Evaluating Contextually Mediated Factual Recall in Multilingual Large Language Models

Yihong Liu[1,2], **Bingyu Xiong**[1], **and Hinrich Schütze**[1,2]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
yihong@cis.lmu.de

Abstract

Large language models (LLMs) can recall a wide range of factual knowledge across languages. However, existing factual recall evaluations primarily assess fact retrieval in isolation, where the queried entity is explicitly named and the fact is requested directly. In natural language use, facts are often accessed through context, where the relevant entity is introduced only *indirectly*. In this work, we study *contextually mediated factual recall*, asking whether LLMs can reliably retrieve factual knowledge when the target entity is embedded in a naturalistic context rather than queried explicitly, across languages. We construct controlled prompts that preserve the underlying fact while introducing referential mediation through contextual sentences. To disentangle contextual effects from name-specific associations, we further compare performance using synthetic names and real names across languages. Evaluating multiple model families in five languages, we find that contextual mediation consistently degrades factual recall, with substantial variation across relations. Larger models are more robust to contextual mediation, exhibiting a reduced performance gap relative to direct queries, while the effect of real names and name origin is mixed and unsystematic. These findings highlight a gap between isolated factual recall and context-dependent language understanding in multilingual LLMs.

1 Introduction

LLMs have demonstrated a strong ability to store and recall factual knowledge spanning various domains gained from their pretraining stage (Petroni et al., 2019; Jiang et al., 2020b). This capability extends beyond English, with multilingual models showing factual recall across languages (Jiang et al., 2020a; Kassner et al., 2021; Yin et al., 2022).

However, most existing evaluations assess factual recall in a highly *decontextualized* setting, where the fact is queried directly and the relevant entity is explicitly named (e.g., "What is the capital of Germany?"). Such formulations minimize linguistic and discourse complexity, but they abstract away from how facts are accessed in natural language, where entities are frequently introduced indirectly and facts must be retrieved through contextual reference (e.g., "After arriving in Germany, Alex heads to the country's political capital. Which city does Alex go to?"). This raises an open question: *does factual knowledge remain equally accessible when it is queried through context?*

In this paper, we evaluate *contextually mediated factual recall* by systematically comparing direct factual queries with contextually mediated counterparts that preserve the same underlying fact while introducing referential mediation (§4). We further examine whether factual recall is influenced by name-specific effects by contrasting synthetic, unseen names with common real names across five languages (§5). Our evaluation spans five languages on three model families, allowing us to analyze how contextual mediation interacts with relation type, model scale, and language. Overall, our results show that contextual mediation consistently degrades factual recall, with substantial variation across relations, while larger models exhibit greater robustness. In contrast, the influence of real names and name origin is mixed and unsystematic, suggesting that factual recall under contextual mediation is driven primarily by surrounding context rather than name surface forms.

2 Related Work

Recent work shows that LLMs are sensitive to contextual information and can be distracted by irrelevant or misleading content (Petroni et al., 2020; Shi et al., 2023). Such effects are amplified in long contexts due to positional biases and competition for attention (Mohtashami and Jaggi, 2023; Liu et al., 2024; Modarressi et al., 2025), and when contex-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 78–87
November 6-10, 2026 ©2026 Association for Computational Linguistics

tual information conflicts with a model's parametric knowledge (Longpre et al., 2021; Chen et al., 2022; Wang et al., 2023; Xu et al., 2024).

LLMs store and recall factual knowledge (Petroni et al., 2019), which has been extensively studied using probing and prompt-based evaluations in English (Roberts et al., 2020; Peng et al., 2022) and in multilingual settings (Jiang et al., 2020a; Kassner et al., 2021; Yin et al., 2022). Several studies show that knowledge acquisition correlates with exposure frequency during pretraining (Liu et al., 2025c; Merullo et al., 2025). Recent work also examines crosslingual consistency and attributes inconsistent recall to representational and entity-level misalignment (Qi et al., 2023; Wang et al., 2025; Lu et al., 2025; Liu et al., 2025d).

Unlike prior work that evaluates factual recall through direct queries or treats context as optional or distracting, we study *contextually mediated factual recall*, where the target entity is introduced indirectly through context. This setting allows us to assess factual robustness under naturalistic contextual mediation rather than isolated fact retrieval.

3 General Experimental Setup

Models We evaluate 9 LLMs of different sizes from 3 model families: **LLaMA** (Grattafiori et al., 2024), **Qwen** (Yang et al., 2025), and **Gemma** (Gemma Team et al., 2025). All models are trained on highly multilingual corpora. See details in §E.

Dataset We use KLAR (Wang et al., 2025), a multilingual factual knowledge probing dataset, in this research. Our evaluation covers **1,742** facts from **9** relation types (cf. Table 2 in §A). We use language-specific prompt templates for each relation provided by KLAR for our direct factual knowledge queries.

Languages. Our evaluation covers five typologically diverse languages spanning multiple writing systems: English (**EN**), Arabic (**AR**), Japanese (**JA**), Korean (**KO**), and Chinese (**ZH**).

Evaluation. We evaluate factual recall using exact-match accuracy. Models are prompted with three in-context examples followed by the test query, and are allowed to generate up to 10 tokens. A prediction is considered correct if the gold answer appears as a prefix of the generated output.

4 Contextually Mediated Factual Recall

This section introduces our setup for evaluating *contextually mediated factual recall* and contrasts it with standard direct factual queries.

4.1 Methodology

In a contextually mediated factual query, factual access is mediated by a short naturalistic context rather than an explicit entity-centric question. Each query consists of two components: (**i**) a *context sentence* that embeds the target entity in a realistic scenario, and (**ii**) a *query sentence* that asks for a factual attribute of that entity. A placeholder *name* appears across the two sentences and serves as a bridge to maintain discourse coherence.

Formally, let e denote the target entity and r a factual relation. A contextually mediated query takes the form: $c(e, n) \parallel q(r, n)$, where c introduces e indirectly through a context involving a name n, and q query the object of relation r by referring to the same name. Correct factual recall requires resolving the referent introduced in the context before recalling the factual relation.

Synthetical Names To minimize name-based biases, we use *synthetic names* that are unlikely to be associated with real individuals. Names can encode demographic, cultural, or frequency-based signals that influence model behavior independently of factual knowledge (Wei et al., 2024; Manchanda and Shivaswamy, 2025). To control for this factor,[1] we query a commercial model (i.e., `ChatGPT-5.2`) to generate candidate synthetic names. We then filter these candidates using Infini-gram (Liu et al., 2025a), which allows querying n-gram frequencies across large corpora.[2] Only names that do not occur in any of the indexed corpora are retained, ensuring minimal prior exposure during pretraining.

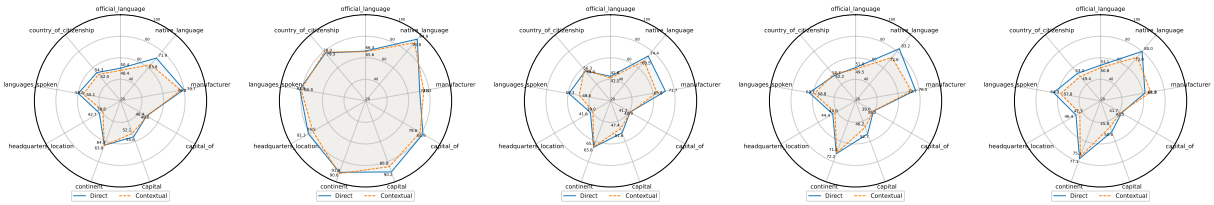
Prompt Format We design two-sentence prompts for each factual relation in English, where the context sentence introduces the target entity implicitly, and the query sentence requests the associated fact. For example, for the *official language* relation, we use:[3]

> "`{name}` *is traveling to* `{subject}` *for a business trip and needs to complete local government forms. Which de jure*

[1] We analyze the effect of real names from different languages separately in §5.

[2] https://infini-gram.readthedocs.io/

[3] See §D for complete templates for all relations.



(a) Arabic (b) English (c) Japanese (d) Korean (e) Chinese

Figure 1: Per-relation factual recall performance for five languages under *direct factual queries* and *contextually mediated queries*. Each radar plot shows accuracy averaged across all models. Across languages, introducing contextual mediation generally leads to a relation-dependent performance degradation compared to direct queries.

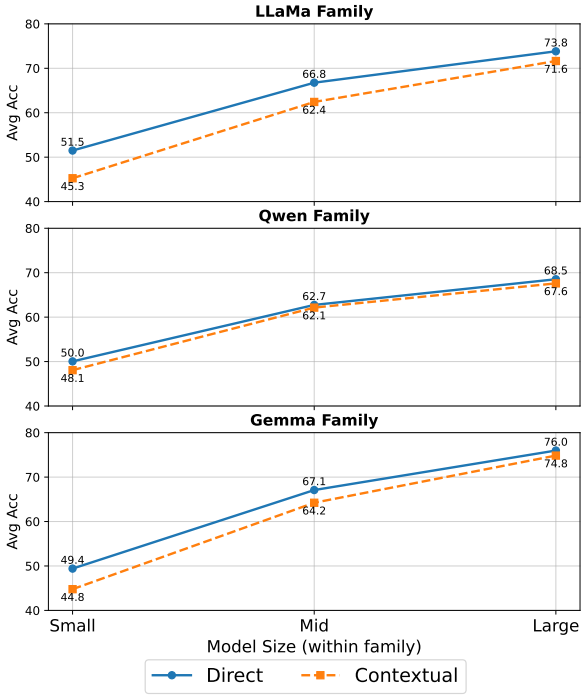


Figure 2: Average factual recall performance across relations as a function of model size for each model family. Larger models in the LLaMA and Gemma families exhibit increased robustness to contextual mediation, while the Qwen family shows a weaker size effect.

> *official language should* `{name}` *use on the paperwork?"*

We construct analogous two-sentence templates for other relations, where the underlying fact remains identical to direct factual queries. The templates are then translated into the other four languages.

4.2 Results and Discussion

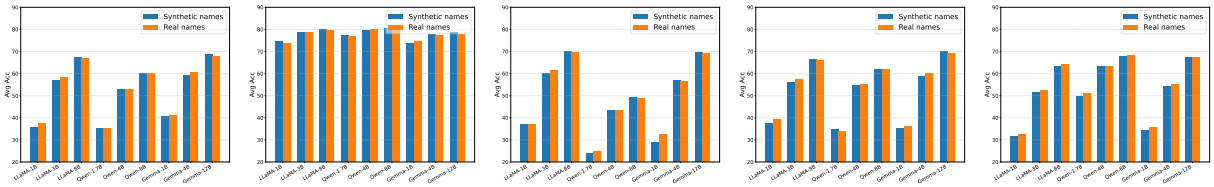
Figure 1 reports per-relation factual recall under direct and contextually mediated queries across languages, allowing us to analyze how introducing indirect contextual references affects factual recall. Figure 2 further examines whether model scale influences robustness to contextual mediation.

Contextual mediation generally reduces factual recall, with strong relation-level variation. Across all languages, contextually mediated queries lead to lower factual recall than direct queries. However, the impact varies substantially by relation. Relations such as `native_language`, `capital`, and `headquarters_location` are consistently the most sensitive, while others, notably `continent`, remain largely stable. We hypothesize that this robustness is related to the limited candidate space of certain relations, which reduces ambiguity even under indirect entity access (Liu et al., 2025b). In some cases, contextual mediation can even improve performance (e.g., `manufacturer` in English and Chinese), suggesting that contextual cues may occasionally aid disambiguation. We further present an error breakdown in §C.

Larger models exhibit increased robustness to contextual mediation. As expected, overall factual recall improves with model scale for both direct and contextually mediated queries. More importantly, larger models tend to be more robust to contextual mediation, as reflected by a reduced performance gap between direct and mediated settings. This trend is clearly observed in the LLaMA and Gemma families. In contrast, the Qwen family shows weaker or less consistent gains in robustness with scale. These findings suggest that increased model capacity improves not only factual knowledge coverage but also the ability to recover relevant facts when entity access is indirect. This may reflect stronger internal representations of entities and relations, or more effective integration of contextual cues in larger models.

5 Influence of Real-Name Bias

In §4, we use synthetic names to minimize potential biases arising from names seen during pretraining. In this section, we explicitly examine whether re-



(a) Arabic (b) English (c) Japanese (d) Korean (e) Chinese

Figure 3: Per-language comparison of contextually mediated factual recall using synthetic versus real names, averaged across relations for each model. The effect of replacing synthetic names with real names is highly mixed across both languages and models, showing no consistent increase or decrease in performance.

Model	EN	AR	JA	KO	ZH
LLaMA-1B	73.9	74.1	73.8	73.4	73.8
LLaMA-3B	78.8	78.0	78.9	78.6	78.8
LLaMA-8B	79.7	79.3	79.8	80.0	79.7
Qwen-1.7B	76.9	76.6	77.2	76.9	76.6
Qwen-4B	80.1	79.0	80.1	79.7	79.9
Qwen-8B	80.1	79.9	80.1	79.9	79.6
Gemma-3-1B	74.5	73.5	74.2	74.3	74.1
Gemma-3-4B	77.2	77.0	77.7	77.5	77.4
Gemma-3-12B	77.7	77.8	78.5	78.1	77.8

Table 1: English contextually mediated factual recall with real names from different language origins. Each cell reports overall accuracy averaged across relations. Across model families and sizes, performance remains largely stable across name origins, indicating no systematic degradation or advantage from non-English names.

placing synthetic names with *real* names further influences contextually mediated factual recall.

5.1 Setup

Real Names. For each considered language, we collect 50 common personal names (25 female and 25 male). Each name is represented both in its *original* script and in a *Latin transliteration*.

Per-Language Queries. We construct contextually mediated factual queries using the same templates as in §4, but replace synthetic names with real names from the corresponding language, written in their original script. We compare the resulting performance against the synthetic-name condition obtained in §4 to assess whether real names introduce additional within-language biases.

English Queries with Crosslingual Names. To further probe the interaction between name origin and contextual mediation, we insert transliterated real names from non-English languages into the English prompt templates used in §4. This setting allows us to test whether names associated with different cultural backgrounds differentially affect English contextually mediated factual recall, where model performance is typically strongest.

5.2 Results and Discussion

Figure 3 compares contextually mediated factual recall when using synthetic versus real names across languages. Table 1 further reports English contextually mediated factual recall using transliterated real names from five language origins.

Real names do not introduce a systematic performance trend. Across models and languages, replacing synthetic names with real names does not lead to a consistent increase or decrease in factual recall. While real names occasionally yield small performance gains (e.g., for Chinese names in the LLaMA family), these effects are neither stable across models nor consistent across languages. Overall, the influence of real names appears highly heterogeneous, However, there is no systematic trend, suggesting that name-related biases interact with model-specific and language-specific factors rather than introducing a uniform effect.

Name origin plays a limited role compared to contextual mediation. Using transliterated real names from different language origins in English prompts results in only marginal performance variation. For example, across LLaMA, Qwen, and Gemma models in Table 1, accuracies remain within a narrow range (typically within $\approx$1%) when replacing English names with Arabic, Japanese, Korean, or Chinese names. Together with the per-language analysis, this suggests that under contextual mediation, factual recall is driven primarily by the *surrounding context* rather than by the specific surface form or linguistic origin of the name.

6 Discussion

Interpreting the recall gap. Contextually mediated queries require connecting the described situation to the relevant entity and factual relation before retrieving an answer. This additional interpretation step may make factual knowledge less accessible. The direct-correct but mediated-incorrect cases in

§C support this interpretation: successful recall under direct prompting does not guarantee successful use of the same fact in context. The mixed effects of real names further suggest that name familiarity alone does not explain the gap. However, our experiments do not isolate the contributions of referential resolution, wording, and other contextual changes.

Practical relevance and improvement directions. Related dependencies arise in conversational QA when questions refer to entities introduced earlier (Anantha et al., 2021). Our templates capture a simplified form of this interaction, although we do not measure its frequency in deployed systems. The observed gap motivates rewriting contextual questions into direct queries that explicitly identify the target entity and relation before answering. A complementary direction is training on paired direct and contextual formulations across languages, using the same factual answer while varying scenarios and referring expressions. Both approaches aim to improve the connection between contextual interpretation and factual retrieval; their effectiveness remains to be tested.

7 Conclusion

We study *contextually mediated factual recall*, a setting where factual knowledge must be retrieved through indirect contextual cues rather than explicit entity mentions. Across five languages and multiple model families, we find that introducing contextual mediation generally degrades factual recall performance, with substantial variation across relations. We further show that larger models are more resilient to contextual mediation. Finally, we find no systematic evidence that real names introduce consistent bias under contextual mediation, indicating that surrounding context plays a more dominant role than name surface forms. Overall, our results highlight important limitations of current factual recall evaluations and motivate the need for more context-aware and multilingual benchmarks.

Limitations

While this study provides insights into contextually mediated factual recall, it has several limitations.

First, we focus on a fixed set of relations and prompt templates, which may not fully capture the diversity and complexity of contextual references encountered in natural language use.

Second, our analysis of name bias is based on a limited set of common names and does not consider sociolinguistic factors such as name frequency or variation across demographic groups.

Finally, although we evaluate multiple open-source model families across different scales, our analysis excludes closed-source models, which may exhibit different or more robust behavior under contextual mediation.

Acknowledgments

This research was supported by the Munich Center for Machine Learning (MCML) and German Research Foundation (DFG, grant SCHU 2246/14-1).

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, 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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 25 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 16 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2025a. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *Preprint*, arXiv:2401.17377.

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.

Yihong Liu, Runsheng Chen, Lea Hirlimann, Ahmad Dawar Hakimi, Mingyang Wang, Amir Hossein Kargaran, Sascha Rothe, François Yvon, and Hinrich Schuetze. 2025b. On relation-specific neurons in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 992–1022, Suzhou, China. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. 2025c. Tracing multilingual factual knowledge acquisition in pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2121–2146, Suzhou, China. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, François Yvon, and Hinrich Schütze. 2025d. On the entity-level alignment in crosslingual consistency. *Preprint*, arXiv:2510.10280.

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

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15077–15107, Suzhou, China. Association for Computational Linguistics.

Sahil Manchanda and Pannaga Shivaswamy. 2025. What is in a name? mitigating name bias in text embedding similarity via anonymization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17759–17781, Vienna, Austria. Association for Computational Linguistics.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. 2025. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: Probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5015–5035, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.

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.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for 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 *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Yumou Wei, Paulo F. Carvalho, and John Stamper. 2024. Uncovering name-based biases in large language models through simulated trust game. *Preprint*, arXiv:2404.14682.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *Preprint*, arXiv:2403.08319.

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

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A KLAR Dataset Details

The statistics of the used subset of the KLAR dataset (Wang et al., 2025) are presented in Table 2. KLAR is based on BMLAMA17 (Qi et al., 2023) that covers 17 languages and contains 20 relation types. We use **1,742** facts grouped into **9** relation categories and **5** typologically different languages.

Relation	Number of Facts
`capital`	336
`capital_of`	212
`continent`	212
`country_of_citizenship`	60
`headquarters_location`	51
`languages_spoken`	104
`manufacturer`	35
`native_language`	130
`official_language`	602
Total	**1,742**

Table 2: Number of facts per relation type.

Model	`native_language`	`capital`	`headquarters_location`
LLaMA-1B	22	33	10
LLaMA-3B	4	14	4
LLaMA-8B	6	13	0
Qwen-1.7B	7	23	3
Qwen-4B	3	10	5
Qwen-8B	2	26	2
Gemma-3-1B	12	30	7
Gemma-3-4B	6	45	4
Gemma-3-12B	2	43	2

Table 3: Counts of *contextual-mediation failures* on English for the most susceptible relations. Each entry reports the number of facts where a model answers the direct query correctly but fails on the corresponding contextually mediated query.

B Complete Results

We present complete per-relation factual recall results covering both direct and contextually mediated queries, for all evaluated models and languages in Table 5, 6, 7, 8, and 9.

C Error Breakdown

We conduct an error analysis to manually check the relations that are most susceptible to contextual mediation in §4, namely `native_language`, `capital`, and `headquarters_location`. Specifically, we focus on facts for which the model correctly answers the direct query but fails under the corresponding contextually mediated formulation, allowing us to isolate errors attributable to contextual mediation rather than missing factual knowledge.

We report the number of such facts for each model and relation in Table 3. Overall, the number of contextual-mediation failures decreases with model scale, most clearly for `native_language` and `headquarters_location`, reinforcing our earlier finding that larger models are generally more robust to contextual mediation (cf. §4). In contrast, the effect of scale on `capital` is less consistent: while LLaMA models exhibit a clear reduction in failures as size increases, both Qwen and Gemma

show non-monotonic trends, with larger models even incurring more errors than their smaller counterparts. This suggests that robustness to contextual mediation is relation-dependent and cannot be explained by model scale alone, pointing to differences in how factual associations are represented and accessed across model families.

D Prompt Templates

We present the prompt template for direct factual recall and contextually mediated factual recall we used in our study for English in Table 4. The prompt templates used for the other four languages (i.e., Arabic, Japanese, Korean, Chinese) are translations of the English prompt templates, which we create using `ChatGPT-5.2` and then have native speakers manually verify.

E Models, Hyperparameters, and Environment

E.1 Models

We evaluate a collection of 9 decoder-only LLMs drawn from four prominent model families: **LLaMA** (Grattafiori et al., 2024), **Qwen** (Yang et al., 2025), and **Gemma** (Gemma Team et al., 2025). LLaMA, Qwen, and Gemma are trained on broadly multilingual corpora. Within the LLaMA family, we evaluate `Llama-3.2-1B`, `Llama-3.2-3B`, and `Llama-3.1-8B`. For Qwen, we include `Qwen3-1.7B-Base`, `Qwen3-4B-Base`, and `Qwen3-8B-Base`. The Gemma models considered are `gemma-3-1b-pt`, `gemma-3-4b-pt`, and `gemma-3-12b-pt`.

E.2 Hyperparameters

All experiments use greedy decoding with a 3-shot prompting setup, where in-context examples are randomly sampled from the same relation. We fix the random seed to 12345 for reproducibility.

E.3 Environment

All experiments are executed using the vLLM framework.[4] Experiments are run on NVIDIA RTX A6000 GPUs with 48 GB of memory.

[4] https://vllm.ai/

Relation	Query Type	Prompt Template
Official language	Direct	What is the official language of {subject}?
	Contextual	{name} is traveling to {subject} for a business trip and needs to complete local government forms. Which de jure official language should {name} use on the paperwork?
Native language	Direct	What is the native language of {subject}?
	Contextual	{name} finds a private letter {subject} wrote to their family as a child. Assuming it was written in the author's native language, what language does {name} see the letter is written in?
Manufacturer	Direct	Which company manufactures {subject}?
	Contextual	{name} picks up {subject} to contact customer service. Which company should {name} contact as the manufacturer?
Capital of	Direct	What is the country or state for which {subject} is the capital?
	Contextual	{name} arrives in the city of {subject}. Which country or state is this city the capital of?
Capital	Direct	Where is {subject}'s capital located?
	Contextual	After arriving in {subject}, {name} heads to the country's political capital. Which city does {name} go to?
Continent	Direct	Which continent is {subject} located in?
	Contextual	{name} is planning a trip to {subject}. Which continent will {name} be traveling to?
Headquarters location	Direct	Where is the headquarters of {subject}?
	Contextual	{name} walks into the headquarters lobby of {subject}. Which city is {name} in right now?
Languages spoken	Direct	What language did {subject} use to communicate?
	Contextual	{name} receives a voice message from {subject}. Which language does {name} hear in the message?
Country of citizenship	Direct	Which country is {subject} a citizen of?
	Contextual	{name} is at border control checking {subject}'s documents. Which country of citizenship does {name} see listed?

Table 4: Prompt templates for direct and contextually mediated factual recall across relations in **English**. {subject} denotes the target entity and {name} denotes a (synthetic or real) person name used for contextual mediation.

Relation	Query	LLaMA-1B	LLaMA-3B	LLaMA-8B	Qwen-1.7B	Qwen-4B	Qwen-8B	Gemma-3-1B	Gemma-3-4B	Gemma-3-12B
official_language	Direct	66.3	67.9	67.8	65.6	64.6	65.0	65.3	67.8	66.3
	Contextual	64.1	65.1	66.3	65.3	65.0	68.4	62.0	67.8	66.6
native_language	Direct	93.8	96.2	95.4	93.8	95.4	95.4	90.0	95.4	98.5
	Contextual	79.2	93.8	92.3	90.0	94.6	95.4	82.3	92.3	96.9
manufacturer	Direct	74.3	74.3	71.4	65.7	74.3	71.4	65.7	71.4	71.4
	Contextual	74.3	77.1	77.1	74.3	77.1	71.4	71.4	77.1	74.3
capital_of	Direct	76.4	81.6	83.0	78.3	84.4	84.0	79.2	82.5	84.9
	Contextual	77.4	76.9	78.3	75.0	83.5	81.6	77.4	77.4	79.7
capital	Direct	90.2	90.8	92.0	88.1	89.6	90.5	87.5	91.1	92.6
	Contextual	83.0	88.7	89.9	85.7	90.5	85.4	81.8	79.5	80.4
continent	Direct	86.8	91.5	93.4	89.6	89.2	93.4	90.6	89.6	91.5
	Contextual	91.5	91.5	93.4	92.0	91.5	93.9	89.6	90.1	93.9
headquarters_location	Direct	76.5	90.2	90.2	70.6	72.5	86.3	66.7	86.3	92.2
	Contextual	64.7	84.3	90.2	70.6	66.7	84.3	56.9	80.4	90.2
languages_spoken	Direct	80.8	83.7	81.7	83.7	84.6	78.8	80.8	82.7	80.8
	Contextual	70.2	81.7	83.7	80.8	84.6	82.7	78.8	82.7	81.7
country_of_citizenship	Direct	71.7	80.0	81.7	73.3	76.7	78.3	80.0	81.7	78.3
	Contextual	70.0	78.3	85.0	80.0	83.3	85.0	70.0	80.0	81.7

Table 5: Per-relation accuracy for direct and contextually mediated queries. Language: **English**.

Relation	Query	LLaMA-1B	LLaMA-3B	LLaMA-8B	Qwen-1.7B	Qwen-4B	Qwen-8B	Gemma-3-1B	Gemma-3-4B	Gemma-3-12B
official_language	Direct	41.4	55.8	60.1	37.2	48.5	55.0	37.9	55.3	62.1
	Contextual	31.7	51.2	60.1	33.6	49.5	55.8	36.2	55.3	62.0
native_language	Direct	60.0	76.9	90.0	49.2	66.9	79.2	55.4	76.9	92.3
	Contextual	35.4	73.1	77.7	36.9	73.1	76.9	43.8	71.5	86.2
manufacturer	Direct	74.3	82.9	85.7	68.6	80.0	82.9	65.7	88.6	88.6
	Contextual	57.1	71.4	82.9	68.6	82.9	88.6	62.9	85.7	91.4
capital_of	Direct	29.7	52.8	59.4	23.1	47.6	55.7	35.4	58.5	69.3
	Contextual	27.8	51.4	64.2	24.1	46.7	58.0	34.0	58.0	71.7
capital	Direct	47.9	64.6	73.8	29.8	48.2	58.6	45.2	61.0	71.4
	Contextual	45.5	61.6	70.5	27.1	48.5	55.1	39.6	56.5	64.3
continent	Direct	49.1	67.9	76.4	50.5	64.6	71.2	50.0	69.8	75.0
	Contextual	43.4	69.3	77.4	54.7	63.7	70.3	56.6	67.9	75.5
headquarters_location	Direct	35.3	41.2	52.9	25.5	37.3	45.1	37.3	45.1	64.7
	Contextual	17.6	43.1	58.8	21.6	35.3	43.1	17.6	49.0	62.7
languages_spoken	Direct	42.3	63.5	73.1	43.3	61.5	70.2	43.3	65.4	76.0
	Contextual	33.7	52.9	72.1	42.3	51.9	62.5	45.2	60.6	75.0
country_of_citizenship	Direct	40.0	50.0	70.0	41.7	51.7	68.3	45.0	51.7	70.0
	Contextual	33.3	43.3	66.7	40.0	51.7	63.3	45.0	53.3	71.7

Table 6: Per-relation accuracy for direct and contextually mediated queries. Language: **Arabic**.

Relation	Query	LLaMA-1B	LLaMA-3B	LLaMA-8B	Qwen-1.7B	Qwen-4B	Qwen-8B	Gemma-3-1B	Gemma-3-4B	Gemma-3-12B
official_language	Direct	35.5	56.5	60.1	15.8	35.9	38.5	29.9	51.3	61.6
	Contextual	29.1	53.8	62.5	15.8	35.4	38.5	27.7	50.0	58.3
native_language	Direct	69.2	83.1	90.8	52.3	74.6	67.7	60.8	78.5	92.3
	Contextual	52.3	78.5	86.9	42.3	70.0	68.5	50.8	76.2	90.8
manufacturer	Direct	60.0	82.9	80.0	57.1	68.6	74.3	54.3	82.9	85.7
	Contextual	51.4	65.7	71.4	48.6	60.0	68.6	51.4	80.0	77.1
capital_of	Direct	25.0	49.1	62.7	10.8	36.8	50.0	16.5	49.1	68.4
	Contextual	22.6	50.5	65.1	10.4	34.4	50.5	15.6	53.3	68.9
capital	Direct	45.8	68.5	76.5	17.6	40.8	52.7	27.4	61.0	74.1
	Contextual	39.3	64.3	76.2	14.0	41.1	47.3	22.3	54.2	68.2
continent	Direct	51.9	79.7	83.5	57.1	59.4	66.0	35.4	74.5	84.9
	Contextual	62.3	76.9	78.3	58.0	56.1	62.7	37.7	70.3	85.4
headquarters_location	Direct	31.4	51.0	60.8	21.6	37.3	45.1	21.6	43.1	62.7
	Contextual	21.6	39.2	54.9	25.5	35.3	45.1	19.6	45.1	64.7
languages_spoken	Direct	51.9	71.2	81.7	31.7	45.2	50.0	51.9	75.0	75.0
	Contextual	37.5	53.8	71.2	19.2	43.3	48.1	34.6	61.5	76.9
country_of_citizenship	Direct	48.3	61.7	63.3	38.3	53.3	68.3	48.3	53.3	71.7
	Contextual	43.3	65.0	68.3	40.0	60.0	71.7	36.7	58.3	78.3

Table 7: Per-relation accuracy for direct and contextually mediated queries. Language: **Japanese**.

Relation	Query	LLaMA-1B	LLaMA-3B	LLaMA-8B	Qwen-1.7B	Qwen-4B	Qwen-8B	Gemma-3-1B	Gemma-3-4B	Gemma-3-12B
official_language	Direct	44.0	58.0	59.3	36.0	49.8	57.8	36.4	56.0	61.5
	Contextual	38.7	52.7	61.3	31.1	54.3	58.3	33.1	54.2	62.3
native_language	Direct	69.2	84.6	93.8	73.1	85.4	88.5	74.6	87.7	92.3
	Contextual	51.5	78.5	80.0	66.9	79.2	80.8	53.1	76.9	86.9
manufacturer	Direct	65.7	82.9	88.6	57.1	74.3	85.7	65.7	82.9	85.7
	Contextual	62.9	74.3	85.7	51.4	71.4	82.9	60.0	82.9	82.9
capital_of	Direct	17.9	44.8	50.9	16.5	37.3	49.5	17.5	48.6	69.3
	Contextual	18.9	41.0	53.8	17.5	38.7	48.6	17.0	47.6	68.4
capital	Direct	41.7	57.7	69.6	26.2	43.5	59.8	36.6	63.7	75.3
	Contextual	31.8	55.1	68.2	22.0	39.6	53.0	28.6	53.3	64.0
continent	Direct	59.4	72.6	84.4	60.4	74.5	78.8	54.2	77.8	87.7
	Contextual	53.8	71.2	81.6	61.8	75.5	84.0	52.8	75.5	86.8
headquarters_location	Direct	23.5	47.1	62.7	21.6	41.2	54.9	29.4	51.0	68.6
	Contextual	21.6	33.3	52.9	27.5	43.1	54.9	23.5	47.1	74.5
languages_spoken	Direct	39.4	64.4	76.9	50.0	68.3	75.0	46.2	73.1	79.8
	Contextual	42.3	58.7	71.2	40.4	63.5	68.3	43.3	68.3	73.1
country_of_citizenship	Direct	36.7	51.7	63.3	36.7	58.3	53.3	33.3	55.0	68.3
	Contextual	31.7	53.3	65.0	31.7	60.0	60.0	36.7	60.0	71.7

Table 8: Per-relation accuracy for direct and contextually mediated queries. Language: **Korean**.

Relation	Query	LLaMA-1B	LLaMA-3B	LLaMA-8B	Qwen-1.7B	Qwen-4B	Qwen-8B	Gemma-3-1B	Gemma-3-4B	Gemma-3-12B
official_language	Direct	35.4	53.2	61.5	50.7	59.5	63.0	39.5	54.5	60.5
	Contextual	27.2	45.8	58.1	50.2	59.6	65.0	36.2	53.5	61.5
native_language	Direct	64.6	78.5	88.5	83.8	92.3	90.8	52.3	80.0	89.2
	Contextual	47.7	70.0	80.0	80.0	90.0	88.5	48.5	65.4	86.9
manufacturer	Direct	45.7	68.6	68.6	54.3	62.9	68.6	60.0	62.9	65.7
	Contextual	62.9	71.4	74.3	62.9	65.7	77.1	57.1	60.0	74.3
capital_of	Direct	21.7	41.0	53.8	32.5	50.9	55.2	22.6	43.9	60.4
	Contextual	19.8	41.5	54.7	28.3	50.5	54.7	17.0	46.2	62.7
capital	Direct	29.8	51.2	67.6	34.5	51.8	65.5	34.5	55.4	64.9
	Contextual	23.2	48.8	63.4	31.5	47.6	58.6	29.2	47.0	62.8
continent	Direct	64.6	77.8	86.3	79.7	88.2	90.6	49.1	74.5	83.0
	Contextual	59.4	73.1	80.2	76.4	87.7	92.0	45.3	73.1	88.2
headquarters_location	Direct	23.5	41.2	54.9	45.1	49.0	58.8	29.4	52.9	62.7
	Contextual	21.6	39.2	51.0	37.3	47.1	51.0	25.5	47.1	60.8
languages_spoken	Direct	43.3	60.6	72.1	65.4	83.7	76.0	43.3	59.6	75.0
	Contextual	26.9	51.9	64.4	61.5	80.8	75.0	35.6	54.8	69.2
country_of_citizenship	Direct	35.0	53.3	61.7	48.3	63.3	70.0	31.7	50.0	66.7
	Contextual	35.0	46.7	56.7	48.3	66.7	66.7	25.0	43.3	56.7

Table 9: Per-relation accuracy for direct and contextually mediated queries. Language: **Chinese**.