Exploring Coherence of LLMs in Multilingual Question Answering

Stefano Campese
University of Trento
Italy
stefano.campese@unitn.it

Ivano Lauriola
Amazon AGI
USA
lauivano@amazon.com

Abstract

Recent studies have highlighted that Large Language Models (LLMs) often exhibit limited coherence, that is the ability to produce consistent responses to semantically equivalent questions. While most prior research has focused exclusively on English, limited investigation has been conducted on other languages. In this work, we study the coherence of LLMs on Question Answering tasks across six languages: English, Italian, German, Chinese, Japanese, and Vietnamese. We evaluate models of varying sizes, ranging from 3.8B to 235B parameters, to examine how coherence scales with model capacity and how it relates to languages. Our findings reveal that (i) coherence is not uniquely related to model size and accuracy and (ii) for some models, coherence varies significantly between languages.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of natural language processing (NLP) tasks, including Question Answering (Li et al., 2024), chatbot (Achiam et al., 2023), coding (Nam et al., 2024; Ugare et al., 2024), and summarization (Jin et al., 2024), to name a few. Despite their success, recent studies have highlighted that LLMs often struggle with coherence, that is the ability to produce consistent outputs for semantically equivalent inputs (Rabinovich et al., 2023; Lauriola et al., 2025). A coherent model should yield similar answers for lexical variations of the same question, reflecting stable knowledge access rather than surface-level pattern matching. Inconsistent responses, instead, suggest that the model's internal representation of knowledge may be brittle or context-dependent. While previous research has explored coherence primarily in English, the extent to which this phenomenon generalizes across languages remains partially underexplored. Multilingual evaluation is crucial, as LLMs are increasingly used in cross-lingual settings and exposed to morphologically diverse inputs.

In this work, we present a systematic multilingual study of LLM coherence. We analyze how coherence varies with model size, architecture, and language, and whether it correlates with traditional measures of accuracy. To this end, we evaluate a diverse set of recent LLMs, ranging from 3.8B to 235B parameters, on semantically equivalent sets of multilingual questions. We measure coherence through embedding-based semantic similarity and an accuracy-driven binary metric, providing complementary perspectives. Our results reveal three key findings. (i) Coherence scales with model size and accuracy, confirming previous findings on English questions; (ii) The degree of coherence differs substantially across model families, suggesting that the way models are trained heavily affects their coherence; (iii) Coherence heavily varies across languages. Moreover, we describe the opportunity of accuracy improvement with more coherent models, without the need of extra information or knowledge. Overall, our study provides new insights into the relationship between language, scale, and knowledge consistency in LLMs, emphasizing that coherence is an intrinsic and underexplored dimension of model reliability.

2 Related work

Recent Large Language Models (LLMs) span a wide range of sizes, from a few billion to hundreds of billions of parameters, and are pre-trained on web-scale corpora. Notable examples include the GPT family (Agarwal et al., 2025; Achiam et al., 2023; Brown et al., 2020), Llama (Touvron et al., 2023), Mixtral (Jiang et al., 2024), Qwen3 (Yang et al., 2025), Apertus (Hernández-Cano et al., 2025), or Phi4 (Abdin et al., 2024).

Despite their success across NLP tasks, LLMs

exhibit several well-documented limitations. They are highly sensitive to input phrasing, and prompt design can significantly impact performance (Voronov et al., 2024; Mizrahi et al., 2024; Arora et al., 2022; Chatterjee et al., 2024; Lu et al., 2024; Raj et al., 2022). Zheng et al. (2023) showed that the option order in multiple-choice QA tasks heavily influence results. Other authors suggested that the order of in-context examples also affects the judgment (Liu et al., 2022; Zhao et al., 2021). Similarly, Raina et al. (2024) pointed out LLM weaknesses with respect to adversarial attacks, e.g., when attempting to manipulate the output. Additionally, Chatterjee et al. (2024) introduced a measure to quantify the sensitivity of an answer for a given prompt, while Lu et al. (2024) highlighted how coherence can be used as an unsupervised proxy for model performance. In order to study the effect of paraphrases, Rabinovich et al. (2023) extended PopQA (Mallen et al., 2022) dataset through variations of input questions. The same authors proposed a metric, based on semantic similarity of the outputs, to quantify the consistency or coherence of models. As further step, Lauriola et al. (2025) showed that many recent LLMs exhibit poor coherence, suggesting a lack of robust understanding, and demonstrated that jointly evaluating multiple paraphrases can improve accuracy. Coherence analyses have been applied beyond modern LLMs, e.g.: Dense Passage Retrieval (Chen et al., 2024; Campese et al., 2025) and Generative Retrieval (Liu et al., 2023).

3 Question-answer coherence

The coherence of an LLM represents its ability to produce the same information for lexical variations of the same input question. From an intuitive point of view, if a model can correctly answer a question, for instance *How many calories in a cucumber?*, the same model is expected to answer other variations of the question with the same informational intent: *Can you tell me how many calories are in an average cucumber?*, or *calories cucumber*. Likewise, when a model fails to answer a question correctly, it is reasonable to expect similar failures across equivalent formulations. However, if the model succeeds on some variations but fails on others, this inconsistency suggests that the model possesses the necessary knowledge to answer correctly but exhibits limitations in its ability to interpret, comprehend, or generalize across lexical variations expressing the same informational intent.

Aligned to Rabinovich et al. (2023); Lauriola et al. (2025), we measure the model Semantic Coherence as the average embeddings similarity between the answers generated from a set of questions semantically equivalent. Formally, let $\mathcal{Q}$ be a set of open-domain well-formed questions and let $\mathcal{C} \subseteq \mathcal{Q}$ be a set of questions such that $\forall (q_i, q_j) \in \mathcal{C}^2 : q_i \equiv q_j$, where $\equiv$ indicates that two questions are semantically equivalent. We used the equivalence definition used by Lauriola et al. (2025). In short, two queries (q_i, q_j) are semantically equivalent iff they have the same information-seeking intent and their answers can be interchanged. In this work, we use the term *cluster* to refer a set of semantically equivalent queries $\mathcal{C}$. Given an embedding model $\mathbf{e} : \mathcal{Q} \rightarrow \mathbb{R}^d$ (query encoder), an LLM δ, and a set of m clusters $\{\mathcal{C}_r\}_{r=1}^{m}$, such that $\mathcal{C}_r = \{q_1, .., q_n\}$, $q_i \equiv q_j \forall i, j$, the Semantic Coherence is defined as follows: $\sum_{r=1}^{m} \left(\frac{2}{mn(n-1)} \sum_{(q_i,q_j) \in \mathcal{C}_r^2} \langle \mathbf{e}(\delta(q_i)), \mathbf{e}(\delta(q_j)) \rangle \right)$, The higher the value, the higher the probability that the model answers two semantically equivalent questions coherently. Note that this approach is tailored to single-answer factual questions. Questions that accept multiple answers or subjective may lead to very different embeddings. However, these questions are outside the scope of this analysis.

4 Experimental setting

Data - We started from queries sampled from two english datasets: SimpleQA and PopQA-TP. PopQA-TP (Rabinovich et al., 2023) is a large-scale open-domain resource consisting of 118k entity-centric question-answer pairs divided into 14k clusters of semantically equivalent variations. Questions expect short, almost entity-level, answers. SimpleQA (Wei et al.) is a benchmark for evaluating factual question answering and reasoning consistency across models. It contains short, single-hop, fact-based questions automatically generated. Each question has a unique verifiable answer, particularly suitable for coherence analysis.

We randomly sampled 500 questions from each dataset to form the seeds queries for constructing multilingual clusters. We generated equivalent variations of the seed questions in: English, Italian, German, Chinese, Japanese, and Vietnamese. For each language, we generated 3 reformulations with 3 different models each: Mistral-v3-large (675B) (Mistral AI, 2025), GLM-4.5 (355 B) (Zeng et al.,

Model	All		English		Italian		German		Chinese		Japanese		Vietnam.	
	Acc	SC	Acc	SC	Acc	SC	Acc	SC	Acc	SC	Acc	SC	Acc	SC
Phi4 3.8B	26.0	41.7	25.1	48.0	27.5	44.9	26.8	45.3	23.7	49.2	23.8	47.8	29.2	42.0
Phi4 14B	27.3	47.3	27.8	53.8	28.5	52.6	30.5	52.9	23.2	51.6	24.0	53.5	29.7	52.2
Qwen3 4B	31.6	43.1	28.4	49.1	27.6	46.7	32.2	47.1	36.4	54.0	34.9	53.2	29.9	56.5
Qwen3 14B	29.4	44.0	27.1	48.4	26.8	47.4	27.0	46.7	34.8	45.2	30.8	45.7	29.7	50.9
Qwen3 80B	35.1	49.1	39.3	57.2	36.6	54.5	38.6	54.9	29.9	59.1	28.8	56.7	37.4	59.9
Qwen3 235B	41.3	46.1	47.0	55.0	42.2	51.9	45.6	51.1	36.0	53.1	32.9	51.8	44.2	55.1
DeepS.R1 32B	28.9	38.6	24.0	44.0	26.4	41.3	26.5	42.1	26.0	39.6	27.3	39.9	42.9	38.6
Apertus 8B	21.5	45.4	20.2	52.9	21.4	51.1	22.1	50.8	19.6	53.5	22.2	51.7	23.8	49.9
Apertus 70B	31.5	51.6	30.7	57.3	32.6	57.7	31.4	57.1	28.4	57.4	32.8	58.6	32.9	59.6
GPT-OSS 20B	29.2	41.3	27.3	44.5	28.7	42.3	27.8	44.0	33.2	41.2	31.1	39.6	27.3	44.4
GPT-OSS 120B	29.9	49.9	31.2	55.0	30.6	51.8	30.4	53.8	26.7	51.5	29.3	49.3	31.2	52.5

Table 1: Accuracy (Acc) and Semantic Coherence (SC) per model across languages. Results based on Mistral-v3-large reformulations.

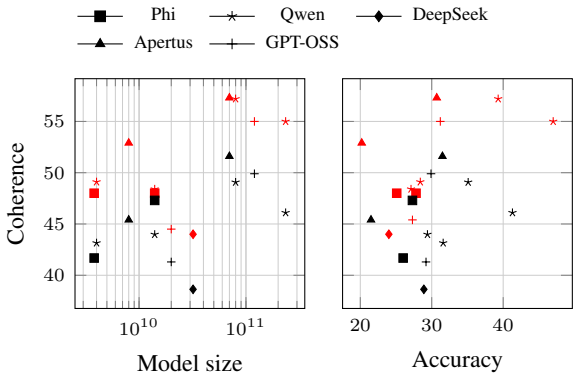

Figure 1: Coherence scaling with model size (left) and accuracy (right). Black: all queries; red: English only.

2025), and Qwen-3.5 (397B) (Qwen Team, 2026). In total, we collected 54 multilingual variations per seed question. To ensure quality, we manually verified a random sample of 150 English and 150 Italian variations, yielding a semantic equivalence rate of approximately 95%. Details on the generation and verification are provided in Appendix A.

Models - We evaluated a diverse set of LLMs, including: Phi4 (3.8B and 14B), Qwen3 (4B, 14B, 80B, 235B), DeepSeekR1 (32B), Apertus (8B, 70B), and GPT-OSS (20B, 120B).

Metrics - In our experiments, we used simple Accuracy, that is the proportion of correctly answered questions over all questions, to measure the correctness of answers. We employed an LLM-as-a-judge approach. In short, we used Mistral-Large-3[1] with the judgment prompt described in SimpleQA dataset. The model compares the generated answer with the reference one provided in each dataset and determines whether the output is correct. The answer generation prompt is detailed in Appendix B. To assess the coherence, we use the Semantic Coherence approach as described in Section 3, defined as the average embedding similarity of answers generated from semantically equivalent queries. To compute the similarity, we used the cosine between embeddings computed through LaBSE (Feng et al., 2022), a multilingual embedding model. Unless otherwise stated, results in the main body are based on questions generated by a single reformulator (Mistral-v3-large). Extended results using all three reformulators (9 variations per language) are reported for the Qwen3 family in Appendix C.

[1] https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512

4.1 Results

Table 1 reports Accuracy and Semantic Coherence for all evaluated models and languages using Mistral-v3-large as reformulator (numbers with all reformulators are available in Table 4 and Figure 2). The relation between Semantic Coherence and model size or accuracy is further summarized in Figure 1. What is striking from the plots is that, aligned to previous work (Lu et al., 2024; Lauriola et al., 2025), we observe a correlation between coherence and accuracy to be 0.393 on all languages and 0.644 for English, measured through Pearson correlation coefficient. The Pearson correlation between coherence and model size is 0.420 across all languages and 0.486 for English, indicating stronger cross-linguistic consistency but a weaker correlation within English. Note that, even within the same family of models, the size does not guarantee high coherence. Amongst other results, we highlight the coherence of Qwen3, which is increasing from 4B to 80B, to drop eventually with the 235B model, from 49.1 to 46.1.

From a multilingual perspective, small models display varying degrees of Semantic Coherence across languages. For instance, Qwen3 4B ranges

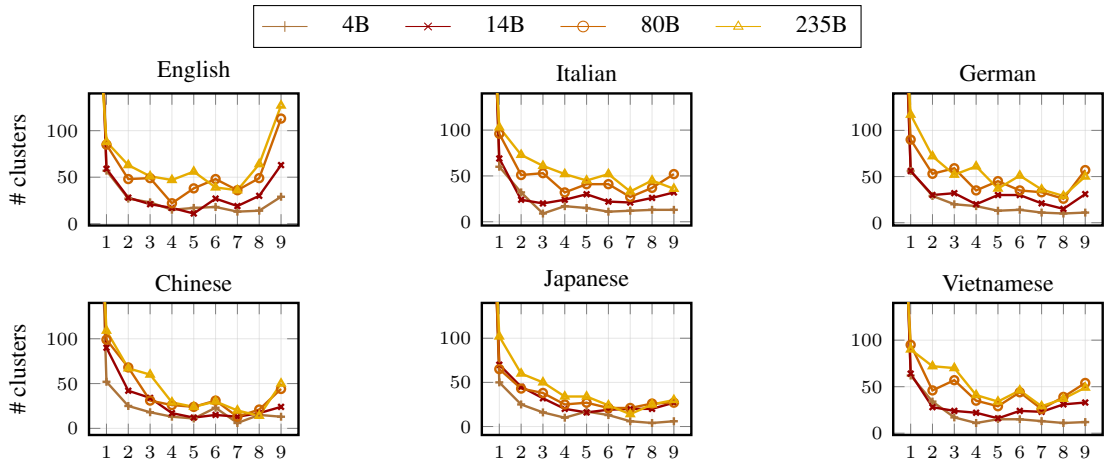

Figure 2: Number of clusters with $0 \ldots 9$ out of 9 correct answers per language. Qwen3 models were used.

from 46.7 (Italian) to 56.5 (Vietnamese), with almost 10% absolute difference. Phi4 3.8B showed similar cross-lingual variability, ranging from 42.0 to 49.2. Phi4 sharply reduces cross-lingual variability from 3.8B to 14B (std from 2.4 to 0.7), while Qwen retains larger discrepancies even at 235B (std 1.6), suggesting that Phi's training is inherently more robust with respect to coherence. Finally, we observed that Qwen3 shows very high coherence on Vietnamese. Even if the model is known to have strong performance on non-english languages, e.g., Chinese (Zhu et al., 2025), we hypothesize this result depends on intrinsic characteristics of the language. Vietnamese is indeed known to be morphologically simpler (Bentz et al., 2016; Coupé et al., 2019) than English and most of European languages (e.g., no tense conjugation, no number/gender inflection). We conjecture this simplifies the understanding of the question, and limited lexical variations improve coherence by definition. However, the same observation is not valid for other models. To further investigate coherence at the cluster level, Figure 2 shows, for each language, the number of clusters with 0 to 9 correct answers out of the 9 reformulations (3 reformulators $\times$ 3 variations). We report Qwen3 models only, as we have all size variants for this family. Full numerical results are in Table 4 (Appendix C). The plots emphasize a complementary, accuracy-driven coherence measure, indicating how often the models answer correctly lexical variations of the same query. Note that, larger versions of the model move the distribution away from the left, where 0/9 questions of the cluster are correctly answered. This aspect indicates that the model starts to answer more *new* questions, with a direct contribution to accuracy improvements. Most important, the large area in the middle of the distribution, where only a fraction of the 9 questions from the same cluster are correctly answered, indicates the *incoherence* (the lower the area, the more coherent the model), or the opportunity of improvement with more coherent models. If a model can answer at least one of the questions of a cluster, then it has the knowledge to answer all of the 9 lexical variations. Training more coherent models means moving these central clusters to the right of the plot. According to this measure on Qwen3 235B, English, Chinese, and Japanese show the highest coherence, while Italian, Vietnamese, and German are the least coherent.

5 Conclusions

We studied the coherence of various LLMs across multiple languages, showing that coherence varies substantially across model families and languages. Aligned with previous work, coherence correlates with accuracy, but model size appears more strongly predictive, suggesting coherence is an intrinsic property not uniquely captured by accuracy. Different families show distinct cross-lingual variability. For instance, Phi4 reduces coherence differences significantly from 4B to 14B, while Qwen requires its 235B version to achieve similar reduction. Our results suggest that coherence is influenced by multiple aspects, including model accuracy, language complexity (e.g., Vietnamese, which is considered morphologically simpler, shows higher coherence), and other unknown aspects of models training. These observations indicate that coherence should be treated as a core dimension of model evaluation, complementary to answer correctness.

Limitations

One remarkable obstacle in coherence analysis is the lack of multilingual resource with annotated lexical variations of queries. On the one hand, most of multilingual resources do not contain multiple lexical variations of the same queries. On the other hand, datasets for paraphrasing or question similarity, thus containing multiple query variations, focus on independent languages. As mitigation, we used an LLM to translated questions as described in Section 4. However, even if the quality of generated queries seems high, we could not control potential bias from the translator. Although results are aligned to the initial hypotheses and our analysis (Appendix A) didn't highlight significant risks, this aspect may have an effect on the study and findings that is nearly impossible to quantify.

Ethics Statement

This work uses publicly available datasets and evaluates open-weight LLMs. No human subjects were involved in the experiments. The multilingual reformulations were generated by LLMs and manually verified on a sample basis. We acknowledge that LLM-generated translations may carry biases from the underlying models. Our study does not involve sensitive or personal data, and all evaluated models were used in accordance with their respective licenses.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 142–153.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Stefano Campese, Alessandro Moschitti, and Ivano Lauriola. 2025. Improving document retrieval coherence for semantically equivalent queries. *arXiv preprint arXiv:2508.07975*.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Haitian Chen, Qingyao Ai, Xiao Wang, Yiqun Liu, Fen Lin, and Qin Liu. 2024. Unsupervised dense retrieval with conterfactual contrastive learning. *arXiv preprint arXiv:2412.20756*.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, 5(9):eaaw2594.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, and 1 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *CoRR*, abs/2403.02901.

Ivano Lauriola, Stefano Campese, and Alessandro Moschitti. 2025. Analyzing and improving coherence of large language models in question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11740–11755, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18608–18616.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. On the robustness of generative retrieval models: An out-of-distribution perspective. *arXiv preprint arXiv:2306.12756*.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are prompts different in terms of sensitivity? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856, Mexico City, Mexico. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint*.

Mistral AI. 2025. Mistral Large 3.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 138–154, Singapore. Association for Computational Linguistics.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517, Miami, Florida, USA. Association for Computational Linguistics.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. In *NeurIPS 2022 ML Safety Workshop*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Shubham Ugare, Tarun Suresh, Hangoo Kang, Sasa Misailovic, and Gagandeep Singh. 2024. Improving llm code generation with grammar augmentation. *arXiv preprint arXiv:2403.01632*.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, and 151 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Shiben Zhu, Wanqin Hu, Zhi Yang, Jiani Yan, and Fang Zhang. 2025. Qwen-2.5 outperforms other large language models in the chinese national nursing licensing examination: Retrospective cross-sectional comparative study. *JMIR Medical Informatics*, 13:e63731.

A Reformulation approach

We selected six languages spanning different typological families and varying levels of representation in LLM training data: English, Chinese, Italian, German, Japanese, and Vietnamese. For each seed question, we generated 10 lexical variations per language (60 total), following the prompt shown in Listing 1. From each model's output, we selected 3 variations per language, resulting in 9 reformulations per language (3 reformulators $\times$ 3 variations). The prompt groups languages into resource tiers to encourage the model to pay equal attention to all languages.

Listing 1: Reformulations generation prompt

```
You are a multilingual query variation generator. Your task
      is to create
lexical variations of an input question across different
      languages and styles.

## Languages (6 required)

Generate TEN variations for each of these languages:

**High-resource languages:**
1. English
2. Chinese - Simplified

**Medium-resource languages:**
3. Italian
4. German

**Low-resource languages:**
5. Vietnamese
6. Japanese

Total output: 60 variations (6 languages  10 variations each)

## Styles Distribution (FIXED per language)

For EACH language, generate EXACTLY:
- **3 Web query style**: Concise, keyword-focused search terms
- **3 Natural question style**: Complete, grammatically
      correct questions
- **3 Chat style**: Informal, friendly, conversational
- **1 Original style**: Direct, faithful translation of the
      input question
  preserving its original structure and formulation (NOT the
       original English
  text, but a translation that maintains the same style as
       the input)

## Rules

- Maintain semantic equivalence across all variations
- Generate exactly 10 variations per language (60 total)
- Follow the EXACT style distribution: 3 web query + 3
      natural question +
  3 chat style + 1 original
- Ensure lexical diversity within each style category
- "Original style" means a faithful translation that
      preserves the input
  question's structure and tone - it is NOT simply copying
       the English input
  text. Each language should have its own translation in
       original style.
- Ensure cultural and linguistic appropriateness
- Output MUST be valid, parseable JSON only
```

To assess whether the reformulated queries preserve the semantic meaning of the original questions, we manually inspected a random sample of 300 variations: 150 in English (95% equivalence) and 150 in Italian (95% equivalence). For all languages, we also employed an automatic verification using the prompt shown in Listing 2

Listing 2: Reformulations verification prompt

```
You are evaluating whether a reformulated question preserves
      the meaning of the original English question.

Original question (English): {original}
Reformulated question ({lang}): {reformulated}

Does the reformulated question ask for the same information
      as the original? Answer ONLY with:
YES - if meaning is preserved (even if phrasing or style
      differs)
NO - if meaning is changed, information is lost, or the
      question asks something different

Answer:
```

English reformulations were nearly always faithful ($\sim$98%). German and Vietnamese also scored highly ($\sim$96%), followed by Italian ($\sim$95%). Chinese and Japanese exhibited slightly lower preservation rates ($\sim$93%), primarily due to three recurring error types: (i) mistranslation of polysemous English words (e.g., river *course* rendered as *academic course* in Chinese), (ii) the reformulation inadvertently containing the answer rather than posing the question, and (iii) loss of specific details in complex multi-part questions. Overall, approximately 95% of the reformulations across all languages were judged to be semantically equivalent to the original, confirming that the reformulation pipeline introduces only a small amount of noise into the experimental setup.

Some examples of reformulations are available in Table 2.

B Answer Generation prompt

All answer-generating models receive the same system prompt instructing them to behave as factual question-answering assistants. The prompt constrains responses to short, unambiguous answers of

Q1:	By 1913, how many stars could astronomer Annie Jump Cannon classify per hour?
Eng	Annie Jump Cannon stars classified per hour 1913
Ger	Annie Jump Cannon Sterne Klassifikation pro Stunde 1913
Ita	stelle classificate all'ora Annie Jump Cannon 1913
Chi	安妮·坎农每小时分类恒星数量1913
Jap	アニー・ジャンプ・キャノン 毎時間分類星 1913
Vie	Annie Jump Cannon phân loại sao mˇi giˋ 1913
Q2:	Which female chemist was awarded the Garvan–Olin Medal in 1952?
Eng	female chemist Garvan-Olin Medal 1952
Ger	weibliche Chemikerin Garvan-Olin Medalle 1952
Ita	chimica donna medaglia Garvan-Olin 1952
Chi	女性化学家 加文-奥林奖章 1952
Jap	女性化学者 ガーヴァン=オリン メダル 1952
Vie	nˇ hóa học gia huy chng Garvan-Olin 1952
Q3:	Who was the producer of Setup
Eng	producer of Setup
Ger	Produzent von Setup
Ita	produttore di Setup
Chi	Setup制片人
Jap	Setup プロデューサー
Vie	nhà san xu´t cua Setup

Table 2: Examples of multilingual query reformulations from GLM-4.5.

Example 1 All models correct

Q: Who was the most prominent figure of the 1986 People Power Revolution?
R: 1986 People Power Revolution prominent figure
Gold: Corazon Aquino

4B : Ferdinand Marcos, Corazon Aquino
14B : Corazon Aquino
80B : Corazon Aquino
235B : Corazon Aquino

Example 2 Only 235B correct

Q: What is written on the right side of the cover art for the Dylan album Planet Waves?
R: Planet Waves album cover right side text
Gold: "Cast-iron songs & torch ballads"

4B : "Live at the Village Vanguard"
14B : "And the wolf shall dwell with the lamb" from Isaiah 11:6.
80B : Bob Dylan and The Band
235B : "Cast-iron songs & torch ballads"

Example 3 All models wrong

Q: What month and year was "Welcome to Paradise" first released?
R: "Welcome to Paradise" release date
Gold: December, 1991

4B : February 13, 2009
14B : June 13, 1988
80B : September 29, 1994
235B : October 4, 1994

Table 3: Examples of generated answers across Qwen3 model sizes. **Q**: original question; **R**: reformulated query (GLM-4.5, English). = correct, = incorrect.

a few words, prohibits hedging language and uncertainty markers, and explicitly forbids replying with Unknown. The full system prompt reads:

Listing 3: Answer generation prompt

```
You are an AI assistant specialized in answering open-domain
     factual questions.
Your goal is to maximize accuracy by giving only short,
     correct answers based on your internal knowledge.

Guidelines:
- Read the question carefully
- Provide a single, factual answer in 1-10 words
- Do not explain, justify, or add extra details
- Never include uncertainty markers (e.g., "maybe", "I think")
- Always provide your best answer - never reply "Unknown"
- Keep the answer concise, unambiguous, and time-stable
```

C Qwen full results

Table 4 reports accuracy and coherence for the Qwen3 model family when answers are generated from nine reformulations per language, obtained by combining three reformulator models (Qwen3.5 397B, GLM-4.5-355B, and Mistral-Large-3 675B) with three paraphrase variations each. Accuracy increases monotonically with model size across all languages, with English consistently achieving the highest scores and Japanese the lowest. Coherence, measured as the average pairwise cosine similarity among the nine answers to each original question, peaks at 80B rather than at 235B, suggesting that the largest model produces more diverse surface forms despite being more accurate. Aligned to previous results, Vietnamese exhibits the highest coherence overall, likely due to its simpler morphology, reducing lexical variation in the generated answers. These trends are consistent with the results reported in the main text using a single reformulator, confirming that the findings generalize across reformulation sources.

Model	All		English		Italian		German		Chinese		Japanese		Vietnam.	
	Acc	**SC**	**Acc**	**SC**	**Acc**	**SC**	**Acc**	**SC**	**Acc**	**SC**	**Acc**	**SC**	**Acc**	**SC**
Qwen3 4B	10.32	50.2	7.53	47.2	7.69	45.5	7.83	48.8	5.54	48.0	7.83	56.6	7.79	41.9
Qwen3 14B	16.15	50.9	13.50	46.7	13.85	45.3	10.79	42.8	12.41	42.5	13.99	48.9	13.45	41.9
Qwen3 80B	29.41	58.1	21.97	52.3	21.84	50.7	16.04	53.0	14.15	51.2	21.82	56.4	20.87	47.0
Qwen3 235B	34.25	56.7	24.12	50.4	24.19	49.8	17.74	49.5	16.07	47.4	23.05	54.4	23.24	44.5

Table 4: Accuracy (Acc) and Semantic Coherence (SC) of Qwen models across languages. Results are averaged over 9 query reformulations per language.