Lab Seminar

2024.01.03

김정환

CogAgent: A Visual Language Model for GUI Agents

Wenyi Hong[1*] Weihan Wang[1*] Qingsong Lv[2] Jiazheng Xu[1*] Wenmeng Yu[2]
Junhui Ji[2] Yan Wang[2] Zihan Wang[1*] Yuxuan Zhang[2*] Juanzi Li[1]
Bin Xu[1] Yuxiao Dong[1] Ming Ding[2†] Jie Tang[1†]
[1]Tsinghua University [2]Zhipu AI
{hwy22@mails, jietang@}.tsinghua.edu.cn, ming.ding@zhipuai.cn

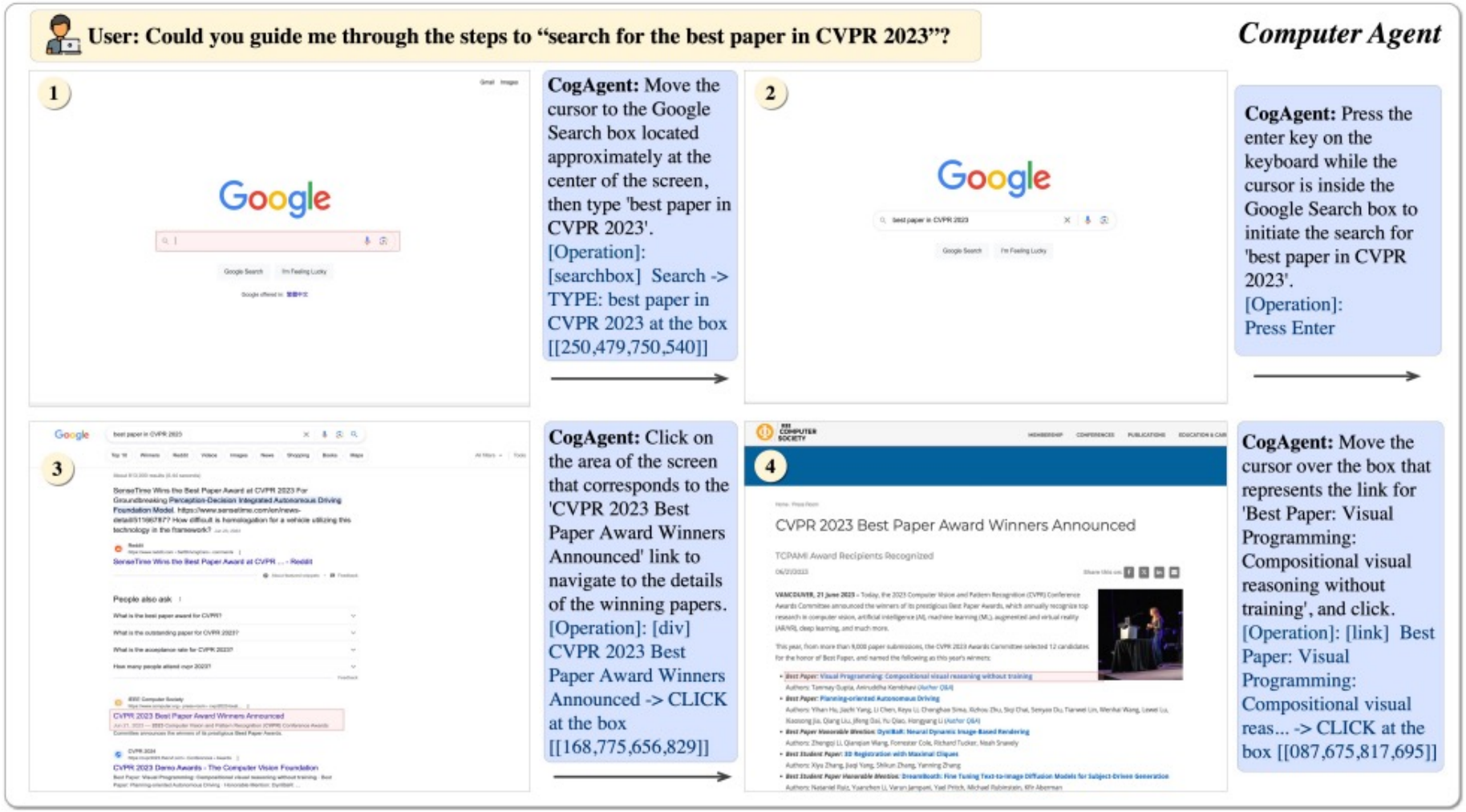
CogAgent : VLM for GUIs

Abstract

- 배경
  - 사람들은 컴퓨터나 스마트폰 화면과 같은 GUI와의 상호작용에 많은 시간을 보내고 있음
- 문제
  - ChatGPT와 같은 LLM은 이메일 작성 같은 작업에서 사람들을 도울 수 있음.
  - GUI를 이해하고 상호작용하는데 어려움을 겪고 있어 자동화 수준을 높이는 데 한계가 있음.
- 방법
  - 본 논문에서는 GUI 이해와 탐색을 전문으로 하는 18B개의 파라미터를 가진 시각 언어 모델(VLM), CogAgent
  - 저해상도 및 고해상도 이미지 인코더를 활용함으로써 CogAgent는 1120x1120 해상도의 입력을 지원
  - 그에 따라 GUI 속 작은 페이지 요소와 그 안의 텍스트를 인식할 수 있음.
- 결과
  - VQAv2, OK-VQA, Text-VQA, ST-VQA, ChartQA, infoVQA, DocVQA, MM-Vet, POPE를 포함한 다섯 개의 텍스트가 풍부한 벤치마크와 네 개의 일반 VQA 벤치마크에서 최고의 성능을 달성
- 의의
  - 스크린샷만을 입력으로 사용하는 CogAgent는 HTML을 입력으로 사용하는 LLM 기반 방법들보다 PC 및 안드로이드 GUI 탐색 작업 벤치마크인 Mind2Web이나 AITW에서 더 뛰어난 성능을 보임

CogAgent : VLM for GUIs

Overview



CogAgent : VLM for GUIs

Overview

CogAgent : VLM for GUIs

Introduction : Problem

- 배경
  - (이상적 시나리오) 작업 설명을 입력하고, 온라인으로 티켓 예매/웹 검색/파일 관리 같은 작업들이 자동으로 완료되는 동안 커피 마시며 휴식
  - AutoGPT(2023) : 사용자가 목표를 설정하면 AI가 자동으로 인터넷을 검색하여 방법을 탐색해 결과물을 내놓음
  - LLM 기반의 Agent와 관련된 연구 다수 존재
- 문제
  - 대부분의 어플리케이션이 GUI를 통해 인간과 상호작용하기 때문에 실제 시나리오에서 상당히 제한적
  - 상호작용을 위한 표준 API의 부족
  - 아이콘, 이미지, 다이어그램, 시공간 정보는 단어로 직접 전달하기 어려움
  - 웹페이지와 같은 텍스트 렌더링 GUI에서조차 캔버스와 iframe과 같은 요소들은 HTML을 통해 기능을 파악할 수 없음
- 해결
  - 시각 언어 모델(VLM)을 기반으로 한 에이전트
  - HTML과 OCR 결과와 같은 텍스트 입력에만 의존하는 대신, VLM 기반 에이전트는 GUI를 직접 인식

CogAgent : VLM for GUIs

Introduction : Challenge

- Dataset
  - 현재 대부분의 VLM은 LAION과 같은 웹상의 자연 이미지로 구성된 데이터셋에서 사전 훈련
  - GUI 이미지는 다른 이미지들과 다른 특징을 가지는 바, GUI와 OCR에 대한 라벨링이 된 데이터셋 구축하였음

- 고해상도 vs. 컴퓨팅
  - GUI에서는 필요와 목적에 따라 작은 아이콘과 텍스트가 흔함
  - 일반적으로 사용되는 224x224의 해상도로는 인식하기 어려움. 그러나 입력 이미지의 해상도를 높이면 긴 시퀀스 길이 발생
  - 1120x1120 이미지는 패치 크기가 14일 때 6400 토큰에 해당하는 시퀀스가 되어 긴 학습시간과 컴퓨팅 리소스 필요

CogAgent : VLM for GUIs

Method

- Base VLM
  - CogVLM-17B : SOTA VLM
- High Resolution Cross-Module
  - GUI 이해에 필수적인 작은 요소 및 텍스트 관련 특징에 중점
- Pre-training
  - Text Recognition Data
  - Visual Grounding Data
  - GUI imagery Data
- Multi-task Fine-tuning and Alignment
  - 스마트폰과 컴퓨터에서 수천 개의 스크린샷을 수동으로 수집
  - Human Labeler에 의해 화면 요소, 잠재적으로 이루어질 수 있는 Task, 질문-답변 형식의 Task Method로 라벨링
  - 웹과 안드로이드 동작에 초점을 맞춘 데이터셋인 Mind2Web과 AITW를 사용 (행동 시퀀스 및 해당 스크린 샷을 포함하며, GPT-4를 사용하여 이를 자연 언어 질문-답변 형식으로 변환)

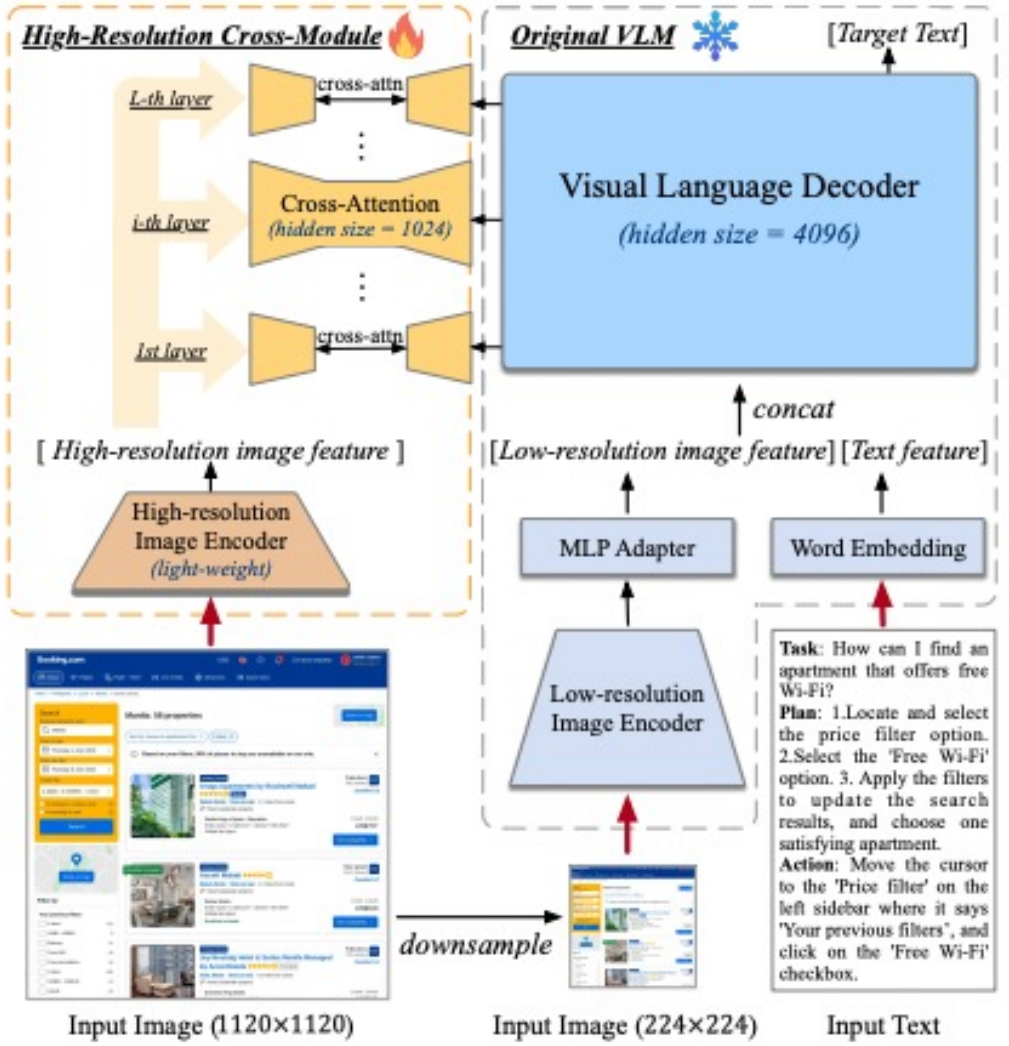
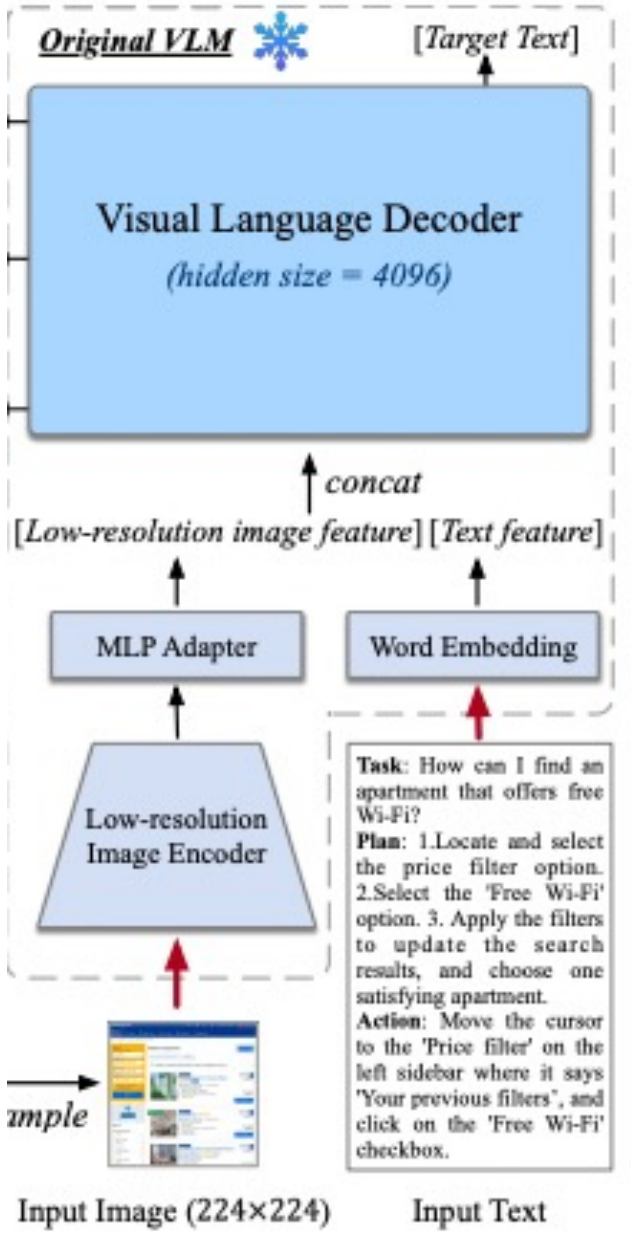


Figure 2. Model architecture of CogAgent. We adopt CogVLM as the original VLM.

CogAgent : VLM for GUIs

Method : Pre-training

- Text Recognition
  - Synthetic Renderings with Text : 80M
  - COYO / LAION-2B의 이미지의 OCR 데이터 : 18M
  - arXiv에서 출시된 소스코드(LaTeX)의 이미지-텍스트 쌍 : 9M
- Visual Grounding
  - GUI 에이전트가 이미지 내의 다양한 요소를 정확하게 이해하고 그 위치를 파악하는 능력을 갖추는 것이 매우 중요
  - LAION-115M에서 샘플링된 이미지-캡션 쌍이 있는 40M 이미지로 구성된 시각적 그라운딩 데이터셋 사용
  - 캡션의 엔티티를 바운딩 박스와 연결하여 그 위치를 나타냄
  - [[x0, y0, x1, y1]]의 형태로 해당 요소의 위치 라벨링
- GUI imagery
  - CCS400K(Common Crawl Screenshot 400k) 데이터셋 구축
  - 모든 보이는 DOM 요소와 해당하는 렌더링된 박스를 컴파일
  - 이 데이터셋을 REC 및 REG 질문-답변 쌍 1억 4천만개로 보완



CogAgent : VLM for GUIs

Experiments

- Foundational Visual Understanding

Method	General VQA		Text-rich VQA					
	VQAv2	OKVQA	OCRVQA	TextVQA	STVQA	ChartQA	InfoVQA	DocVQA
task-specific fine-tuning models								
Pix2Struct [16]	-	-	-	-	-	58.6	40.0	76.6
BLIP-2 [18]	82.2	59.3	72.7	-	-	-	-	-
PALI-X-55B [8]	86.0	66.1	75.0	71.4	79.9	70.9	49.2	80.0
$\text{CogVLM}_{\text{task-specific}}$ [38]	84.7	64.7	74.5	69.7	-	-	-	-
generalist models								
UReader [40]	-	57.6	-	-	-	59.3	42.2	65.4
Qwen-VL [2]	79.5	58.6	**75.7**	63.8	-	65.7	-	65.1
Qwen-VL-chat [2]	78.2	56.6	70.5	61.5	-	66.3	-	62.6
Llava-1.5 [20]	80.0	-	-	61.5	-	-	-	-
Fuyu-8B [3]	74.2	60.6	-	-	-	-	-	-
$\text{CogVLM}_{\text{generalist}}$ [38]	83.4	58.9	74.1	68.1	-	-	-	-
CogAgent (Ours)	**83.7**	**61.2**	75.0	**76.1**	**80.5**	**68.4**	**44.5**	**81.6**

Table 1. **Performance on Visual Question Answering benchmarks.** Bold text indicates the best score among the generalist category, and underlined text represents the best score across both generalist and task-specific categories.

CogAgent : VLM for GUIs

Experiments

- GUI Agent
  - Computer Interface : Mind2Web
  - Smartphone Interface : Android in the Wild(AITW)

Method	cross-task	cross-website	cross-domain	overall
Representations of screen inputs: HTML				
GPT-3.5[29](few-shot)	18.6	17.4	16.2	17.4
GPT-4[30]† (few-shot)	36.2	30.1	26.4	30.9
Flan-T5XL [10]	52.0	38.9	39.6	43.5
LLaMA2-7B[37]	52.7	47.1	50.3	50.1
LLaMA2-70B[37]	55.8	51.6	55.7	54.4
Representations of screen inputs: Image				
CogAgent (Ours)	**62.3**	**54.0**	**59.4**	**58.2**

Table 3. **Performance on Mind2Web.** † denotes element selection from top-10 element candidates, others from top-50, following Deng et al. [10]. Results for GPT-3.5 and GPT-4 are from Deng et al. [10].

Method	GoogleApp	Install	WebShop	General	Single	Overall
Representations of screen inputs: textual description (OCR+icon)						
GPT-3.5[29](few-shot)	10.47	4.38	8.42	5.93	9.39	7.72
LLaMA2-7B[37]†	30.99	35.18	19.92	28.56	27.35	28.40
Representations of screen inputs: image						
Auto-UIunified[43]	71.37	76.89	70.26	**68.24**	84.58	74.27
CogAgent (Ours)	**74.95**	**78.86**	**71.73**	65.38	**93.49**	**76.88**

Table 4. **Performance on Android in the Wild (AITW) dataset.** † represents models individually fine-tuned on each subset, while others are unified models across all subsets. The results of LLaMA2 and GPT-3.5 are from Zhan and Zhang [43].

CogAgent : VLM for GUIs

Conclusion

- Contribution
  - CogAgent는 고해상도 입력을 위한 향상된 사전 훈련 데이터 구축과 효율적인 아키텍처를 갖춘 VLM 기반 GUI 에이전트
  - CogAgent는 다양한 VQA 및 GUI 벤치마크에서 SOTA를 달성하며, 오픈 소스로 제공될 예정

- Future Works
  - CogAgent는 VLM 기반 GUI 에이전트의 초기 탐구로서, 여전히 몇 가지 단점 존재
  - 정확하지 않은 출력 좌표와 여러 이미지를 처리하는 능력 부족 등이 있으며, 이는 추가 연구가 필요함을 시사

- Demo
  - http://36.103.203.44:7861/

CogAgent : VLM for GUIs

Ideation

- GUI Agent 구현
  - 현재는 매번 스크린샷을 입력하여 다음 액션을 Bounding Box 형식으로 반환 받는 방식
  - 입력에 대한 출력 이후 다음 화면 스크린샷 촬영 및 Bounding Box 형식의 반환에 대해 해당 동작 자동화 가능

- Human-GUI Agent Interaction
  - 위에서 구현한 GUI Agent에 대한 상호작용 User Research 가능
  - 컴퓨터 인터페이스 : Mind2Web / 안드로이드 : AITW 데이터셋 존재

HCC LAB SNU

INSTRUCTION TUNING WITH HUMAN CURRICULUM

Bruce W. Lee[*,1,3] **Hyunsoo Cho**[*,2,3] **Kang Min Yoo**[2,3]
[1]University of Pennsylvania [2]Seoul National University [3]NAVER Cloud
brucelws@seas.upenn.edu
johyunsoo@europa.snu.ac.kr
kangmin.yoo@navercorp.com

Corgi : Instruction Tuning with Human Curriculum

Abstract

- 배경 : Instruction Tuning

OpenAI는 GPT-3.5를 학습시킬 때
학습 데이터의 양과 출처의 확대에 중점을 두었다.

이들은 텍스트, 코드 등 다양한 유형의 데이터 뿐만 아니라,
CSV, JSON 같은 여러 포맷을 포함하는 인터넷 데이터를
학습에 활용하는 것을 우선시하였다.

GPT-4의 개발과정에서 OpenAI는 학습데이터의 양 뿐만 아니라
그 질 또한 중요하다는 것을 인식하게 되었다.

특히 지시사항과 그에 상응하는 출력이 적절하게 짝지어진 것,
다양한 형식의 지시사항을 반영하는 것이 모델의 성능에
중대한 영향을 미친다는 사실을 발견한 것이다.

Corgi : Instruction Tuning with Human Curriculum

Abstract

- 문제
  - 최근 Instruction Tuning을 위한 데이터가 전체 성능에 중요한 영향을 끼친다는 연구 다수
  - Instruction Data의 품질과 다양한 형식의 지시 포함 여부, 단계별 추론 포함 여부가 성능 결정
  - 데이터를 보다 구체적이고 추적 가능한 방식으로 정리하고 학습하는 방법론에 대한 연구는 난제로 남아있음
- 방법 : CORGI Framework
  - LLM을 고등학생처럼 여기며 고등학교와 대학교와 같은 교육기관에서 지적 지식을 점진적/단계적으로 습득하도록 함
  - 간단한 것에서부터 복잡한 것으로 학습하는 기본 원칙에 따라 LLM 학습
  - 실제 교육과정을 반영한 합성 데이터셋 CORGI 구축
  - 교육과정 별 단계적 학습을 반영하여 LLM 학습
- 결과
  - 학습 데이터의 순서를 최적화한 점진적 학습이 무작위 학습보다 성능 우위
  - CORGI의 작은 데이터셋 크기(66K)에도 불구하고 지식 벤치마크(MMLU)에서 성능 개선
  - 상식 추론 벤치마크, 언어 이해 벤치마크에서도 성능 향상

Corgi : Instruction Tuning with Human Curriculum

Overview

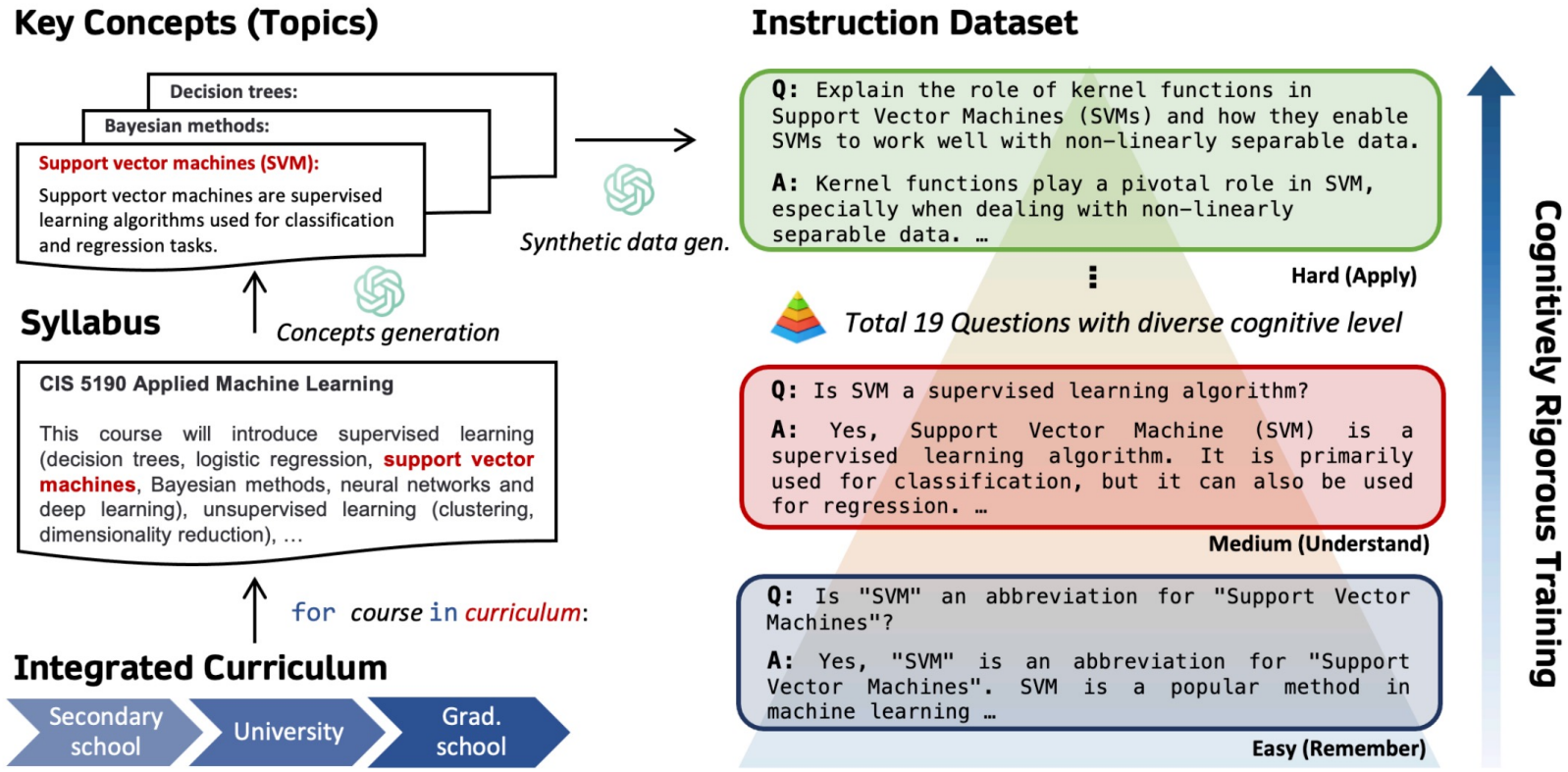


Figure 1: Overview of our educational framework. We create a dataset based on a continuum from secondary school to grad school, extracting multiple concepts from each course. For every concept, we formulate 19 questions of varied cognitive levels using Bloom's taxonomy.

Corgi : Instruction Tuning with Human Curriculum

Dataset Construction

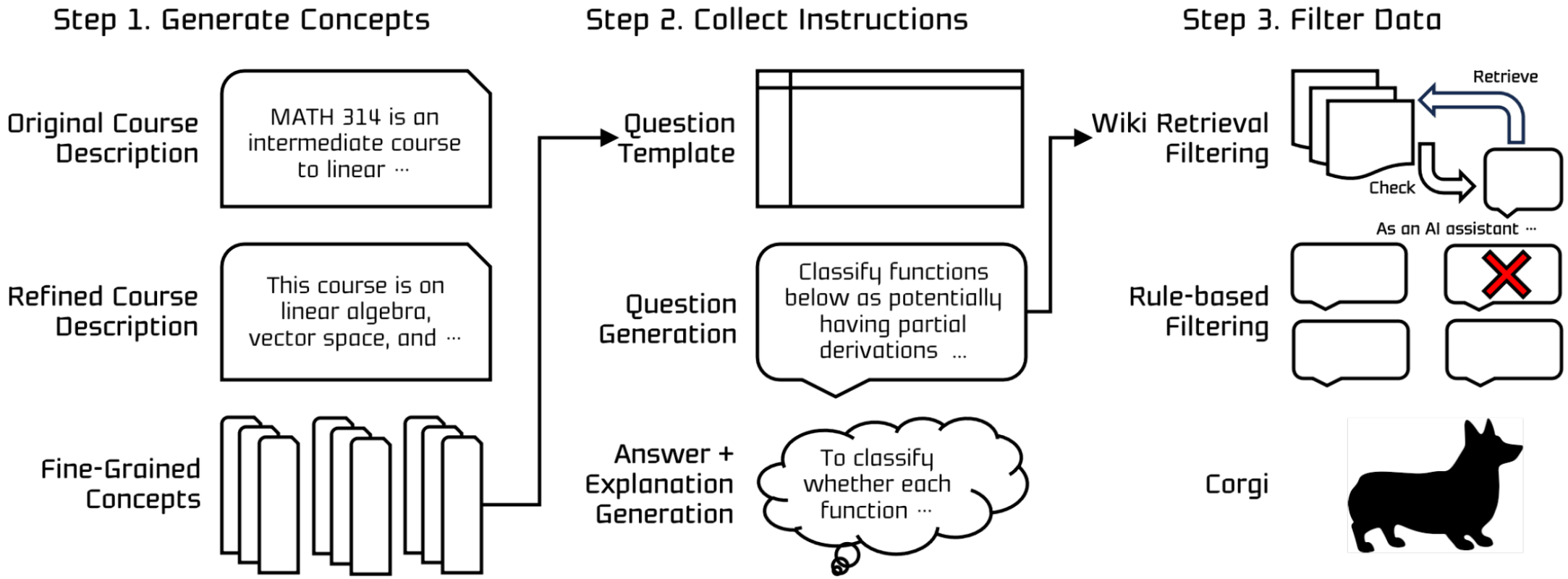


Figure 7: A visual description of the dataset construction steps.

Corgi : Instruction Tuning with Human Curriculum

Dataset Construction : Generate Concept

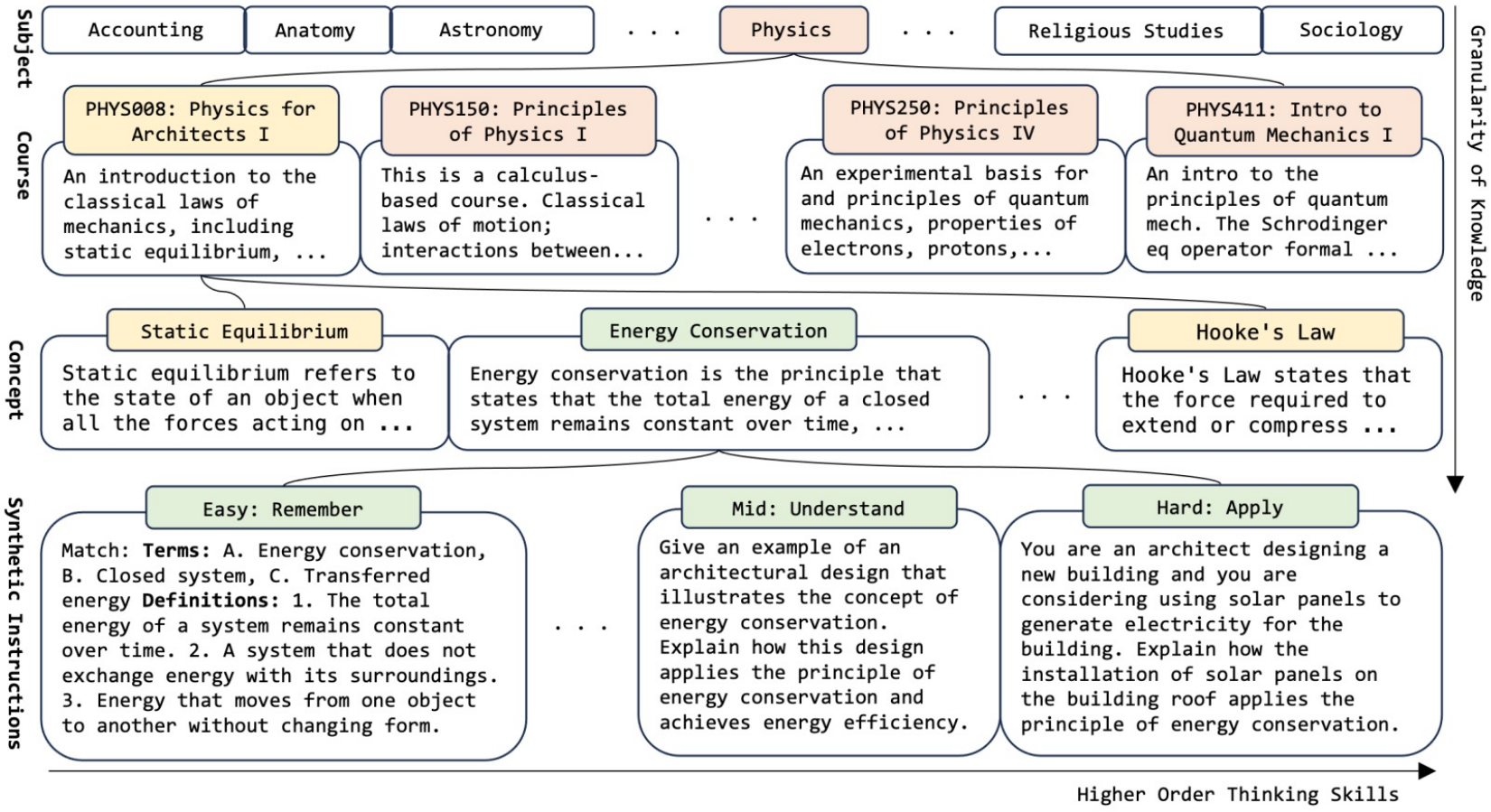


Figure 8: A hierarchical description and example of CORGI dataset.

Corgi : Instruction Tuning with Human Curriculum

Dataset Construction : Question Generation

Prompt Template

You are a {subject} professor teaching "{subject}, {course_title}, {concept}"
You are making questions for a test that questions student's various levels of thinking. The current question tests students on {cognitive_process} ({cognitive_load}), out of remembering (easy), understanding (medium), and applying (hard).
Come up with an exam question to assess student's ability to {cognitive_process_definition}
Question Format:
- {question_format}
Test Constraints:
- All questions should be self-contained (answerable using the provided information)
- All questions must have a clear, defined answer
- All questions must not use graphics
- Follow Question Format!
- Print only question only!! (Don't print the answer)

Example Prompt

You are a Higher Education - Astronomy professor teaching "Higher Education - Astronomy, A Survey of the Universe, Solar System: The solar system refers to the collection of celestial bodies, including the sun, planets, satellites, asteroids, comets, and other smaller objects that are bound together by gravitational forces. This concept involves the study of the formation, characteristics, and dynamics of these objects within the system, as well as their interactions with each other."
You are making questions for a test that questions student's various levels of thinking. The current question tests students on understanding (medium), out of remembering (easy), understanding (medium), and applying (hard).
Come up with an exam question to assess student's ability to construct a cause-and-effect model of a system (e.g., Explain the causes of important 18th-century events in France)
Question Format:
- a redesigning task where one is asked to change the system to accomplish some goal (such as, "How could you improve a bicycle tire pump so that it would be more efficient?")
Test Constraints:
- All questions should be self-contained (answerable using the provided information)
- All questions must have a clear, defined answer
- All questions must not use graphics
- Follow Question Format!
- Print only question only!! (Don't print the answer)
- equations in plain text
- no MCQ, don't provide options
- make questions have as high educational value as possible
- do NOT duplicate your previous question
Previous Question:
- Suppose you are studying the solar system, and you observe that a comet is moving in a highly elliptical orbit around the Sun. Construct a cause-and-effect model to explain the factors that could have influenced the comet's orbit.
Question
Question: ...

Corgi : Instruction Tuning with Human Curriculum

Experiments

- Training Sequences

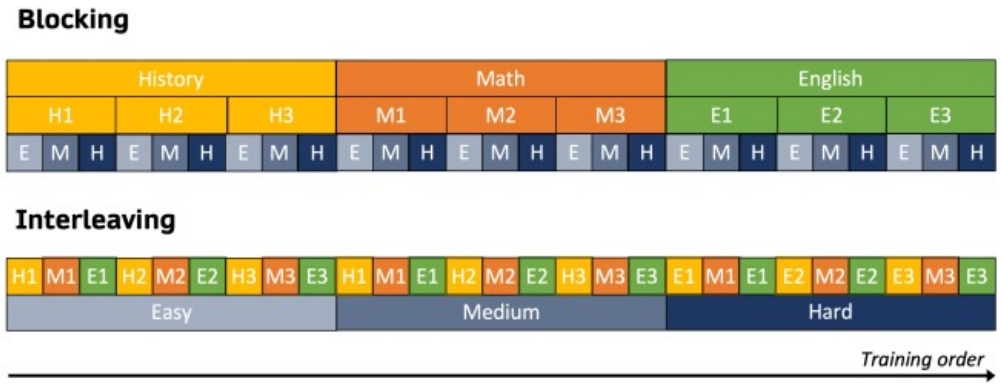


Figure 2: A comparison of two training sequences. Small blocks (e.g., H1, M1) stand for fine-grained concepts per subject. *Blocking* naively stacks hierarchical blocks per subject, while *interleaving* cyclically revisits each subject, adhering to the cognitive hierarchy from Bloom's taxonomy.

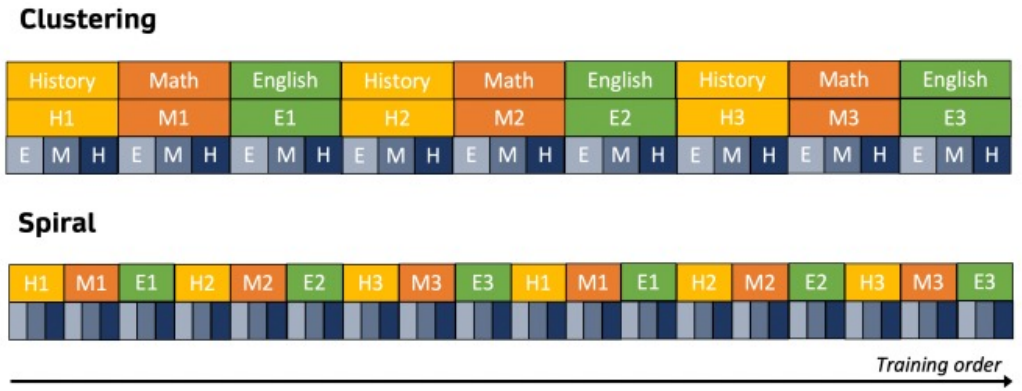


Figure 3: (Continued from Figure 2) **More examples of local progressions**. A comparison of clustering and spiral training sequences. The *clustering* stacks hierarchical blocks for each concept, while the *spiral* cyclically revisits each concept and alternates cognitive difficulty from Bloom's taxonomy.

Corgi : Instruction Tuning with Human Curriculum

Experiments

- Training Details
  - Base Model : LLaMA2 13B (bf16 precision)
  - Global Batch Size : 256
  - 1 batch per GPU, 16 gradient accumulation
  - 16 x A100 GPUs
  - Learning Rate : 2e-5
  - Sequence Length : 2048
  - Epochs : 5
- Evaluation
  - MMLU (Multi-domain Knowledge through exam question)
  - HellaSwag (Adversarial commonsense natural language inference)
  - ARC (Scientific reasoning on grade-school questions)
  - TruthfulQA (Adversarial Facts)
  - PIQA (Physical commonsense reasoning on atypical situations)
  - CommonsenseQA (Commmonsense reasoning abilites on Real-world concepts)
  - OpenBookQA
  - Lambada

Corgi : Instruction Tuning with Human Curriculum

Results

Table 1: Performances of LLaMA 2 13B based models on 6 different benchmarks.

Model	# Data	MMLU	ARC	PIQA	CSQA	OBQA	HellaSwag†
		General Knowledge	Sci. Exams - Hard Set	Physical Objects	Real-World Concepts	Science Textbooks	Real-World Activities
		5-shot	25-shot	10-shot	10-shot	5-shot	10-shot
CORGI†	66K	**57.74**	**58.70**	**81.99**	**70.19**	**51.80**	**82.98**
CORGI- Blocking		55.63	56.57	80.20	69.53	48.60	81.89
Vicuna v1.5	125K	56.50	55.80	81.56	**70.19**	47.40	80.21
WizardLM v1.2	250K	55.26	55.97	81.45	68.30	49.60	80.91
LLaMA 2 13B	-	54.99	56.31	80.85	68.30	48.00	80.80

†The default CORGI model uses an interleaved sorting approach as described in Section 2.2.

Table 2: Performances of respective datasets on LLaMA 2 13B on three different categories of tasks. This table is a breakdown of Figure 4

Curriculum	MMLU	TriviaQA	TruthfulQA	ARC	CSQA	OBQA	PIQA	HellaSwag	Lambada
	World Knowledge			Commonsense Reasoning				Language Understanding	
	5-shot	64-shot	0-shot	25-shot	10-shot	5-shot	10-shot	10-shot	0-shot
Interleaving	**57.74**	**64.34**	**47.44**	**58.70**	**70.19**	**51.80**	**82.0**	**83.0**	**76.1**
Blocking	55.63	61.95	43.27	56.57	69.53	48.60	80.20	81.89	75.99
Clustering	55.24	58.75	42.12	57.42	67.65	49.00	80.31	81.89	75.65
Spiral	54.46	61.92	41.25	56.66	68.96	49.00	80.52	81.89	76.13
Random Shuffle	54.76	62.44	42.57	57.42	68.63	49.40	80.3	79.31	75.0
LLaMA 2 13B	54.99	62.44	39.91	56.31	68.30	48.00	80.85	80.80	76.56

Corgi : Instruction Tuning with Human Curriculum

Results

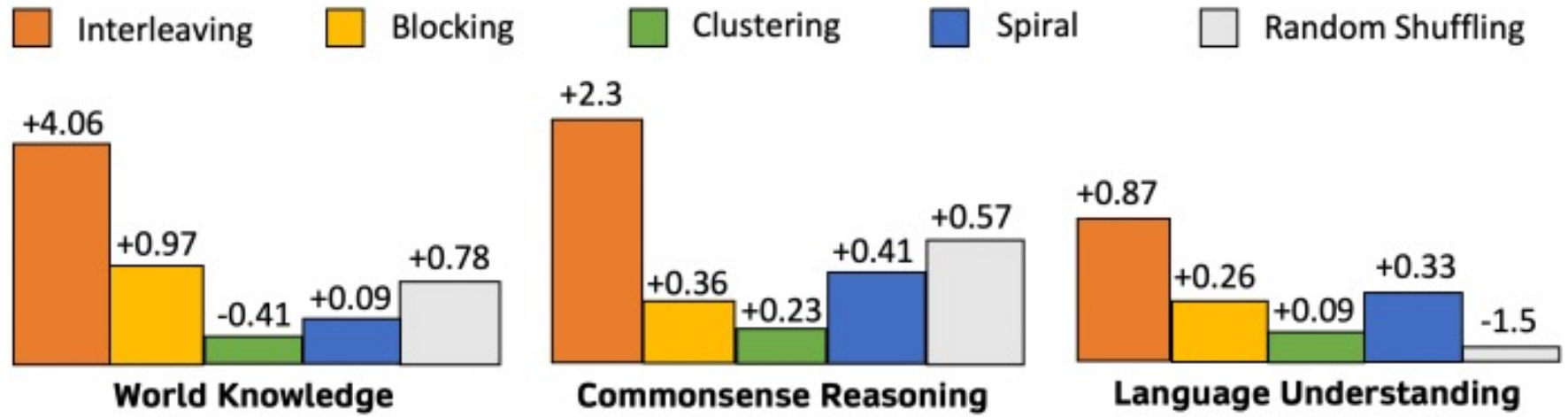


Figure 4: **Local curriculum diminishes performance improvement.** The figure shows a macroscopic, averaged performance comparison of several benchmark improvements with respect to the base model (LLaMA 2 13B) performance. *World Knowledge:* MMLU, TruthfulQA, TriviaQA, *Commonsense Reasoning:* OpenBookQA, ARC, PIQA, CommonsenseQA, *Language Understanding:* HellaSwag, and Lambada. A full breakdown of this chart is given in the Appendix H.

Corgi : Instruction Tuning with Human Curriculum

Conclusion

- Instruction Tuning을 위한 새로운 방법론인 CORGI 소개
- 구조화되고 교육학적 영감을 받은 데이터셋 구축 방법론
- 추론 및 지식 기반 Task에서 기존 벤치마크 뛰어 넘을 뿐만 아니라, 컴퓨팅 리소스 많이 쓰지 않으면서 Outperform
- 인간의 경험 기반으로 구조화된 고품질의 데이터가 모델 성능에 있어 중요한 역할을 한다는 것을 강조

- 연구 Ideation
  - 인간의 교육과정으로 Finetuning 하였을 때 성능이 유의미하게 향상 확인
  - “나라 별 교육과정”으로 데이터셋을 구축하여 해당 데이터로 Finetuning하였을 때 모델 성능 차이는 어떨까?
  - MMLU (Exam base Model Evaluation Benchmark)에 해당 실험 진행

HCC
LAB
SNU

QnA