented in influences the chances of that information to make it into the final LLM answer (RQ2). Based on pairs of haystack configurations that differ only in the languages that each piece of information is presented in, we saw that while language was typically not the decisive factor when selecting among conflicting information, in the cases where this happened there is in many cases a strong tendency for models to prefer one language over another. Furthermore, we found that the language preferences are to a large extent consistent across LLMs (RQ3), and even for LLMs trained in different politico-cultural spheres (RQ4). The strongest pattern is the nearly universal dispreference for information presented in Russian, even though there is no corresponding failure of retrieval in *monolingual* Russian haystacks, which means that results can not simply be explained by models being worse at processing information in Russian. Given our relatively small sample of languages (5), we are unable to tell whether this is unique to Russian or applies to a larger category of languages. One might hypothesize that orthography contributes, since Russian is written in Cyrillic script unlike the majority of Latin-orthography languages, but this is contradicted by the fact that the other universal pattern is a *preference for* the other non-Latin orthography in our data, Chinese. Similarly, our results can hardly be explained by the size of training data alone, when Turkish is often preferred over English. This raises important questions that we leave to future work: what is the pattern of language preferences at a larger scale, and what are the mechanisms that cause this bias? In order to fully answer these questions, one would need a larger sample of languages and of models trained in different parts of the world, as well as the application of methods for investigating the internal representations of models (e.g., Fraser-Taliente et al., 2026).

Our findings mostly agree with those of Wen-Yi et al. (2025), who found no major differences in the multilingual abilities of PRC-trained LLMs, and those trained elsewhere. This stands in contrast to other work (Buyl et al., 2026) which found considerable differences with respect to politically important content. We hypothesize that this difference is mainly due to a general neglect during model training of question answering under conflicting information, especially in multilingual settings. Another important direction of future work is to investigate how changes in model training can either strengthen or weaken the types of language bias we have identified, as well as pinpointing the mechanisms that contribute to unbalanced weighing of information.

Limitations

All articles were originally in English, and predominantly from English-speaking areas. The first names are all typical of males from 1940s Liverpool, and the surnames are chosen to be English-sounding. A broader selection of both needle and haystack data would have allowed investigating the interaction between the topics discussed and the language used for the discussion.

The sample of languages and models are both relatively small, compared to the total number of languages and LLMs. This makes it difficult or impossible to generalize findings beyond our particular sample of languages. In addition, the samples are not as diverse as one would like, with three out of five languages being Indo-European and the majority of LLMs in the 3–9B parameter range.

For comparability with general LLMs and because of the limited compute budget typical for information retrieval tasks, our experiments set 'reasoning' (chain-of-thought, CoT) models to use a minimal amount of CoT effort. However, Fu et al. (2026) found in a broader study on LLM safety that increasing CoT tokens was a strong predictor of identifying harmful intent. It is therefore important to note that our findings may not hold for state-of-the-art models with maximum CoT effort. If a large compute budget is available and focus is set on identifying contradictions, the results by Lovering et al. (2025) indicate that state-of-the-art models can have a reasonably high rate (somewhat over 50%) of identifying conflicting information.

Although we investigate the effect of prompting language as part of our answer to RQ2, we abstain from performing the rest of the analyses in this paper separately for the five different prompting languages. By instead averaging over all prompting languages, we obtain a somewhat conservative estimate of the amount of language bias, since biases due to the prompting language are averaged out.

We note that the results of our research could potentially be applied to harmful activities such as promoting disinformation. By careful selection of which language the disinformation is presented in, and considering the languages likely used for conflicting (factual) information and LLM prompting, it is possible to increase the likelihood of disinformation being chosen in LLM-generated summaries or answers. Yet, we believe that the advantages of publishing this research outweigh the disadvantages, because only by bringing attention to the problem can we spur research on effective countermeasures.

References

01AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, and 13 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, and 1 others. 2026. Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2(1):7.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. 2026. Natural language autoencoders produce unsupervised explanations of llm activations. *Transformer Circuits Thread*.

Yu Fu, Haz Sameen Shahgir, Huanli Gong, Zhipeng Wei, N. Benjamin Erichson, and Yue Dong. 2026. Is reasoning capability enough for safety in long-context language models? *Preprint*, arXiv:2602.08874.

GemmaTeam, 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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

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.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms. `https://github.com/gkamradt/needle-in-a-haystack`. GitHub repository.

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.

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.

Charles J. Lovering, Seth Ebner, Brandon Smock, Michael Krumdick, Saad Rabbani, Ahmed Muhammad, Varshini Reddy, and Chris Tanner. 2025. On finding inconsistencies in documents. *Preprint*, arXiv:2512.18601.

Jakob Schuster, Vagrant Gautam, and Katja Markert. 2026. Whose facts win? llm source preferences under knowledge conflicts.

Sagi Shaier, Ari Kobren, and Philip V. 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, Miami, Florida, USA. Association for Computational Linguistics.

The Stan Development Team. 2026. Stan reference manual, 2.38.0.

Nelvin Tan, Yaowen Zhang, James Asikin Cheung, Fusheng Liu, Yu-Ching Shih, and Dong Yang. 2026. Improved evidence extraction for document inconsistency detection with llms. *Preprint*, arXiv:2601.02627.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024. Needle in a multimodal haystack. In *Advances in Neural Information Processing Systems*, volume 37, pages 20540–20565. Curran Associates, Inc.

Andrea W Wen-Yi, Unso Eun Seo Jo, Lu Jia Lin, and David Mimno. 2025. How chinese are chinese language models? the puzzling lack of language policy in china's llms. *Preprint*, arXiv:2407.09652.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Models used in the study

A summary of the LLMs included in our study is given in Table 9. These include three GPT-5 family models[4], two sizes of Gemma 3 models (GemmaTeam et al., 2025), one Llama-3.1 model (Grattafiori et al., 2024), Command A (Cohere et al., 2025), Mistral NeMo[5] and Ministral [6], GLM-4 (GLM et al., 2024), Qwen 3 (Yang et al., 2025), and Yi-1.5 (01AI et al., 2025).

B Monolingual haystack retrieval summary

In Table 10 we show the retrieval rates in monolingual haystacks. This complements Table 3 in the main text and demonstrates that results are qualitatively similar to the multilingual setting.

C Full tables of language bias results

Here we present the full tables of results for the largest (25 000 words) haystack size configuration:

- Number of wins for each pair of (L_1, L_2) summed over all prompting languages (Table 11).
- Ratio of wins for each L_1 summed over all L_2, shows separately for each prompting language (Table 12).

D Results with reduced prompt

Table 13 presents outcome statistics for a selection of representative configurations using a prompt that omits the final sentence ("Answer with only the full name of [person]." and its translations into the other languages). The intention is to show that even state-of-the-art models in the easiest setting (short haystack) confidently assert a single answer in the face of contradictory information.

E Additional language bias values

Here we present the estimates of b_{l_1,l_2} for intermediate haystack sizes: 2500 words (Table 14, 5000 words (Table 15, and 10000 words (Table 16. The estimates for the smallest (1000) and largest (25 000) are found in the main text.

[4] https://openai.com/index/introducing-gpt-5/
[5] https://mistral.ai/news/mistral-nemo
[6] https://mistral.ai/news/ministraux/

F Additional pairwise bias tables

Here we present tables of the number of wins for each model and language pair, for intermediate haystack sizes: 1000 words (Table 17), 2500 words (Table 18), 5000 words (Table 19), 10000 words (Table 20).

G Non-conflicting monolingual retrieval

Table 21 shows the retrieval rates for the 12 attempts per needle language and prompt language in the easiest of our settings, non-conflicting monolingual haystacks. All models are able to achieve a perfect result in at least one language (typically English and/or Chinese), although some models struggle with certain needle and/or prompt languages.

Model	Size	Origin
gpt-5.2	?	USA
gpt-5-mini	?	USA
gpt-5-nano	?	USA
gemma-3-4b-it	4B	USA
gemma-3-27b-it	27B	USA
Llama-3.1-8B-Instruct	8B	USA
c4ai-command-r7b-12-2024	7B	Canada
Ministral-8B-Instruct-2410	8B	France
Ministral-Nemo-Instruct-2407	12B	France
GLM-4-9B-0414	9B	PRC
Qwen3-4B	4B	PRC
Yi-1.5-9B-32K	9B	PRC

Table 9: Large Language Models (LLMs) used in this study.

	1 000 words			25 000 words		
Model	Both	None	One	Both	None	One
Monolingual haystacks						
GEMMA-3-27B-IT	21	0	579	0	49	551
GEMMA-3-4B-IT	8	7	585	0	189	411
LLAMA-3.1-8B-INSTRUCT	7	15	578	0	6	594
GPT-5.2-2025-12-11	2	0	598	10	8	582
GPT-5-MINI-2025-08-07	4	0	596	12	2	586
GPT-5-NANO-2025-08-07	0	0	600	0	4	596
C4AI-COMMAND-R7B-12-2024	1	6	593	1	10	589
MINISTRAL-8B-INSTRUCT-2410	0	3	597	0	40	560
MISTRAL-NEMO-INSTRUCT-2407	0	16	584	0	289	311
GLM-4-9B-0414	0	3	597	0	171	429
QWEN3-4B	2	19	579	0	52	548
YI-1.5-9B-32K	0	29	571	0	327	273

Table 10: Summary of outcomes from all conflicting haystack retrievals in the monolingual setting. We report the number of times that the model correctly identifies **both** answers, or retrieval fails so that **none** of the answers is identified, or reports only **one** of the two possible answers without mentioning the other.

	cmn	deu	eng	rus	tur	Σ
GEMMA-3-27B-IT						
cmn	–	20	**35**	**64**	23	142
deu	6	–	9	31	6	52
eng	9	14	–	**53**	7	83
rus	5	13	7	–	12	37
tur	11	15	14	**43**	–	83
GEMMA-3-4B-IT						
cmn	–	5	17	**36**	13	71
deu	9	–	13	**26**	12	60
eng	6	7	–	**32**	7	52
rus	2	2	3	–	2	9
tur	2	2	9	**27**	–	40
LLAMA-3.1-8B-INSTRUCT						
cmn	–	24	**48**	**65**	19	156
deu	14	–	**18**	**49**	13	94
eng	7	3	–	**49**	5	64
rus	16	9	20	–	6	51
tur	11	23	**25**	**53**	–	112
GPT-5.2-2025-12-11						
cmn	–	11	5	39	26	81
deu	**38**	–	13	**60**	**37**	148
eng	**72**	**50**	–	**80**	**71**	273
rus	23	13	6	–	25	67
tur	27	14	5	34	–	80
GPT-5-MINI-2025-08-07						
cmn	–	17	5	**62**	21	105
deu	40	–	12	**70**	36	158
eng	**51**	**37**	–	**93**	**40**	221
rus	14	5	4	–	15	38
tur	26	15	15	**57**	–	113
GPT-5-NANO-2025-08-07						
cmn	–	**42**	**54**	**64**	7	167
deu	7	–	25	30	5	67
eng	11	19	–	19	5	54
rus	1	21	9	–	5	36
tur	21	**36**	**45**	**52**	–	154

	cmn	deu	eng	rus	tur	Σ
C4AI-COMMAND-R7B-12-2024						
cmn	–	12	15	44	9	80
deu	14	–	8	37	10	69
eng	28	19	–	29	26	102
rus	23	23	31	–	18	95
tur	13	12	10	**56**	–	91
MINISTRAL-8B-INSTRUCT-2410						
cmn	–	19	**30**	**52**	18	119
deu	6	–	15	**33**	4	58
eng	6	3	–	**51**	8	68
rus	10	8	14	–	9	41
tur	11	9	19	**49**	–	88
MISTRAL-NEMO-INSTRUCT-2407						
cmn	–	9	12	**27**	9	57
deu	3	–	1	8	12	24
eng	12	12	–	18	21	63
rus	1	3	7	–	2	13
tur	5	6	6	9	–	26
GLM-4-9B-0414						
cmn	–	19	**32**	**49**	15	115
deu	4	–	18	**23**	6	51
eng	9	10	–	14	6	39
rus	2	4	6	–	0	12
tur	6	9	22	**18**	–	55
QWEN3-4B						
cmn	–	11	42	**49**	7	109
deu	19	–	19	**48**	20	106
eng	18	6	–	**33**	16	73
rus	1	0	3	–	0	4
tur	12	15	19	**51**	–	97
YI-1.5-9B-32K						
cmn	–	9	**22**	**16**	17	64
deu	2	–	13	1	0	16
eng	5	12	–	7	1	25
rus	0	0	0	–	0	0
tur	4	0	10	0	–	14

Table 11: Total number of times, over all prompt languages, where language L_1 (row) is chosen over L_2 (column) in *both* orders of a contrastive pair. Bold: row language wins over column language. Light gray: non-significant difference (Bonferroni-corrected $p < 0.05$). Haystack size is 25 000.

	cmn	deu	eng	rus	tur	Mean
GEMMA-3-27B-IT						
cmn	**92**	85	76	60	88	80
deu	30	**64**	60	10	52	43
eng	52	63	**69**	39	50	54
rus	8	4	0	**84**	0	19
tur	60	57	72	30	**87**	61
GEMMA-3-4B-IT						
cmn	**92**	67	73	60	94	77
deu	71	**95**	90	54	70	76
eng	40	59	**77**	38	53	54
rus	0	0	0	**56**	0	11
tur	58	53	54	38	**64**	53
LLAMA-3.1-8B-INSTRUCT						
cmn	**98**	77	78	52	78	77
deu	56	**86**	75	26	63	61
eng	38	48	**58**	11	42	39
rus	2	2	2	**96**	0	20
tur	74	69	67	52	**94**	71
GPT-5.2-2025-12-11						
cmn	**69**	26	33	20	22	34
deu	58	**82**	68	48	56	62
eng	88	90	**97**	89	87	90
rus	9	15	20	**67**	11	24
tur	28	25	23	20	**70**	33
GPT-5-MINI-2025-08-07						
cmn	**80**	40	42	19	40	44
deu	61	**94**	67	57	57	67
eng	92	87	**95**	84	73	86
rus	0	5	3	**60**	3	14
tur	40	38	55	23	**92**	49
GPT-5-NANO-2025-08-07						
cmn	**92**	73	74	77	84	80
deu	33	**57**	41	31	18	36
eng	21	28	**40**	23	31	29
rus	5	19	17	**36**	14	18
tur	90	76	85	87	**100**	88

	cmn	deu	eng	rus	tur	Mean
C4AI-COMMAND-R7B-12-2024						
cmn	**97**	31	32	13	70	49
deu	42	**77**	65	14	44	48
eng	31	75	**100**	48	40	59
rus	27	22	2	**100**	24	35
tur	62	52	66	19	**86**	57
MINISTRAL-8B-INSTRUCT-2410						
cmn	**100**	68	79	50	71	74
deu	48	**92**	63	33	50	57
eng	34	64	**82**	8	40	46
rus	2	2	0	**89**	0	19
tur	64	62	64	32	**100**	64
MISTRAL-NEMO-INSTRUCT-2407						
cmn	**100**	70	53	64	67	71
deu	38	**67**	47	27	20	40
eng	50	85	**96**	40	62	67
rus	0	0	0	**87**	0	17
tur	25	31	30	29	**82**	39
GLM-4-9B-0414						
cmn	**94**	86	73	71	96	84
deu	35	**71**	64	65	12	49
eng	24	33	**58**	23	15	31
rus	12	5	0	**31**	0	10
tur	71	100	61	65	**75**	74
QWEN3-4B						
cmn	**92**	58	51	48	86	67
deu	61	**93**	78	77	73	76
eng	27	57	**71**	40	33	46
rus	0	0	2	**16**	0	4
tur	72	57	67	60	**87**	68
YI-1.5-9B-32K						
cmn	**95**	71	75	100	83	85
deu	33	**88**	33	29	29	42
eng	29	29	**78**	27	32	39
rus	0	0	0	**0**	0	0
tur	38	50	20	0	**70**	36

Table 12: Percentage of times, over all contrastive pairs and languages L_2, where language L_1 (row) is chosen over L_2 in *both* orders of a contrastive pair, when prompting using L_p (column). Bold marks the diagonal, and thus represents the proportion of times when the information in the prompting language is chosen. Black marks significant difference from 50%, light gray non-significant difference (Bonferroni-corrected $p < 0.05$). Haystack size is 25 000.

Model	Haystack	Both	None	One
GPT-5.2-2025-12-11	1000	1113	0	1887
GPT-5-NANO-2025-08-07	1000	367	1	2632
GPT-5-NANO-2025-08-07	25000	12	16	2972

Table 13: Summary of outcomes from all conflicting haystack retrievals.

Language pair l_1, l_2	$P(> 0)$	95% CI
German vs Chinese	59.8%	$[-0.62, 0.79]$
English vs Chinese	68.6%	$[-0.54, 0.86]$
Chinese vs Russian	100.0%	$[2.26, 3.71]$
Chinese vs Turkish	61.1%	$[-0.61, 0.76]$
English vs German	88.1%	$[-0.30, 1.15]$
German vs Russian	100.0%	$[2.42, 3.92]$
German vs Turkish	93.5%	$[-0.15, 1.20]$
English vs Russian	100.0%	$[2.43, 3.88]$
English vs Turkish	99.3%	$[0.14, 1.48]$
Turkish vs Russian	100.0%	$[2.52, 4.05]$

Table 14: Estimates of the log-odds bias parameter b_{l_1,l_2}, representing the overall bias (over all prompting languages and models) of language l_1 over language l_2. Haystack size is 2500.

Language pair l_1, l_2	$P(> 0)$	95% CI
German vs Chinese	75.6%	$[-0.54, 1.06]$
English vs Chinese	50.4%	$[-0.76, 0.78]$
Chinese vs Russian	100.0%	$[2.33, 3.96]$
Chinese vs Turkish	64.8%	$[-0.63, 0.94]$
English vs German	67.5%	$[-0.61, 0.98]$
German vs Russian	100.0%	$[2.99, 4.82]$
German vs Turkish	95.6%	$[-0.09, 1.52]$
English vs Russian	100.0%	$[2.43, 4.18]$
English vs Turkish	97.1%	$[-0.03, 1.49]$
Turkish vs Russian	100.0%	$[2.17, 3.80]$

Table 15: Estimates of the log-odds bias parameter b_{l_1,l_2}, representing the overall bias (over all prompting languages and models) of language l_1 over language l_2. Haystack size is 5000.

Language pair l_1, l_2	$P(> 0)$	95% CI
German vs Chinese	72.5%	$[-0.51, 0.97]$
Chinese vs English	74.6%	$[-0.49, 0.98]$
Chinese vs Russian	100.0%	$[2.61, 4.32]$
Chinese vs Turkish	69.9%	$[-0.56, 0.99]$
German vs English	75.4%	$[-0.51, 1.06]$
German vs Russian	100.0%	$[2.48, 4.21]$
German vs Turkish	78.0%	$[-0.46, 1.09]$
English vs Russian	100.0%	$[2.60, 4.32]$
English vs Turkish	51.8%	$[-0.73, 0.74]$
Turkish vs Russian	100.0%	$[2.66, 4.33]$

Table 16: Estimates of the log-odds bias parameter b_{l_1,l_2}, representing the overall bias (over all prompting languages and models) of language l_1 over language l_2. Haystack size is 10000.

	cmn	deu	eng	rus	tur	Σ
GEMMA-3-27B-IT						
cmn	–	5	2	**39**	3	49
deu	11	–	2	16	9	38
eng	4	6	–	**34**	10	54
rus	2	3	2	–	5	12
tur	17	4	3	19	–	43
GEMMA-3-4B-IT						
cmn	–	20	**46**	**76**	30	172
deu	11	–	13	**78**	30	132
eng	6	18	–	**71**	25	120
rus	9	2	4	–	7	22
tur	10	17	31	**75**	–	133
LLAMA-3.1-8B-INSTRUCT						
cmn	–	5	8	**43**	9	65
deu	6	–	4	**41**	5	56
eng	10	4	–	**51**	9	74
rus	12	9	13	–	18	52
tur	6	0	1	**61**	–	68
GPT-5.2-2025-12-11						
cmn	–	7	8	30	22	67
deu	**59**	–	11	**57**	**48**	175
eng	**70**	**49**	–	**76**	**60**	255
rus	26	10	5	–	20	61
tur	34	16	5	40	–	95
GPT-5-MINI-2025-08-07						
cmn	–	24	16	**81**	26	147
deu	40	–	18	**75**	39	172
eng	**54**	39	–	**101**	**52**	246
rus	7	7	3	–	15	32
tur	34	20	9	**59**	–	122
GPT-5-NANO-2025-08-07						
cmn	–	**45**	44	**60**	23	172
deu	15	–	16	**48**	6	85
eng	22	30	–	**59**	21	132
rus	11	13	17	–	7	48
tur	21	**50**	46	**64**	–	181

	cmn	deu	eng	rus	tur	Σ
C4AI-COMMAND-R7B-12-2024						
cmn	–	15	11	46	**35**	107
deu	14	–	1	41	18	74
eng	**40**	**31**	–	35	**47**	153
rus	31	21	33	–	20	105
tur	4	10	1	44	–	59
MINISTRAL-8B-INSTRUCT-2410						
cmn	–	**14**	13	**59**	11	97
deu	1	–	1	**54**	17	73
eng	10	11	–	**58**	24	103
rus	3	2	3	–	13	21
tur	3	13	7	**42**	–	65
MISTRAL-NEMO-INSTRUCT-2407						
cmn	–	**32**	30	**86**	21	169
deu	6	–	1	**56**	21	84
eng	17	**18**	–	**75**	34	144
rus	1	4	1	–	2	8
tur	12	16	13	**80**	–	121
GLM-4-9B-0414						
cmn	–	**32**	24	**84**	27	167
deu	8	–	5	**61**	14	88
eng	7	9	–	**79**	10	105
rus	7	10	15	–	3	35
tur	11	14	25	**70**	–	120
QWEN3-4B						
cmn	–	17	25	**84**	16	142
deu	16	–	12	**62**	20	110
eng	10	21	–	**83**	22	136
rus	1	5	4	–	0	10
tur	25	22	24	**83**	–	154
YI-1.5-9B-32K						
cmn	–	31	19	**82**	23	155
deu	14	–	7	**75**	29	125
eng	20	13	–	**72**	**33**	138
rus	3	7	7	–	5	22
tur	7	15	8	**72**	–	102

Table 17: Total number of times, over all prompt languages, where language L_1 (row) is chosen over L_2 (column) in *both* orders of a contrastive pair. Bold: row language wins over column language. Light gray: non-significant difference (Bonferroni-corrected $p < 0.05$). Haystack size is 1000.

	cmn	deu	eng	rus	tur	Σ
GEMMA-3-27B-IT						
cmn	–	2	8	**44**	11	65
deu	**16**	–	8	31	7	62
eng	20	11	–	**51**	15	97
rus	13	13	4	–	12	42
tur	15	7	12	**34**	–	68
GEMMA-3-4B-IT						
cmn	–	23	26	**78**	**33**	160
deu	18	–	28	**63**	**28**	137
eng	18	12	–	**72**	30	132
rus	11	7	10	–	11	39
tur	4	5	15	**63**	–	87
LLAMA-3.1-8B-INSTRUCT						
cmn	–	14	14	**62**	7	97
deu	12	–	14	**69**	9	104
eng	5	7	–	**62**	12	86
rus	17	13	21	–	13	64
tur	11	8	8	**68**	–	95
GPT-5.2-2025-12-11						
cmn	–	9	2	34	13	58
deu	**56**	–	7	**52**	**47**	162
eng	**71**	**54**	–	**79**	**69**	273
rus	20	12	3	–	23	58
tur	**41**	12	7	37	–	97
GPT-5-MINI-2025-08-07						
cmn	–	18	13	**70**	20	121
deu	**55**	–	21	**85**	**44**	205
eng	**64**	32	–	**96**	**52**	244
rus	11	6	5	–	14	36
tur	**47**	14	10	**64**	–	135
GPT-5-NANO-2025-08-07						
cmn	–	**29**	**42**	**52**	7	130
deu	9	–	35	**41**	4	89
eng	16	34	–	**52**	11	113
rus	5	12	12	–	0	29
tur	19	**33**	**58**	**60**	–	170

	cmn	deu	eng	rus	tur	Σ
C4AI-COMMAND-R7B-12-2024						
cmn	–	15	8	39	15	77
deu	20	–	0	45	20	85
eng	**42**	**48**	–	41	**53**	184
rus	39	34	36	–	30	139
tur	17	9	1	**61**	–	88
MINISTRAL-8B-INSTRUCT-2410						
cmn	–	15	11	**67**	14	107
deu	15	–	2	**66**	15	98
eng	14	5	–	**69**	11	99
rus	4	7	8	–	9	28
tur	8	15	17	**58**	–	98
MISTRAL-NEMO-INSTRUCT-2407						
cmn	–	**29**	32	**80**	15	156
deu	7	–	2	**64**	15	88
eng	19	11	–	**60**	15	105
rus	1	5	3	–	4	13
tur	12	22	19	**62**	–	115
GLM-4-9B-0414						
cmn	–	**30**	**26**	**77**	**34**	167
deu	6	–	20	**66**	26	118
eng	0	11	–	**69**	18	98
rus	4	1	13	–	2	20
tur	9	19	10	**69**	–	107
QWEN3-4B						
cmn	–	13	24	**73**	13	123
deu	21	–	23	**63**	27	134
eng	13	9	–	**70**	16	108
rus	6	2	5	–	2	15
tur	29	24	29	**75**	–	157
YI-1.5-9B-32K						
cmn	–	21	27	**79**	**45**	172
deu	20	–	24	**76**	29	149
eng	22	11	–	**72**	36	141
rus	2	5	8	–	3	18
tur	6	22	22	**80**	–	130

Table 18: Total number of times, over all prompt languages, where language L_1 (row) is chosen over L_2 (column) in *both* orders of a contrastive pair. Bold: row language wins over column language. Light gray: non-significant difference (Bonferroni-corrected $p < 0.05$). Haystack size is 2500.

	cmn	deu	eng	rus	tur	Σ
GEMMA-3-27B-IT						
cmn	–	9	8	**41**	12	70
deu	11	–	13	**37**	19	80
eng	24	18	–	**37**	18	97
rus	8	11	6	–	21	46
tur	8	6	10	31	–	55
GEMMA-3-4B-IT						
cmn	–	14	25	**56**	18	113
deu	17	–	10	**30**	**18**	75
eng	19	8	–	**57**	12	96
rus	11	9	9	–	13	42
tur	6	3	13	27	–	49
LLAMA-3.1-8B-INSTRUCT						
cmn	–	**19**	**33**	**64**	12	128
deu	3	–	**21**	**70**	11	105
eng	2	4	–	**61**	7	74
rus	19	15	19	–	11	64
tur	11	17	12	**68**	–	108
GPT-5.2-2025-12-11						
cmn	–	7	4	26	15	52
deu	**61**	–	14	**52**	**47**	174
eng	**73**	**42**	–	**78**	**75**	268
rus	22	11	2	–	23	58
tur	**40**	8	8	34	–	90
GPT-5-MINI-2025-08-07						
cmn	–	17	10	**67**	26	120
deu	**48**	–	19	**70**	41	178
eng	**61**	**48**	–	**91**	38	238
rus	12	9	6	–	16	43
tur	38	26	16	**75**	–	155
GPT-5-NANO-2025-08-07						
cmn	–	**39**	**44**	**51**	6	140
deu	10	–	24	**41**	13	88
eng	10	21	–	30	8	69
rus	1	1	12	–	1	15
tur	**37**	30	**45**	**69**	–	181

	cmn	deu	eng	rus	tur	Σ
C4AI-COMMAND-R7B-12-2024						
cmn	–	15	12	43	15	85
deu	27	–	0	48	27	102
eng	**49**	**29**	–	35	**47**	160
rus	35	27	39	–	24	125
tur	3	9	5	**53**	–	70
MINISTRAL-8B-INSTRUCT-2410						
cmn	–	16	24	**79**	8	127
deu	10	–	10	**65**	10	95
eng	11	8	–	**72**	8	99
rus	4	8	10	–	7	29
tur	6	10	14	**46**	–	76
MISTRAL-NEMO-INSTRUCT-2407						
cmn	–	22	24	**61**	25	132
deu	9	–	6	**64**	7	86
eng	8	7	–	**57**	17	89
rus	5	9	9	–	7	30
tur	7	7	13	**63**	–	90
GLM-4-9B-0414						
cmn	–	14	28	**69**	20	131
deu	22	–	18	**43**	12	95
eng	12	9	–	**47**	16	84
rus	8	3	9	–	4	24
tur	9	14	8	**39**	–	70
QWEN3-4B						
cmn	–	15	18	**64**	12	109
deu	23	–	18	**63**	23	127
eng	20	5	–	**57**	18	100
rus	3	1	2	–	8	14
tur	22	15	33	**77**	–	147
YI-1.5-9B-32K						
cmn	–	28	20	**67**	**34**	149
deu	9	–	10	**46**	11	76
eng	14	9	–	**55**	**34**	112
rus	0	1	4	–	1	6
tur	4	10	12	**64**	–	90

Table 19: Total number of times, over all prompt languages, where language L_1 (row) is chosen over L_2 (column) in *both* orders of a contrastive pair. Bold: row language wins over column language. Light gray: non-significant difference (Bonferroni-corrected $p < 0.05$). Haystack size is 5000.

	cmn	deu	eng	rus	tur	Σ
GEMMA-3-27B-IT						
cmn	–	10	14	**44**	15	83
deu	8	–	13	26	9	56
eng	8	8	–	**39**	6	61
rus	7	8	5	–	9	29
tur	18	8	13	26	–	65
GEMMA-3-4B-IT						
cmn	–	8	15	**47**	9	79
deu	8	–	12	**31**	**17**	68
eng	9	6	–	**59**	22	96
rus	5	3	4	–	4	16
tur	3	0	9	**32**	–	44
LLAMA-3.1-8B-INSTRUCT						
cmn	–	10	19	**69**	4	102
deu	13	–	15	**67**	4	99
eng	7	2	–	**59**	2	70
rus	20	12	22	–	12	66
tur	6	8	**20**	**81**	–	115
GPT-5.2-2025-12-11						
cmn	–	8	2	21	18	49
deu	**61**	–	5	**51**	**40**	157
eng	**77**	**47**	–	**78**	**57**	259
rus	19	7	3	–	20	49
tur	33	13	10	35	–	91
GPT-5-MINI-2025-08-07						
cmn	–	15	10	**65**	18	108
deu	37	–	18	**78**	42	175
eng	**57**	34	–	**93**	**46**	230
rus	13	4	5	–	17	39
tur	26	19	15	**58**	–	118
GPT-5-NANO-2025-08-07						
cmn	–	**27**	**57**	**63**	13	160
deu	7	–	26	**30**	4	67
eng	11	17	–	**39**	10	77
rus	2	5	5	–	0	12
tur	28	**35**	**57**	**70**	–	190

	cmn	deu	eng	rus	tur	Σ
C4AI-COMMAND-R7B-12-2024						
cmn	–	11	14	41	17	83
deu	20	–	1	44	24	89
eng	30	**38**	–	41	**42**	151
rus	25	24	38	–	22	109
tur	10	7	8	**57**	–	82
MINISTRAL-8B-INSTRUCT-2410						
cmn	–	11	24	**68**	9	112
deu	17	–	7	**66**	15	105
eng	13	2	–	**58**	13	86
rus	9	8	11	–	10	38
tur	13	8	15	**53**	–	89
MISTRAL-NEMO-INSTRUCT-2407						
cmn	–	**31**	26	**69**	26	152
deu	10	–	4	**47**	18	79
eng	17	8	–	38	24	87
rus	9	14	16	–	7	46
tur	12	18	17	**57**	–	104
GLM-4-9B-0414						
cmn	–	30	**41**	**58**	15	144
deu	15	–	**28**	**30**	13	86
eng	4	6	–	**41**	7	58
rus	1	4	11	–	4	20
tur	5	12	20	**44**	–	81
QWEN3-4B						
cmn	–	14	13	**68**	9	104
deu	29	–	20	**59**	11	119
eng	14	9	–	**61**	23	107
rus	5	0	1	–	1	7
tur	27	23	27	**91**	–	168
YI-1.5-9B-32K						
cmn	–	16	13	**57**	**47**	133
deu	13	–	4	**33**	4	54
eng	8	2	–	**42**	19	71
rus	1	3	4	–	1	9
tur	1	14	11	**45**	–	71

Table 20: Total number of times, over all prompt languages, where language L_1 (row) is chosen over L_2 (column) in *both* orders of a contrastive pair. Bold: row language wins over column language. Light gray: non-significant difference (Bonferroni-corrected $p < 0.05$). Haystack size is 10000.

Prompt	cmn	deu	eng	rus	tur
GEMMA-3-27B-IT					
cmn	10	10	12	10	12
deu	12	12	12	11	12
eng	12	11	12	12	12
rus	11	11	12	12	10
tur	11	11	12	9	12
GEMMA-3-4B-IT					
cmn	9	5	9	4	9
deu	10	7	11	6	12
eng	11	10	12	7	10
rus	8	9	10	10	11
tur	9	6	8	5	9
LLAMA-3.1-8B-INSTRUCT					
cmn	12	12	12	12	12
deu	12	12	12	10	12
eng	12	12	12	11	12
rus	12	12	11	12	12
tur	12	12	12	12	12
GPT-5.2-2025-12-11					
cmn	12	12	12	12	12
deu	12	12	12	12	12
eng	12	12	12	12	12
rus	12	12	12	12	12
tur	12	12	12	12	12
GPT-5-MINI-2025-08-07					
cmn	12	12	12	12	12
deu	12	12	12	12	12
eng	12	12	12	12	12
rus	12	12	12	12	12
tur	12	12	12	12	12
GPT-5-NANO-2025-08-07					
cmn	12	12	12	12	12
deu	12	12	12	12	12
eng	12	12	12	12	12
rus	12	12	12	12	12
tur	12	12	12	12	12

Prompt	cmn	deu	eng	rus	tur
C4AI-COMMAND-R7B-12-2024					
cmn	12	12	12	11	12
deu	12	12	12	12	12
eng	12	12	12	12	12
rus	12	12	12	12	12
tur	12	11	12	9	12
MINISTRAL-8B-INSTRUCT-2410					
cmn	12	12	12	12	12
deu	12	12	12	12	12
eng	12	12	12	12	12
rus	12	12	12	12	12
tur	12	12	12	12	12
MISTRAL-NEMO-INSTRUCT-2407					
cmn	7	5	8	1	5
deu	10	8	11	2	3
eng	9	9	12	4	8
rus	7	9	10	9	4
tur	8	5	10	1	9
GLM-4-9B-0414					
cmn	10	10	12	10	10
deu	12	12	12	6	0
eng	11	11	12	8	11
rus	12	12	12	8	8
tur	10	1	8	4	2
QWEN3-4B					
cmn	12	12	12	12	12
deu	12	12	12	7	12
eng	12	12	12	6	12
rus	12	12	12	12	12
tur	12	12	12	8	12
YI-1.5-9B-32K					
cmn	12	4	12	1	3
deu	12	6	9	0	5
eng	12	6	12	0	5
rus	12	2	10	1	4
tur	12	4	10	0	4

Table 21: Total number of times a given **non-conflicting** was successfully retrieved in a **monolingual** haystack, by prompt language. Haystack size is 25 000.