ReTRE: Benchmarking LLM Transfer Robustness with Structure-Preserving Variants

ZhongDong Li[1,2*] **Weijie Shi**[3] **Yue Cui**[4] **Haolun Ma**[3] **Yuanjun Liu**[2]
Jiawei Li[3] **An Liu**[2] **Jia Zhu**[5] **Jiajie Xu**[1,2†]

[1]Laboratory for Big Data Research and Intelligent Decision Making on Graduate Employment, Soochow University
[2]School of Computer Science and Technology, Soochow University
[3]Hong Kong University of Science and Technology [4]Alibaba Group
[5]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University

Abstract

Large language models (LLMs) have achieved strong performance on standard benchmarks, yet their performance is not robust across different task manifestations. It remains unclear how performance changes under controlled task rewrites that preserve the original solution structure, while varying the rewrite type and level. To address this question, we introduce ReTRE (Rewrite-based Transfer Robustness Evaluation), an evaluation benchmark inspired by learning transfer theory that probes transfer robustness along two rewrite levels: Near Transfer and Far Transfer. ReTRE employs a multi-agent system to construct textual and visual variants while preserving the structure of the original solution. Evaluations on mathematical and science tasks across state-of-the-art multimodal LLMs reveal a consistent transfer gap: performance exhibits a general declining trend as transfer similarity drops and strong text performance can face performance decline under cross-modal transfer. Crucially, we identify a divergence between post-training paradigms: reinforcement learning preserves transfer robustness, whereas supervised fine-tuning tends to overfit the training distribution, leading to severe degradation in far-transfer performance despite strong in-distribution accuracy.

1 Introduction

Large language models (LLMs) have achieved strong performance across a wide range of benchmarks (Cao et al., 2025), and such performance is often taken as evidence that a model has acquired the targeted knowledge or skills. However, learning transfer theory emphasizes that applying acquired knowledge to novel contexts is a key signal of deep understanding[1]. This motivates a practical evaluation question: Do LLMs remain robust when facing novel manifestations of the same problem?

One common approach is to rewrite existing evaluation datasets (Wu et al., 2024; Wang et al., 2024b; Huang et al., 2025a,b; Kirtane et al., 2025). These studies construct new test instances through operations such as data perturbation, character substitution, and semantic rewriting, revealing that high benchmark scores can stem from data contamination or sensitivity to surface expressions. However, such approaches seldom control for the degree of divergence between original and rewritten tasks, nor do they systematically vary the transformation type. As LLMs are increasingly expected to generalize across diverse and unfamiliar contexts (Mumuni and Mumuni, 2025), a more principled question arises: How does model performance change under controlled, structure-preserving rewrites that vary in both type and degree?

Inspired by learning transfer theory (Perkins et al., 1992; Barnett and Ceci, 2002; Hilton and Pellegrino, 2012), we propose ReTRE (Rewrite-based Transfer Robustness Evaluation), a benchmark that evaluates transfer robustness using structure-preserving rewrites, as shown in Figure 1. ReTRE generates variants along two complementary dimensions: Knowledge Domain (KD) and Modality Context (MC). Each instantiated at two discrete transfer levels (near and far). In the Knowledge Domain dimension, variants are designed to change the semantic background of a problem while keeping the intended reasoning procedure aligned with the original. Near-transfer variants remain within related STEM domains (e.g., reframing a physics story in a chemistry context), whereas far-transfer variants move the same problem structure into non-STEM domains such as economics, law, or social sciences. In the Modality Context dimension, variants are designed to change how task information is presented while keeping the information needed to solve the problem aligned with the original. Near-

*Email: zdli123@stu.suda.edu.cn
†Corresponding author.
[1]https://poorvucenter.yale.edu/transfer-of-knowledge-to-new-contexts

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44257–44268
July 2-7, 2026 ©2026 Association for Computational Linguistics

transfer variants apply within-text reformats, such as converting a prose description into a structured Markdown table. Far-transfer variants shift from text to the visual channel by representing task-relevant structure in diagrams or plots.

In practice, both types of rewrites preserve the core solution structure. Specifically, KD rewrites keep the abstract variables and relations at the level of the solution template while modifying domain-specific entities and contextual framing. In contrast, MC rewrites maintain the problem's information content while altering the input modality and its organization. To construct these variants at scale, we design a three-layer agentic pipeline that pairs a generator with a corresponding verifier at each stage. The pipeline extracts a transferable solution structure from the original task, formulates context-appropriate transfer strategies, and synthesizes the final variant. Iterative validation across stages ensures structure preservation and correctness.

We evaluate a suite of mainstream multimodal LLMs on mathematical reasoning tasks using ReTRE. Results reveal a clear transfer gap: near-transfer variants yield performance close to the original, whereas far-transfer variants incur substantial degradation. Notably, Modality Context (MC) variants cluster tightly with original problems in embedding space, and MC-Near occasionally yields slight gains from reduced parsing ambiguity due to structured formatting. In contrast, Knowledge Domain (KD) variants form distinct clusters far removed from the original distribution, with KD-Far posing a consistent challenge across all models. We also find that thinking-mode models demonstrate enhanced transfer robustness. For instance, claude-haiku-4-5 (thinking) improves over its standard counterpart by 3.6 percentage points in average accuracy.

Motivated by these findings, we conduct controlled experiments to isolate the effects of post-training paradigms on transfer robustness. We observe a striking divergence: full Supervised Fine-Tuning achieves perfect in-distribution accuracy (100%) but collapses catastrophically on transfer variants, whereas Reinforcement Learning, like Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025), improves both in-distribution and transfer performance simultaneously. We further investigate the impact of model scale and generational evolution on robustness. Scaling from 2B to 235B generally enhances robustness, but the improvement is not strictly monotonic; mid-scale gains show diminishing returns. Meanwhile, knowledge domain transfer remains a persistent challenge in generational evolution. Finally, we extend ReTRE to natural science reasoning on GPQA, where MC-Far emerges as the most challenging setting, causing performance drops exceeding 20 percentage points for the top model. Our contributions are as follows:

- We propose ReTRE, a learning-transfer-inspired diagnostic benchmark that profiles transfer robustness via controlled task transformations along two axes—Knowledge Domain and Modality Context—each instantiated at near and far transfer distances.
- On MATH500, we observe a consistent transfer gap across mainstream multimodal LLMs: performance is relatively stable under near-transfer settings but degrades substantially under far-transfer settings. Moreover, strong performance on text-based settings does not necessarily translate to cross-modal variants, where models can still exhibit notable failures. Thinking-mode models consistently outperform their non-thinking counterparts in transfer robustness.
- In a controlled study on Qwen3-VL-8B-Instruct, we find that post-training paradigms strongly affect transfer robustness: full SFT achieves perfect in-distribution accuracy but collapses on transfer variants, whereas GSPO improves both in-distribution and transfer performance. Model scaling and generational evolution generally enhance robustness, though gains are not monotonic, and distant-domain transfer remains a bottleneck.

2 Related Work

2.1 LLM Benchmark

In recent years, a large number of benchmarks (Ni et al., 2025) have been proposed to evaluate LLMs, covering diverse tasks such as natural language understanding, machine translation, and text generation. Advanced LLMs have achieved remarkable performance across these benchmarks. Recently, the evaluation paradigm has shifted from assessing task-specific performance to measuring more general capabilities (Cao et al., 2025), including reasoning ability, instruction following, and knowledge understanding. Nevertheless, several studies

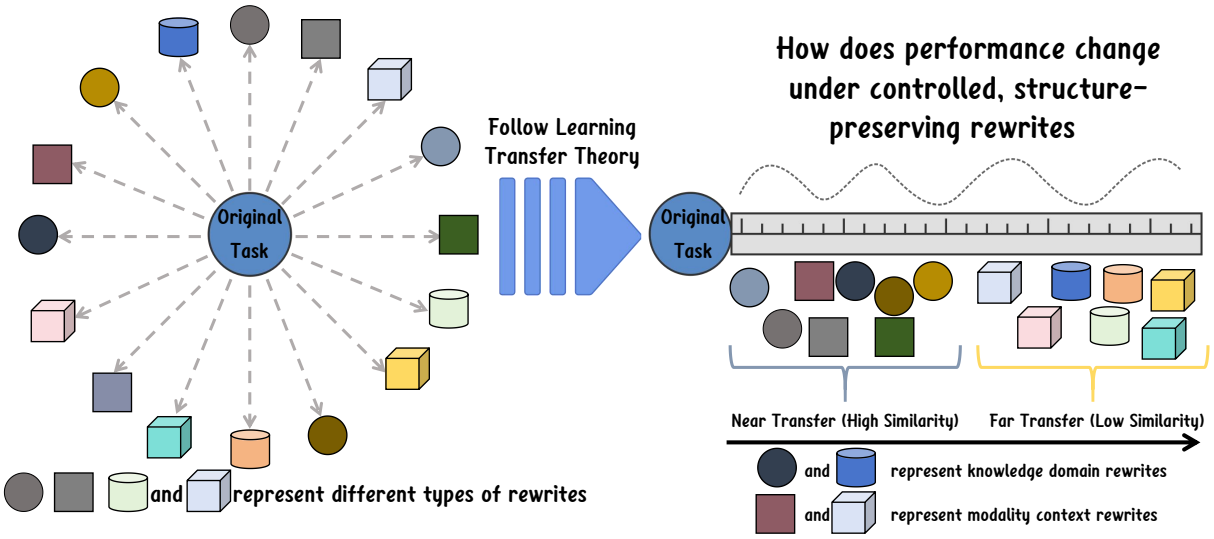

Figure 1: Motivation for ReTRE. (Left) Existing rewrite-based evaluations often overlook the similarity between rewritten variants and the original task, and do not consider how performance changes under controlled, structure-preserving rewrites. (Right) ReTRE categorizes rewrites into distinct transfer level (near and far) within knowledge domain and modality context types. This discrete setting provides a solution to evaluate transfer robustness.

have revealed that the high scores achieved by models often result from surface-level pattern matching rather than genuine comprehension. This issue is largely attributed to data leakage between training and evaluation datasets (Wu et al., 2024). To mitigate this problem, one common approach is data perturbation (Lunardi et al., 2025; Kirtane et al., 2025; Huang et al., 2025b), which introduces variations to test the robustness of model reasoning. However, in practical scenarios, the same problem can manifest in diverse forms, and existing perturbation methods do not systematically consider how these forms differ in type or degree. This gap motivates a more controlled evaluation framework.

2.2 Learning Transfer Theory

Learning transfer is a central concept in educational psychology, commonly defined as the ability to apply acquired knowledge or skills beyond the original learning context (Perkins et al., 1992). A widely adopted distinction in this literature is between near transfer and far transfer. Near transfer typically refers to situations where the transfer context remains highly similar to the original learning context, whereas far transfer involves applying the same underlying knowledge to contexts that differ substantially in surface characteristics or situational framing. Barnett and Ceci (Barnett and Ceci, 2002) further formalize transfer by proposing a six dimensional taxonomy. While this taxonomy provides a rich conceptual framework for analyzing human learning, several dimensions (e.g., social context and physical context) are not directly applicable to model evaluation.

3 Method

3.1 Problem Formulation

We define transfer robustness as the performance change under controlled shifts in task manifestation, while the underlying reasoning procedure is preserved. Given a source dataset $\mathcal{D}_{\text{src}} = \{(x, y)\}$, we construct a target dataset $\mathcal{D}_{\text{tar}} = \{(x', y')\}$ using a multi-agent system $\mathcal{M}$ that applies a transformation $\mathcal{T}$ under a transfer setting s:

$$(x', y') = \mathcal{T}(x, y; s) \text{ via } \mathcal{M}, \text{ s.t. } \mathcal{L}_{\text{core}}(x') \approx \mathcal{L}(x). \quad (1)$$

Here $\mathcal{L}(\cdot)$ extracts the reasoning procedure required to solve the original task, and $\mathcal{L}_{\text{core}}(\cdot)$ denotes the core reasoning component of the transferred task. The constraint enforces that solving x' relies on the same solution procedure as x, while permitting limited auxiliary scaffolding to keep the transferred instance self-contained and to prevent answer leakage (e.g., the solution appearing verbatim in the problem text).

For mathematical reasoning tasks, we typically preserve the original answer since domain transfer maintains numerical relationships directly. For natural science question answering, however, directly re-contextualizing domain-specific concepts (e.g., chemical reactions, physical laws) into unrelated

domains would require external prerequisite knowledge. To address this, we wrap the transferred content within a fictional scenario and provide explicit rule descriptions with minimal demonstrations, enabling the model to solve the problem using the same reasoning procedure without relying on real-world domain knowledge. In such cases, the answer may differ in surface form, but the underlying solution procedure remains equivalent.

We apply this formulation to two task families: (1) mathematical reasoning, where the output is a short-form answer derived from quantitative relations, and (2) natural science question answering, where the model selects the correct option from multiple choices. The same transfer framework applies to both families, though the concrete rewrite operations differ across tasks and transfer axes.

3.2 Data Transfer Pipeline

Figure 2 illustrates our three-layer Generator-Validator pipeline. Each layer pairs a generator agent with a validator agent. The pipeline employs layer-wise retry: if a candidate fails validation, we regenerate at that layer without discarding validated outputs from earlier layers. Instances that exceed the maximum retry limit are discarded from the final dataset.

Layer 1 focuses on structure extraction. Agent1 (Extractor) extracts a transferable representation from the source instance. For mathematical reasoning, this representation records the essential quantitative entities, constraints, and the minimal solution skeleton needed to compute the answer. For natural science question answering, it captures the information required to determine the correct option while preserving the original question intent and decision structure. When the transfer axis is modality context, the extractor additionally produces a structured split: Part A contains the question frame and options, while Part B contains the content to be converted to another modality. The extractor also assigns a content type label for Part B, such as numerical values, process descriptions, or relational mappings. Agent2 (Validator) then checks that the extraction is faithful to the original and contains no hallucinated information. For modality context transfers, it additionally verifies complementarity: neither Part A nor Part B alone should be sufficient to solve the problem, yet their combination must provide all information necessary for solving.

Layer 2 handles transfer design. Agent3 (Designer) converts the extracted structure into a concrete transfer plan under the chosen axis and level. For knowledge domain transfers, the designer specifies a target context and constructs a mapping that rewrites domain entities and terminology while keeping the abstract variables, relations, and solution skeleton fixed. Near transfer uses a context within a closely related domain family, while far transfer targets a substantially different domain. For natural science question answering, since domain-specific concepts cannot be directly re-contextualized without introducing external prerequisites, the designer constructs a fictional scenario with explicit rule descriptions and minimal demonstrations, ensuring the instance remains solvable from the provided information alone. For modality context near transfer, the designer produces a table schema to organize the data. For modality context far transfer, the design strategy differs by task family: for mathematical reasoning, the designer produces a visualization specification to guide subsequent code generation; for natural science question answering, the designer produces a detailed image generation prompt that a text-to-image model can follow directly.

We adopt different strategies because mathematical problems involve explicit numerical data that can be precisely rendered through programmatic plotting, whereas science problems often involve complex conceptual structures (e.g., molecular diagrams, reaction pathways) that are more naturally produced by generative image models. In all cases, the strategy must avoid embedding the question text or answer options in the visual artifact and must introduce meaningful non-textual visual structure. Agent4 (Validator) verifies that the plan is compatible with the extracted structure, preserves solvability, and introduces no unstated external dependencies.

Layer 3 performs problem generation. Agent5 (Generator) synthesizes the final transferred instance. For knowledge domain transfers, the generator instantiates the mapping and rewrites the task into the target context while preserving the core structure. For modality context near transfer, it renders the designated content into a structured format such as a Markdown table and merges it with the question frame. For modality context far transfer, the generation process depends on the task family: for mathematical reasoning, the generator produces executable plotting code based on the visualization specification and executes it in a sandboxed environment to render the diagram; for natural science

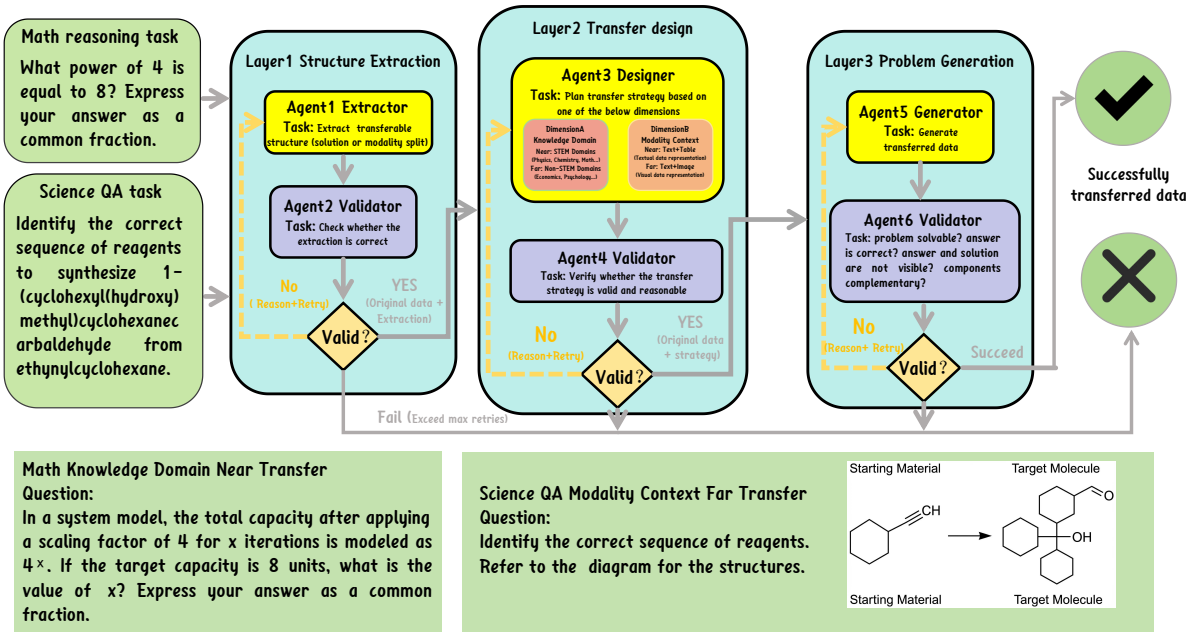

Figure 2: The ReTRE data transfer framework. Each layer forms a Generator-Validator pair with iterative refinement until the validation criteria are satisfied or the retry limit is reached. The bottom panel shows representative transfer examples for a math reasoning task and a science QA task. For space reasons, we omit the multiple-choice options in the science QA example; the options remain identical to the original after MC transfer.

question answering, the generator invokes a text-to-image model to produce the visual artifact. In both cases, the visual component is paired with the textual question and options. Agent6 (Validator) performs end-to-end validation, accepting a candidate only when it satisfies structure preservation, information sufficiency, answer consistency, and leakage prevention. For modality context transfers, it additionally verifies that the textual and visual components remain complementary, ensuring that solving the problem requires integrating both rather than reading the answer from either component alone. Only instances passing all validation criteria are added to the final dataset.

4 Experiments

Our experiments mainly investigate how the performance of the model changes when it is applied to variant data as the transfer distance increases. If the model performance remains stable, it indicates that the model has transfer robustness.

4.1 Setup

We select MATH500 (Lightman et al., 2023) and GPQA (Rein et al., 2024) as the baseline datasets, which are typical evaluation datasets in the fields of mathematics and natural sciences. We use the multi-agent framework to construct the Near and Far transfer data in Knowledge Domain and Modality Context dimensions. Due to the involvement of image input, the models we choose all support multimodal input. We select the currently advanced models for examination, including gpt-5-mini (OpenAI, 2025a), o4-mini (OpenAI, 2025b), claude-haiku-4-5-20251001 and claude-haiku-4-5-20251001-thinking (Anthropic, 2025), gemini-3-pro-all, gemini-2.5-flash-lite-nothinking and gemini-2.5-flash-lite-thinking (Comanici et al., 2025). To enhance the reliability of the experimental results, we set the temperature to 0 for all models and use pass@1 as the metric.

4.2 Main Results

Table 1 reports the comprehensive evaluation on MATH500. Across all evaluated multimodal LLMs, we observe a consistent transfer gap: performance is generally preserved under near-transfer settings but drops under far-transfer settings, with the largest degradations appearing in the most challenging shifts. Among the models, claude-haiku-4-5-20251001-thinking achieves the highest accuracy on the Original setting (93.5%), while gemini-3-pro-all attains the best average accuracy across all settings, indicating strong performance.

What's more, thinking-mode models tend to be more transfer-robust than their non-thinking counterparts. For example, claude-haiku-4-5-20251001-thinking outperforms claude-haiku-4-5-20251001 across all transferred settings. This pattern suggests that explicit deliberation mechanisms can mitigate degradation when task manifestations change while the underlying solution structure is preserved.

To contextualize these performance trends, Figure 3 visualizes the embedding distribution of original and transferred instances using representations extracted by qwen2.5-vl-embedding (Bai et al., 2025b). The distribution highlights a clear contrast between transfer types. Modality-context variants remain close to the Original cluster, suggesting that they preserve semantic proximity in the embedding space. However, semantic proximity does not guarantee cross-modal correctness: despite clustering near the Original, MC-Far introduces visual inputs that can still trigger failures, and several models exhibit marked drops from their text-based performance (e.g., gemini-3-pro-all: 93.1% to 84.6%; gpt-5-mini: 89.9% to 76.8%). These results indicate that strong text performance does not necessarily translate to cross-modal variants, even when the task semantics remain closely aligned.

Importantly, rewrites do not uniformly reduce performance. In MC-Near, multiple models slightly improve over the Original setting (e.g., gemini-3-pro-all: 93.1% to 93.5%; gpt-5-mini: 89.9% to 91.2%; gemini-2.5-flash-lite-nothinking: 85.6% to 90.2%). A plausible explanation is a structured-formatting effect: converting free-form text into well-organized tables reduces parsing ambiguity and facilitates information extraction. This benefit weakens for MC-Far, where the shift to visual representations introduces additional perception and grounding challenges.

In contrast, knowledge-domain variants form separated clusters that are farther from the Original distribution, reflecting substantial semantic drift under KD transfer. Consistent with this shift, all models show degradation on KD-Far (e.g., gemini-3-pro-all: 93.1% to 84.9%; gpt-5-mini: 89.9% to 80.0%). Overall, the results suggest that current models are comparatively more resilient to presentation changes within text (MC-Near) than to large semantic shifts across domains (KD-Far), while cross-modal transfer (MC-Far) remains a distinct failure mode where strong text performance can mask significant weaknesses.

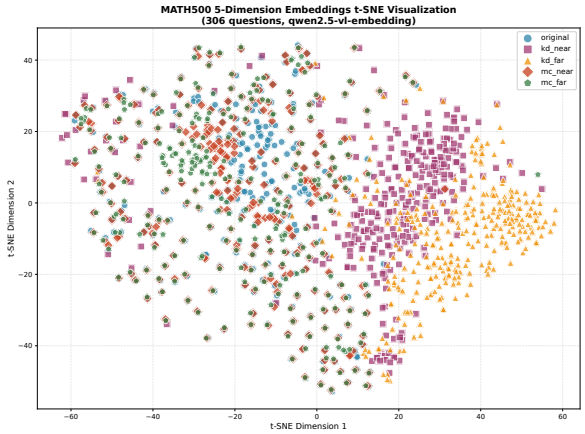

Figure 3: t-SNE visualization of problem embeddings extracted by qwen2.5-vl-embedding. The distribution reveals a fundamental distinction: Modality Context variants cluster closely with the Original problems, indicating semantic preservation. In contrast, Knowledge Domain variants form distinct, isolated clusters, reflecting significant semantic drift.

4.3 Do Different Post-training Paradigms Have an Impact on the Transfer Robustness of the Model?

To verify the impact of different training paradigms on the robustness of model transfer, we conducted a controlled experiment. We trained the Qwen3-VL-8B-Instruct model on the MATH500 dataset using full-scale SFT and GSPO methods, and then evaluated it on the corresponding four sets of transfer data.

The experimental results present a striking contrast between the two post-training paradigms, as shown in Table 2. We first observe that SFT induces a robustness collapse. While the SFT model achieves a perfect accuracy of 100% on the Original dataset, its performance collapses catastrophically across all transfer variants. For instance, accuracy on the MC-Far and KD-Near datasets drops to 43.4% and 47.1% respectively, significantly lower than the Base model (68.6% and 64.4%). This suggests that the SFT paradigm drives the model towards rote memorization of specific training instances, sacrificing generalization for in-distribution performance.

In stark contrast, GSPO enhances generalization and demonstrates superior capability. It not only outperforms the Base model on the Original dataset (85.3% vs 77.5%) but also significantly improves transfer robustness (e.g., 74.2% on KD-Near vs Base 64.4% on KD-Near). Unlike SFT, GSPO improves the model's alignment with the target

Model	Original	KD-Near	KD-Far	MC-Near	MC-Far
gemini-3-pro-all	93.1%	87.5%	84.9%	93.5%	84.6%
claude-haiku-4-5-20251001-thinking	93.5%	85.9%	81.6%	91.8%	89.5%
o4-mini	92.2%	84.2%	80.7%	91.2%	81.7%
claude-haiku-4-5-20251001	91.2%	80.6%	79.0%	88.2%	85.3%
gpt-5-mini	89.9%	82.9%	80.0%	91.2%	76.8%
gemini-2.5-flash-lite-thinking	89.5%	82.6%	76.1%	87.6%	82.7%
gemini-2.5-flash-lite-nothinking	85.6%	78.3%	75.4%	90.2%	85.3%

Table 1: Main evaluation results on the MATH500 benchmark.

Model	Original	KD-Near	KD-Far	MC-Near	MC-Far
Qwen3-VL-8B-Instruct	77.5%	64.4%	59.5%	75.2%	68.6%
Qwen3-VL-8B-Instruct(SFT)	100.0%	47.1%	48.0%	55.6%	43.4%
Qwen3-VL-8B-Instruct(GSPO)	85.3%	74.2%	69.0%	85.3%	76.8%

Table 2: Performance comparison of Base, SFT, and GSPO settings.

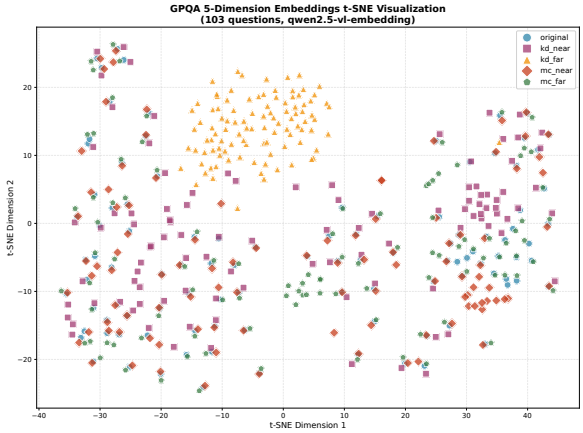

Figure 4: t-SNE visualization of GPQA embeddings. KD-Far instances separate distinctly from the original distribution, while MC variants remain entangled, indicating different degrees of semantic shift.

domain without overfitting to the specific phrasing or parameters of the training set.

4.4 How Robust are Models under Transfer in Natural Science Reasoning?

We evaluate model transfer robustness on GPQA, a natural science multiple-choice benchmark. Adopting the same two-axis design as in our main experiment (Knowledge Domain and Modality Context), we construct four transfer variants for 103 GPQA problems. Table 3 reports the accuracy across the original questions and all transfer settings.

Overall, transfer shifts consistently degrade performance for state-of-the-art models, with MC-Far presenting the most significant challenge. For instance, gemini-3-pro-all achieves 91.3% on the original questions but drops 21.3 percentage points to 70.0% on MC-Far. Similarly, o4-mini declines from 82.5% to 64.1% in the same setting. Interestingly, gpt-5-mini shows balanced degradation across both axes, with comparable performance on KD-Far (68.9%) and MC-Far (68.0%).

To interpret these patterns from a representation perspective, we visualize the embedding distribution in Figure 4. This highlights a critical insight: although MC-Far causes the most severe performance degradation, the problems remain semantically close to the source in the text embedding space. This suggests that the difficulty of Modality Context transfer stems not from semantic drift, but from the challenge of cross-modal grounding required to interpret the visual transformation.

In contrast, near-transfer settings yield performance largely consistent with the original. Most models exhibit only minor fluctuations, which we attribute to the inherent similarity between near-transfer variants and the original questions, combined with the limited granularity of pass@1 evaluation on 103 items (each problem corresponds to approximately 0.97% accuracy). We also note that some weaker models score higher on certain transfer variants than on the original questions. For example, gemini-2.5-flash-lite-nothinking improves from 49.5% to 53.4% on KD-Near and 54.4% on MC-Near. This suggests that the expert-level phrasing of GPQA may pose additional linguistic challenges for models with limited reasoning capacity, which the rewritten variants inadvertently alleviate.

Model	Original	KD-Near	KD-Far	MC-Near	MC-Far
gemini-3-pro-all	91.3%	82.5%	81.6%	88.3%	70.0%
gpt-5-mini	84.5%	79.6%	68.9%	81.6%	68.0%
o4-mini	82.5%	81.5%	70.9%	74.8%	64.1%
claude-haiku-4-5-20251001-thinking	70.9%	72.8%	66.0%	71.8%	62.1%
gemini-2.5-flash-lite-thinking	63.1%	62.1%	52.4%	59.2%	51.6%
claude-haiku-4-5-20251001	58.3%	66.0%	54.4%	57.3%	54.4%
gemini-2.5-flash-lite-nothinking	49.5%	53.4%	52.4%	54.4%	43.7%

Table 3: Evaluation results on GPQA (103 problems). We report accuracy (%) on the original questions and four transfer settings (KD/MC $\times$ Near/Far).

5 Conclusion

This work examines the transfer robustness of large language models under controlled, structure-preserving task transformations. We introduce ReTRE, a transfer-oriented evaluation framework that varies task manifestations along Knowledge Domain and Modality Context axes at near and far levels, while preserving the underlying solution structure through a multi-agent pipeline. Experiments on mathematical reasoning and natural science QA reveal a consistent transfer gap: performance remains relatively stable under near transfer but degrades substantially under far transfer, especially for knowledge-domain shifts. Notably, strong performance on text-based settings does not necessarily translate to cross-modal variants, where models can struggle even when the underlying reasoning structure is preserved. Further analyses show that RL-based post-training improves transfer robustness, whereas supervised fine-tuning can overfit in-distribution data and collapse under transfer.

Limitations

Our work has certain limitations, and these limitations should be acknowledged. Firstly, we constructed transfer variants for the MATH500 and GPQA-diamond datasets. However, not all of them were successfully constructed. For instance, there were only 306 transfer variants for MATH500. Additionally, considering economic benefits, for the image construction of MATH500, we used code to draw the pictures. For GPQA, we attempted to use the same configuration models (gemini-2.5-flash and gpt-5-mini) from MATH500 to generate data, but the success rate was too low. Therefore, we replaced them with gemini-3-pro-all and gpt-5.2 to act as agents, and used gemini-3-pro-image-preview to generate images. We considered that this was because gemini-2.5-flash itself had difficulties in understanding the task (due to budget considerations, we did not provide sufficient reasoning configurations). However, the focus of our work is on evaluating the model, not the way of data construction. In the future, we will consider how to utilize open-source models to construct data for evaluating SOTA models.

Acknowledgements

We acknowledge the support of the National Natural Science Foundation of China under Grant (No. 62572336, No. 62272334, No. 62577050), the Laboratory for Big Data Research and Intelligent Decision Making on Graduate Employment at Soochow University, the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2026C02A1236).

References

Anthropic. 2025. Claude 4.5 and claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5. Accessed 2025-11-03.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Susan M Barnett and Stephen J Ceci. 2002. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi

Wang, Dan Huang, and 1 others. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Margaret L Hilton and James W Pellegrino. 2012. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, and 1 others. 2025a. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*.

Shulin Huang, Linyi Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, and 1 others. 2025b. Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning. *arXiv preprint arXiv:2502.16268*.

Neeraja Kirtane, Yuvraj Khanna, and Peter Relan. 2025. Mathrobust-lv: Evaluation of large language models' robustness to linguistic variations in mathematical reasoning. *arXiv preprint arXiv:2510.06430*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Riccardo Lunardi, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. On robustness and reliability of benchmark-based evaluation of llms. *Preprint*, arXiv:2509.04013.

Alhassan Mumuni and Fuseini Mumuni. 2025. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. *arXiv preprint arXiv:2501.03151*.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A survey on large language model benchmarks. *Preprint*, arXiv:2508.15361.

OpenAI. 2025a. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed 2025-11-03.

OpenAI. 2025b. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed 2025-11-03.

David N Perkins, Gavriel Salomon, and 1 others. 1992. Transfer of learning. *International encyclopedia of education*, 2(2):6452–6457.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. 2024b. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. 2024. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. *arXiv preprint arXiv:2412.13670*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2024. Natural plan: Benchmarking llms on natural language planning. *Preprint*, arXiv:2406.04520.

A Additional Experiments

A.1 Does the Size of the Model Affect the Transfer Robustness?

To analyze whether scaling up model parameters naturally confers transfer robustness, we conducted a comparative experiment on the instruct model of Qwen3-VL (Bai et al., 2025a) from 2B to 235B. As shown in Table 4, we observe a clear scaling law: increasing model size from 2B to 235B yields overall performance gains across transfer settings, though improvements are not strictly monotonic at

every scale. Notably, the largest model (235B) exhibits the strongest resistance to domain shifts, particularly in the challenging KD-Far setting (72.5%), significantly outperforming the 8B model (59.5%).

However, a granular inspection reveals that transfer robustness does not scale linearly with parameter count. Specifically, we observe diminishing returns at mid-scale. The transition from 2B to 4B yields a substantial gain in average robustness (+13.8%), suggesting that the 2B model is severely under-parameterized for complex reasoning tasks. In contrast, scaling from 4B to 8B results in a robustness plateau: average improvement narrows to 1.7%, and notably, KD-Near accuracy even exhibits a marginal decline (64.7% → 64.4%). This stagnation suggests that mid-scale parameter increases, without concurrent advances in architecture or data quality, may saturate the model's capacity for surface-level pattern generalization.

A.2 Does Generational Evolution Enhance Transfer Robustness?

To determine whether generational advancements, encompassing model architecture, data scaling, and training methodologies, confer inherent improvements in robustness, we conducted a longitudinal evaluation of the Qwen-VL lineage: Qwen2-VL-7B (Wang et al., 2024a), Qwen2.5-VL-7B, and Qwen3-VL-8B. As evidenced in Table 5, we observe a strict monotonic improvement across all evaluated dimensions. Most notably, Qwen3-VL-8B achieves state-of-the-art performance with a substantial elevation in Original accuracy (77.5%), representing a significant leap from the 43.5% baseline established by its predecessor, Qwen2-VL. Beyond these aggregate gains, a granular analysis reveals a fundamental dichotomy in how evolving models contend with distinct perturbation types. On one hand, sensitivity to surface-level formatting appears to be largely resolved. While the Qwen2 architecture suffered a precipitous 11.1% performance degradation when shifting context formats (dropping from 43.5% in Original to 32.4% in MC-Near), Qwen3-VL demonstrates remarkable resilience, narrowing this "Context Gap" to a negligible 2.3% (77.5% vs. 75.2%). This trajectory suggests that recent architectural optimizations have effectively mitigated vulnerability to modal context variations.

In stark contrast, the challenge of Knowledge Domain (KD) transfer has intensified. Although Qwen3-VL secures a higher absolute baseline, it incurs a more severe penalty when generalizing to distant knowledge domains (KD-Far), exhibiting an 18.0% drop compared to the 15.1% decline observed in Qwen2. This widening "Knowledge Gap" underscores that while scaling parameters and data enhances general capabilities, it does not automatically bestow the reasoning flexibility necessary to bridge significant semantic domain shifts.

A.3 Extension to Planning Tasks

To further verify the generalization ability of ReTRE, we implemented it on the Natural Plan (Zheng et al., 2024). We selected 10 questions from each of the Trip Planning, Meeting Planning, and Calendar Scheduling categories to generate four variant datasets. Additionally, we recruited two master's students for human evaluation at a cost of $1.50 per question. The experimental results are shown in Table 6. The transfer robustness challenges observed in previous experiments also manifest in the planning domain, with all models showing significant degradation across transfer conditions.

A.4 Human Evaluation

To validate the quality of our generated transfer variants, we conducted a human evaluation study. We sampled 50 questions from MATH500 (10 per difficulty level) and 50 from GPQA-Diamond (at least 10 per subject type), evaluating all four transfer variants (KD-Near, KD-Far, MC-Near, MC-Far) for each question. For MATH500, we recruited two graduate students with a mathematics background; for GPQA, we recruited two graduate students with backgrounds in natural sciences and artificial intelligence. Each annotator was compensated at $1.50 per problem.

Annotators assessed each variant along the following dimensions: (1) Answer Correctness, whether the reference answer is correct; (2) Core Reasoning Step Similarity, rated on a 0–2 scale, where higher scores indicate greater similarity; (3) Modality Necessity (MC only), whether solving the problem requires integrating both modalities; (4) Image Quality (MC-Far only), rated on a 0–2 scale, where higher scores indicate better clarity; (5) Transfer Validity (KD only), whether the transformation constitutes a valid near/far transfer (MC transfer validity is self-evident from the modality change and thus not separately evaluated); and (6) Data Leakage, whether the question contains the answer or solution.

Table 7 reports the results averaged across two annotators. Answer correctness ranges from 96% to 100% across all conditions, with 0% data leakage, confirming that generation errors affect at most 2% to 4% of problems. Modality necessity ranges from 98% to 100% on both datasets, and image quality ranges from 84% to 94%. Core reasoning step similarity is consistently high for MATH500, ranging from 95.4% to 99.5%. GPQA KD variants show lower similarity because the fictional-scenario wrapping introduces extra steps alongside the core reasoning, diluting the similarity score.

Model	Original	KD-Near	KD-Far	MC-Near	MC-Far
Qwen3-VL-2B-Instruct	62.7%	50.3%	45.4%	56.9%	52.3%
Qwen3-VL-4B-Instruct	71.6%	64.7%	59.5%	70.9%	69.9%
Qwen3-VL-8B-Instruct	77.5%	64.4%	59.5%	75.2%	68.6%
Qwen3-VL-235B-A22B-Instruct	89.9%	80.6%	72.5%	88.9%	86.6%

Table 4: Impact of model scale on transfer robustness.

Model	Original	KD-Near	KD-Far	MC-Near	MC-Far
Qwen2-VL-7B-Instruct	43.5%	26.5%	28.4%	32.4%	31.7%
Qwen2.5-VL-7B-Instruct	56.2%	48.0%	44.4%	55.6%	52.9%
Qwen3-VL-8B-Instruct	77.5%	64.4%	59.5%	75.2%	68.6%

Table 5: Longitudinal comparison of robustness across Qwen-VL generations.

Model	Original	KD-Near	KD-Far	MC-Near	MC-Far
gpt-5-mini	60.0%	53.3%	53.3%	53.3%	36.7%
o4-mini	56.7%	40.0%	36.7%	56.7%	30.0%
gemini-2.5-flash-lite-thinking	53.3%	40.0%	40.0%	50.0%	26.7%
claude-haiku-4-5-20251001-thinking	33.3%	13.3%	23.3%	26.7%	20.0%
gemini-2.5-flash-lite-nothinking	23.3%	26.7%	6.7%	30.0%	3.3%

Table 6: Model evaluation on Natural Plan transfer variants.

Transfer	Dataset	Ans. Acc.	Step Sim.	Leakage	Modal. Nec.	Img. Qual.	Trans. Val.
MC-Near	MATH500	100%	99.5%	0%	99.0%	–	–
MC-Far	MATH500	100%	99.0%	0%	100%	84.0%	–
KD-Near	MATH500	96.0%	95.4%	0%	–	–	99.0%
KD-Far	MATH500	98.0%	95.5%	0%	–	–	94.0%
MC-Near	GPQA	100%	96.0%	0%	98.0%	–	–
MC-Far	GPQA	100%	100%	0%	100%	94.0%	–
KD-Near	GPQA	100%	53.0%	0%	–	–	100%
KD-Far	GPQA	100%	54.0%	0%	–	–	100%

Table 7: Human evaluation results averaged across two annotators. For binary metrics (0/1), we report the proportion scoring 1. For scaled metrics (0–2), we report the proportion scoring ≥ 1. "–" = not applicable.