CONFLICT BLINDNESS: HOW LLMS HANDLE CONTRADICTORY EVIDENCE

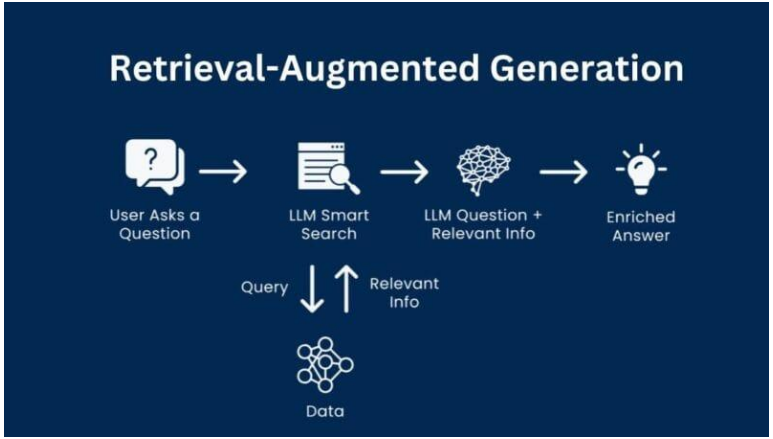
Murathan Kurfalı
RISE Research Institutes of Sweden
Joint work with Robert Östling, Stockholm University

STAYING UP-TO-DATE IS DIFFICULT

- A language model is always pre-trained on yesterday's world.
- Retraining the model every time the world changes (?) is infeasible.
  - GPT-4 training was reported to cost **more than $100M.**
  - Grok 4 training was estimated at **~$490M**.

RAG / LONG CONTEXT TO THE RESCUE

- Instead of putting every fact into the parameters, append relevant new information to the prompt.
- Long-context models make it possible to provide more information at once.
- Together, these are now a standard recipe for evidence-based QA, summarization, and decision support (Lewis et al., 2020; Izacard & Grave, 2021; Gao et al., 2023).



IS PROVIDING CONTEXT ENOUGH?

A samurai is found dead in a forest.

- A bandit,
- the samurai's wife,
- the dead man (through a medium),
- and the woodcutter

give incompatible accounts of what happened.

EVEN WIKIPEDIA IS NOT PERFECT

Wikipedia contains corpus-level inconsistencies:

*"With 99% confidence, approximately **3.3% ± 1.7%** of English Wikipedia facts contradict other information in the same corpus" (*Semnani et al., 2025)

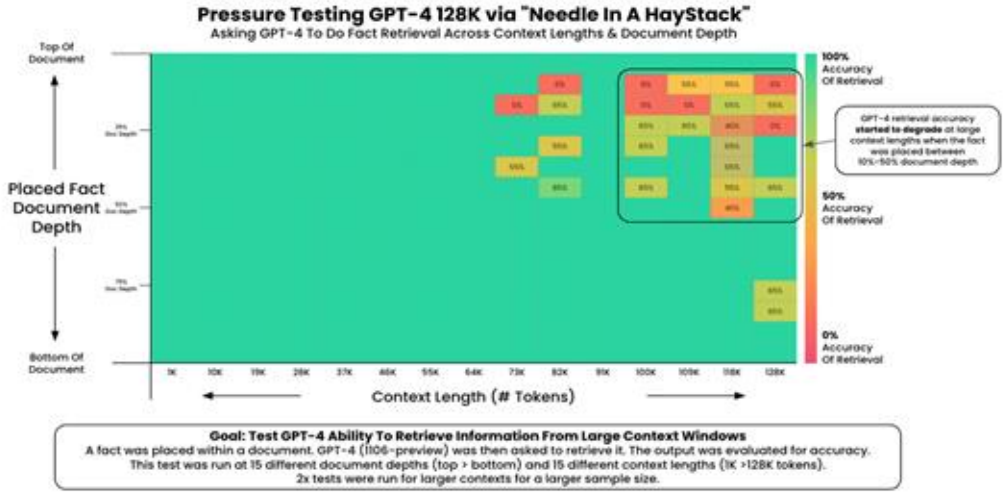
- News changes as events unfold: NewsEdits contains 1.2M articles and 4.6M versions.
- Retrieved contexts can contain outdated, duplicated, noisy, translated, adversarial, or mutually incompatible claims (Spangher et al., 2022; Semnani et al., 2025).

BACKGROUND

BACKGROUND: THE NEEDLE-IN-A-HAYSTACK (NIAH)

NIAH tests whether LLMs can retrieve a specific piece of information (the "needle") embedded inside a large body of text (the "haystack").

- Objective: evaluate whether models can pinpoint relevant facts within lengthy contexts.
- "*The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day.*"



LONG CONTEXT IS NOT A FREE LUNCH

Even when models can perfectly retrieve all relevant evidence, task performance still drops as input length increases (Du et al., 2025). Across math, QA, and coding tasks, accuracy drops by **13.9%–85%.**

The degradation remains even when:

- irrelevant text is replaced with whitespace,
- irrelevant tokens are masked out,
- relevant evidence is placed immediately before the question.

[QUESTION-PART1]
Jack has 16 apples. He gives 3 to her brother and buys 8 more.

[LONG IRRELEVANT TEXT]
… thousands of tokens from unrelated essays …

[QUESTION]
How many apples does Jack have now?

MODELS DO NOT READ CONTEXT UNIFORMLY

The **"Lost in the Middle"** effect: Models are sensitive to where relevant evidence appears (Liu et al., 2024):

- Beginning of context: often easier
- End of context: often easier
- Middle of context: often weaker

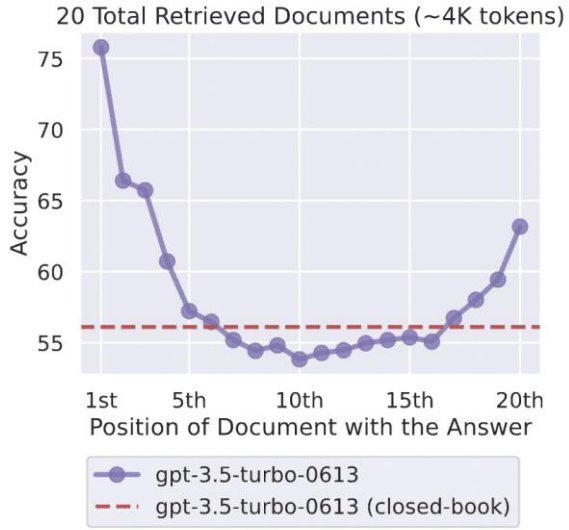


Figure 1: Changing the location of relevant information (in this case, the position of the passage that answers an input question) within the language model's input context results in a U-shaped performance curve—models are better at using relevant information that occurs at the very beginning (primacy bias) or end of its input context (recency bias), and performance degrades significantly when models must access and use information located in the middle of its input context.

KNOWLEDGE CONFLICT: CONTEXT VS. MODEL MEMORY

- Another line of work studies conflict between model memory and provided context (Neeman et al., 2023; Zhang et al., 2025).
- Example structure:
  - Model memory: "The capital of France is Paris."
  - Context: "The capital of France is Berlin."
  - Question: "What is the capital of France?"
- Findings point out that models often over-rely on parametric knowledge instead of following context.

CONTRADICTION DETECTION: WHEN THE MODEL IS WARNED

- A nearby line of work asks models to detect or verify contradictions directly (NoCha / ContraDoc-style settings; Karpinska et al., 2024; Li et al., 2024; Lovering et al., 2025).
- Typical setup:
  - the task is contradiction detection or claim verification,
  - the model is explicitly prompted to look for inconsistency

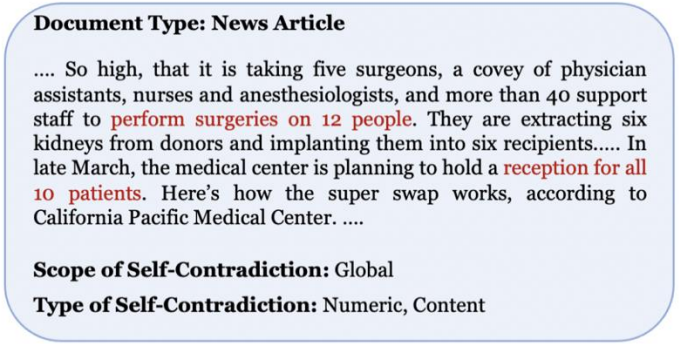


Figure 1: Example of a self-contradictory document from CONTRADOC. The highlighted parts in green show the evidence for the self-contradiction. Additionally, information about the scope and type of the contradiction is also present.

CONTRADICTION DETECTION: WHEN THE MODEL IS WARNED

- A nearby line of work asks models to detect or verify contradictions directly (NoCha / ContraDoc-style settings; Karpinska et al., 2024; Li et al., 2024; Lovering et al., 2025).
- Typical setup:
  - the task is contradiction detection or claim verification,
  - the model is explicitly prompted to look for inconsistency

Model	Accuracy	Precision	Recall	F1
GPT3.5	50.1%	100.0%	0.2%	0.4 %
GPT4	53.8%	97.0%	8.0%	15.6%
PaLM2	52.0%	61.0%	13.4%	22.0%
LLaMAv2	50.5%	51.0%	38.3%	43.7%

Table 2: Performance of different LLMs on **Binary Judgement** experiment.

CONFLICT BLINDNESS

- **Conflict blindness:** when a model is given incompatible contextual evidence but returns a single answer without making the conflict visible.

- *Hallucination:* the answer is unsupported by the context.
- *Conflict blindness:* the answer is supported by part of the context, while incompatible supported evidence is ignored.

Conflicting needles in a haystack: how LLMs behave when faced with contradictory information

Kurfali & Östling, EMNLP 2025

RESEARCH QUESTIONS

- To what extent are current LLMs *able to identify inconsistent information* in an information retrieval setting?
- When models fail to identify inconsistent information, *which factors influence their choice* of which information to rely on?
- If we model one LLM's behavior, to what extent does it *predict the behaviors of other LLMs*—within and across families/sizes?

EXTENDING NIAH TO CONFLICTING KNOWLEDGE SCENARIOS

- The original NIAH benchmark checks whether LLMs can retrieve a single, known fact from a long document.
- We extend this setup by embedding three mutually exclusive pseudo-facts into each haystack.
- There is no single correct answer.

NEEDLE-LEVEL VARIABLES

Repetition Count: How many times a candidate statement appears in the haystack
– Possible values: 1×, 2×, 5×

Position Index: The order in which each needle is inserted
– Possible values: 1st, 2nd, 3rd (relative insertion order)

Needle Identity: Surface form of the candidate answer
– Possible values: Fyrundell, Zequarith, Dralveth

HAYSTACK-LEVEL VARIABLES

Layout: Spatial arrangement of the three needles within the text

- **Beginning-Sequential:** all needles consecutively at start
- **Beginning-5%-Apart:** at start, spaced every 5% of document

- **End-Sequential:** all needles consecutively at end
- **End-5%-Apart:** near end, spaced every 5%

- **Centered@50:** all around midpoint (~50%)
- **Centered@45,50,55:** tightly grouped at 45%, 50%, and 55%

- ***Even:*** *at 25%, 50%, and 75% of the text*

Question Type: Domain of the factual query being tested
- Geographical: "What is the capital of Xaltruvia?"
- Historical: "Which treaty resolved the Xaltruvian conflict?"
- Biology: "In which genus is Xaltruvia classified?"

Semantic Closeness: Thematic relatedness of surrounding content to the query
- Related: Capital cities, peace treaties, botanical taxonomy
- Unrelated: Formula 1, rock music, economic crises

HAYSTACK EXAMPLE

Domain	Question	Candidate Needles
Geography	What is the capital of Xaltruvia?	*The capital of Xaltruvia is Dralveth.* *The capital of Xaltruvia is Fyrundell.* *The capital of Xaltruvia is Zequarith.*
History	Which treaty resolved the Xaltruvian conflict?	*The Xaltruvian conflict was resolved by the Treaty of Dralveth.* *The Xaltruvian conflict was resolved by the Treaty of Fyrundell.* *The Xaltruvian conflict was resolved by the Treaty of Zequarith.*
Biology	In which genus is Xaltruvia classified?	*Xaltruvia is classified in the genus Dralveth.* *Xaltruvia is classified in the genus Fyrundell.* *Xaltruvia is classified in the genus Zequarith.*

Configuration: *Beginning-5%-apart, (2,1,1), (Dralveth, Fyrundell, Zequarith), Geographical, Semantically related*

The capital of Xaltruvia is Dralveth. The capital of Xaltruvia is Dralveth. A capital city, or just capital, is the municipality holding primary status in a country, state, province, department, or other subnational division, usually as its seat of the government. *[..10 sentences..]* The capital of Xaltruvia is Fyrundell. *[..12 sentences..]*. The capital of Xaltruvia is Zequarith. The convergence of political and economic or cultural power is by no means universal. Traditional capitals may be economically eclipsed by provincial rivals as is the case with Nanjing by Shanghai, Quebec City by Montreal, and several US state capitals. The decline of a dynasty or culture could also mean the extinction of its capital city, as occurred at Babylon and Cahokia. [..174 sentences..]

What is the capital of Xaltruvia?

INFERENCE

1) Probability-based:
   - Compute the normalized likelihood of each candidate answer given the haystack and question.

$$p(x \mid c) = \frac{p_{\text{LLM}}(x \mid c)}{\sum_{x' \in X} p_{\text{LLM}}(x' \mid c)}$$

2) Generation-based:
   - Prompt the model and take the greedy decoded answer, then label the output:
     - single: mentions exactly one needle.
     - mixed: mentions multiple needles as plausible.
     - refused: declines, expresses uncertainty, or says the information is unavailable.

EXPERIMENTAL SETUP

- ~5,000 words per haystack (~7,000 tokens).
- Each haystack contains 3 mutually exclusive candidate facts.
- 1,764 unique configurations cover permutations of the experimental variables.
- Selection behavior is modeled with conditional logistic regression.
- Models evaluated:
  - 8 open-source LLMs, 2B–70B.
  - 2 Commercial (OpenAI) models (GPT-4.1-nano; GPT-4.1-mini)

RESULTS

Odds ratio shows how much more (or less) likely a needle is to be selected by an LLM when a specific factor (e.g. repetition or position) changes compared to the reference value, while holding other variables constant.

Hyp.	Variable	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Gemma 2-9b	Gemma 2-27b	Llama 3.2-3B	Llama 3.1-8B	Llama 3.3-70B
H1	2x	17.86**	5.04**	10.48**	14.83**	3.33**	5.79**	26.42**	3.63**
	5x	21.19**	5.64**	8.00**	25.89**	7.66**	12.29**	63.05**	2.34**
H2	2nd	3.63**	0.78	7.86**	0.83	2.20**	4.22**	1.45	0.51**
	3rd	0.77	1.02	6.23**	0.12**	1.45	0.97	0.12**	0.22**
H3	Fyrundell	0.42**	0.79	0.41**	0.94	1.48**	0.62**	0.42**	2.20**
	Zequarith	1.70**	0.57**	0.36**	0.96	0.99	0.24**	0.08**	1.05
H4	Fyrundell × HIST	5.46**	0.58**	0.22**	0.20**	0.42**	1.15	1.16	1.11
	Fyrundell × BIO	0.38**	0.24**	0.23**	1.37	0.71	0.73	5.67**	1.88**
	Zequarith × HIST	0.70*	0.22**	2.67**	0.15**	0.36**	0.46**	2.77**	2.15**
	Zequarith × BIO	0.11**	0.41**	4.25**	1.76**	0.95	0.17**	1.51	2.28**
H5	2x × Clust@45-55	0.19**	0.67	0.26**	1.00	1.13	0.93	1.41	0.94
	2x × Clust@50	0.23**	0.68	1.08	0.97	1.05	0.58	1.44	0.44*
	2x × End-Seq	0.34**	1.02	1.29	1.19	1.30	2.30*	2.31*	1.17
	2x × Even	0.34**	1.44	0.33**	1.80	0.71	0.87	0.61	0.66
	2x × beg-5%-apart	0.46*	1.18	0.48	0.91	1.22	0.95	0.95	0.89
	2x × end-5%-apart	0.19**	0.97	0.41*	1.44	0.91	1.17	0.94	0.80
	5x × Clust@45-55	0.19**	1.14	0.46*	0.51	0.66	0.64	0.75	1.74
	5x × Clust@50	0.33**	0.90	2.06	1.59	1.11	1.13	8.80**	0.83
	5x × End-Seq	1.18	1.18	3.42**	6.57**	0.77	0.92	1.15	1.76
	5x × Even	0.42*	1.89	0.59	1.20	0.43*	0.54	0.31**	0.67
	5x × beg-5%-apart	0.67	1.40	0.79	0.47	0.57	0.70	0.89	0.54
	5x × end-5%-apart	0.34**	0.75	0.91	0.75	0.18*	0.57	0.34*	0.76
H6	2nd × Clust@45-55	1.65	1.53	0.91	3.04**	6.15**	0.56	0.36**	0.95
	2nd × Clust@50	0.67	0.88	1.25	1.05	0.62	1.08	0.70	0.70
	2nd × End-Seq	1.93	3.12**	0.36**	0.37**	0.20**	1.12	0.27**	0.34**
	2nd × Even	1.17	1.23	0.43*	7.59**	6.53**	1.21	1.55	0.81
	2nd × beg-5%-apart	1.16	2.79**	0.93	1.57	1.42	2.26*	3.72**	0.35**
	2nd × end-5%-apart	0.51*	1.79*	0.30**	0.59	0.23*	1.25	0.45*	0.19**
	3rd × Clust@45-55	6.96**	0.55*	1.82	12.53**	1.18	5.81**	5.70**	0.32**
	3rd × Clust@50	2.72**	1.02	2.41*	2.07	0.27**	4.56**	2.07	1.14
	3rd × End-Seq	31.88**	2.20**	0.72	6.19**	4.30**	15.86**	12.73**	1.21
	3rd × Even	7.74**	1.49	0.56	67.08**	3.76**	7.55**	48.10**	1.73*
	3rd × beg-5%-apart	0.77	2.32**	1.24	1.20	4.36**	6.42**	18.58**	0.52*
	3rd × end-5%-apart	13.01**	2.62**	1.66	64.92**	28.04**	79.24**	114.27**	6.04**
H7	2nd × 2x	0.33**	1.20	0.94	0.37**	1.54	0.53*	0.20**	0.71
	2nd × 5x	0.36**	1.32	1.22	0.63	0.76	0.77	0.38**	1.10
	3rd × 2x	0.81	0.92	2.92**	0.77	2.19*	1.34	0.32**	1.16
	3rd × 5x	0.76	0.99	2.07*	1.38	1.68	1.00	0.61	1.79*
H8	2x × Related	0.64*	0.93	0.54**	0.33**	0.79	1.21	0.67*	1.88**
	5x × Related	0.81	1.51*	0.53**	0.32**	0.88	1.18	0.45**	2.12**

DISCUSSION

- Repetition reliably increases selection across all models.
- The boost is largest in smaller/mid models (e.g., Llama-3.1-8B) and modest in some larger ones (e.g., Llama-3.3-70B)
- Layout has only a limited effect on the effect of repetition.
- Semantic relatedness typically narrows the repetition advantage, with a few model-specific exceptions.

Hyp.	Variable	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Gemma 2-9b	Gemma 2-27b	Llama 3.2-3B	Llama 3.1-8B	Llama 3.3-70B
H1	2x	17.86**	5.04**	10.48**	14.83**	3.33**	5.79**	26.42**	3.63**
	5x	21.19**	5.64**	8.00**	25.89**	7.66**	12.29**	63.05**	2.34**
H5	2x × Clust@45-55	0.19**	0.67	0.26**	1.00	1.13	0.93	1.41	0.94
	2x × Clust@50	0.23**	0.68	1.08	0.97	1.05	0.58	1.44	0.44*
	2x × End-Seq	0.34**	1.02	1.29	1.19	1.30	2.30*	2.31*	1.17
	2x × Even	0.34**	1.44	0.33**	1.80	0.71	0.87	0.61	0.66
	2x × beg-5%-apart	0.46*	1.18	0.48	0.91	1.22	0.95	0.95	0.89
	2x × end-5%-apart	0.19**	0.97	0.41*	1.44	0.91	1.17	0.94	0.80
	5x × Clust@45-55	0.19**	1.14	0.46*	0.51	0.66	0.64	0.75	1.74
	5x × Clust@50	0.33**	0.90	2.06	1.59	1.11	1.13	8.80**	0.83
	5x × End-Seq	1.18	1.18	3.42**	6.57**	0.77	0.92	1.15	1.76
	5x × Even	0.42*	1.89	0.59	1.20	0.43*	0.54	0.31**	0.67
	5x × beg-5%-apart	0.67	1.40	0.79	0.47	0.57	0.70	0.89	0.54
	5x × end-5%-apart	0.34**	0.75	0.91	0.75	0.18*	0.57	0.34*	0.76
H7	2nd × 2x	0.33**	1.20	0.94	0.37**	1.54	0.53*	0.20**	0.71
	2nd × 5x	0.36**	1.32	1.22	0.63	0.76	0.77	0.38**	1.10
	3rd × 2x	0.81	0.92	2.92**	0.77	2.19*	1.34	0.32**	1.16
	3rd × 5x	0.76	0.99	2.07*	1.38	1.68	1.00	0.61	1.79*
H8	2x × Related	0.64*	0.93	0.54**	0.33**	0.79	1.21	0.67*	1.88**
	5x × Related	0.81	1.51*	0.53**	0.32**	0.88	1.18	0.45**	2.12**

DISCUSSION

- No universal primacy/recency rule.
- End-position needles usually spikes.

Hyp.	Variable	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Gemma 2-9b	Gemma 2-27b	Llama 3.2-3B	Llama 3.1-8B	Llama 3.3-70B
H1	2x	17.86**	5.04**	10.48**	14.83**	3.33**	5.79**	26.42**	3.63**
	5x	21.19**	5.64**	8.00**	25.89**	7.66**	12.29**	63.05**	2.34**
H2	2nd	3.63**	0.78	7.86**	0.83	2.20**	4.22**	1.45	0.51**
	3rd	0.77	1.02	6.23**	0.12**	1.45	0.97	0.12**	0.22**
H6	2nd × Clust@45-55	1.65	1.53	0.91	3.04**	6.15**	0.56	0.36**	0.95
	2nd × Clust@50	0.67	0.88	1.25	1.05	0.62	1.08	0.70	0.70
	2nd × End-Seq	1.93	3.12**	0.36**	0.37**	0.20**	1.12	0.27**	0.34**
	2nd × Even	1.17	1.23	0.43*	7.59**	6.53**	1.21	1.55	0.81
	2nd × beg-5%-apart	1.16	2.79**	0.93	1.57	1.42	2.26*	3.72**	0.35**
	2nd × end-5%-apart	0.51*	1.79*	0.30**	0.59	0.23*	1.25	0.45*	0.19**
	3rd × Clust@45-55	6.96**	0.55*	1.82	12.53**	1.18	5.81**	5.70**	0.32**
	3rd × Clust@50	2.72**	1.02	2.41*	2.07	0.27**	4.56**	2.07	1.14
	3rd × End-Seq	31.88**	2.20**	0.72	6.19**	4.30**	15.86**	12.73**	1.21
	3rd × Even	7.74**	1.49	0.56	67.08**	3.76**	7.55**	48.10**	1.73*
	3rd × beg-5%-apart	0.77	2.32**	1.24	1.20	4.36**	6.42**	18.58**	0.52*
	3rd × end-5%-apart	13.01**	2.62**	1.66	64.92**	28.04**	79.24**	114.27**	6.04**
H7	2nd × 2x	0.33**	1.20	0.94	0.37**	1.54	0.53*	0.20**	0.71
	2nd × 5x	0.36**	1.32	1.22	0.63	0.76	0.77	0.38**	1.10
	3rd × 2x	0.81	0.92	2.92**	0.77	2.19*	1.34	0.32**	1.16
	3rd × 5x	0.76	0.99	2.07*	1.38	1.68	1.00	0.61	1.79*

DISCUSSION

- Not all needles are created equally: Surface forms carry stable, model-specific biases.
- "Fyrundell" is down-weighted in several models; "Zequarith" is disfavored in Llama but preferred in Qwen-2.5-3B.
- Domains flip these tendencies—history, biology, and geography can turn a penalty into a boost (and vice versa).

Hyp.	Variable	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Gemma 2-9b	Gemma 2-27b	Llama 3.2-3B	Llama 3.1-8B	Llama 3.3-70B
H1	2x	17.86**	5.04**	10.48**	14.83**	3.33**	5.79**	26.42**	3.63**
	5x	21.19**	5.64**	8.00**	25.89**	7.66**	12.29**	63.05**	2.34**
H2	2nd	3.63**	0.78	7.86**	0.83	2.20**	4.22**	1.45	0.51**
	3rd	0.77	1.02	6.23**	0.12**	1.45	0.97	0.12**	0.22**
H3	Fyrundell	0.42**	0.79	0.41**	0.94	1.48**	0.62**	0.42**	2.20**
	Zequarith	1.70**	0.57**	0.36**	0.96	0.99	0.24**	0.08**	1.05
H4	Fyrundell × HIST	5.46**	0.58**	0.22**	0.20**	0.42**	1.15	1.16	1.11
	Fyrundell × BIO	0.38**	0.24**	0.23**	1.37	0.71	0.73	5.67**	1.88**
	Zequarith × HIST	0.70*	0.22**	2.67**	0.15**	0.36**	0.46**	2.77**	2.15**
	Zequarith × BIO	0.11**	0.41**	4.25**	1.76**	0.95	0.17**	1.51	2.28**

DISCUSSION

Model similarity is measured using symmetrized KL divergence between predicted answer distributions.

- Lower values mean more similar selection behavior.
- Llama and Gemma models show stronger within-family agreement.
- Qwen2.5 models vary more internally.
- Selection strategies are therefore partly family-specific and partly model-specific.

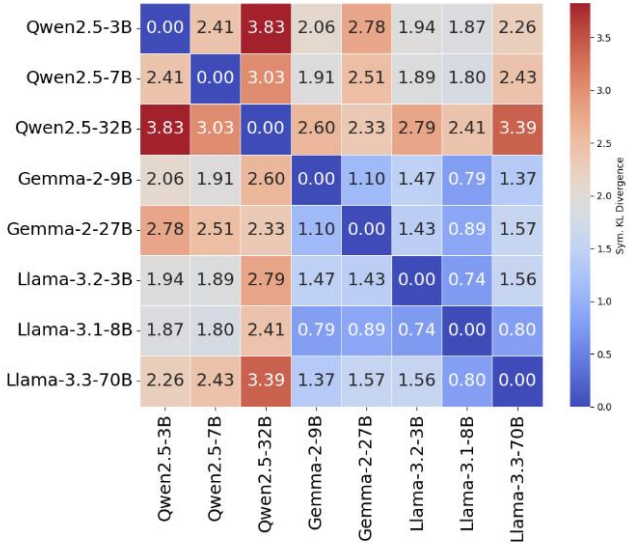


Figure 3: Inter-model similarity via symmetrized KL across all configurations; lower values indicate higher similarity.

GENERATION-BASED RESULTS

- Generations overwhelmingly produce a single answer: 98.2% in Gemma-2-9B, 92.5% in Gemma-2-27B, and ~80% on average.
- Mixed outputs are uncommon (≈1–30% by model), and refusals are rare (<1% for half the models).
- GPT-4o-mini and GPT-4.1-nano also rarely surface the conflict

Model	Single	Mixed	Refused
Qwen 2.5-3B	94.0%	5.4%	0.6%
Qwen2.5-7B	76.4%	14.6%	9.0%
Qwen2.5-32B	64.6%	22.2%	13.2%
Llama-3.2-3B	69.6%	24.6%	5.8%
Llama-3.1-8B	76.9%	22.2%	0.9%
Llama-3.3-70B	69.0%	30.7%	0.3%
Gemma-2-9B	98.2%	0.7%	1.1%
Gemma-2-27B	92.5%	7.3%	0.2%
GPT-4o-mini	88.3%	11.7%	—
GPT-4.1-nano	97.0%	3.0%	—
Overall	80.1%	16.0%	3.9%

CONCLUSIONS

- LLMs do not just retrieve; they make **silent choices.**
  - Generations are mostly single-answer outputs, so uncertainty is rarely surfaced.
- Repetition consistently boosts selection.
- Position and layout matter, but there is no universal primacy/recency rule.
- Surface forms carry stable, model-specific biases.

LANGUAGE BIAS UNDER CONFLICTING INFORMATION IN MULTILINGUAL LLMS

ÖSTLING & KURFALI, UNDER REVIEW FOR EMNLP.

INTRODUCTION

- In real RAG systems, retrieved documents are often multilingual: news, social media, historical records, and translated summaries.
- If English and Chinese sources support different answers, does the model treat them symmetrically?
- This study asks whether the language affects which claim is selected.

MULTILINGUAL NEWS HAYSTACKS

- Haystacks contain up to 34 authentic news articles, about 25,000 English words before translation.
- Five languages: Chinese, German, English, Russian, Turkish.
- Each haystack uses two languages. Non-needle articles are assigned pseudo-randomly 50/50, and article order is controlled across contrastive pairs.
- Needles are inserted into topically fitting articles to simulate realistic retrieval.

RESEARCH QUESTIONS

- **RQ1**. How do multilingual LLMs behave when faced with conflicting information in a naturalistic retrieval setting?
- **RQ2**. Are they biased with respect to which language they prioritize under contradiction?
- **RQ3**. Are any languages consistently favoured or disfavoured across LLMs?

METHOD: NEEDLES

- Four question categories, each with two sentence templates:
  - original lead vocalist of a band
    editor of a newspaper
    long-time CEO of a company
    chairman of an organization
- Pseudo-surnames: Delcroft, Quellman, Pikehart.
- Pseudo-entities: Cinderfax, Noiseweld, Motelvine, Brovencia, Clevantra, Teraluxis.

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.

METHOD: CONTRASTIVE LANGUAGE PAIRS

- The same two conflicting names are inserted, but their languages are swapped across paired haystacks.
  - Haystack A: name X appears in English; name Y appears in Chinese.
  - Haystack B: name X appears in Chinese; name Y appears in English.
- If the language wins after swapping the names, language is a plausible deciding factor rather than surname identity.

[...] 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

WHAT COUNTS AS A "WIN"?

Possible outcomes

- L1 wins
  - The name presented in L1 is retrieved in both swapped haystacks.
- L2 wins
  - The name presented in L2 is retrieved in both swapped haystacks.
- Same surname wins
  - The model follows the surface form, not the language.
- Retrieval failure
  - One or both haystacks return neither valid name.

[...] 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

LANGUAGE-BIAS ANALYSIS

Compare L1 wins vs. L2 wins after discarding surname-driven cases and retrieval failures.

- Null model: P(L1 wins) = P(L2 wins) = 0.5.
- A significant imbalance means that one language is preferred for that model and language pair.

EXPERIMENTAL SCALE

Component	Value
Haystacks per model	**480 bilingual conflicting**: language-bias analysis **120 monolingual conflicting**: conflict-detection analysis **60 monolingual non-conflicting**: retrieval controls
Context lengths	1,000 · 2,500 · 5,000 · 10,000 · 25,000 words
Models	12 multilingual LLMs, including GPT-5 family, Gemma-3, Llama-3.1, Mistral, GLM, Qwen, Yi, and Command-R

RESULTS: CONFLICT DETECTION STILL FAILS

- ~1K-word haystacks: LLMs overwhelmingly give one answer.
- ~25K-word haystacks: LLMs sometimes fail to give either alternative.

	1 000 words			25 000 words		
Model	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.

RESULTS: CONFLICT DETECTION STILL FAILS

- The retrieval rates in monolingual haystacks are also very similar.

Model	1 000 words			25 000 words		
	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.

RESULT: RUSSIAN IS STRONGLY DISFAVORED; CHINESE FAVORED AT LONG CONTEXTS

The intervals are on a log-odds scale, e.g.,

- exp(2.54) ≈ 12.7
- exp(4.17) ≈ 64.7

Above zero means the first language is preferred.

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.

RESULT: RUSSIAN IS STRONGLY DISFAVORED; CHINESE FAVORED AT LONG CONTEXTS

The intervals are on a log-odds scale, e.g.,

- exp(2.54) ≈ 12.7
- exp(4.17) ≈ 64.7

Above zero means the first language is preferred.

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.

RESULT: CAPABILITY AND BIAS ARE SEPARABLE

- Monolingual controls show retrieval is possible for Russian.
- The disadvantage appears when Russian-presented evidence competes with another language.
- This points to conflict-resolution bias, not just basic language capability.

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.

RESULT: PROMPT LANGUAGE ALSO MATTERS

- Models often prefer information presented in the same language as the prompt.
  - A Russian prompt can partly counteract the general Russian disadvantage.
  - A Chinese prompt increases the chance that Chinese-presented evidence is reported.
- For multilingual RAG, the user's query language can influence which source language appears in the answer.

CONCLUSION

- Multilingual models usually collapse contradictory evidence to one answer.
- When contexts get longer, some models increasingly retrieve neither answer rather than surface the conflict.
- The language of the evidence can affect which answer survives.
- The language of the prompt can also bias the selected evidence.
- Multilingual RAG needs conflict evaluation, not only retrieval accuracy by language.

Models often collapse contradictory evidence into a single answer.

The selected answer can be shaped by repetition, position, layout, surface form, and language.

A RAG answer can therefore be grounded in one source and still be misleading because incompatibility inside evidence is hidden.

FUTURE WORK

WP1: BROADER EVALUATION FRAMEWORK

- Current work focuses mainly on direct categorical factual conflicts.
- The benchmark will also vary repetition, position, *source credibility*, language, number of needles, and context size.

Conflict type	Short definition	Example needles
Categorical factual	conflict in a **categorical fact** such as identity, location etc.	**A:** The CEO of Nordvale was Anna Mercer. **B:** The CEO of Nordvale was David Rho.
Numeric factual	conflict in a **quantitative fact** such as number, date, age, duration	**A:** The report says 18 patients were admitted. **B:** The report says 27 patients were admitted.
Quantifier	Conflict depends on words like all, some, none, only, at least.	**A:** All students handed in their projects on time. **B:** Certain students failed the course because they missed the project deadline.
Temporal	Conflict depends on temporal reference, or the relative timing of events.	**A:** As of April, 2026, the Lakehurst airport had been closed for a year. **B:** In December, 2025, the largest aircraft landed at Lakehurst Airport.
Inferential	Conflict requires an inferential analysis.	**A:** The team won its last cup in 2020. **B:** As of 2026, the team has not won anything in 30 years.

WP2: MECHANISMS OF CONFLICT BLINDNESS

- Behavioural results do not tell us why the conflict disappears.
- Three hypotheses:
  - A. The conflict is never represented.
  - B. The conflict is represented early but lost later.
  - C. The conflict is represented but overridden by repetition, position, surface form, or language.

- Methods: layerwise representation analysis, probing classifiers, activation patching, and cross-model transfer tests.

WP3: MITIGATION ACROSS THE PIPELINE

- How best to overcome conflict blindness?
  - Retrieval-side: cluster related passages and surface disagreements before generation.
  - Inference-time: use self-consistency, verifier-guided generation, and explicit source comparison.
  - Post-training: conflict-aware instruction tuning and preference learning / DPO.

TAKE-HOME MESSAGE

- RAG and long context help with outdated knowledge, but context is not automatically reliable.
- Evidence can be noisy, outdated, multilingual, duplicated, or contradictory.
- Current models often fail to surface that conflict.
- Conflict handling should be evaluated directly, not inferred from retrieval accuracy.

THANK YOU!