Scaling Human-Centric Trustworthy Foundation Model Reasoning

Jiayu Liu

On Behalf of Yi R. (May) Fung, Assistant Professor **HKUST**

yrfung@ust.hk

Spring 2025

What We Do: Extending Language Models from Shallow Textual Understanding to Richer Capabilities as the Connecting Bridge Across Modalities and Tasks

Deeper Understanding of Building Blocks

The Foundations of Human-Centric Trustworthy AI/NLP Reasoning

(fundamental principles for scalable alignment, robust knowledge understanding, generalizable intelligence, etc.)

Novel Paradigms

Adapting Language + X to the Open-World Ecosystem

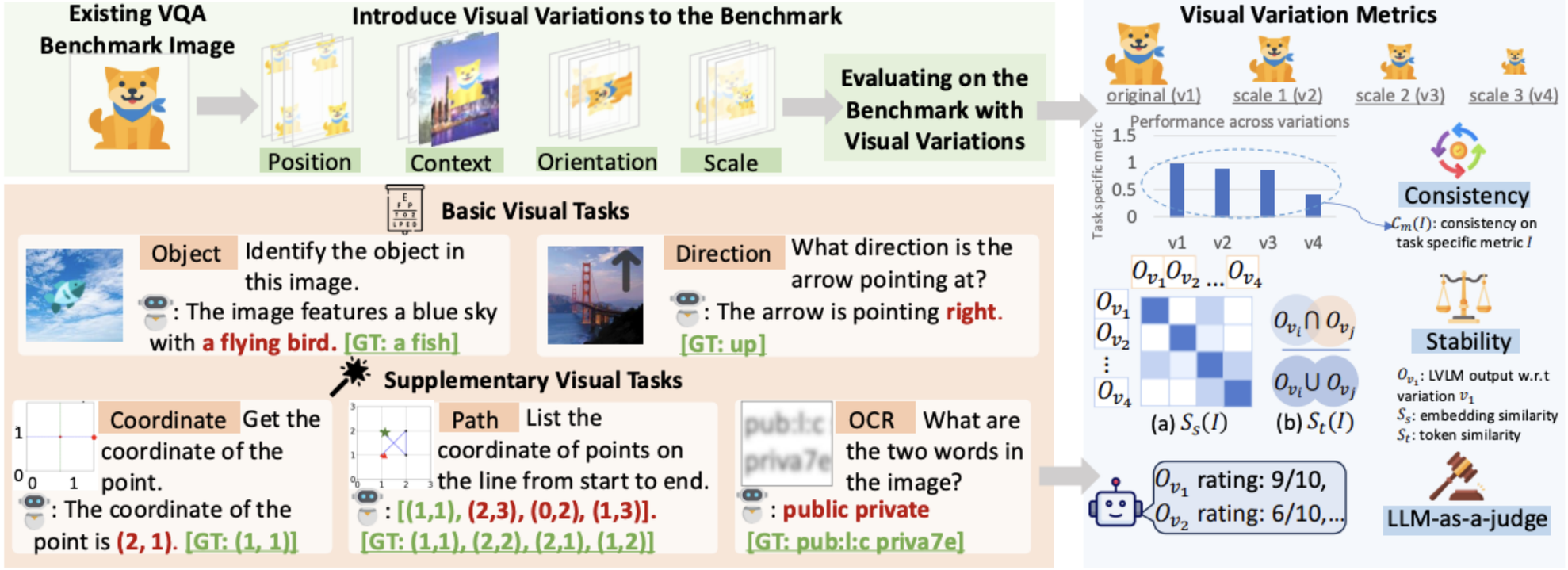
(novel problem formulations, benchmark assessments, agentic frameworks, adaptive learning styles)

Applications

AI for Coding | **AI for Health** | **AI for Business** | **AI for Science**

Let's Start with an Acuity Test on LVLM Reasoning Robustness to Simple Visual Variations, which is Key for Real-World Generalization

❖ Our automated data generation framework to conduct a holistic sweep of visual variations on LVLMs, along with our metric definitions.



Fan et al. V2R-Bench: Holistically Evaluating LVLM Robustness to Fundamental Visual Variations. Preprint, 2025.

Impact of direction, position and scale variation on LVLM

- (a)(b): The accuracy of the LVLM is higher in the peripheral than the center, meaning that LVLMs have the tendency to infer from the context, rather than focus on the objects.
- -(c) The LVLMs experience a sharp decline on a visual threshold, resembling the human visual acuity

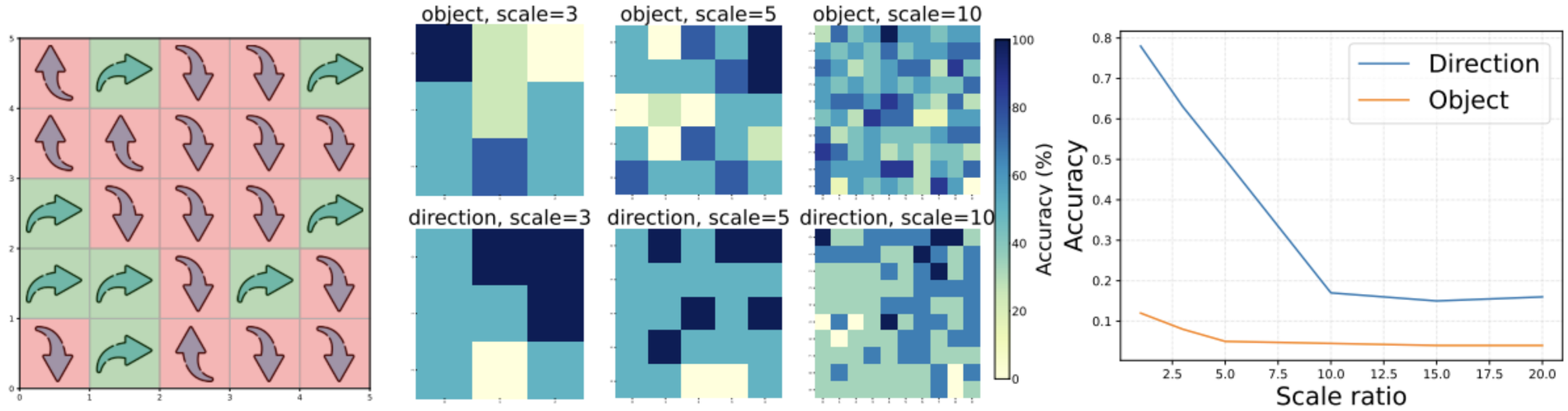


Figure 2: *(a)* Demonstration of position bias effects. *(b)* Accuracy heatmaps for object recognition and direction recognition, across object scales and position variations. *(c)* Model accuracy as a function of relative object scale.

Fan et al. V2R-Bench: Holistically Evaluating LVLM Robustness to Fundamental Visual Variations. Preprint, 2025.

The LVLM component analysis - multimodal projector

- (a)Feature space of the aligned feature does not align with that of the text embedding.
- -(b)We directly decode the output of mm-projector using the language embedding matrix. The output is not coherent natural languages.
- -(c) Demonstration of the features varies as the object position changes. This lead to error accumulation in the LVLM pipeline

Fan et al. V2R-Bench: Holistically Evaluating LVLM Robustness to Fundamental Visual Variations. Preprint, 2025.

In addition to Perception Robustness, LVLM Cross-Context Reasoning is also a Largely Overlooked Challenge!

人类秒懂，AI却懵圈：VLM²-Bench揭示视觉语言模型「视觉关联」能力短板

机器之心 2025年03月14日 15:30 北京

我们在**浏览不同的照片时可以找到出现在多张照片的同一个人**，但是我们并不需要在之前就见过这个人，叫得出名字或者对这个人很了解，而是简单的在不同的图片间通过脸部特征在视觉上的比对和关联。**同理我们还会拿着喜欢球鞋的图片去线下门店比对挑选出一样的款式（如下图）**，而不需要知道这个鞋的具体产品型号，只需要把鞋的花纹这一视觉特征给关联起来即可。**这种视觉关联的能力显然是不依赖于先验知识，是纯粹基于视觉侧的关联。**

日常生活中我们经常利用“视觉关联”，比如图中这个男孩正拿着手机上的图片去线下门店一一比对，来挑选出一样的球

- ***Cross-context visual reasoning*** is extremely simple and straightforward for the human cognitive process…
- But it is **quite challenging for current large vision language models** (**LVLMs**), especially across multiple images and videos!

Why, and ***how can we improve***?

Zhang et al. VLM^2-Bench: A Closer Look at How Well VLMs Implicitly Link Explicit Matching Visual Cues. Preprint, 2025.

How Well VLMs Implicitly Link Explicit Matching Visual Cues: VLM²-Bench

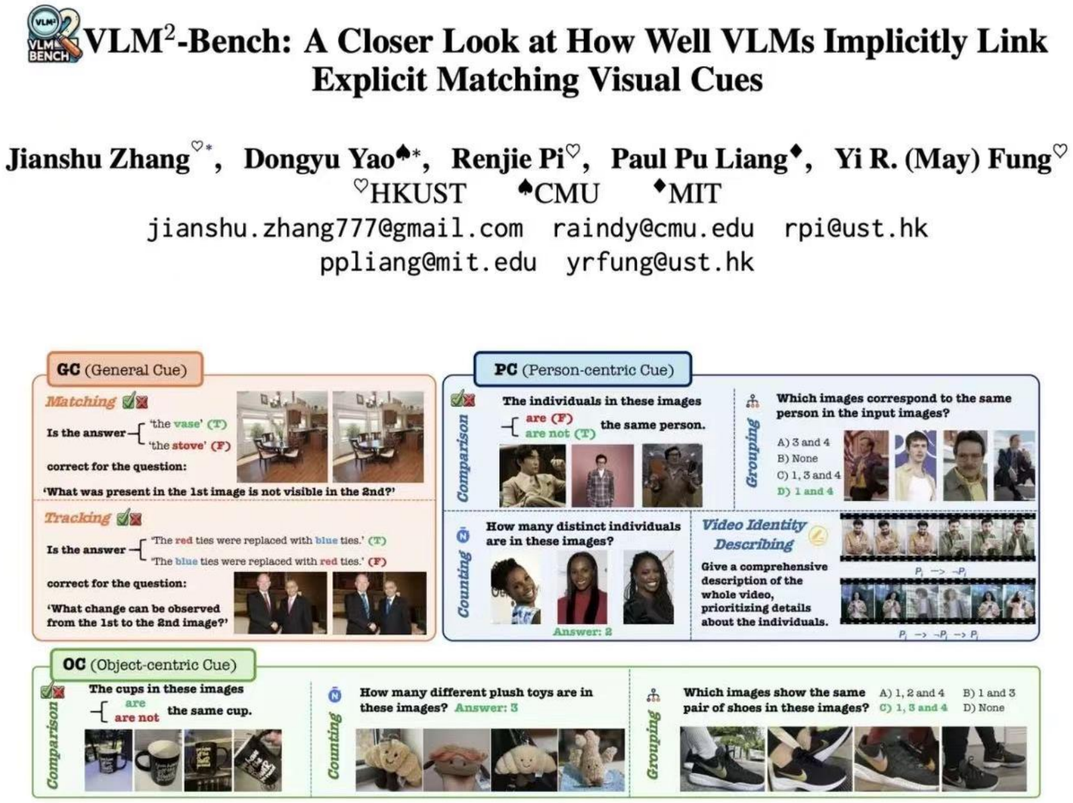
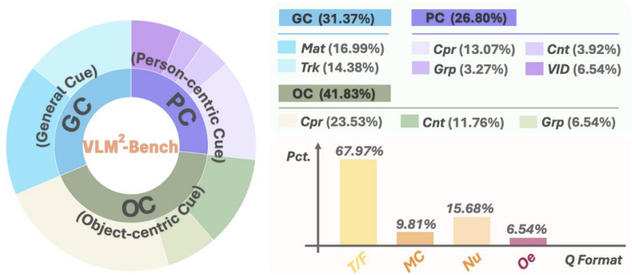


Figure 2: Overview of **VLM²-Bench**. The benchmark is categorized into three subsets based on visual cues: GC (General Cue), OC (Object-centric Cue), and PC (Person-centric Cue), each comprising multiple subtasks. To comprehensively evaluate VLMs' ability to visually link matching cues, the benchmark includes diverse question formats—T/F, multiple-choice, numerical, and open-ended—ensuring a comprehensive evaluation.

- Statistical overview: 9 subtasks across the 3 main categories of visual cues.
- Inter-annotator agreement of over 0.98 in Kappa score



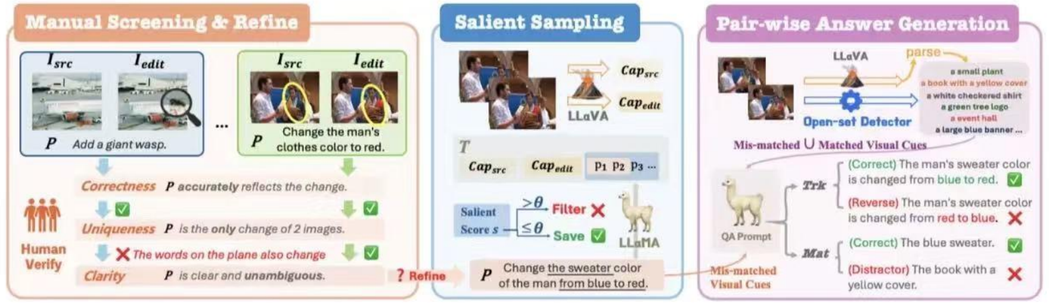


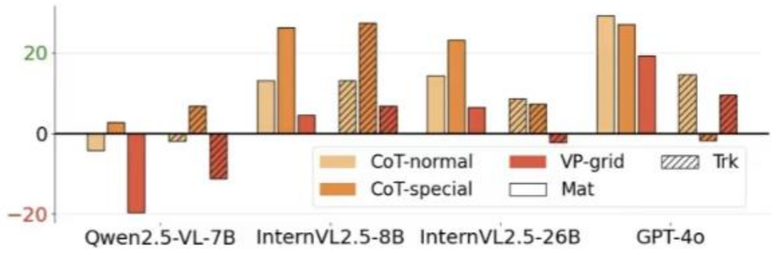
Figure 3: Construction of **GC**: (i) We start by manually verifying the edited image data based on three key criteria. (ii) A VLM is then prompted to generate captions for each image, followed by salient score-based filtering to retain the challenging cases. (iii) Finally, visual cues are extracted from two sources and incorporated into a QA prompt, guiding an LLM to generate both positive and negative answer pairs.

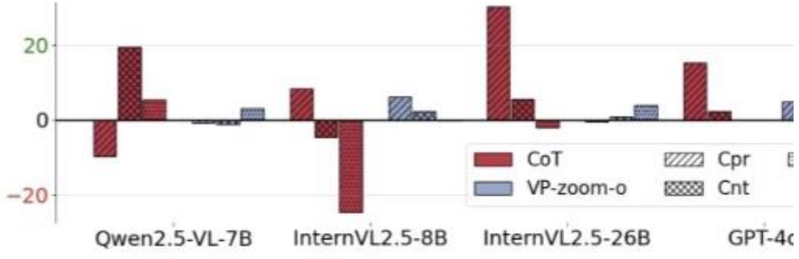
SOTA LVLMs Still Lag Far Behind Human Performance

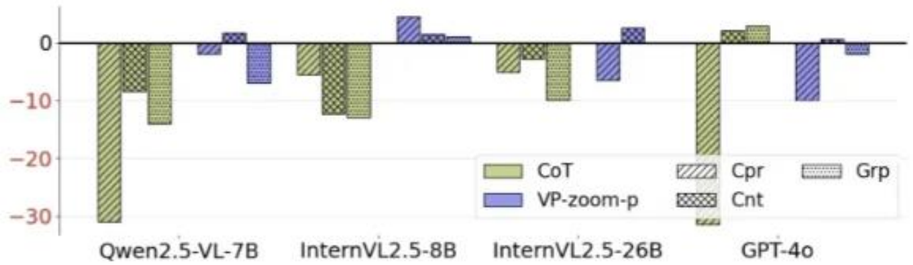
Baselines or Models	GC		OC			PC				Overall*	
	Mat	*Trk*	*Cpr*	*Cnt*	*Grp*	*Cpr*	*Cnt*	*Grp*	*VID*	Avg	Δ_{human}
Chance-Level	25.00	25.00	50.00	34.88	25.00	50.00	34.87	25.00	-	33.72	-61.44
Human-Level	95.06	98.11	96.02	94.23	91.92	97.08	92.87	91.17	100.00	95.16	0.00
LLaVA-OneVision-7B	16.60	13.70	47.22	56.17	27.50	62.00	46.67	37.00	47.25	39.35	-55.81
LLaVA-Video-7B	18.53	12.79	54.72	62.47	28.50	62.00	66.91	25.00	59.00	43.32	-51.84
LongVA-7B	14.29	19.18	26.67	42.53	18.50	21.50	38.90	18.00	3.75	22.59	-72.57
mPLUG-Owl3-7B	17.37	18.26	49.17	62.97	31.00	63.50	58.86	26.00	13.50	37.85	-57.31
Qwen2-VL-7B	27.80	19.18	68.06	45.99	35.00	61.50	58.59	49.00	16.25	42.37	-52.79
Qwen2.5-VL-7B	35.91	43.38	71.39	41.72	47.50	80.00	57.98	69.00	46.50	54.82	-40.34
InternVL2.5-8B	21.24	26.03	53.33	55.23	46.50	51.50	60.00	52.00	5.25	41.23	-53.93
InternVL2.5-26B	30.50	30.59	43.33	51.48	52.50	59.50	59.70	61.00	21.75	45.59	-49.57
GPT-4o	37.45	39.27	74.17	80.62	57.50	50.00	90.50	47.00	66.75	60.36	-34.80

❖ Note that models perform better in linking person-centric (PC) cues than object-centric (OC) cues.

❖ Interestingly, reasoning in language via COT helps, but **visual prompting yields mixed results**.







Zhang et al. VLM^2-Bench: A Closer Look at How Well VLMs Implicitly Link Explicit Matching Visual Cues. Preprint, 2025.

The First Step is Benchmark Assessments for Gleaning Insights… Then We Also Propose Novel Solutions to Train Models Better

MACAROON: Training Vision-Language Models To Be Your Engaged Partners

Shujin Wu[1,2*] **May Fung**[1] **Sha Li**[1] **Yixin Wan**[3] **Kai-Wei Chang**[3] **Heng Ji**[1]
[1]University of Illinois Urbana-Champaign
[2]University of Southern California [3]University of California, Los Angeles
{shujinwu}@usc.edu {yifung2, hengji}@illinois.edu

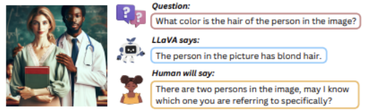


Figure 1: Existing LVLMs fail to ask clarifying questions or acknowledge their knowledge boundary, resulting in biased and hallucinated responses.

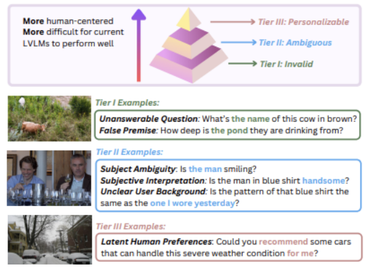


Figure 2: Typical examples for each question type within our defined hierarchy.

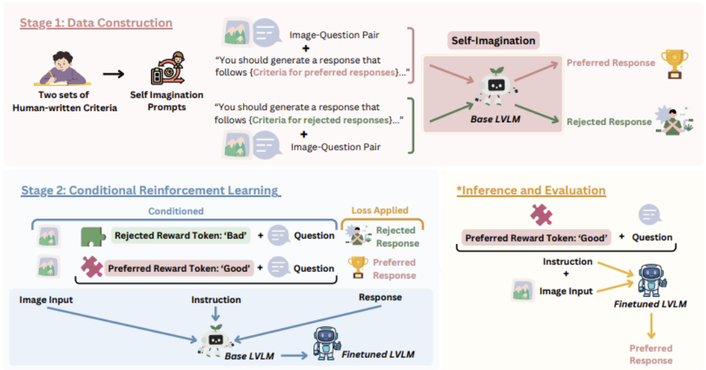


Figure 6: Overview of MACAROON. In the data construction stage, MACAROON avoids using extensive human or teacher model supervision via self-imagined desirable and undesirable responses based on human-written criteria. The contrastive response pairs, together with general vision-language instruction tuning samples, are effectively utilized through conditional reinforcement learning.

CALM: Unleashing the Cross-Lingual Self-Aligning Ability of Language Model Question Answering

Yumeng Wang[2] **Zhiyuan Fan**[2] **Qingyun Wang**[1] **Yi R. (May) Fung**[2*] **Heng Ji**[1*]
[1]University of Illinois Urbana-Champaign, [2]HKUST
ywanglu@connect.ust.hk yrfung@ust.hk hengji@illinois.edu

Abstract

Large Language Models (LLMs) are pretrained on extensive multilingual corpora to acquire both language-specific cultural knowledge and general knowledge. Ideally, while LLMs should provide consistent responses to culture-independent questions across languages, we observe significant performance disparities. To address this, we explore the **C**ross-Lingual Self-**A**ligning ability of **L**anguage **M**odels (**CALM**) to align knowledge across languages. Specifically, for a given question, we sample multiple responses across different languages, and select the most self-consistent response as the target, leaving the remaining responses as negative examples. We then employ direct preference optimization (DPO) to align the model's knowledge across different languages. Evaluations on the MEDQA and X-CSQA datasets demonstrate CALM's effectiveness in enhancing cross-lingual knowledge question answering, both in zero-shot and retrieval-augmented settings. We also found that increasing the number of languages involved in CALM training leads to higher accuracy and consistency. We offer a qualitative analysis of how cross-lingual consistency can enhance knowledge alignment and explore the method's generalizability[1].

1 Introduction

LLMs have been pre-trained on various knowledge domains in multiple languages, capturing extensive world knowledge (Yu et al., 2024). This knowledge can be either sociocultural-dependent (Sun et al., 2023; Liu et al., 2025) or sociocultural-independent

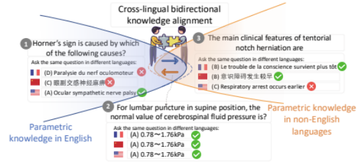


Figure 1: Knowledge is not well-aligned across languages. (1) represents knowledge encoded in English that is difficult to retrieve from other languages. (2) is the knowledge that is already well-aligned across languages. (3) is the knowledge encoded in other languages that is difficult to retrieve in English. Ideally, we want all the culture-independent knowledge to fall into (2).

et al., 2024; Wu et al., 2025a). Research indicates that LLMs exhibit varying proficiency when addressing the same task across different languages (Xu et al., 2024; Huang et al., 2024b). This variability stems from the difficulty of accessing knowledge encoded in one language while using others.

To bridge the gap, recent papers introduced cross-lingual consistency (Qi et al., 2023), which pertains to the capacity to provide consistent responses across different languages when presented with the same query. The ultimate goal is to achieve language-agnostic question-answering proficiency in LLMs, enabling them to generalize effectively in multilingual environments. Gao et al. (2024) highlighted the positive impact of multilingual pre-training and instruction tuning on enhancing cross-

Scaling Laws of Synthetic Data for Language Models

Zeyu Qin[*] **Qingxiu Dong** **Xingxing Zhang** **Li Dong** **Xiaolong Huang** **Ziyi Yang** **Mahmoud Khademi** **Dongdong Zhang** **Hany Hassan Awadalla** **Yi R. Fung** **Weizhu Chen** **Minhao Cheng** **Furu Wei**

Abstract

Large language models (LLMs) achieve strong performance across diverse tasks, largely driven by high-quality web data used in pre-training. However, recent studies indicate this data source is rapidly depleting. Synthetic data emerges as a promising alternative, but it remains unclear whether synthetic datasets exhibit predictable scalability comparable to raw pre-training data. In this work, we systematically investigate the scaling laws of synthetic data by introducing SYNTHLLM, a scalable framework that transforms pre-training corpora into diverse, high-quality synthetic datasets. Our approach achieves this by automatically extracting and recombining high-level concepts across multiple documents using a graph algorithm. Key findings from our extensive mathematical experiments on SYNTHLLM include: (1) SYNTHLLM generates synthetic data that reliably adheres to the *rectified scaling law* across various model sizes; (2) Performance improvements plateau near 300B tokens; and (3) Larger models approach optimal performance with fewer training tokens. For instance, an 8B model peaks at 1T tokens, while a 3B model requires 4T. Moreover, comparisons with existing synthetic data generation and augmentation methods demonstrate that SYNTHLLM achieves superior performance and scalability. Our findings highlight synthetic data as a scalable and reliable alternative to organic pre-training corpora, offering a viable path toward continued improvement in model performance.

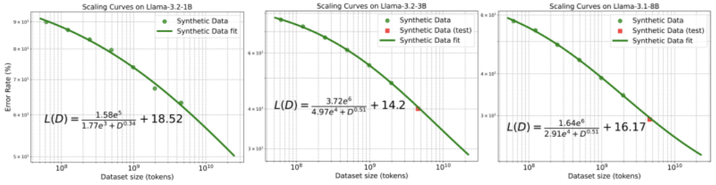


Figure 1: Scaling laws on different model sizes. The x axis denotes the number of training tokens. The y axis represents the models' error rates on MATH. The green points represent the data sizes used to fit the scaling laws, while the red points are used to test the prediction performance of the fitted curves.

Our Team and Research Mission

Yi R. (May) Fung, PI

Zhitao He

Shijue Huang

Zhaochen Su

Zhiyuan Fan

Dadi Guo

Zeyu Qin

Rui Min

Yuchen Huang

Yumeng Wang

Overarching Goal: Advance **human-centered trustworthy AI** with **multimedia knowledge reasoning** capability and **scalable alignment** principles for helping **solve real-world problems.**