# Rephrasing to Remember: Test-Time Scaling for Knowledge Recall in LLMs

Guo Yibo　Xin Zhao
The University of Tokyo
{y-guo,xzhao}@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

## Abstract

Large language models (LLMs) can recall factual knowledge, but such recall is often unreliable and sensitive to query phrasing. Recent test-time scaling methods improve reliability without additional training, yet typically operate on a single query formulation. We propose a test-time approach that ensembles self-paraphrased questions via logit averaging to improve factual recall, using multiple semantically equivalent rewrites for each query to aggregate the model's predictions. Experimental results on MyriadLAMA and HotpotQA benchmarks show consistent gains across multiple LLMs, and analyses indicate that paraphrase quality, quantity, and diversity all affect ensemble performance, with quality being the most important.

## 1 Introduction

Large language models (LLMs) encode substantial factual knowledge acquired during pretraining and can often be queried as implicit knowledge bases [1]. However, the reliability of factual recall can be limited, particularly for facts that are sparsely represented in the training data [2, 3, 4, 5]. As a result, LLMs may produce inconsistent or incorrect answers to semantically equivalent questions that differ only in phrasing.

Test-time scaling has emerged as an effective paradigm for maximizing LLM performance in such settings by allocating additional computation in inference [6, 7]. Existing approaches include Chain-of-Thought [8], iterative self-correction [9], and self-consistency [10]. These methods can be viewed as instances of **multi-path inference**, which aggregates predictions from diverse reasoning paths to improve reliability [10, 11]. While effective on reasoning- and knowledge-intensive tasks [12, 13], these approaches retain the original query, and thus do not directly address the model's sensitivity to variations in query phrasing during factual recall.

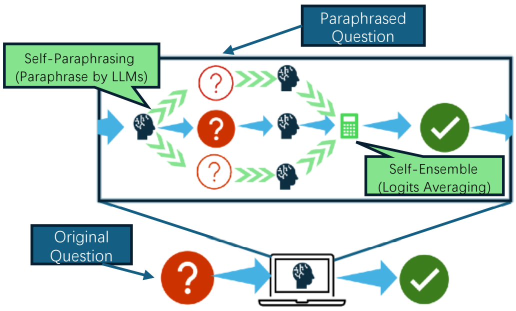
Paraphrased Question
Self-Paraphrasing (Paraphrase by LLMs)
Self-Ensemble (Logits Averaging)
Original Question

**Figure 1** The proposal of SELF-COLLAB, a self-collaborative inference paradign to improve knowledge recall through a two-step procedure: **Self-Paraphrase** and **Self-Ensemble**.

In this study, we propose a self-collaborative inference paradigm, **SELF-COLLAB**, to maximize the performance of LLMs as implicit knowledge sources; it asks the model to rephrase queries before recalling facts. For each question, the model generates a small set of semantically equivalent paraphrases. We then aggregate model predictions across these paraphrases by averaging their output logits to produce a single, more reliable final answer.

We evaluate our approach on the MyriadLAMA [14] and HotpotQA [15] benchmarks. Across five Qwen3 models of different sizes, **SELF-COLLAB** consistently improves accuracy compared with answering the original query alone. We further analyze factors affecting ensemble effectiveness, including the number, quality, and diversity of paraphrases. Our results show that quality is the main driver of gains, while increasing the number provides additional improvements if quality is sufficient. Diversity analysis reveals that certain combinations amplify ensemble benefits, whereas others can reduce performance, clarifying when paraphrase ensembling helps and when it can hurt.

## 2 Self-Collaborative Inference

In this section, we propose **SELF-COLLAB**, a self-collaborative inference designed to enhance LLM's knowledge recall. This is motivated by the observation that successful knowledge recall in LLMs relies on the robustness to query phrasing [14]; we aim to let LLMs read a query from multiple perspectives for better knowledge recall.

Specifically, we formalize a knowledge query as a latent intent $\alpha$ (*e.g.*, $\langle$Japan, capital, ?$\rangle$) with a desired answer $\beta$ (*e.g.*, Tokyo). The observed input $x_i$ is a linguistic realization of this intent, i.e., $x_i = \mathrm{Lang}_i(\alpha)$ (*e.g.*, Where is the capital of Japan?), and the LLM predicts according to $\mathrm{LLM}_\theta(\beta \mid x)$. Crucially, a single linguistic realization often provides an incomplete and biased projection of $\alpha$, as different surface forms may emphasize different semantic aspects and activate different internal retrieval pathways.

### 2.1 Self-Collaborative Inference

SELF-COLLAB approximates conditioning on the latent intent $\alpha$ by marginalizing over multiple semantically equivalent but linguistically diverse realizations of the same query at inference time.

$$\mathrm{LLM}_\theta(\beta \mid \alpha) \approx \sum_{i=1}^{K} \mathrm{LLM}_\theta(x_i \mid \alpha) \cdot \mathrm{LLM}_\theta(\beta \mid x_i) \quad (1)$$

Formally, SELF-COLLAB enables a target model $\mathrm{LLM}_\theta$ to improve its own inference through a two-stage procedure: **Self-Paraphrasing** followed by **Self-Ensembling**.

**Step 1: Self-Paraphrasing** Given an original query $x_0(\alpha)$ that expresses latent intent $\alpha$, we first prompt $\mathrm{LLM}_\theta$ to generate multiple paraphrases of $x_0$. Crucially, SELF-COLLAB relies exclusively on the target model itself for paraphrase generation, without introducing external paraphrasing models or resources. Concretely, we prompt $\mathrm{LLM}_\theta$ with a paraphrasing instruction and sample candidate paraphrases $x_1, \ldots, x_N$. In practice, raw generations may include failures such as duplications, ill-formed outputs. To ensure paraphrase quality, we apply lightweight filtering rules, including duplication checks and structured-output parsing, and retain only valid paraphrases within a bounded number of attempts. As a result, we obtain a set of high-quality paraphrases $x_i{}_{i=1}^{K}$, where each represents a distinct linguistic realization of the same intent.

**Step 2: Self-Ensembling** After generating paraphrases, for each $x_i$, we inference them by $\mathrm{LLM}_\theta$ and obtain the next-token probability vector $\mathbf{z}_t^i \in \mathbb{R}^{|\mathcal{V}|}$ at decoding step $t$. We then aggregate the probability by uniform averaging to obtain the fused probability $\bar{\mathbf{z}}_t$: $\bar{\mathbf{z}}_t = \sum_{k=1}^{K} w_k \, \mathbf{z}_t^k$. We decode greedily from this distribution, selecting the token the highest probability and concatenate it to all the paraphrases. At each decoding step $t+1$, we repeat the whole process until the end of generation.

### 2.2 Experimental Settings

**Datasets.** We conduct experiments on two knowledge-centric benchmarks: **MyriadLAMA** [14] [1], which evaluates factual knowledge recall, and **HotpotQA** [15], which targets multi-hop factual recall and reasoning. For each dataset, we randomly sample 200 questions for test.

**Models.** We evaluate 5 LLMs from the same family, Qwen3, covering diverse model scales from 0.6 to 14B.

**Generation Settings.** During the self-paraphrasing stage, we generate up to two paraphrases per query, with a maximum of 16 generation attempts. We set the temperature to 1.5 and top-$p$ to 0.95 to encourage linguistic diversity. If fewer than two valid paraphrases are produced after filtering, we retain all available paraphrases. For the self-ensemble stage, we apply greedy decoding across all methods and fix all random seeds to ensure reproducibility.

**Evaluation Protocol.** We evaluate all methods under both zero- and five-shot settings. Performance is measured using normalized exact match (EM), where both predictions and references are normalized prior to comparison.

**Baselines.** Given the same paraphrase set, we compare SELF-COLLAB against the following baselines: (i) ParaAvg, which performs independent decoding for each paraphrase and reports the mean EM across paraphrases; and (ii) Origin, which evaluates performance using only the original query provided in the dataset.

### 2.3 Evaluation Result

Table 1 reports the result. Across most settings, SELF-COLLAB outperforms both Origin and ParaAvg, demonstrating the effectiveness of self-collaboration inference.

**Parapharse performs worse but integrating them helps** On most evalution scenorio, ParaAvg fails to surpass the original prompt and in some cases slightly degrades performance. This shows that the auto-generated paraphrases

[1] MyriadLAMA is a multi-prompt factual recall dataset, but we only adopt one prompt per query as seed prompt.

| **Model** | **HotpotQA (O / P / S)** | | **MyriadLAMA (O / P / S)** | |
|---|---|---|---|---|
| | **0-shot** | **5-shot** | **0-shot** | **5-shot** |
| Qwen3-0.6B | 0.010 / 0.017 / **0.075** | 0.105 / 0.092 / **0.130** | **0.010** / 0.003 / 0.000 | **0.270** / 0.153 / 0.125 |
| Qwen3-1.7B | 0.050 / 0.047 / **0.085** | 0.110 / 0.093 / **0.220** | 0.005 / 0.005 / **0.055** | **0.365** / 0.359 / 0.310 |
| Qwen3-4B | 0.075 / 0.072 / **0.165** | 0.160 / 0.152 / **0.280** | 0.140 / 0.152 / **0.245** | 0.360 / 0.349 / **0.375** |
| Qwen3-8B | 0.055 / 0.054 / **0.185** | 0.230 / 0.237 / **0.345** | 0.255 / 0.253 / **0.305** | **0.390** / 0.355 / 0.385 |
| Qwen3-14B | 0.105 / 0.097 / **0.245** | 0.335 / 0.300 / **0.395** | 0.235 / 0.213 / **0.280** | 0.365 / 0.374 / **0.460** |

**Table 1** Normalized EM with datasets and shot settings shown in parallel. (O) Original, (P) ParaAvg, (S) SELF-COLLAB.

have lower quality than the original prompts in the dataset. However, despite of the issue, ensembling the original one with paraphrases by SELF-COLLAB could largely boost the knowledge recall, indicating that the fusion of multiple inference paths enables effective information integration.

**Impact of Model Scale.** The effectiveness of SELF-COLLAB increases with model scale. Across both datasets, larger models (8B and 14B) exhibit more consistent performance gains, suggesting that sufficient representational capacity is essential for effective self-collaborative inference. In contrast, smaller models display higher variance and even suffer performance degradation in certain settings, such as the 5-shot MyriadLAMA evaluation with the 0.6B and 1.7B models. An inspection of intermediate outputs reveals that these smaller models often generate low-quality paraphrases, including semantically meaningless tokens or repetitions of few-shot demonstrations. Such noisy paraphrases propagate into the self-ensemble stage, ultimately undermining the benefits of self-collaboration.

# 3 Ablation Analysis on Paraphrases

This section investigates how SELF-COLLAB improves knowledge recall, and under which conditions it fails to provide benefits. We systematically analyze the impact of paraphrase **quantity**, **quality**, and **linguistic diversity** on the performance gains through prompt ensemble.

## 3.1 Evaluation Settings

We consider four model scales, Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, and focus our analysis on the 5-shot prompting setting with greedy decoding. To enable a systematic investigation of how paraphrases influence self-ensemble performance, we leverage the MyriadLAMA, which provides 100 paraphrases per query spanning diverse lexical choices, syntactic structures, and semantic variations. This corpus includes both manually crafted high-quality paraphrases and automatically generated paraphrases produced by GPT-based models.

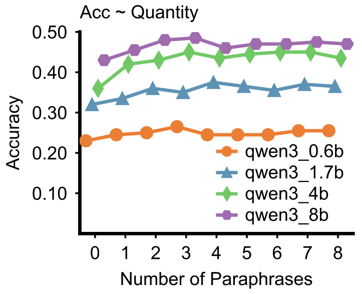

**Figure 2** Quantity analysis to ensemble performance.

For each MyriadLAMA query, we uniformly sample $K$ paraphrases without duplication to form a paraphrase set. Unless otherwise specified, we set $K = 5$ and sample five paraphrase sets per query, treating each set as an individual evaluation instance. We randomly select 200 queries, resulting in 1,000 paraphrase-set instances in total. Results are reported separately for each model scale, using mean accuracy aggregated across all sampled paraphrase sets.

## 3.2 Quantity

We vary the number of paraphrases $K$ in a set and examine how accuracy changes when the number of paraphrases becomes larger. Concretely, for each question we sample $K \in \{1, ..., 8\}$ paraphrases from datasets. Form 1 to 8, we add more and more paraphrases to the original question and evaluate them under the same 5-shot greedy-decoding setup. The results are shown in Figure 2. When we use only a few paraphrases, adding more clearly improves accuracy. But once we already have many paraphrases, adding extra ones brings little benefit.

## 3.3 Quality

To measure paraphrase set quality, we measure the correctness of each paraphrase by evaluating when the paraphrase is evaluated along and whether it can predict the answer. We denote the set of such paraphrases as $S_{\text{correct}} \subseteq S$.

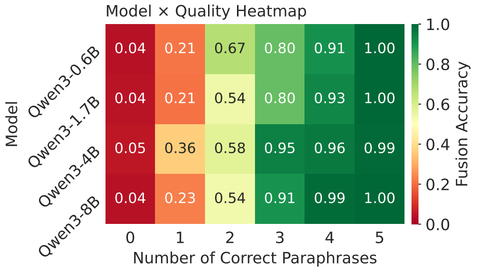

**Figure 3** Paraphrase quality analysis to ensemble performance.

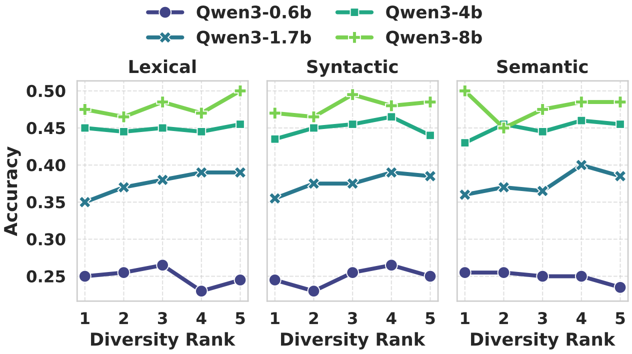

**Figure 4** Diversity comparison from 3 linguistic dimensions.

Then, we measure the single-prompt correctness ratio within a paraphrase set as:

$$r_{\text{correct}}(S) = \frac{|S_{\text{correct}}|}{|S|}. \quad (2)$$

To compute the single-paraphrase correctness ratio, we evaluate each paraphrase on its own to obtain $S_{\text{correct}}$, and then compute $|S_{\text{correct}}|$ for each sampled set. In this experiment, we fix the set size to $|S| = 5$, so we use the number of correct paraphrases instead of $r_{\text{correct}}(S)$.

The results are shown in Figure 3; it shows that Self-Ensembling does not behave like a simple linear mixture. When most paraphrases are good, the ensemble is quite robust: adding a few bad paraphrases only slightly reduces accuracy, much less than what a proportional dilution intuition would predict. For example, replacing 1 out of 5 good paraphrases with a bad one lowers accuracy by less than 10%, instead of the roughly 20% drop expected from a linear view. In contrast, when most paraphrases are bad, adding a few good ones can greatly improve accuracy. This suggests that good paraphrases can have a stronger effect than their fraction in the set.

## 3.4 Linguistic Diversity

We examine how linguistic diversity of paraphrase sets influence the ensemble result, from three dimensions.

**Lexical Diversity**: we normalize each sentence into a set of unique tokens by tokenizing, stemming and lemmatization, compute all pairwise Jaccard similarities, and define lexical diversity as one minus the mean similarity.
**Syntactic Diversity**: we obtain a dependency parse for each sentence using spaCy, and compute the mean pairwise distance among the corresponding dependency trees.
**Semantic Diversity**: we encode each sentence with ModernBERT, apply attention-mask-aware mean pooling followed by L2 normalization, compute all pairwise Euclidean distances between normalized embeddings, and take the mean distance as the semantic diversity score.

For each paraphrase set, we computed three distinct diversity metrics. We ranked all sampled sets along each dimension and partitioned them into five quintiles. This ensures that sets within the same group represent a comparable level of diversity (*e.g.*, the *i*-th lowest diversity quintile) for that specific dimension.

As shown in Figure 4, we report the accuracy across these different diversity dimensions and levels for four Qwen3 models. The result demonstrates that increasing diversity from three dimensions can all improve the recall accuracy after ensemble. However, we notice such gains exhibit a trend of diminishing returns. The performance improvement tends to plateau or even fluctuate slightly once the diversity reaches the highest quintiles (Group 4 and 5). This suggests that while introducing diversity effectively breaks dominance loops or coverage gaps, excessive divergence may introduce noise or irrelevant candidates that the ensemble method cannot fully leverage.

# 4 Conclusion

We propose **SELF-COLLAB**, a self-collaborative inference paradigm that improves the reliability of factual knowledge recall in large language models. SELF-COLLAB expands linguistic diversity via self-paraphrasing and integrates evidence from multiple prompts through logits-level ensembling. Experiments across multiple model scales and two knowledge recall benchmarks demonstrate consistent improvements in factual recall. Further ablation studies systematically analyze the effects of paraphrase quality, quantity, and diversity, providing new insights into the mechanisms underlying robust knowledge recall under linguistic variation. In future work, we will extend the work to reasoning and logic tasks.

# References

[1] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 5506–5521, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, **Advances in Neural Information Processing Systems**, Vol. 37, pp. 60626–60668. Curran Associates, Inc., 2024.

[5] Xin Zhao, Naoki Yoshinaga, Yuma Tsuta, and Akiko Aizawa. Tracing multilingual knowledge acquisition dynamics in domain adaptation: A case study of english-japanese biomedical adaptation, 2025.

[6] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. What, how, where, and how well? a survey on test-time scaling in large language models, 2025.

[7] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837. Curran Associates, Inc., 2022.

[9] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 1417–1440, 11 2024.

[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[11] Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Hengchao Shang, Zongyao Li, Jinlong Yang, Zhanglin Wu, Zhiqiang Rao, Shimin Tao, and Hao Yang. M-ped: Multi-prompt ensemble decoding for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 16693–16711, Suzhou, China, November 2025. Association for Computational Linguistics.

[12] Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 6904–6917, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[13] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-math: Small LLMs can master math reasoning with self-evolved deep thinking. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, **Proceedings of the 42nd International Conference on Machine Learning**, Vol. 267 of **Proceedings of Machine Learning Research**, pp. 20640–20661. PMLR, 13–19 Jul 2025.

[14] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 13186–13214, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[15] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.