

Near-Miss Context Contamination and Language-Asymmetric Evidence Arbitration in LLMs

Jebish Purbey^{1,2}, Sanjeeb Prasad Panday²,
Barnnita Shrestha³, Bikash Balami⁴, Aakash Pandey¹, Bibek Khanal⁵

¹Dogma Group ²Pulchowk Campus, Tribhuvan University

³Asian Institute of Technology ⁴CDCSIT, Tribhuvan University

⁵IoE, Tribhuvan University

Correspondence: jebishpurbey@gmail.com, sanjeeb@ioe.edu.np

Abstract

Retrieval-augmented generation assumes that language models can use relevant context while rejecting semantically related but non-evidential context. Further, multilingual retrieval adds a second assumption that when evidence appears in different languages, models arbitrate by factual content instead of language cues. We test both these assumptions as generator-side failures to understand how models bind, compare, and trust evidence across languages. We first study *near-miss contamination* where we add true, relation-matched context about a nearby entity that does not answer the target question. Across Mintaka and PopQA, this context raises distractor adoption before any direct contradiction is introduced, and the effect persists across every tested model and language.

We then study *language evidence arbitration* where English and local-language evidence disagree. We find that language preference is model-family dependent. GPT-5.4-mini leans toward English across all four languages, while DeepSeek models lean toward local-language evidence for all tested languages. We also run a label-only control where both snippets are written in English and only the source label varies and find that evidence labeled as English is over-trusted, with 95% bootstrap confidence intervals excluding zero across all 16 model-language pairs. Our results thus show that context use is shaped by entity proximity, language preference, and source-language metadata.

1 Introduction

Retrieval-augmented generation (RAG) is usually evaluated through the lens of retrieved-passages quality and final-answer accuracy, leaving unexamined how models handle plausible context that is not actually evidence for the target question. When exact evidence is unavailable, retrieval

systems often fetch passages about nearby entities, same-relation facts, or topically related records. A question about Nepal may return a sentence about Bhutan, or a question about one river may return context about a neighboring one. Such passages can be true and semantically close to the question while still belonging to the wrong entity. Where prior work on knowledge conflicts focuses on explicit contradiction (Longpre et al., 2021b; Xie et al., 2024; Chen et al., 2022), we study a different failure mode here which we term as **near-miss contamination**, or interchangeably **near-miss context contamination**. Near-miss contamination occurs when true, relation-matched context about the wrong entity leads the model to adopt the wrong object, with no explicit contradiction present. Near-miss contamination can exist across languages.

Multilingual RAG also introduces a related but different generator-side problem. When a system retrieves conflicting English and local-language evidence for the same question, a model has to decide which source to trust, which thus becomes a language evidence arbitration task. An ideal choice should track the correct evidence irrespective of the language. Yet, we observe that models can show content-independent preference for one language’s evidence over the other, which we term as **language-asymmetric evidence arbitration**.

We evaluate both of these failures within a single factual QA framework. Our near-miss probe measures distractor adoption under six conditions of increasing semantic proximity, moving from unrelated context to direct contradiction, and treats near-neighbor context as the key manipulation. Our arbitration probe measures evidence selection between conflicting English and local-language passages, and then removes comprehension difficulty via a label-only condition fixing the language and varying only the label. Our two probes together isolate entity grounding and source-label authority as separate properties of how generators use context.

Code and data available at <https://github.com/jebish/nmcc>

Our contributions are:

1. A controlled near-miss contamination probe for factual QA that uses true, relation-matched context about the wrong entity to isolate entity misbinding as a distinct failure mode.
2. A consistent dose-response pattern across four models, two datasets, and five languages. We show that near-neighbor context is more harmful than unrelated or same-relation-random context in every condition tested, and a known-only analysis restricted to initially correct responses confirms the effect.
3. A language arbitration probe that separates content preference from label trust. We find that model families diverge in which language’s content they prefer, an instance of language-asymmetric evidence arbitration, but a label-only test, where both snippets are written in English and only the source tag differs, shows all four still favor the English label regardless of family.

2 Related Work

Retrieval-augmented generation. RAG systems condition generation on retrieved evidence instead of using stored knowledge alone (Lewis et al., 2020b). Dense passage retrieval (Karpukhin et al., 2020), retrieval-augmented pretraining (Guu et al., 2020), and multi-passage generative readers (Izacard and Grave, 2021) form the core RAG infrastructure, evaluated on knowledge-intensive benchmarks such as KILT (Petroni et al., 2021). Recent RAG benchmarks additionally probe noise robustness and negative rejection (Chen et al., 2024), hallucination under retrieved evidence (Niu et al., 2024), and citation attribution in generation (Gao et al., 2023). Our experiments test whether generators reject plausible context that matches the target relation while belonging to the wrong entity.

Knowledge conflicts and misleading context. The tension between parametric memory and retrieved evidence is well documented. Longpre et al. (2021b) show that entity substitution in retrieved context degrades question answering, while Xie et al. (2024) find that models differ in whether they defer to retrieved context or rely on parametric memory. Conflicting evidence from multiple sources can compound factual errors (Chen et al.,

2022), and the resulting conflict can become a tug-of-war between internal and external knowledge (Jin et al., 2024). Distraction from irrelevant context is a related failure mode (Shi et al., 2023; Amiraz et al., 2025), as is non-uniform use of evidence across long contexts (Liu et al., 2024). Our distractor context is true and relation-compatible but belongs to a nearby entity, separating near-miss contamination from explicit contradiction and arbitrary irrelevance.

Parametric knowledge and entity frequency.

Petroni et al. (2019) frame language models as implicit knowledge stores, and subsequent work shows that long-tail factual knowledge remains difficult for LLMs (Kandpal et al., 2023). Entity popularity is a strong predictor of QA accuracy in PopQA (Mallen et al., 2023), and dense retrievers struggle on simple Wikidata-derived entity queries in EntityQuestions (Sciavolino et al., 2021). Both findings motivate our known/unknown decomposition and near-neighbor construction.

Multilingual factual QA. Multilingual QA benchmarks cover aligned open-domain questions (Longpre et al., 2021a), typologically diverse information-seeking questions (Clark et al., 2020), cross-lingual open retrieval (Asai et al., 2021), and extractive cross-lingual QA (Lewis et al., 2020a). We use Mintaka (Sen et al., 2022) as the primary multilingual evaluation resource, augmented with a Nepali translation. Recent work studies language preference in multilingual RAG systems (Park and Lee, 2025), and MEMERAG provides end-to-end multilingual meta-evaluation for RAG (Cruz Blandón et al., 2025). Observed accuracy gaps conflate language preference with translation quality, script handling, entity aliasing, and comprehension differences. Our arbitration setup holds the answer space fixed by requiring a direct choice between conflicting snippets, and the label-only condition isolates the source-language tag as the sole variable once language is controlled.

3 Methodology

We study two failure modes in how models use retrieved evidence. Section 3.2 introduces near-miss context, where a true, relation-matched fact about the wrong entity is inserted as evidence for a factual question. Section 3.3 introduces language evidence arbitration, where conflicting English and local-language evidence force a model to choose which

source to trust. For both studies, we construct evidence directly instead of retrieving it, so there is no search index or retriever in our methodological pipeline, and every distractor is guaranteed to be factually true, relation-matched, and tied to one tracked incorrect answer. We use controlled factual QA throughout because it supplies a specific distractor object to track alongside a measurable binary context-following decision. We query four models from two different families: GPT-5.4 (2026-03-05) (Singh et al., 2026), GPT-5.4-mini (2026-03-17), DeepSeek-V4-Pro (2026-04-24) (DeepSeek-AI et al., 2026), and DeepSeek-V4-Flash (2026-04-24) at temperature 0. Our primary metric is the adoption of the distractor object from the manipulated evidence. This does not depend on any overall accuracy changes. The symmetric paired conditions isolate language preference from confounds such as answer position and prior parametric knowledge. All confidence intervals reported are 95% bootstrap intervals over items.

3.1 Datasets

We use Mintaka (Sen et al., 2022) and PopQA (Mallen et al., 2023) as the datasets for this study. For Mintaka, we use 2,757 items spanning the train, development, and test splits, across English, Hindi, Arabic, German, and Nepali. Nepali is not in Mintaka’s original release and thus, we translate it from English using DeepSeek-V4-Pro. All translations are verified and corrected by native speakers. For PopQA, we use 3,000 stratified items in English, Hindi, and Nepali, with Hindi and Nepali translated and verified by the same process, and we extend its answer aliases with target-language entity names so translated answers match correctly during scoring. The near-miss probe uses both datasets, and the language arbitration probe uses Mintaka only.

3.2 Near-Miss Context

Near-miss context is a true, relation-matched fact about the wrong entity for a given question. We use it to test whether a model adopts this wrong entity’s fact as the answer, even though that entity is not the one the question asks about.

Context conditions. We present each question under six context conditions. Table 1 shows all six conditions for a single Mintaka item. We use a no-context baseline to establish parametric performance and gold context to verify that correct

evidence suppresses distractor adoption. The four distractor conditions grow more relevant to the question, from unrelated context, which shares no entity or relation with the question, through same-relation-random context and near-neighbor context, to direct contradiction, which names the same tracked distractor as the answer to the exact same question. Adoption close to the unrelated baseline would mean proximity adds little distractor influence, while adoption close to direct-contradiction rates would suggest nearby facts are nearly as misleading as explicit false claims.

The near-neighbor condition is our core experimental configuration. We embed each question with all-MiniLM-L6-v2 (Reimers and Gurevych, 2019; Wang et al., 2020), and rank candidates sharing the target’s Wikidata category and question type by cosine similarity. Each candidate must also refer to a different entity and produce a different answer than the target. When the combined pool is too small, we fall back to the same-category pool. We verify that all Mintaka and PopQA items receive a candidate this way. We select the highest-similarity eligible candidate with no threshold. The mean cosine similarity on Mintaka is 0.64 and on PopQA is 0.69. We pair, for instance, a question about which capital city has the highest elevation (gold: La Paz) with a near-neighbor question about which US city has the highest population (distractor: New York City), at a cosine similarity of 0.626. The resulting context is factually true and semantically plausible but belongs to the wrong subject entity. We run this selection once, over English question embeddings, before translation, so the same candidate carries over to every language, and the embedding model only ever has to work in English.

Metric. DOAR measures the probability that the model outputs the distractor object o' under context condition c :

$$\text{DOAR}_c = P(\hat{o}_c = o') \quad (1)$$

The near-miss effect Δ_{NN} subtracts no-context adoption to isolate the proximity-specific contribution:

$$\Delta_{\text{NN}} = \text{DOAR}_{\text{NN}} - \text{DOAR}_{\text{NC}}. \quad (2)$$

We report both metrics over all items, and additionally split by whether an item is known, meaning answered correctly without context, or not. This isolates the effect from each item’s own no-context baseline, and the gap tests whether the model is more vulnerable with no prior knowledge.

Condition	Context shown to the model
No context	None
Gold	“The Ender’s Game series was written by Orson Scott Card.”
Unrelated	“Katy Perry’s first hit song was ‘I Kissed a Girl,’ which became a breakthrough ...”
Same-relation random	“The Hannibal Lecter series was written by Thomas Harris.”
Near-neighbor (similarity 0.55)	“The Hunger Games series was written by Suzanne Collins.”
Direct contradiction	“The Ender’s Game series was written by Suzanne Collins.”

Table 1: The six context conditions for a single Mintaka item for the question “Who wrote the Ender’s Game series?” Suzanne Collins is the distractor tracked in the near-neighbor and direct-contradiction conditions.

3.3 Language Evidence Arbitration

We run the arbitration probe on Mintaka across Hindi, Arabic, German, and Nepali, restricting to model-calibrated unknown items so conflicting evidence has an opportunity to shift the response. We test two symmetric conflict conditions to isolate directional language preference. In the first, we place the wrong answer in the English snippet and the correct answer in the local-language snippet. In the second, we reverse the assignment and place the wrong answer in the local-language snippet. EOR (English Override Rate) is the rate of following wrong English evidence in the first condition, and LOR (Local Override Rate) is the rate of following wrong local evidence in the second. Their signed difference gives the Language Arbitration Bias Score:

$$\text{LABS} = \text{EOR} - \text{LOR}. \quad (3)$$

A positive LABS means a model follows false English evidence more often than false evidence in the local language, an English-leaning arbitration. A negative LABS means it follows false local evidence more often, a local-language preference. We run both conflict conditions twice, once with the question posed in the target language and once with the question posed in English, so local-language evidence does not systematically coincide with the question’s own language. If the question were always posed in the target language, a model that simply trusts evidence matching the question’s language would look identical to one with a genuine content-language preference. We pool the two question-language variants when computing EOR and LOR, so the reported rates measure evidence-language preference on its own, separate from this confound.

The label-isolation condition tests whether source-language metadata carries authority independent of the evidence text itself. Both snippets are written in English, so unlike EOR and LOR, the

two differ only in their source-language tag and not in the language of the underlying text. One snippet is tagged English and states the wrong answer, and the other is tagged with the local language and states the correct answer. For a question asking where the Little Bighorn National Monument is located (gold: Montana), the English-tagged snippet claims Washington, D.C. and the Hindi-tagged snippet, also written in English, states Montana. The rate of following the mislabeled English-tagged snippet is the Language-Label Authority Bias (LLAB). This isolates how much trust shifts from the label alone, with comprehension difficulty removed. EOR and LOR are symmetric by construction, so any LABS signal that remains after controlling for position and evidence text reflects directional language preference instead of comprehension asymmetry.

We use the embedding model only to build near-neighbor candidates for the study in Section 3.2. Here, EOR, LOR, LABS, and LLAB come entirely from the four models under evaluation.

4 Near-Miss Results

4.1 Near-Miss Context Drives Dose-Response

Near-miss context drives a dose-response effect on distractor adoption. Dose-response describes how a measured effect grows as the dose of a substance increases. We borrow this term from pharmacology to describe the near-miss results. Here the dose is the semantic distance between a distractor and the target entity, moving from no context through unrelated, same-relation, near-neighbor, and direct-contradiction conditions. Figure 1 shows this across the five conditions for English. Unrelated and same-relation-random distractors stay near the no-context baseline, while near-neighbor adoption rises markedly before direct contradiction drives it further. A true fact about a nearby entity is more misleading than one about an unrelated entity, with no explicit contradiction present. This is near-

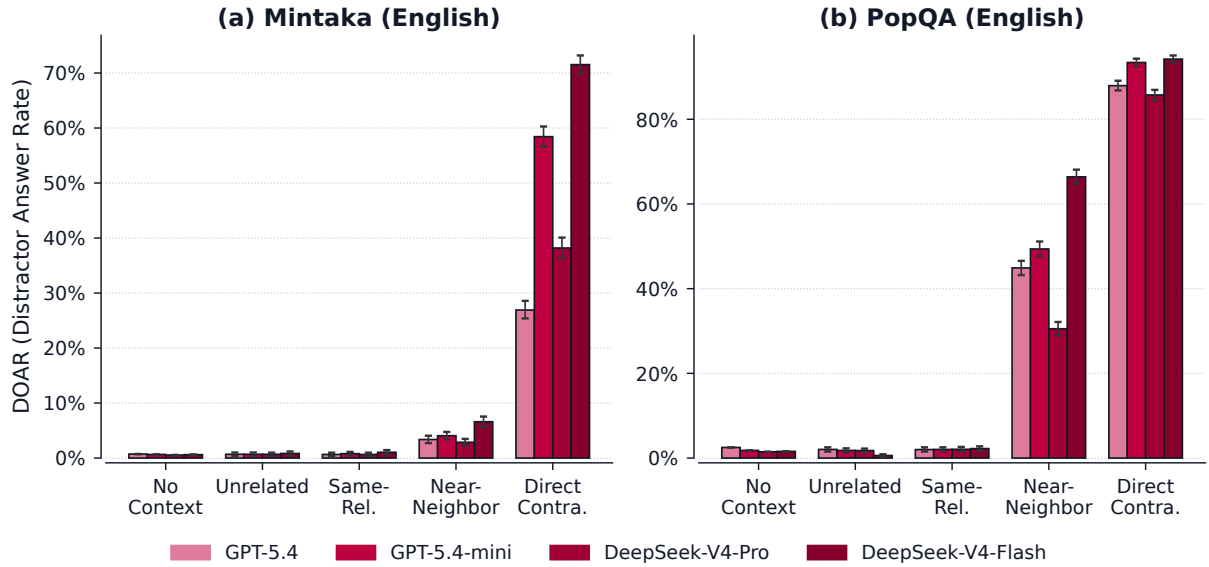


Figure 1: Dose-response for distractor adoption on English Mintaka and PopQA. Near-neighbor context produces a clear step before direct contradiction.

Model	En	Hi	Ar	De	Ne	Avg.
GPT-5.4	2.6	6.5	5.5	3.8	4.4	4.6
GPT-5.4-mini	3.4	6.5	6.3	4.1	5.5	5.2
DS-V4-Pro	2.3	6.6	5.6	2.5	5.4	4.5
DS-V4-Flash	6.0	20.5	19.9	10.2	19.0	15.1

Table 2: Near-miss effect Δ_{NN} (%) on Mintaka. Avg. is the mean across five languages.

Model	En	Hi	Ne	Avg.
GPT-5.4	42.4	55.2	61.4	53.0
GPT-5.4-mini	47.6	62.9	65.6	58.7
DS-V4-Pro	29.1	40.9	50.1	40.0
DS-V4-Flash	64.8	81.8	83.8	76.8

Table 3: Near-miss effect Δ_{NN} (%) on PopQA. Hindi and Nepali exceed English in every row.

miss contamination. Within the near-neighbor condition itself, distractor similarity to the target varies continuously across items, and adoption tracks this similarity as well (Figure 3). Appendix A.3 reports the full similarity distribution, correlation statistics, and per-language breakdown.

Across Mintaka model-language pairs, near-neighbor and direct-contradiction adoption rates are positively correlated ($r = 0.81$, Figure 2). On Mintaka English, where many questions fall within parametric knowledge, Δ_{NN} ranges from 2.3% for DeepSeek-V4-Pro to 6.0% for DeepSeek-V4-Flash, a modest but consistent perturbation. On PopQA English, where the questions are harder and the unknown fraction is larger, Δ_{NN} ranges

Model	Mintaka	PopQA
GPT-5.4	3.0	30.1
GPT-5.4-mini	3.4	35.1
DeepSeek-V4-Pro	3.1	18.4
DeepSeek-V4-Flash	9.8	63.5

Table 4: Known-only near-neighbor effect Δ_{NN} (%), averaged across languages within each dataset.

from 29.1% for DeepSeek-V4-Pro to 64.8% for DeepSeek-V4-Flash. This suggests near-miss contamination scales with context-dependence for correct answers.

4.2 Near-Miss Contamination Persists Across Languages

The near-miss effect is positive across every model-language pair in both the datasets (Tables 2–3). DeepSeek-V4-Flash is the most susceptible model throughout, peaking at 20.5% on Mintaka Hindi and 19.9% on Mintaka Arabic.

On PopQA, Hindi and Nepali adoption exceeds English for every model. The 95% bootstrap confidence intervals for the Hindi-minus-English delta range from 10.5 to 20.4 percentage points across models, all above zero. For DeepSeek-V4-Flash, Δ_{NN} reaches 64.8% in English, 81.8% in Hindi, and 83.8% in Nepali. Hindi and Nepali are both translated from English in PopQA, so the near-miss effect grows even when the underlying evidence is a direct translation of the same English sentence.

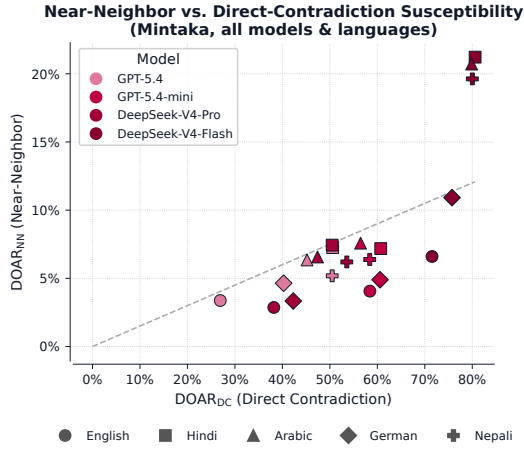


Figure 2: DOAR under near-neighbor context vs. direct contradiction for all Mintaka model-language pairs. Near-miss contamination is smaller than direct contradiction but strongly related to it.

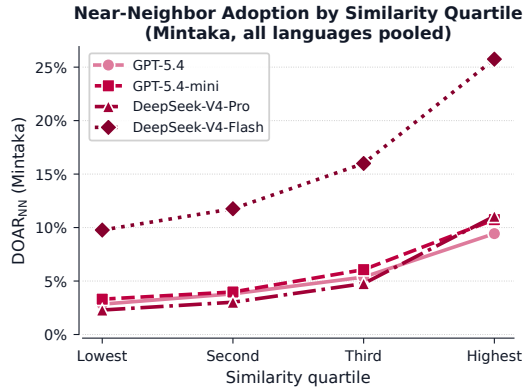


Figure 3: Near-neighbor adoption by similarity quartile (Mintaka, all languages pooled). Adoption rises with how similar the near-neighbor distractor is to the target.

4.3 Near-Miss Adoption Is Higher on Unknown Items

Unknown items are systematically more vulnerable to near-miss context than known items (Figure 4). On Mintaka Arabic, GPT-5.4 near-neighbor adoption rises from 3.8% on known items to 10.2% on unknown items, and DeepSeek-V4-Flash rises from 12.2% to 30.4%. Models that already know the correct answer still show positive adoption, but the effect is substantially larger where context is the primary guide. Model uncertainty and near-miss susceptibility compound each other. We restrict the sample to items each model already answers correctly without context and repeat the near-neighbor check on this known-only subset. Table 4 shows that adoption stays positive in every model-dataset cell for known-only conditions. DeepSeek-V4-Flash shows the largest known-only adoption, at

9.8% on Mintaka and 63.5% on PopQA. Thus, near-miss context can flip an otherwise correct answer into an incorrect one.

5 Language Arbitration Results

5.1 Language Preference Reverses by Family

Evidence arbitration direction depends on model family and target language (Table 6). GPT-5.4-mini leans English across all four languages, from +3.6 on Nepali (95% CI [0.3, 7.1]) to +8.6 on Hindi (95% CI [5.5, 11.8]), with confidence intervals entirely above zero. GPT-5.4 is mostly near zero for Hindi, German, and Nepali, with a supported local preference for Arabic alone (−3.0, 95% CI [−5.4, −0.6]). DeepSeek models lean local in all tested pairs, most strongly for Arabic, where DeepSeek-V4-Pro reaches −14.2% and DeepSeek-V4-Flash reaches −13.3% (full confidence intervals in Appendix Table 9).

We can separate a model’s multilingual accuracy and its evidence preference. Strong English performance is compatible with local-language evidence following when the two conflict, and a directional LABS value in one family member does not transfer to the other, since GPT-5.4 itself stays near zero for three of the four languages even though GPT-5.4-mini shows a clear English-leaning preference throughout.

The EOR–LOR scatter plot (Figure 6) shows the same family split described above. GPT-5.4 sits near the diagonal, where EOR and LOR are close, while GPT-5.4-mini sits clearly below the diagonal, having a clear English preference. DeepSeek models sit above the diagonal in every tested pair, so DeepSeek follows local-language evidence more than English evidence throughout. Within DeepSeek, this local preference is strongest for Arabic (−14.2% for DeepSeek-V4-Pro and −13.3% for DeepSeek-V4-Flash). This pattern is consistent with a learned, source-level preference for which language’s evidence to trust, rather than a general difficulty parsing non-English text. Together, this is language-asymmetric evidence arbitration, a preference for one language’s evidence that varies by model family.

5.2 English Source Labels Are Systematically Over-Trusted

Across all 16 model-language cells, every model over-trusts English-labeled evidence in the label-isolation test (Figure 5), with LLAB ranging from

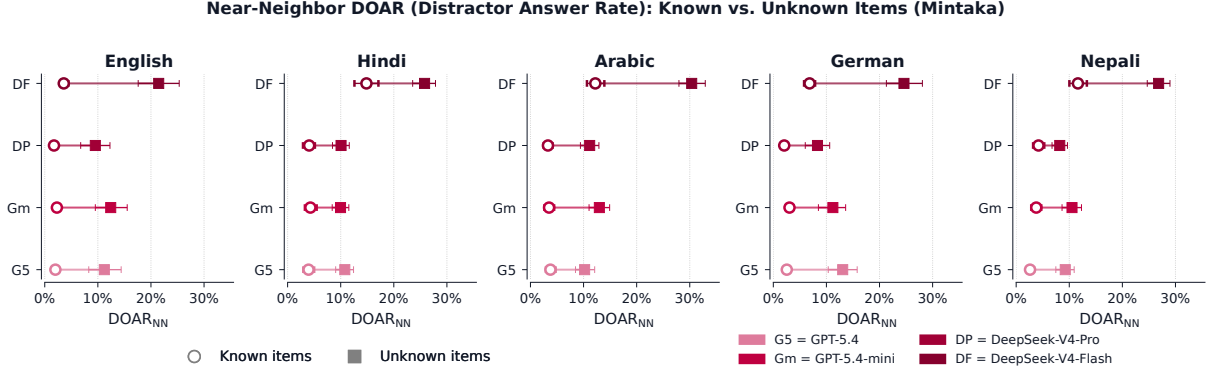


Figure 4: Near-neighbor adoption for known and unknown Mintaka items. Unknown items are usually more vulnerable, so near-miss contamination is most damaging where models depend on context most.

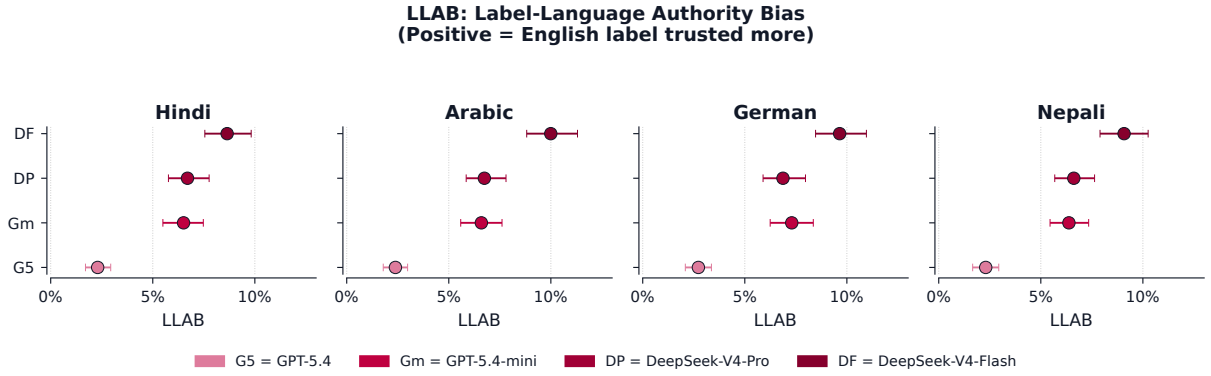


Figure 5: LLAB for all model-language pairs. Every value is positive, showing that an English source label is trusted more even when both snippets are written in English.

Model	English first	English second
GPT-5.4	4.9	10.2
GPT-5.4-mini	10.1	12.8
DS-V4-Pro	2.2	7.6
DS-V4-Flash	3.1	10.5

Table 5: EOR by evidence position, averaged across target languages, with the question posed in the target language.

2.3% to 10.0% (Table 6) and bootstrap confidence intervals uniformly above zero (Appendix Table 9). The label effect coexists with local-language content preference in DeepSeek models. In the multilingual arbitration (Section 5.1), where the two snippets are actually written in different languages, DeepSeek favors the local-language evidence. In the label-isolation test, where both snippets are written in English, DeepSeek still favors the English-tagged one. Thus, a model’s preference for local-language text does not carry over to its trust in an English label.

For RAG systems, the label effect exposes a metadata sensitivity in the generator. Labels such

as “English source”, “translated source”, or a language name can change the model’s answer regardless of the content language. This points to a source-authority heuristic instead of content comparison.

5.3 Evidence Order Does Not Explain the Family Split

We perform a position-swap control, placing English-wrong evidence second instead of first, to test whether LABS reflects presentation order instead of language preference in the target language probe. Table 5 reports EOR with the question posed in the target language only, and its “English first” column is not directly comparable to the pooled EOR in Table 6. English-wrong adoption increases for every model, so position affects absolute override rates. The increase is largest for DeepSeek-V4-Flash, from 3.1% with English first to 10.5% with English second, and GPT-5.4 rises from 4.9% to 10.2%. DeepSeek models nonetheless remain substantially more local-leaning than GPT models in the symmetric EOR–LOR comparison regardless

Model	Language	EOR	LOR	LABS	LLAB
GPT-5.4	Hindi	7.8	8.4	-0.6	2.3
	Arabic	6.7	9.8	-3.0	2.4
	German	6.7	8.9	-2.2	2.7
	Nepali	6.7	9.8	-3.0	2.3
GPT-5.4-mini	Hindi	18.5	9.9	+8.6	6.5
	Arabic	18.0	11.5	+6.5	6.6
	German	15.4	8.5	+6.9	7.3
	Nepali	14.8	11.2	+3.6	6.4
DeepSeek-V4-Pro	Hindi	8.5	14.4	-5.9	6.7
	Arabic	9.4	23.6	-14.2	6.7
	German	5.8	17.3	-11.5	6.9
	Nepali	9.3	18.2	-8.9	6.6
DeepSeek-V4-Flash	Hindi	11.4	20.0	-8.6	8.6
	Arabic	12.4	25.8	-13.3	10.0
	German	8.5	20.8	-12.3	9.6
	Nepali	9.3	22.3	-13.0	9.1

Table 6: Evidence arbitration metrics on Mintaka (%). LABS = EOR − LOR. Positive values indicate English bias, and negative values indicate local-language bias. Full confidence intervals appear in Appendix Table 9.

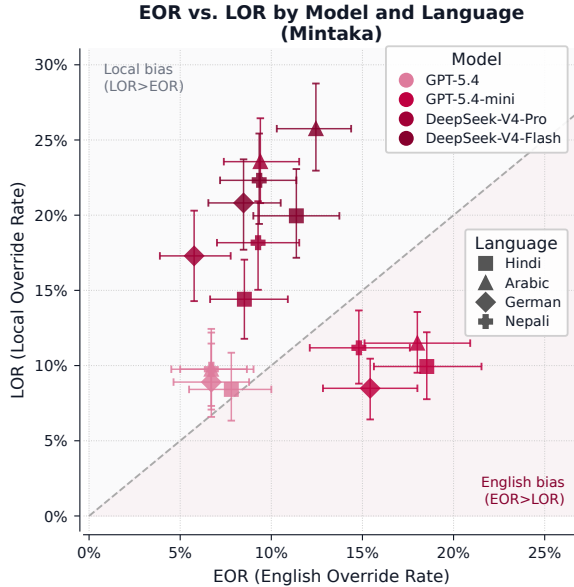


Figure 6: EOR vs. LOR by model and language. Points above the diagonal follow local-language evidence, while points below follow English evidence.

of order, so the family split is not a position artifact.

6 Discussion

6.1 Entity Recognition and Proposition Understanding Can Diverge

The paired probes separate entity recognition from proposition comparison. Surface-level entity recognition is sufficient for near-miss contamination, since a model may recognize or transliterate a Nepali entity name well enough to influence the answer without parsing the full proposition. Evi-

dence arbitration requires comparing claims across languages, a strictly harder task.

This divergence appears in both directions across the two probes. In the near-miss probe (Section 3.2), Table 4 restricts to items each model already answers correctly without context, so the correct proposition is already known, yet distractor adoption stays positive in every model-dataset pair, reaching 9.8% on Mintaka and 63.5% on PopQA for DeepSeek-V4-Flash. Entity-level contamination survives intact propositional knowledge. In the arbitration probe, both snippets carry full propositional content in different languages, and correctness is balanced across the two conflict conditions by design, so a model comparing propositions on their merits should show no consistent pull toward either language. Yet, DeepSeek-V4-Flash follows the local-language claim over the English claim by a wide margin, and GPT-5.4-mini shows the reverse pull toward English (Table 6). Proposition-level comparison gives way to a language-identity preference tied to model family, not to which claim is true. The two probes thus target genuinely separate failure surfaces, and a model’s robustness on one does not predict its robustness on the other.

6.2 Implications for RAG Evaluation

RAG evaluation that measures only answer accuracy under gold retrieval misses generator-side failure modes that our probes expose. A system can retrieve a true passage and still fail when the passage describes the wrong entity, and direct-contradiction tests do not exercise this error because the dis-

tractor is flagrantly wrong rather than plausibly related. Near-miss tests better approximate realistic retrieval mistakes. They are semantically close, relation-matched, and factually true about a different subject. Near-miss contamination effects are generally smaller than direct-contradiction effects but substantial enough to matter in any retrieval setting where passages are not tightly entity-grounded. The same principle extends to language arbitration. A model can retrieve good multilingual evidence and still follow source labels inconsistently across model families, so multilingual RAG systems should test generator-side arbitration separately from retrieval quality.

A practical evaluation suite should thus test near-neighbor contamination, English/local-language conflict arbitration, and label-isolation sensitivity to source tags. Together, these three checks cover entity grounding, evidence-language preference, and source-label authority as separate points where a generator can fail even when retrieval succeeds.

6.3 Generator-Side Grounding Failures

The near-miss setting is easy to misclassify as a retrieval-only problem. If the retriever returns Bhutan for a Nepal question, retrieval has failed to locate the exact supporting passage. The generator still must decide whether the retrieved sentence answers the question, whether the entity in the context matches the entity in the question, and whether the object in the context should be copied into the response. Copying the object from a true but mismatched passage is a generator-side grounding failure.

Improving retrievers will reduce the frequency of near misses, but it will not eliminate them. Entity-centric retrieval tends to surface the closest matching record when the exact one is missing or poorly indexed, and this happens more often for long-tail or low-resource-language queries, where an exact match is less likely to be indexed at all. PopQA is built around exactly these long-tail entities, so real deployments answering PopQA-like questions face this kind of retrieval slippage most often, on top of the model’s own weaker prior knowledge for the same entities. Without subject-identity checks, a plausible retrieved passage can become evidence by default. The same point applies to multilingual conflict. A retrieval system can return two passages that are each reasonable under their own language or translation pipeline, while the generator must resolve disagreement between them. When the

model follows one language’s evidence over the other regardless of which passage is actually correct, the gap belongs to the generator and not the retriever.

Each failure calls for different mitigation methods, some of which we propose here as a direction for future work. An explicit grounding check would compare the subject named in a retrieved passage against the subject named in the question before the passage is used as evidence. Every near-miss error in our results involves exactly this kind of mismatch, so a check at this level would catch it directly. Cross-language disagreement is harder, since a passage can name the correct subject and still be the wrong one to trust. Conflict-aware prompting or evidence aggregation that surfaces disagreement between passages instead of silently picking one is one direction worth exploring. Our LABS results show that this preference reverses across families and is not even guaranteed to hold within one, so a single fixed trust ordering between languages will not generalize across different generator models.

7 Conclusion

We studied how context can mislead factual QA models after retrieval. Near-miss contamination shows that true context about a nearby but incorrect entity can still lead a model away from the correct answer. Models also differ in which language’s evidence they trust when two sources disagree, a language-asymmetric evidence arbitration that does not track how well a model understands either language. Even when the two evidence texts are both written in English and differ only in which language they are tagged with, models consistently trust the English-tagged one. Context use depends on relevance, entity grounding, and which language’s evidence a model chooses to trust. RAG evaluations should include controlled near-miss retrieval errors and multilingual arbitration tests alongside gold-context accuracy.

Limitations

Methodological control. We use templated evidence for both studies. The templates give causal control over what a passage says, where it appears, and what language it is written in. Real RAG systems return longer, noisier, and more variable context than what is constructed here. All experiments use a single prompt template at temperature

zero, applied identically to all four models, so a result reflects how each model behaves under one shared instruction and not the best answer it could reach under a prompt tuned specifically for it. Effect magnitudes could shift under different prompt phrasing, chain-of-thought instructions, or a differently aligned variant of the same underlying model. These are controlled failure probes. They tell us whether a failure mode exists and how it responds to controlled changes in context. Estimating how often such failures occur in a live production system needs a different kind of study, with naturally retrieved passages and varied prompting.

Scope. We test entity-centric factual QA across English, German, Arabic, Hindi, and Nepali. Similar mechanisms may appear in multi-hop reasoning or in longer, discursive responses, but those settings introduce harder evaluation problems. Many other language families and scripts are not tested here. The conclusions should be tested on a wider multilingual set and on harder task types before being treated as general typological or task-level claims.

Translation quality. We translated Mintaka Nepali and PopQA Hindi and Nepali using DeepSeek-V4-Pro. These were subsequently reviewed and corrected by native speakers. Residual code mixing in proper names, titles, awards, and genre labels is domain-typical and expected, and it is not a translation defect. The translation and verification process does not cover all possible idiomatic variants. Results should be interpreted as covering only the tested language conditions. They are not exhaustive typological claims.

Model transparency and drift. Closed and hosted models can change over time, and their parameter counts are not always public. We report the evaluated model names, prompt templates, decoding settings, and aggregate outputs so the experiments can be repeated under the closest available deployment. Exact reproduction may still be affected by provider-side updates.

Potential risks. The probes in this paper create misleading evidence for diagnostic evaluation of RAG systems. The same construction could be misused to make systems more persuasive with wrong evidence. The released artifacts should be used to measure and reduce such failures. They should not be used to deploy misleading retrieval behavior.

Acknowledgements

We thank Ari Holtzman of the University of Chicago for the hypothesis of *similarity eclipsing* (Holtzman, 2026), which motivated our early interest in how information that overlaps with but does not directly contradict a model’s existing knowledge can still shift its answers. We also thank the reviewers and the Area Chair for their helpful feedback throughout the review process.

References

- Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. [The distracting effect: Understanding irrelevant passages in RAG](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18228–18258, Vienna, Austria. Association for Computational Linguistics.
- Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. [XOR QA: Cross-lingual open-retrieval question answering](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.
- Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. [Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. [Benchmarking large language models in retrieval-augmented generation](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages](#). *Transactions of the Association for Computational Linguistics*, 8:454–470.
- María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, and Marcello Federico. 2025. [MEMERAG: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

- (Volume 1: Long Papers), pages 22577–22595, Vienna, Austria. Association for Computational Linguistics.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). Preprint, arXiv:2606.19348.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. [Retrieval augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Ari Holtzman. 2026. [Similarity eclipsing](#).
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. [Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. [Large language models struggle to learn long-tail knowledge](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Patrick Lewis, Barlas Oguz, Rutu Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. [MLQA: Evaluating cross-lingual extractive question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021a. [MKQA: A linguistically diverse benchmark for multilingual open domain question answering](#). *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021b. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.
- Jeonghyun Park and Hwanhee Lee. 2025. [Investigating language preference of multilingual RAG systems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5647–5675, Vienna, Austria. Association for Computational Linguistics.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. [KILT: a benchmark for knowledge](#)

- intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. [Simple entity-centric questions challenge dense retrievers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. [Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts](#). In *International Conference on Learning Representations*, volume 2024, pages 35623–35646.

A Additional Near-Miss Analysis

A.1 Known and Unknown Items

Near-neighbor adoption is consistently larger on unknown items than on known items. The known/unknown split is model-specific and comes from the no-context pass, such that an item is known if the model answers it correctly without context and unknown if it does not. Across all dataset-model-language cells, the unknown-over-known ratio ranges from 1.20x to 6.00x.

At the model level (Figure 7), DeepSeek-V4-Flash shows the largest Mintaka gap, with near-neighbor DOAR averaging 9.8% on known items and 25.8% on unknown items. On PopQA, both bars rise further, with DeepSeek-V4-Flash reaching 63.5% on known items and 85.3% on unknown items, and the other models also showing larger adoption on unknown items.

A.2 Category-Level Heterogeneity

Near-miss effects vary substantially by category (Figure 8). On Mintaka, sports (11.5 points) and geography (7.9 points) show the largest average effects, followed by movies. PopQA category-level effects are considerably larger overall, with genre at 67.7 points and occupation at 52.4 points.

To check whether this variation simply reflects how often the model already knows the answer in each category, we compare each category’s effect size to its unknown share (Table 7, Table 8). On Mintaka, the two show no reliable relationship. Pearson correlation is $r = 0.49$ ($p = 0.22$) and Spearman correlation is $\rho = 0.14$ ($p = 0.74$) across the eight categories, and dropping sports alone collapses the correlation to near zero. On PopQA, a positive correlation appears across the four categories, but it is dependent almost entirely on genre, which has both the highest unknown share and the largest effect. Dropping genre alone collapses this correlation. So, category-level effect size cannot simply be defined by how well the model already knows the category.

A.3 Near-Neighbor Similarity and Adoption

Similarity between the near-neighbor distractor and the target varies continuously across items, span-

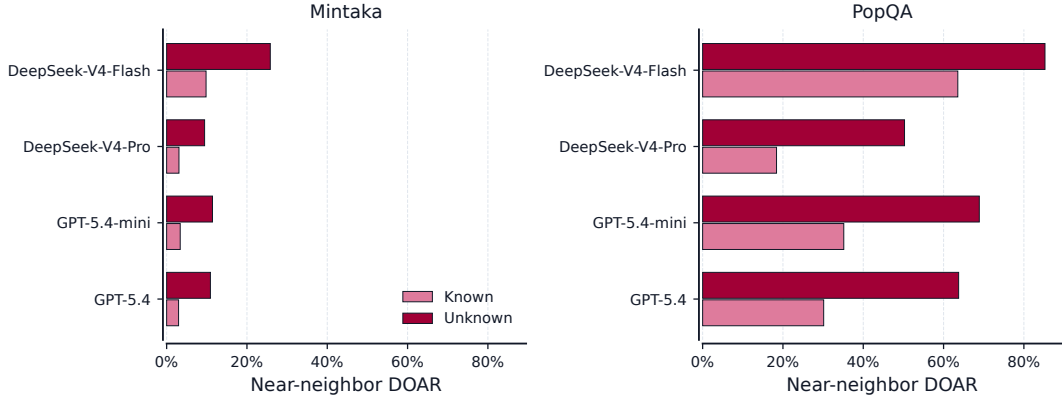


Figure 7: Near-neighbor adoption on known and unknown items, averaged across languages within each dataset. Unknown items are consistently more vulnerable, while the absolute rate depends strongly on the dataset and model.

Category	Known (%)	Unknown (%)	Effect (pts)
Sports	57.7	42.3	11.5
Geography	70.0	30.0	7.9
Movies	66.1	33.9	7.3
Politics	56.4	43.6	7.2
Music	65.0	35.0	6.8
Videogames	69.4	30.6	6.5
History	64.1	35.9	6.1
Books	66.4	33.6	4.9

Table 7: Known/unknown share and near-miss effect size by category (Mintaka).

Category	Known (%)	Unknown (%)	Effect (pts)
Genre	18.7	81.3	67.7
Occupation	34.9	65.1	52.4
Place of birth	24.4	75.6	43.2
Father	36.4	63.6	32.8

Table 8: Known/unknown share and near-miss effect size by category (PopQA).

ning from low similarity to near-identical. Across the 2,757 Mintaka items, cosine similarity has mean 0.639, median 0.637, and interquartile range 0.519–0.759. Across the 3,000 PopQA items, mean similarity is 0.691, median 0.669, and interquartile range 0.610–0.748.

Quartile-binned adoption by model shows that the adoption rises with similarity across every model on both datasets (Figure 9). The pooled point-biserial correlation between similarity and adoption is $r = 0.14$ on Mintaka and $r = 0.28$ on PopQA. Per model-language pair, the correlation ranges from 0.10 to 0.21 on Mintaka and 0.13 to 0.41 on PopQA, and it stays positive throughout. This same rise appears within every tested language as well (Figures 10–11). On Mintaka, DeepSeek-V4-Flash’s curve starts higher than the other three

models at the lowest similarity quartile in every tested language, then rises the same way the other models’ curves do.

This similarity-adoption relationship can be understood better with two examples at different ends of the similarity range. At similarity 0.43, the target question asks who is known as the Father of Medicine (gold: Hippocrates), and the near-neighbor context states that Christiaan Barnard performed the first heart transplant. The two entities share little beyond a loose medical connection, and adoption on this item is rare. At similarity 0.94, the target question asks which actor has the most Oscar nominations for his performances (gold: Jack Nicholson, the record among male actors), and the context answers a differently scoped question about the most Academy Award nominations by any actor, naming Meryl Streep, the overall record holder. The two entities are easy to confuse, and adoption on this item is common.

B Additional Language Arbitration Analysis

B.1 LABS Confidence Intervals

Confidence intervals by model-language pair (Figure 12) show which arbitration estimates are directional. GPT-5.4-mini’s English-leaning estimates for every language are confidence-supported, while GPT-5.4 has one confidence-supported estimate, a local-language preference for Arabic. DeepSeek-V4-Pro and DeepSeek-V4-Flash are local-leaning with intervals excluding zero in all four languages.

B.2 Full Evidence Arbitration Table

Table 9 reports the full confidence intervals behind Table 6, along with the LLAB magnitude for each

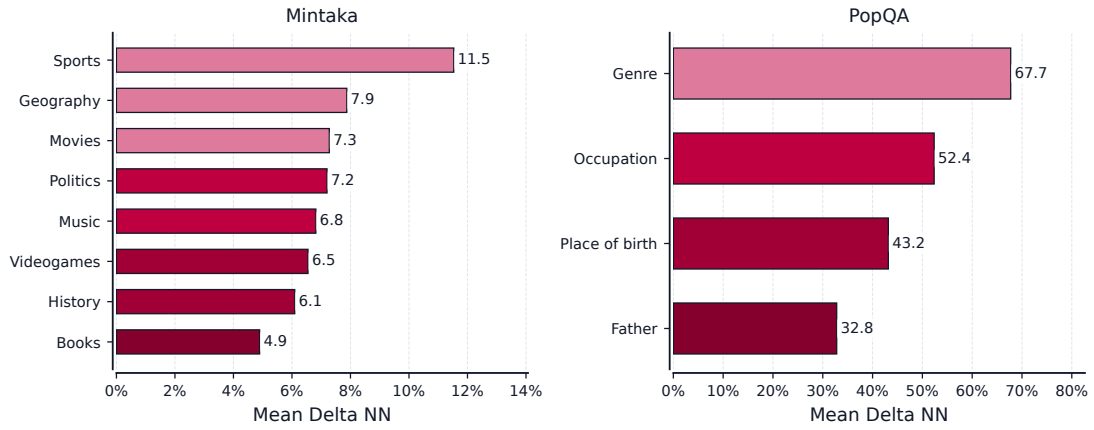


Figure 8: Mean near-miss effect by dataset category, averaged across models and tested languages. Mintaka effects vary across topical categories, while PopQA effects are largest for genre and occupation relations.

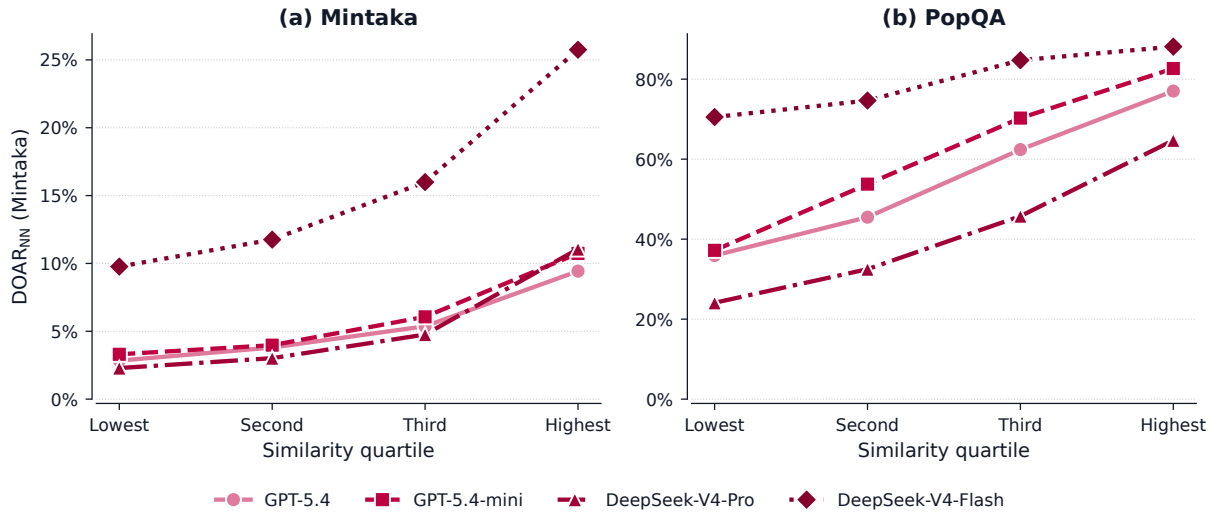


Figure 9: Near-neighbor adoption by similarity quartile, broken down by model, for Mintaka and PopQA.

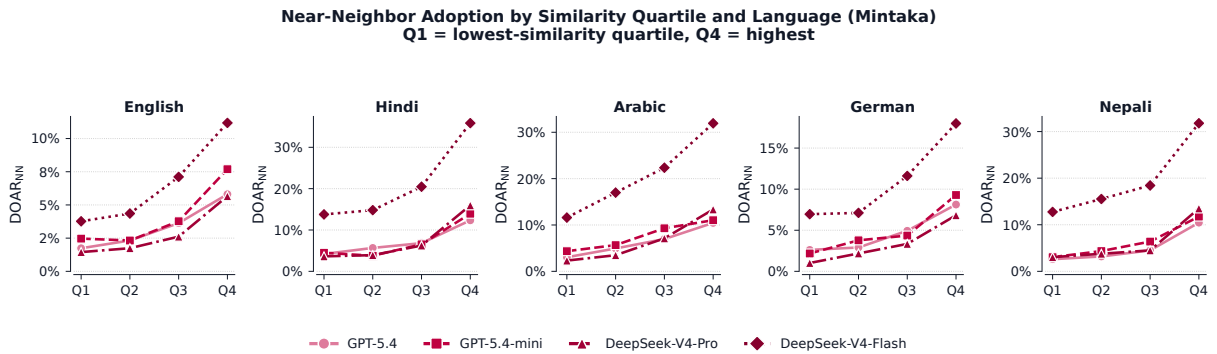


Figure 10: Near-neighbor adoption by similarity quartile and language (Mintaka).

Near-Neighbor Adoption by Similarity Quartile and Language (PopQA)
Q1 = lowest-similarity quartile, Q4 = highest

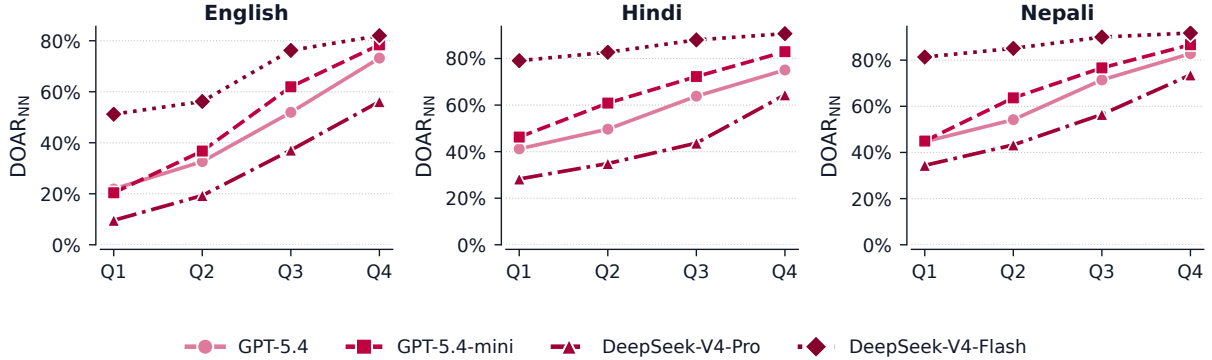


Figure 11: Near-neighbor adoption by similarity quartile and language (PopQA).

model-language pair. LLAB magnitude scales with model, staying near 2.5 points for GPT-5.4, rising to roughly 6.4 to 7.3 points for GPT-5.4-mini and DeepSeek-V4-Pro, and reaching 8.6 to 10.0 points for DeepSeek-V4-Flash. This scaling does not track each model’s own LABS magnitude, or even its sign. GPT-5.4-mini’s LABS spans +3.6 to +8.6, an English-leaning preference, while DeepSeek-V4-Pro’s spans −5.9 to −14.2, a local-leaning preference in the opposite direction, yet the two models land at almost the same LLAB magnitude. Trusting a snippet more because it carries an English label is a separate behavior from the direction or strength of a model’s actual local-content preference.

C Artifacts, Data Quality, and Reproducibility

C.1 Scientific Artifacts and Intended Use

Our experiments use two public QA artifacts, Mintaka and PopQA, and produce derived diagnostic artifacts for near-miss contamination and language evidence arbitration. Mintaka serves as a multilingual entity-centric QA benchmark, and PopQA provides long-tail entity-centric factual questions. The derived artifacts are designed for research evaluation of generator grounding failures. They stress-test whether a generator accepts wrong-entity evidence, follows conflicting evidence, or treats source-language labels as authority cues. Training on them to make models more persuasive with misleading evidence falls outside their intended use.

For context construction, we apply the same evidence-building pipeline to both datasets. Neighbor

construction uses sentence embeddings from all-MiniLM-L6-v2 (Reimers and Gurevych, 2019; Wang et al., 2020). The derived artifacts retain the original question-answer structure and add controlled evidence fields, distractor metadata, model outputs, and aggregate metrics.

C.2 Licensing and Access Conditions

We release the Nepali translations for both datasets and the Hindi translation for PopQA, along with the evaluation code. The translated datasets are released under the same license as the original dataset. The Mintaka translations are released under Mintaka’s own Creative Commons Attribution 4.0 license, and the PopQA translations are released under PopQA’s MIT license. The sentence embedding model used for near-neighbor selection, all-MiniLM-L6-v2, is available under the Apache 2.0 license.

C.3 Dataset Coverage

Mintaka contributes 2,757 items after combining train, development, and test portions used in this evaluation. Each item has a target question, target answer, near-neighbor distractor, unrelated distractor, and evidence fields for English, Hindi, Arabic, German, and Nepali. Items with severe gaps were identified and repaired before experimental runs. PopQA contributes 3,000 stratified items with English, Hindi, and Nepali evidence fields. The near-miss probe evaluates 4 models across 5 Mintaka languages and 3 PopQA languages, with 6 context conditions per item. The arbitration probe evaluates 4 models across 4 target languages on model-calibrated unknown items, plus a label-isolation

Model	Lang	EOR (lo–hi)	LOR (lo–hi)	LABS	LABS (lo–hi)	LLAB (lo–hi)
GPT-5.4	Hindi	7.8 (5.5–10.0)	8.4 (6.3–10.9)	−0.6	−3.3–+2.1	2.3 (1.7–2.9)
	Arabic	6.7 (5.0–8.7)	9.8 (7.3–12.2)	−3.0	−5.4–−0.6	2.4 (1.8–3.0)
	German	6.7 (4.6–8.8)	8.9 (6.6–11.5)	−2.2	−4.5–+0.4	2.7 (2.1–3.4)
	Nepali	6.7 (4.5–9.0)	9.8 (7.1–12.4)	−3.0	−6.0–+0.0	2.3 (1.7–2.9)
GPT-5.4-mini	Hindi	18.5 (15.6–21.5)	9.9 (7.8–12.2)	+8.6	+5.5–+11.8	6.5 (5.5–7.5)
	Arabic	18.0 (15.1–20.9)	11.5 (9.5–13.6)	+6.5	+3.6–+9.4	6.6 (5.6–7.6)
	German	15.4 (12.8–18.0)	8.5 (6.4–10.5)	+6.9	+4.1–+9.9	7.3 (6.2–8.4)
	Nepali	14.8 (12.1–17.6)	11.2 (8.8–13.7)	+3.6	+0.3–+7.1	6.4 (5.5–7.3)
DS-V4-Pro	Hindi	8.5 (6.6–10.9)	14.4 (11.8–17.0)	−5.9	−8.9–−3.1	6.7 (5.8–7.8)
	Arabic	9.4 (7.4–11.5)	23.6 (20.8–26.4)	−14.2	−17.0–−11.2	6.7 (5.9–7.8)
	German	5.8 (3.9–7.8)	17.3 (14.3–20.3)	−11.5	−14.4–−8.8	6.9 (5.9–8.0)
	Nepali	9.3 (7.0–11.5)	18.2 (15.0–20.9)	−8.9	−12.0–−5.6	6.6 (5.7–7.6)
DS-V4-Flash	Hindi	11.4 (9.0–13.7)	20.0 (17.2–23.1)	−8.6	−11.9–−5.4	8.6 (7.6–9.8)
	Arabic	12.4 (10.3–14.4)	25.8 (23.0–28.8)	−13.3	−16.6–−10.3	10.0 (8.8–11.3)
	German	8.5 (6.5–10.5)	20.8 (17.7–23.7)	−12.3	−15.1–−9.5	9.6 (8.5–11.0)
	Nepali	9.3 (7.2–11.4)	22.3 (19.4–25.4)	−13.0	−16.4–−9.7	9.1 (7.9–10.3)

Table 9: Evidence arbitration metrics on Mintaka with 95% bootstrap confidence intervals. LABS is computed using unrounded rates, so it can differ by 0.1 if computed using rounded EOR–LOR from the columns above.

condition.

C.4 Translation Verification and Text Quality

Mintaka Nepali, PopQA Hindi, and PopQA Nepali evidence were translated using DeepSeek-V4-Pro. All translated items were then reviewed by native speakers, who corrected errors in fluency, entity transcription, and grammatical agreement. Mintaka’s Hindi, Arabic, and German evidence is native content, though back-translation was still performed as a quality check during evidence construction. Structural validation for each sample across languages confirmed that each item had the required evidence fields, distractor fields, and result rows. Native-script checks flagged strings with low Devanagari or Arabic-script coverage. Flagged cases were re-examined to distinguish genuine defects from expected Latin-script names, titles, awards, and genre labels.

A small number of Mintaka items had residual defects including missing evidence fields, filtered placeholders, and a spurious offensive suffix. These were repaired before the experiments. The final dataset has zero occurrences of the filtered-content placeholder and the spurious offensive suffix. Residual code mixing in entity names, genre labels, and titles is domain-typical and expected.

C.5 Personally Identifying and Offensive Content

Our probe uses public entity-centric QA resources and generated evidence templates. It does not collect user data, recruit human subjects, or add private

personal information. Some questions and answers refer to public people, places, creative works, and organizations because those entities are part of the underlying public QA datasets. We checked the generated evidence and result records for visible offensive strings and filtered placeholders. The severe offensive-string defect found during auditing was removed from the dataset.

C.6 Experimental Settings and Reproducibility

All model calls use deterministic decoding with temperature 0. The answer budget is short because the prompt asks for only the answer entity. We do not perform any hyperparameter search. Context condition, language, source label, evidence order, and model are the main intervention parameters which we fix by design. Reasoning effort is set to none. Exact reproduction may require the same provider deployment or a dated snapshot, because closed hosted models can change after evaluation.

All confidence intervals are 95% bootstrap intervals over items. The reported tables use means over the relevant item sets, and the appendix figures state when values are averaged across languages, categories, or models. The near-neighbor selection uses cosine similarity over English question embeddings, computed within each dataset separately. The highest-scoring eligible candidate is selected after entity QID and answer-string filtering.

The final evaluation contains 612,048 model-response records. The near-miss probe contributes 546,840 records from 4 models, 2 datasets, 32

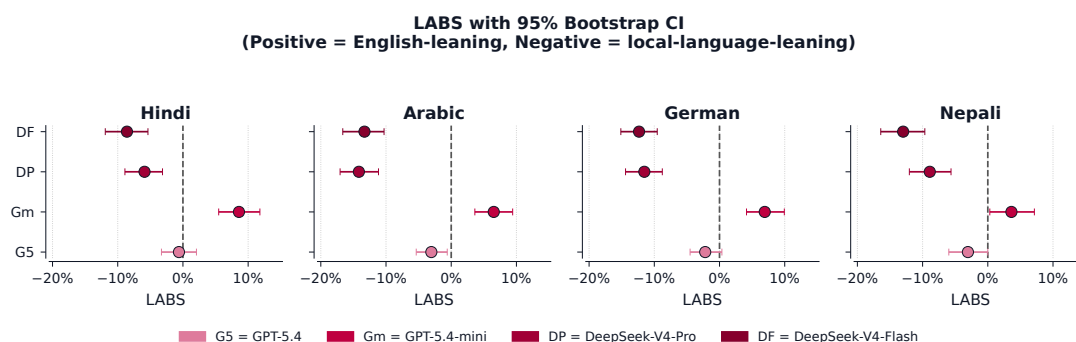


Figure 12: LABS with 95% bootstrap confidence intervals for every model-language cell. Positive values indicate English-leaning arbitration, and negative values indicate local-language-leaning arbitration.

model-language pairs, 6 context conditions, and the full item set for each dataset. Mintaka contributes 2,757 items in English, Hindi, Arabic, German, and Nepali and PopQA contributes 3,000 items in English, Hindi, and Nepali. The arbitration probe contributes 65,208 Mintaka records: 28,128 records for symmetric English/local-language conflict conditions and 37,080 records for label isolation.

Parameter counts are not disclosed by OpenAI and are not reported here. Per its HuggingFace model card, DeepSeek-V4-Pro is a 1.6T-parameter model with 49B activated, and DeepSeek-V4-Flash is a 284B-parameter model with 13B activated.

D Prompt Templates

All prompts are sent as a single user-role message through each provider’s standard chat completions interface, with no system-prompt override. The same message structure is used identically across all four models. Across every template below, only the question, context, and evidence text are shown in the target language, and the instruction text itself is always in English.

Evidence arbitration template

You are given a question and two evidence snippets. The snippets may disagree. Answer the question using the evidence you judge to be correct. Respond with ONLY the answer entity.

Evidence A ({lang_a}):
{text_a}

Evidence B ({lang_b}):
{text_b}

Question:
{question}

Answer:

Near-miss standard template

Answer the question. The context below may or may not be relevant - use your judgement. Respond with only the answer entity, nothing else.

Context:
{context}

Question: {question}

Answer:

Near-miss no-context template

Answer the question. Respond with only the answer entity, nothing else.

Question: {question}

Answer:

Label-isolation template

You are given two evidence snippets and a question.

Evidence A (labeled as: {label_a}):
{text_a}

Evidence B (labeled as: {label_b}):
{text_b}

Question: {question}

Answer with only the answer entity.