WenMind: A Comprehensive Benchmark for Evaluating LLMs in Chinese Classical Literature and Language Arts

Jiahuan Cao[†1,3], **Yang Liu**[†1,3], **Yongxin Shi**[1,3], **Kai Ding**[2,3], **Lianwen Jin**[*1,3]

[1]South China University of Technology

[2]INTSIG Information Co., Ltd

[3]INTSIG-SCUT Joint Lab on Document Analysis and Recognition

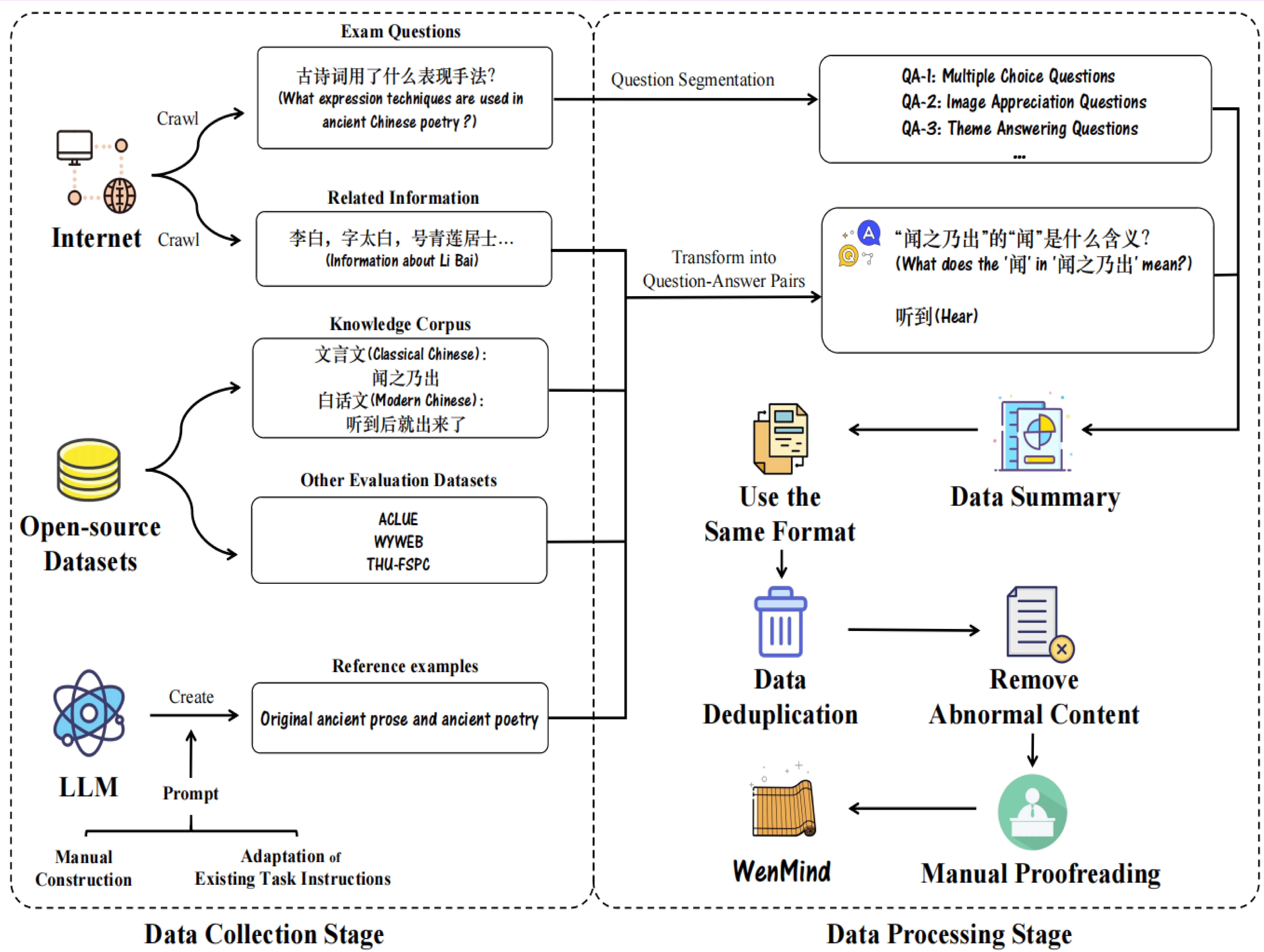
INTRODUCTION

NEURAL INFORMATION PROCESSING SYSTEMS

WenMind, a benchmark for evaluating Large Language Models (LLMs) in Chinese Classical Literature and Language Arts (CCLLA). It covers Ancient Prose, Ancient Poetry, and Ancient Literary Culture, featuring 4,875 question-answer pairs across 42 tasks.

DATA CONSTRUCTION

NEURAL INFORMATION PROCESSING SYSTEMS



Data Collection:

(a) Internet: Curated exam questions and CCLLA texts for Q&A pairs.

(b) Open-Source Datasets: Utilized resources like C2MChn and ACLUE; pro-cessed and standardized question-answer pairs.

(c) LLM: Generated reference answers for open-ended tasks using ERNIE-3.5, with manual refinement.

Data Processing:

Ensured data quality through question segmentation, standar-dization, deduplication, cleaning, and manual proofreading.

DATA STATISTICS

NEURAL INFORMATION PROCESSING SYSTEMS

The statistics of the WenMind Benchmrak. "Q" represents "Question" and "A" represents "Answer".

Domain	Tasks	#Q	Max. #Q	Min. #Q	Avg. Q Tokens	Avg. A Tokens
Ancient Prose	15	1,900	200	7	107.51	62.12
Ancient Poetry	16	1,845	200	20	73.42	94.93
Ancient Literary Culture	11	1,130	100	100	26.68	14.26
Overall	42	4,875	200	7	75.87	63.44

(a) (b) (c)

Data statistics of WenMind:
Distributions of (a) sentence length, (b) sub-domains and (c) capabilities.

Results and Analysis

NEURAL INFORMATION PROCESSING SYSTEMS

Results of all evaluated models on different domains and capabilities.

Model	Overall	Domain			Capability		
		Ancient Prose	Ancient Poetry	Ancient Literary Culture	Understanding	Generation	Knowledge
Baichuan2-7B-Chat [42]	41.2	49.5	33.6	39.5	47.8	58.2	27.7
Baichuan2-13B-Chat [42]	45.5	53.4	39.8	41.6	53.7	58.4	31.2
Firefly-Baichuan2-13B [54]	38.7	44.7	33.1	37.8	45.2	50.2	26.9
ChatGLM2-6B [43]	35.4	43.9	29.9	30.0	43.8	52.3	19.6
ChatGLM3-6B [43]	39.5	50.9	32.4	32.0	50.9	55.7	20.0
InternLM2-Chat-7B [55]	50.2	53.4	47.5	49.3	54.7	63.3	40.8
Qwen1.5-0.5B-Chat [41]	26.1	36.7	17.0	23.4	37.2	43.4	6.7
Qwen1.5-4B-Chat [41]	39.6	48.5	32.5	36.1	48.0	52.5	24.9
Qwen1.5-7B-Chat [41]	50.3	55.5	48.2	44.7	57.9	65.0	36.2
Qwen1.5-14B-Chat [41]	54.9	60.5	52.8	49.1	62.5	65.3	42.0
Qwen1.5-32B-Chat [41]	57.0	63.3	52.6	53.4	64.6	65.7	44.4
Qwen1.5-72B-Chat [41]	58.5	64.0	55.6	54.0	65.9	67.4	46.3
Yi-1.5-6B-Chat [52]	47.2	53.4	42.9	43.7	54.7	61.9	33.3
Yi-1.5-9B-Chat [52]	51.7	58.4	46.6	48.6	59.1	65.0	38.1
Yi-1.5-34B-Chat [52]	57.4	63.0	52.0	56.6	63.2	69.6	46.4
ERNIE-3.5-8K-0329 [10]	62.2	63.5	55.7	70.7	64.4	74.8	55.9
ERNIE-4.0-8K-0329 [10]	**64.3**	**66.3**	**56.6**	73.4	**66.8**	**76.1**	57.8
Spark-3.5 [56]	60.9	59.8	54.1	**73.7**	60.2	66.9	**60.2**
Gemma-1.1-7B-IT [57]	25.2	32.4	21.8	18.6	34.9	47.7	6.2
Ziya-LLaMA-13B-v1.1 [58]	34.1	42.5	28.2	29.5	43.5	50.2	17.2
LLaMA2-7B-Chat [40]	13.0	14.0	14.3	9.2	16.8	26.9	4.2
LLaMA2-13B-Chat [40]	23.7	29.7	21.6	17.1	32.2	40.5	7.9
LLaMA2-Chinese-7B-Chat [45]	18.1	29.6	11.2	10.0	27.5	25.1	3.6
LLaMA2-Chinese-13B-Chat [46]	23.7	36.4	15.3	16.0	35.7	35.3	4.5
LLaMA3-8B-Instruct [59]	34.7	45.0	27.5	29.1	46.1	57.4	13.4
LLaMA3-Chinese-8B-Chat [60]	37.3	49.9	30.1	27.7	50.2	55.7	15.2
GPT-3.5 [61]	35.3	46.1	30.5	25.1	47.1	50.7	15.6
GPT-4 [62]	50.2	60.3	44.2	43.1	61.3	61.7	32.4
Ancient-Chat-LLM-7B [51]	32.7	42.6	23.9	30.5	41.1	39.1	19.9
Bloom-7B-Chunhua [48]	32.5	42.7	24.0	29.3	42.2	41.4	17.3
Xunzi-Qwen1.5-7B [47]	37.0	44.8	29.4	36.2	44.9	46.8	23.8
Average	41.2	48.5	35.6	38.0	49.2	54.5	27.1

(a) Performance Gaps: ERNIE-4.0 leads with a score of 64.3, while most models score between 20-60, indicating significant room for improvement in CCLLA tasks.

(b) Data Matters: Pre-training on large, high-quality Chinese datasets plays a critical role in performance, surpassing fine-tuned English models even when supplemented with Chinese data.

(c) Incremental Pre-training Limitations: Models specifically pre-trained on CCLLA data underperform, likely due to insufficient data coverage and catastrophic forgetting of general knowledge.

(d) Knowledge Deficit: LLMs struggle with knowledge-focused tasks, particularly in Ancient Poetry and Literary Culture, performing better in generation and understanding.

(e) Scaling Law: Larger models with more parameters show better performance, consistent with the scaling law in the CCLLA domain.

DLVCLAB

Conclusion

Our WenMind evaluation reveals significant gaps in LLM performance within the CCLLA domain, with the top model scoring only 64.3. These results highlight the need for better pre-training data and strategies to improve knowledge retention. Moving forward, expanding training datasets and refining model fine-tuning will be key to advancing LLM capabilities in CCLLA. WenMind offers a strong foundation for future research and development in this field.